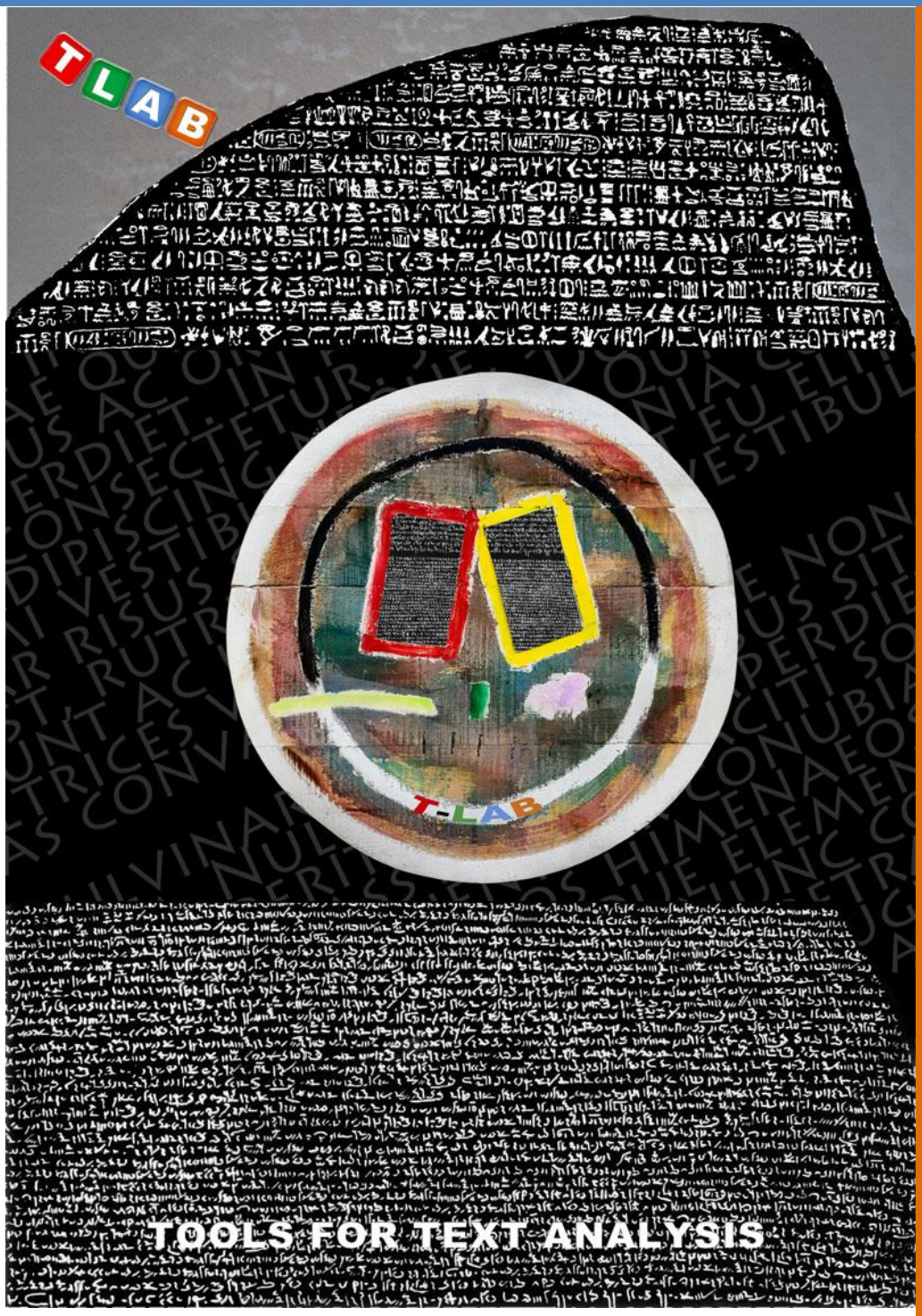
T-LAB 10

Manuale d'Uso



The above artwork has been realized for T-LAB
by Claudio Marini (http://www.claudiomarini.it/)
in collaboration with Andrea D'Andrea.

Strumenti per l'Analisi dei Testi

Website: https://www.tlab.it/
E-mail: info@tlab.it

INDICE

Installazione e requisiti di sistema

Configurazione minima richiesta:

- Sistema operativo: Windows 7 o successivi
- RAM: 4 Gb
- Full HD (risoluzione raccomandata 1920 x 1080).

Installazione:

- Doppio click sul file Setup.exe
- Seguire le istruzioni che appaiono sullo schermo
- Uscire dal programma
- Attendere una comunicazione con la chiave di attivazione
- - Per maggiori informazioni: https://www.mytlab.com/T-LAB_Installation.pdf

Cosa fa e cosa consente di fare

T-LAB è un software costituito da un insieme di **strumenti linguistici**, **statistici e grafici per l'analisi dei testi** che possono essere utilizzati nelle seguenti pratiche di ricerca: Analisi di Contenuto, Sentiment Analysis, Analisi Semantica, Analisi Tematica, Text Mining, Perceptual Mapping, Analisi del Discorso, Network Text Analysis, Document Clustering, Text Summarization.

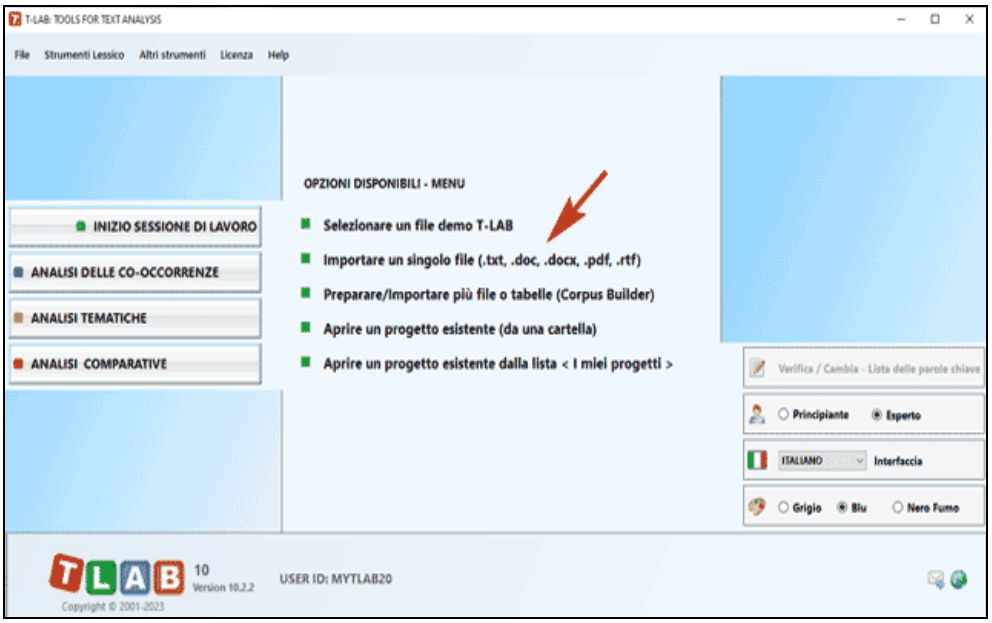
In effetti, tramite gli strumenti **T-LAB** i ricercatori possono gestire agevolmente attività di analisi come le seguenti:

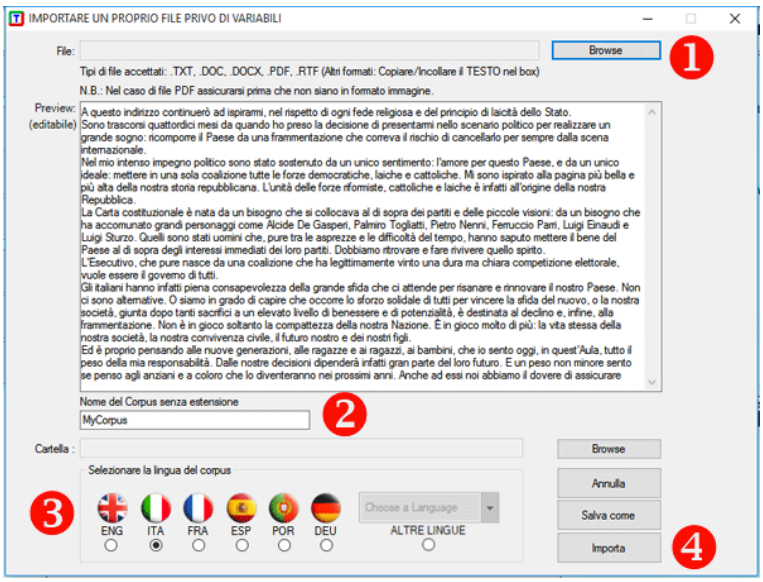
- esplorare, misurare e mappare la **relazioni di co-occorrenza** tra parole-chiave;
- realizzare una **classificazione automatica** di unità testuali o documenti, sia tramite un approccio **bottom-up** (cioè che tramite l'analisi dei **temi emergenti**), sia tramite un approccio **top-down** (cioè tramite l'uso di **categorie predefinite**);
- verificare quali **unità lessicali** (cioè parole o lemmi), quali **unità di contesto** (cioè frasi o paragrafi) e quali **temi** sono 'caratteristici' di specifici sottoinsiemi di testi (ad es., discorsi di specifici leader politici, interviste con specifiche categorie di persone, etc.);
- applicare categorie per la **sentiment analysis**;
- eseguire vari tipi di **analisi delle corrispondenze** e **cluster analysis**;
- creare **mappe semantiche** che rappresentano **aspetti dinamici** del discorso (cioè relazioni sequenziali tra parole o temi);
- rappresentare ed esplorare un qualsiasi testo come una **rete** di relazioni;
- ottenere misure e rappresentazioni grafiche relative a **testi e discorsi** trattati come **sistemi dinamici**;
- personalizzare e applicare **vari tipi di dizionari**, sia per l'analisi lessicale che per l'analisi di contenuto;
- verificare i contesti di occorrenza (ad es., **concordanze**) di parole e lemmi;
- analizzare tutto il **corpus** o solo alcuni dei suoi **sottoinsiemi** (ad esempio gruppi di documenti) utilizzando varie liste di parole-chiave;
- creare, esplorare ed esportare vari tipi di **tabelle di contingenza** e **matrici di co-occorrenze**.

L'interfaccia del software è particolarmente **user friendly** e i testi analizzabili possono essere i più vari:
- un singolo testo (es. un'intervista, un libro, etc.);
- un insieme di testi (es. più interviste, pagine web, articoli di giornale, risposte a domande aperte, messaggi Twitter etc.).

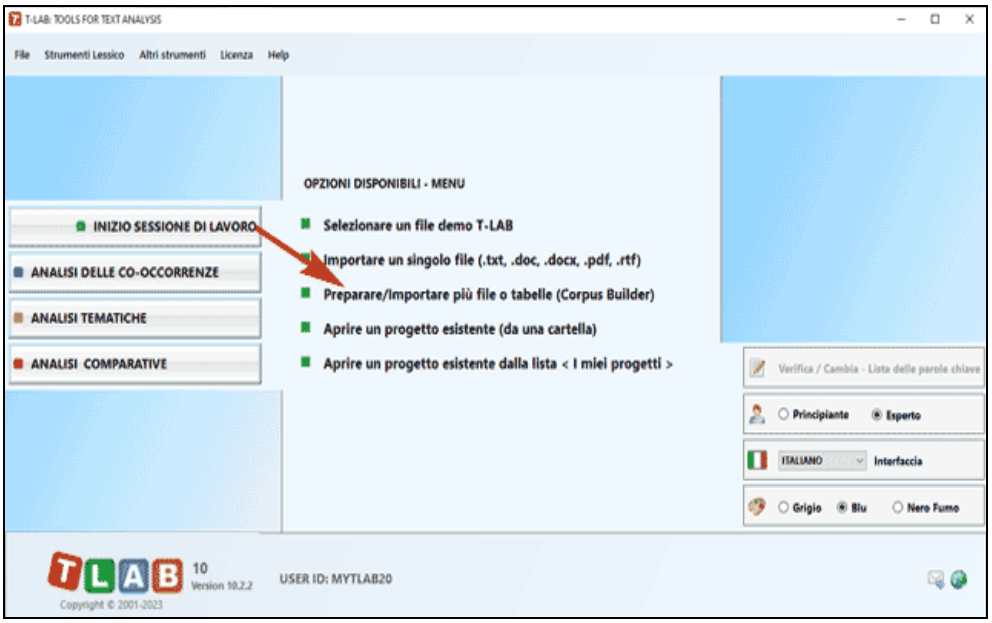
Tutti i testi possono essere codificati con variabili categoriali e/o con un identificativo (**Unique Identifier**) che corrisponde a unità di contesto o a casi (es. risposte a domande aperte).

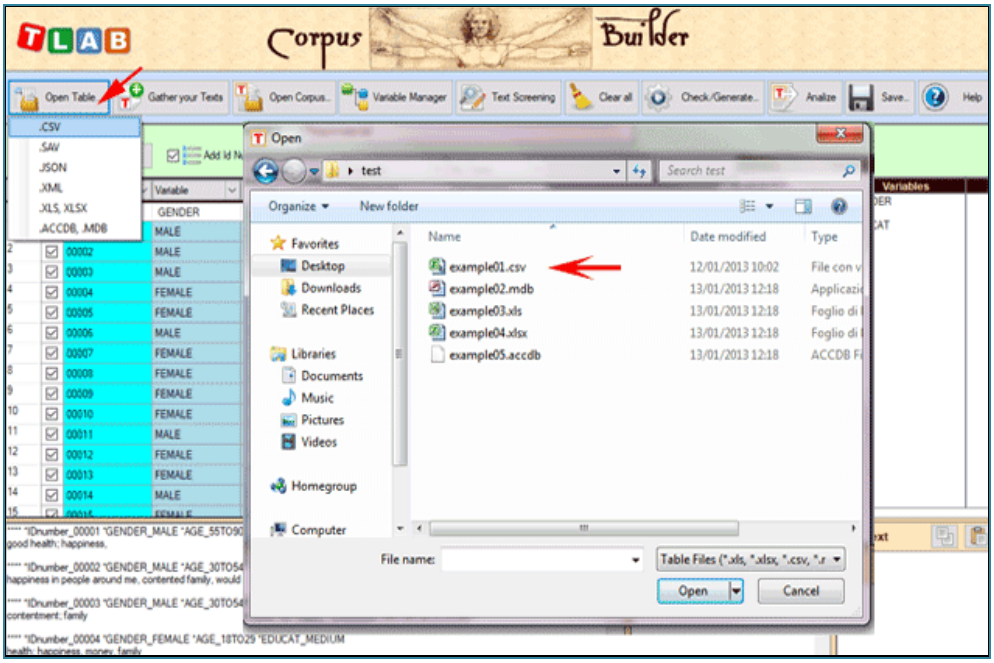
Nel caso di un singolo documento (o di un corpus trattato come unico testo) **T-LAB** non richiede ulteriori accorgimenti: basta selezionare l'opzione 'Importare un singolo file…' e procedere (vedi sotto).

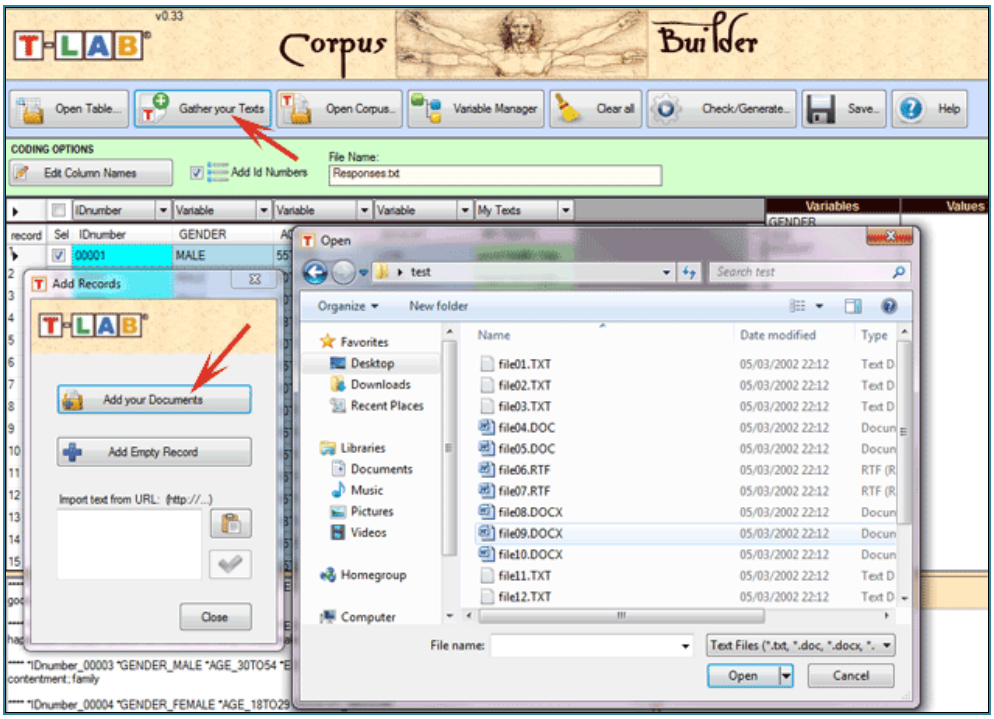




Diversamente, negli altri casi va usato il modulo **Corpus Builder** che – in modo automatico - facilita la trasformazione di vari tipi di materiali testuali e vari tipi di file in un **corpus** pronto per essere importato da **T-LAB** (vedi sotto)**.**



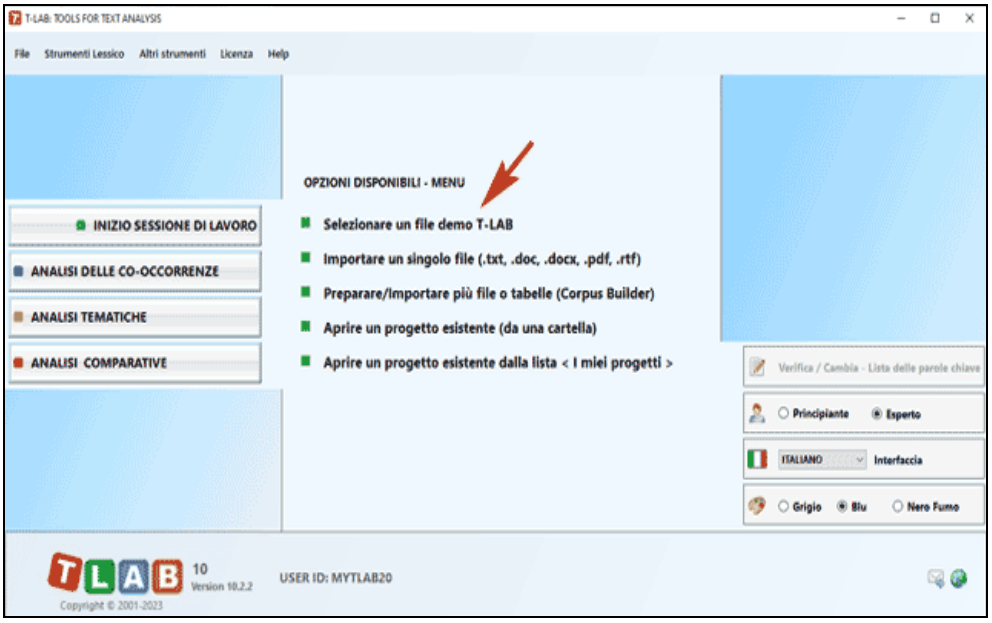




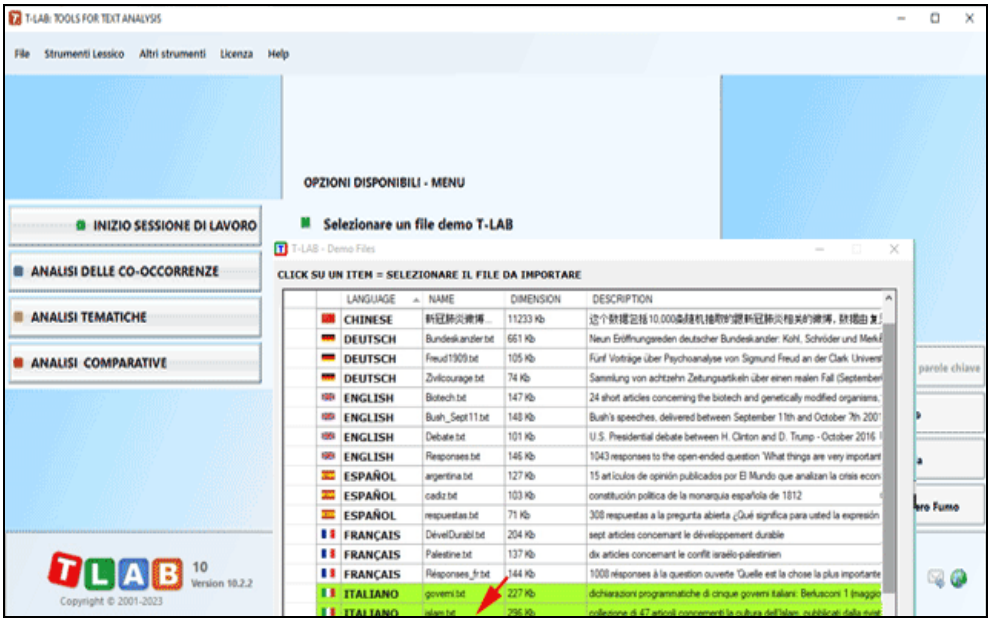
N.B.: Al momento - per garantire l'uso integrato dei vari strumenti - ogni file/corpus da analizzare non deve superare i 90 Mb (cioè circa 55.000 pagine in formato testo). Per ulteriori informazioni, vedere la sezione 'Requisiti e prestazioni' dell'Help / Manuale.

Per verificare rapidamente le funzionalità del software sono sufficienti i seguenti passi:

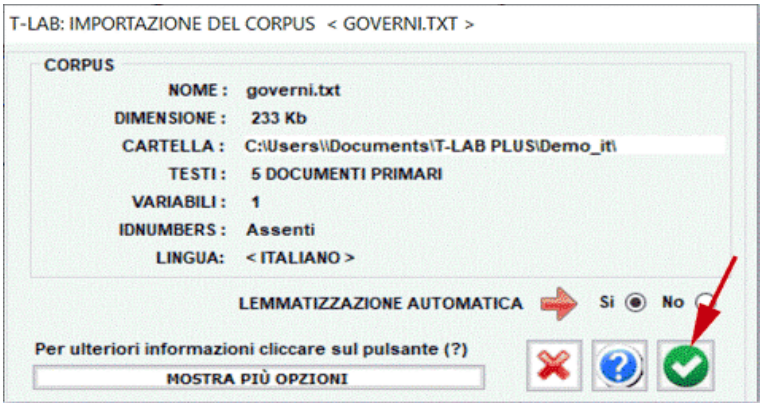
1 - **Selezionare l'opzione 'Selezionare un file demo T-LAB"**



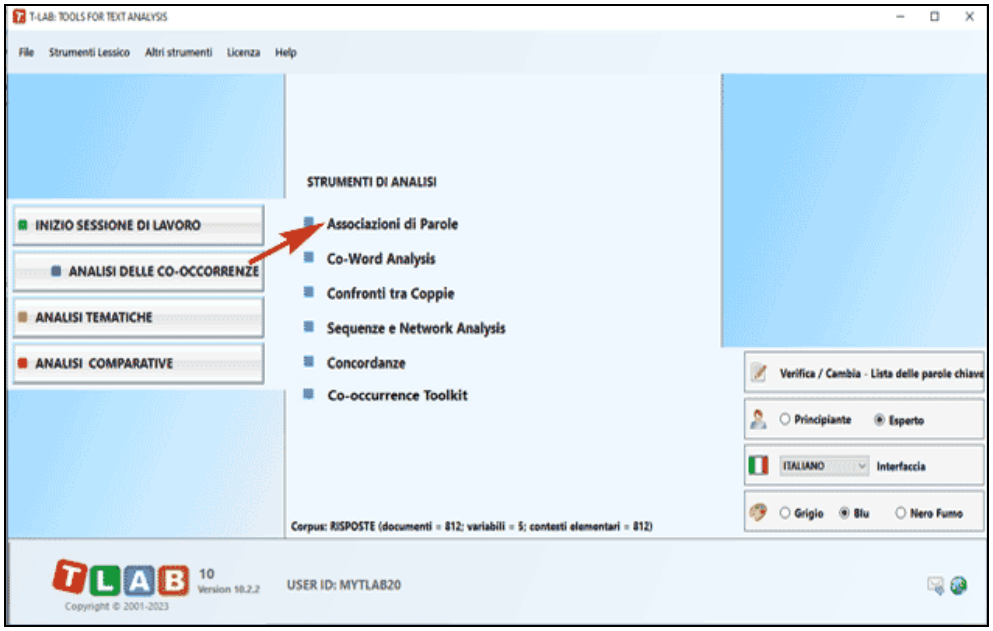
2 - Selezionare un corpus da analizzare



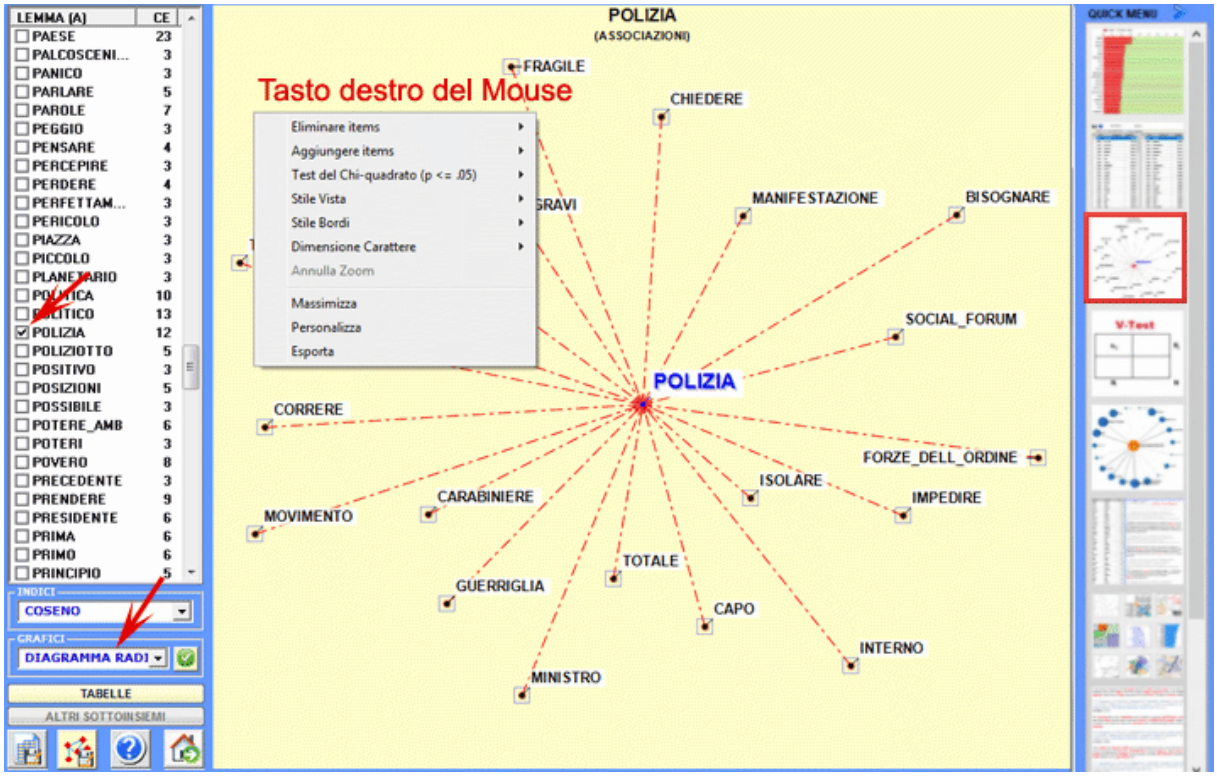
3 - Cliccare su "ok" nella prima finestra di Setup

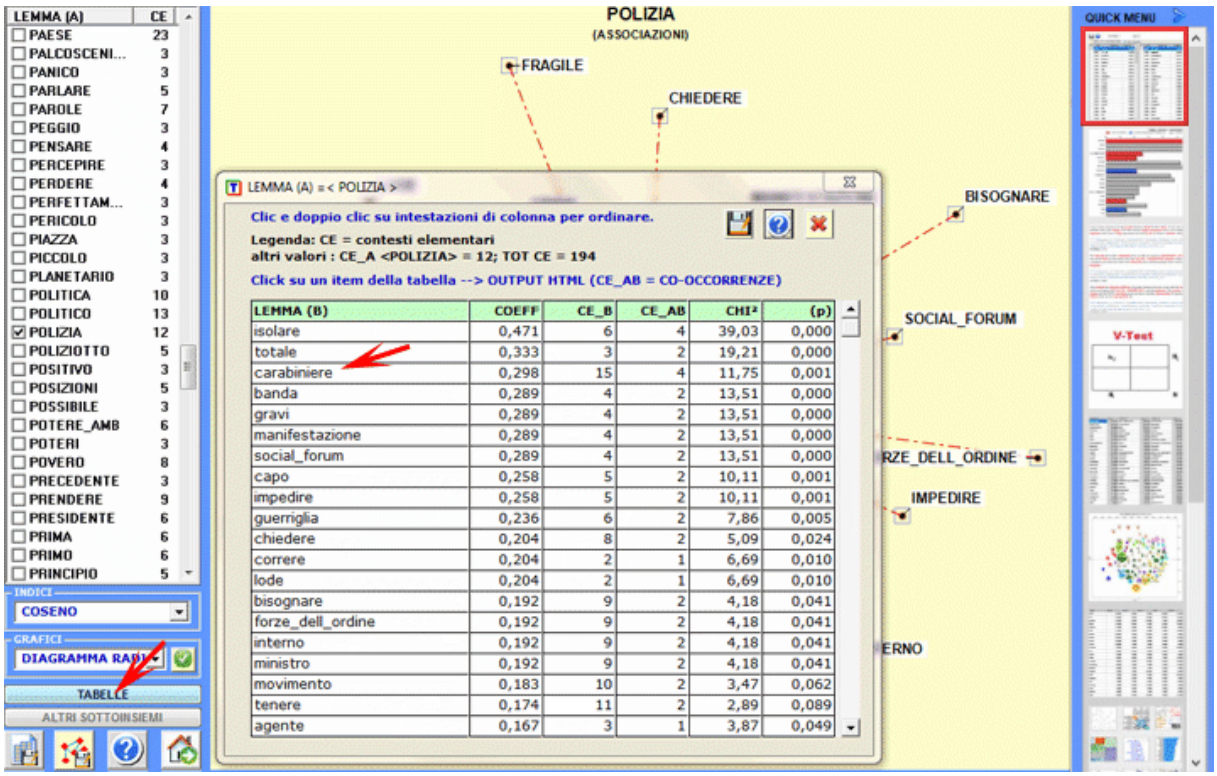


4 - Scegliere uno strumento all'interno di uno dei sub-menu ‘Analisi’



5 - **Verificare i risultati**





6 - **Utilizzare l'help contestuale per interpretare grafici e tabelle**

Tutti i grafici possono essere personalizzati e salvati in vari formati.

Di seguito vengono fornite le informazioni essenziali per capire cosa **T-LAB** fa e come può essere usato.

Dal punto di vista esterno, l'uso del software è organizzato dall'**interfaccia**, cioè dal **menu principale**, dai **sub-menu** e dalle **funzioni** (strumenti) che li compongono.

Da un punto di vista logico, oltre che dall'interfaccia utente, il sistema **T-LAB** è organizzato da due componenti principali:

- il **database**, cioè è il "luogo" informatico in cui il corpus in input (cioè il testo o l'insieme dei testi da analizzare) è rappresentato come un insieme di **tabelle** in cui sono registrate le **unità di analisi**, le loro caratteristiche e le loro reciproche relazioni;
- gli **algoritmi**, cioè sottoinsiemi di **istruzioni** che consentono di usare l'interfaccia utente, di consultare e modificare il database, di costruire ulteriori tabelle con in dati in esso contenuti, di effettuare **calcoli statistici** e di produrre **output** che rappresentano le relazioni tra i dati analizzati.

Per capire come **T-LAB** funziona e come può essere usato, è di fondamentale importanza aver chiaro quali unità di analisi sono archiviate nel suo database e quali algoritmi statistici vengono usati nelle varie analisi. Infatti, le tabelle dati analizzate sono sempre costituite da righe e colonne le cui intestazioni corrispondono alle unità di analisi archiviate nel database, mentre gli algoritmi regolano i processi che consentono di individuare relazioni significative tra i dati e di estrarre utili informazioni.

Le **unità di analisi** di **T-LAB** sono di due tipi: **unità lessicali** e **unità di contesto**.

A - le **UNITA' LESSICALI** sono parole, singole o multiple, archiviate e classificate in base a un qualche criterio. Più precisamente, nel database **T-LAB** ogni unità lessicale costituisce un record classificato con due campi: forma e lemma. Nel primo campo, denominato **forma**, sono elencate le parole così come compaiono nel corpus, mentre nel secondo, denominato **lemma**, sono elencate le label attribuite a gruppi di unità lessicali classificate secondo criteri linguistici (es. lemmatizzazione) o tramite dizionari e griglie semantiche definite dall'utilizzatore.

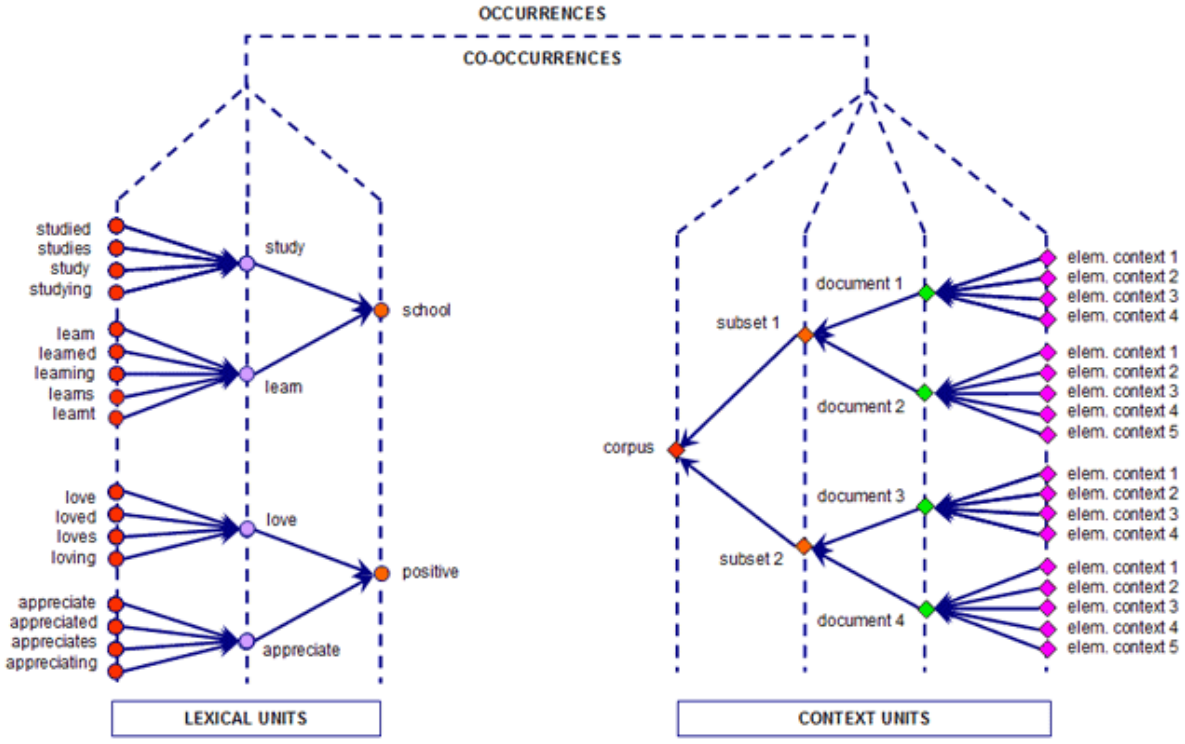
B - le **UNITA' DI CONTESTO** sono porzioni di testo in cui può essere suddiviso il corpus. Più esattamente, nella logica **T-LAB**, le unità di contesto possono essere di tre tipi:

B.1 **documenti primari**, corrispondenti alla suddivisione "naturale" del corpus (es. interviste, articoli, risposte a domande aperte, etc.), ovvero ai **contesti iniziali** definiti dall'utilizzatore;

B.2 **contesti elementari**, corrispondenti alle unità sintagmatiche (frammenti di testo, frasi, paragrafi) in cui può essere suddiviso ogni contesto iniziale;

B.3 **sottoinsiemi del corpus**, corrispondenti a gruppi di documenti primari riconducibili alla stessa "categoria" (es. interviste di "uomini" o di "donne", articoli di un particolare anno o di una particolare testata, etc.) o a cluster tematici ottenuti con specifici strumenti **T-LAB**.

Il diagramma seguente illustra le possibili relazioni tra unità lessicali e unità di contesto che **T-LAB** ci permette di analizzare.

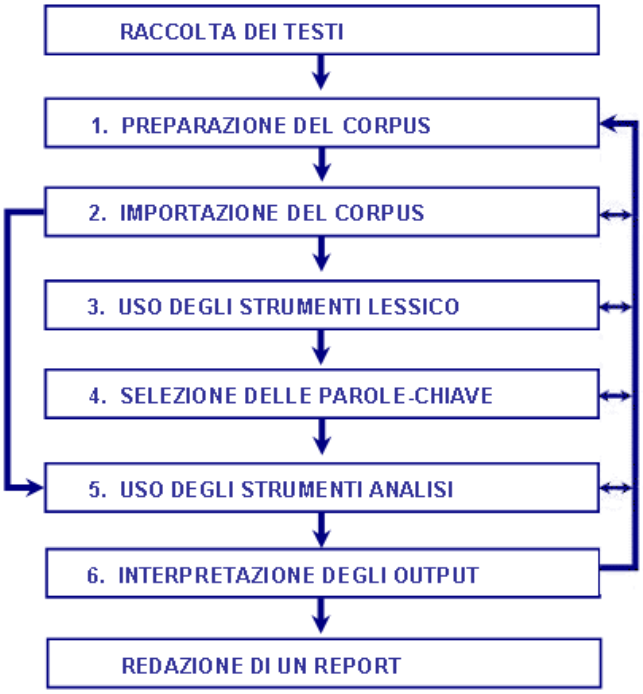


A partire da questa organizzazione del database, **T-LAB** consente - in modo automatico - di esplorare e di analizzare le relazioni tra le unità di analisi di **tutto il corpus** o di suoi **sottoinsiemi**.

In **T-LAB**, la selezione di un qualsivoglia strumento di analisi (click del mouse) attiva sempre un processo semiautomatico che, con poche e semplici operazioni, genera qualche tabella input, applica qualche algoritmo di tipo statistico e produce alcuni output.

In ipotesi, un tipico **progetto** di lavoro in cui viene usato **T-LAB** è costituito dall'insieme delle attività analitiche (operazioni) che hanno per oggetto il medesimo **corpus** ed è organizzato da una **strategia** e da un **piano** dell'utilizzatore. Quindi, inizia con la **raccolta dei testi** da analizzare e termina con la **redazione di un report**.

La successione delle varie fasi è illustrata nel diagramma seguente:



N.B.:
- Le sei fasi numerate, dalla preparazione del corpus all'interpretazione degli output, sono supportate da strumenti **T-LAB** e sono sempre reversibili;

- Tramite le impostazioni automatiche è possibile evitare due fasi (3 e 4); tuttavia, ai fini della **qualità** dei risultati, si raccomanda l'uso delle funzioni **Personalizzazione del Dizionario** (strumento del menu "Lessico") e **Impostazioni Personalizzate** (cioè selezione delle parole-chiave).

Proviamo ora a commentare le varie fasi una dopo l'altra:

1 - La **PREPARAZIONE DEL CORPUS** consiste nella trasformazione dei testi da analizzare in un file (**corpus**) che può essere elaborato dal software.

Nel caso di un unico testo (o di un corpus trattato come unico testo) **T-LAB** non richiede ulteriori accorgimenti.

Quando, invece, il corpus è costituito da più testi e vengono utilizzate **codifiche** che rinviano all'uso di qualche **variabile**, nella fase di preparazione bisogna utilizzare il modulo **Corpus Builder** che – in maniera automatica – procede alla trasformazione di vari materiali testuali in un file corpus pronto per essere importato da **T-LAB**.

N.B.:

- Al termine della fase di preparazione si raccomanda di creare una nuova cartella di lavoro con al suo interno il solo file corpus da importare.

- Durante le analisi, si raccomanda di tenere il file corpus e la relativa cartella di lavoro su un hard disk dello stesso computer su in cui è installato **T-LAB**. Diversamente, l'esecuzione delle varie procedure potrebbe risultare rallentata e il software potrebbe segnalare degli errori.

2 - **L'IMPORTAZIONE DEL CORPUS** consiste in una serie di **processi automatici** che trasformano il corpus in un insieme di tabelle integrate nel **database T-LAB**.

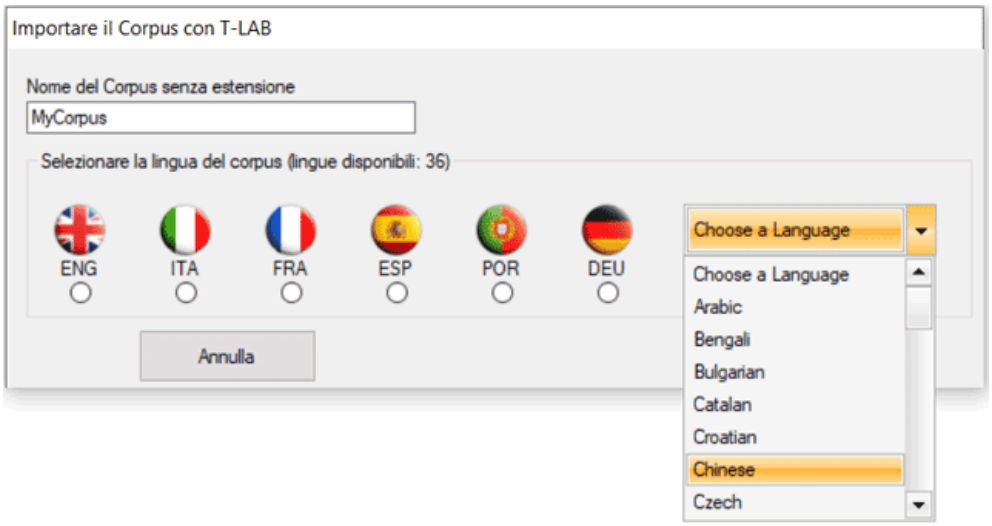
Nella fase di **pre-processing T-LAB** realizza i seguenti trattamenti: **normalizzazione** del testo; riconoscimento di **multi-words** e **stop-words**; **segmentazione** in contesti elementari; **lemmatizzazione** automatica o **stemming**; costruzione del **vocabolario** del corpus; selezione delle **parole chiave**.

Di seguito la lista complete delle trenta (30) lingue per le quali **T-LAB** supporta la lemmatizzazione automatica o lo stemming.

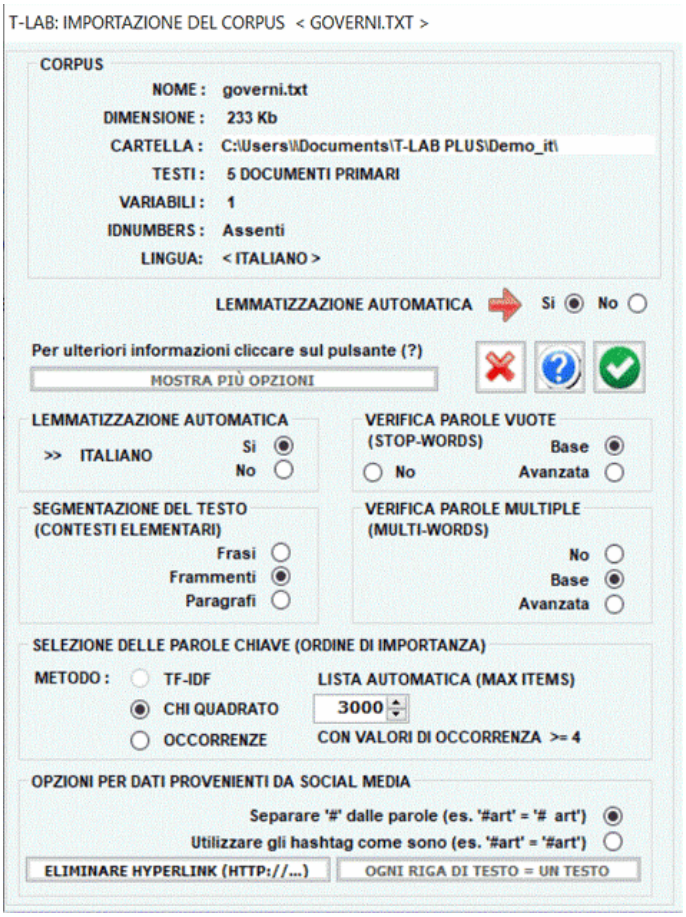
LEMMATIZZAZIONE: catalano, croato, francese, inglese, italiano, latino, polacco, portoghese, rumeno, russo, serbo, slovacco, spagnolo, svedese, tedesco, ucraino.

STEMMING: arabo, bengali, bulgaro, ceco, danese, finlandese, greco, hindi, indonesiano, marathi, norvegese, olandese, persiano, turco, ungherese.

In ogni caso, senza lemmatizzazione automatica e/o usando dizionari personalizzati, possono essere analizzati testi in **tutte le lingue** le cui parole siano separate da spazi e/o da punteggiatura.



A partire dalla selezione della lingua, l'intervento dell'utilizzatore è richiesto per definire le scelte indicate nella finestra seguente:

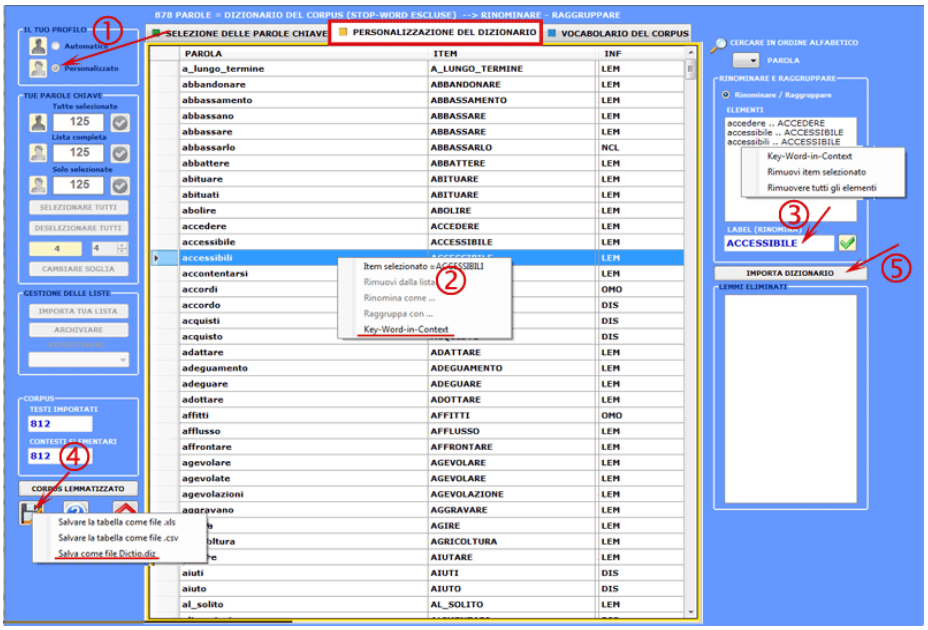


N.B.: Poiché i trattamenti preliminari determinano il tipo e la quantità delle unità di analisi (cioè quali e quante unità di contesto e quali e quante unità lessicali), scelte diverse in questa fase comportano risultati diversi delle successive analisi. Per questa ragione, tutti gli output **T-LAB** mostrati nel manuale e nell'help hanno solo valore indicativo.

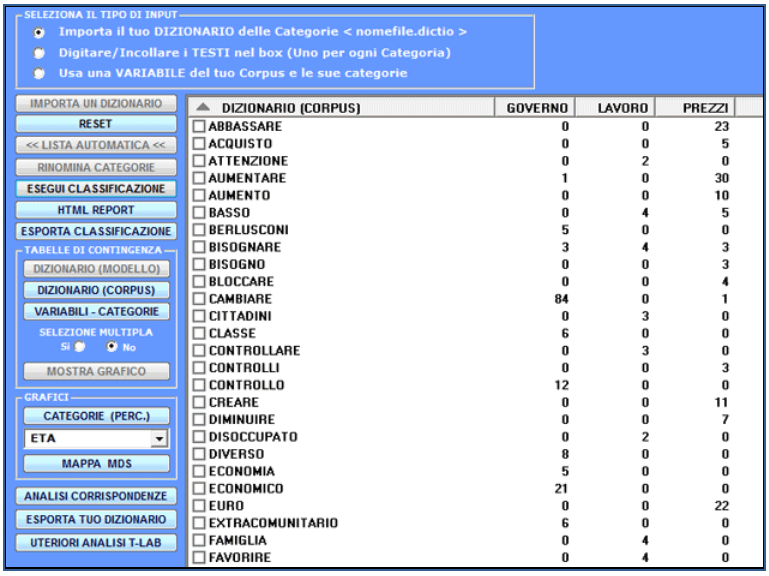
3 - L'USO DEGLI STRUMENTI LESSICO è finalizzato a verificare il corretto **riconoscimento** delle unità lessicali e a personalizzare la loro **classificazione**, cioè a verificare e a modificare le scelte automatiche fatte da **T-LAB**.

Le modalità dei vari interventi sono illustrate nelle corrispondenti voci dell'help (e del manuale).

In particolare si rinvia alla corrispondente voce dell'help (e del manuale) per una dettagliata descrizione del processo **Personalizzazione del Dizionario** (vedi sotto). Infatti, qualsiasi modifica relativa alle voci del dizionario (es., raggruppamento di due o più item) incide sia sul calcolo delle occorrenze che su quello delle co-occorrenze.

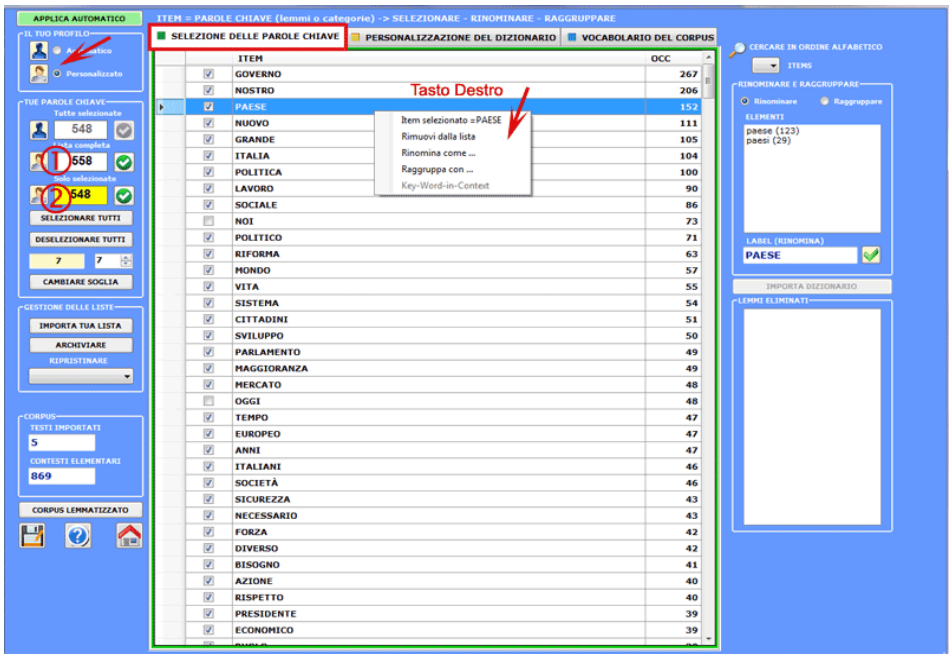


N.B.: Quando l'utilizzatore, senza perdere alcuna informazione lessicale, intende applicare schemi di codifica che raggruppano più parole o lemmi in poche categorie (da 2 a 50) è consigliabile utilizzare lo strumento **Classificazione Basata su Dizionari** incluso nel sottomenu **Analisi Tematiche** (vedi sotto).



4 - LA SELEZIONE DELLE PAROLE-CHIAVE consiste nella predisposizione di una o più liste di unità lessicali (parole, lemmi o categorie) da utilizzare per costruire le tabelle dati da analizzare.

L'opzione **impostazioni automatiche** rende disponibile liste di **parole chiave** selezionate da **T-LAB**; tuttavia, poiché la scelta delle unità di analisi è estremamente rilevante ai fini delle successive elaborazioni, si consiglia vivamente l'uso delle **impostazioni personalizzate**. In questo modo l'utilizzatore potrà scegliere di modificare la lista suggerita da **T-LAB** e/o di costruire liste che meglio corrispondono ai suoi obiettivi di indagine.



In ogni caso, nella costruzione di queste liste, valgono i seguenti criteri:

- verificare la **rilevanza** quantitativa (totale delle occorrenze) e qualitativa (non banalità del significato) dei vari item;

- verificare le **limitazioni** degli strumenti analitici che si intendono utilizzare (vedi nota a fine di questo capitolo);

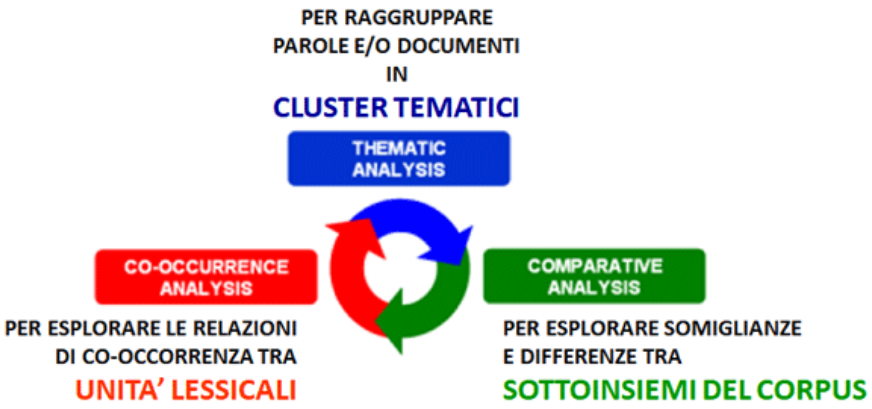
- verificare se l'insieme degli item è compatibile con la propria **strategia** di indagine (vedi punto seguente: 5).

5 - L'USO DEGLI STRUMENTI D'ANALISI è finalizzato alla produzione di output (tabelle e grafici) che rappresentano **relazioni significative** tra le unità di analisi e che consentono di fare **inferenze**.

Attualmente **T-LAB** include venti diversi strumenti di analisi, ciascuno dei quali funziona con una sua specifica logica; cioè, usa specifici algoritmi e produce specifici output.

Di conseguenza, a seconda della tipologia di testi che intende analizzare e degli obiettivi che intende perseguire, l'utilizzatore deve di volta in volta decidere quali strumenti sono più appropriati per la sua **strategia di analisi**.

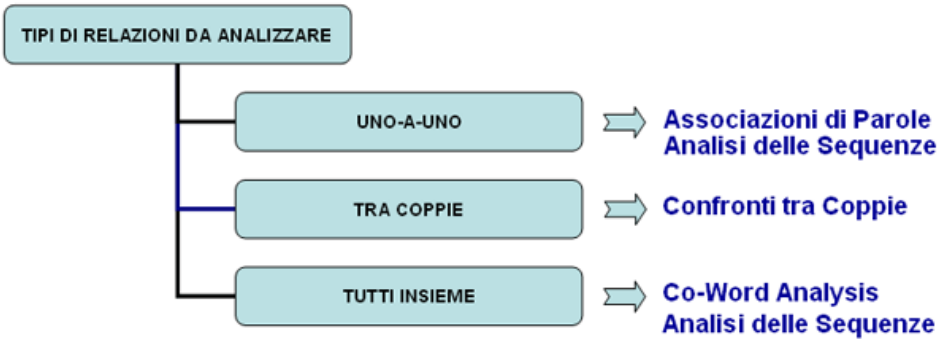
A questo proposito, oltre alla distinzione fra strumenti per **analisi delle co-occorrenze**, per **analisi comparative** e per **analisi tematiche**, è utile considerare che alcuni di questi ultimi consentono di ottenere ulteriori sottoinsiemi del corpus basati su similarità di contenuto.



In generale, anche se l'uso degli strumenti **T-LAB** può essere circolare e reversibile, possiamo individuare tre punti di avvio (start points) che corrispondono ai tre sub-menu ANALISI:

A : **STRUMENTI PER ANALISI DELLE CO-OCCORRENZE**

Questi strumenti consentono di analizzare vari tipi di relazioni tra le unità lessicali (parole, lemmi o categorie).

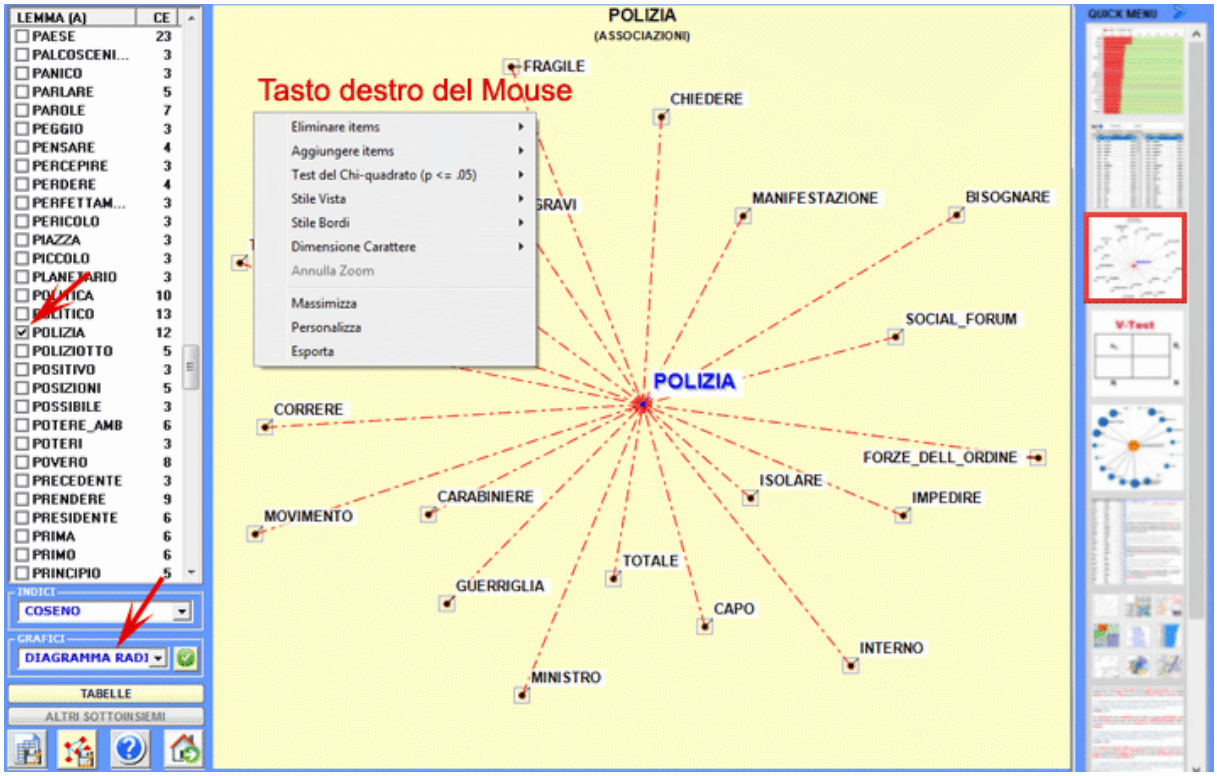


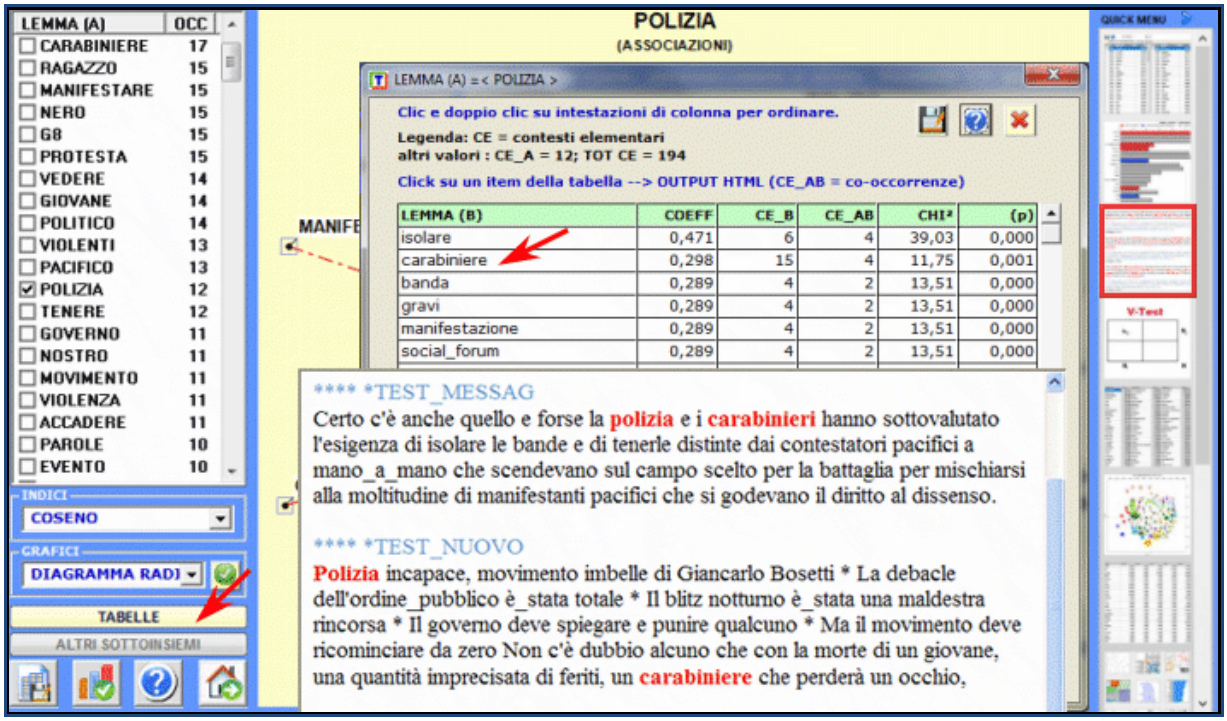
A seconda dei tipi di relazioni da analizzare, le funzioni **T-LAB** indicate in questo diagramma (box colorati) usano uno o più dei seguenti strumenti statistici: **Indici di Associazione**, **Test del Chi Quadro**, **Cluster Analysis**, **Multidimensional Scaling, Analisi delle Componenti Principlali, t-SNE** e **Catene Markoviane**.

Ecco alcuni esempi di output (N.B.: per ulteriori informazioni sulla interpretazione degli output si rimanda alle corrispondenti sezioni della guida / manuale):

- Associazioni di Parole

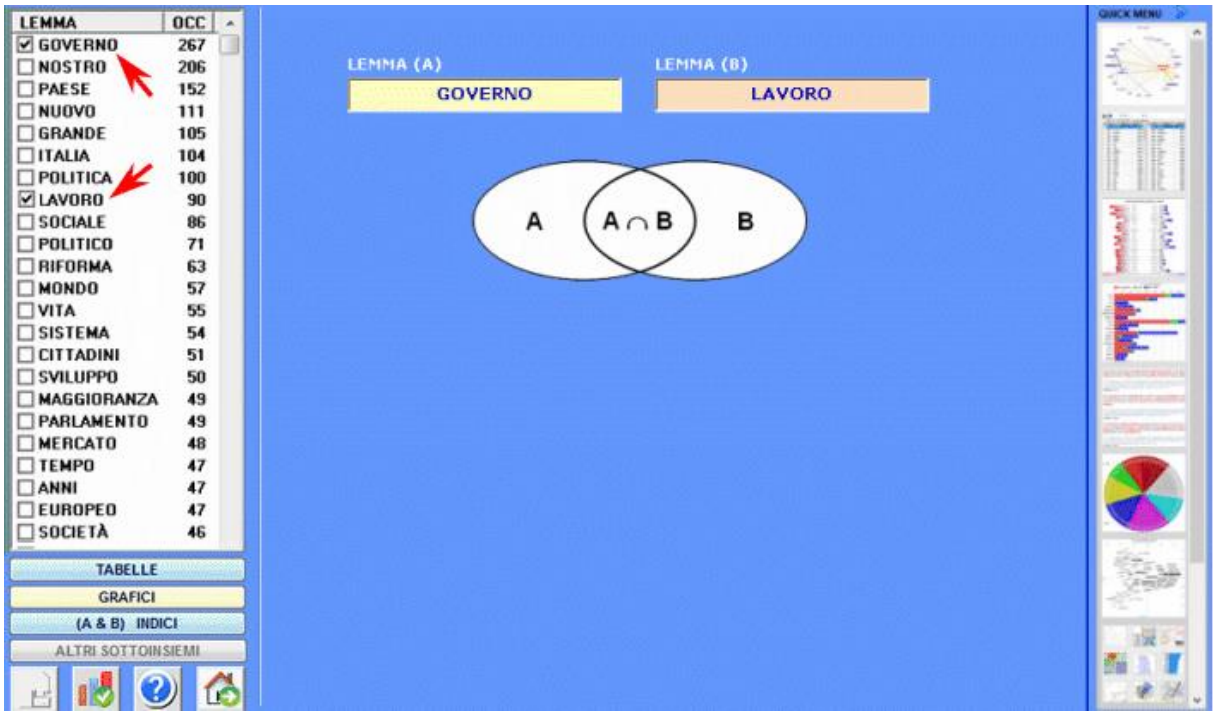
Questo strumento **T-LAB** ci consente di verificare come i contesti di **co-occorrenza** determinano il significato locale delle **parole chiave**.

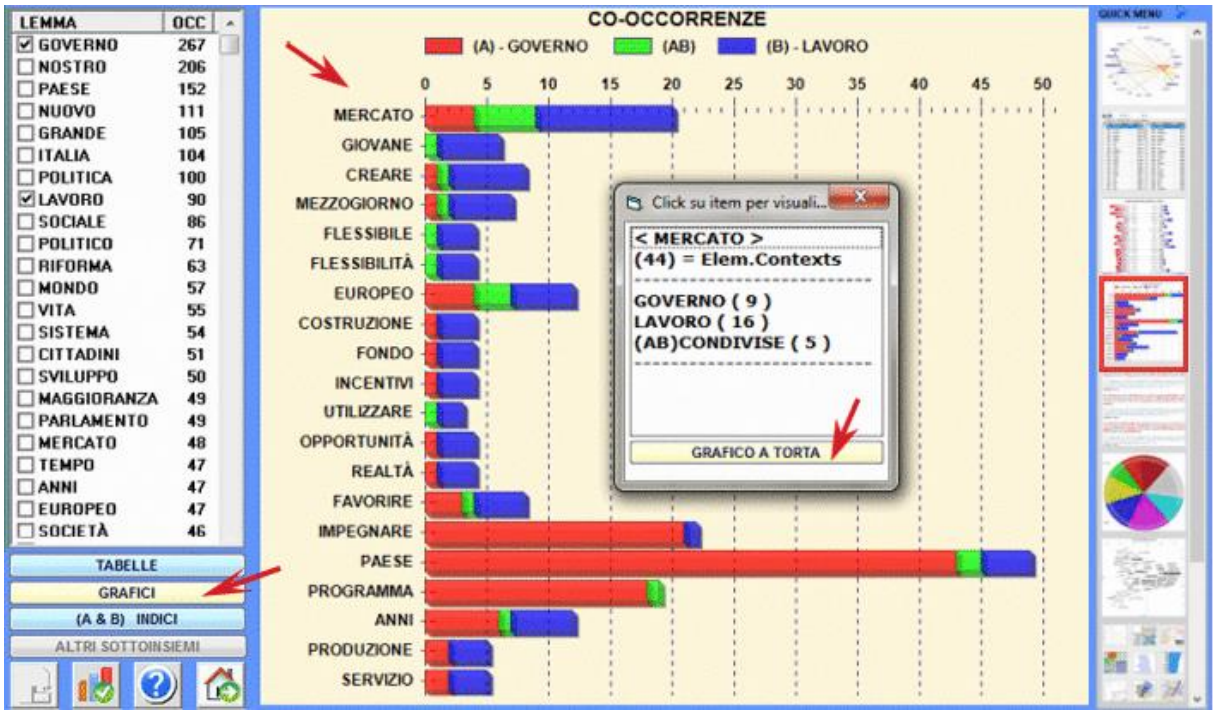




- Confronti tra Coppie

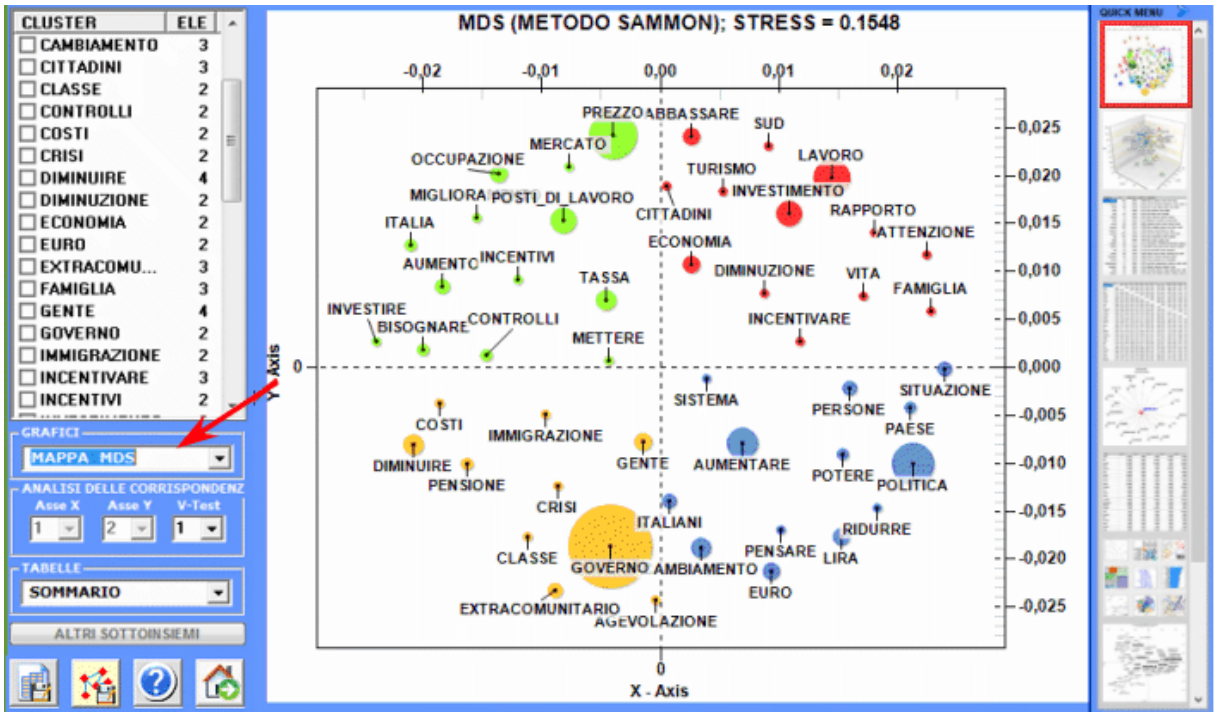
Questo strumento **T-LAB** consente di confrontare insiemi di **contesti elementari** (cioè contesti di co-occorrenza) in cui sono presenti gli elementi di una coppia di **parole chiave**.

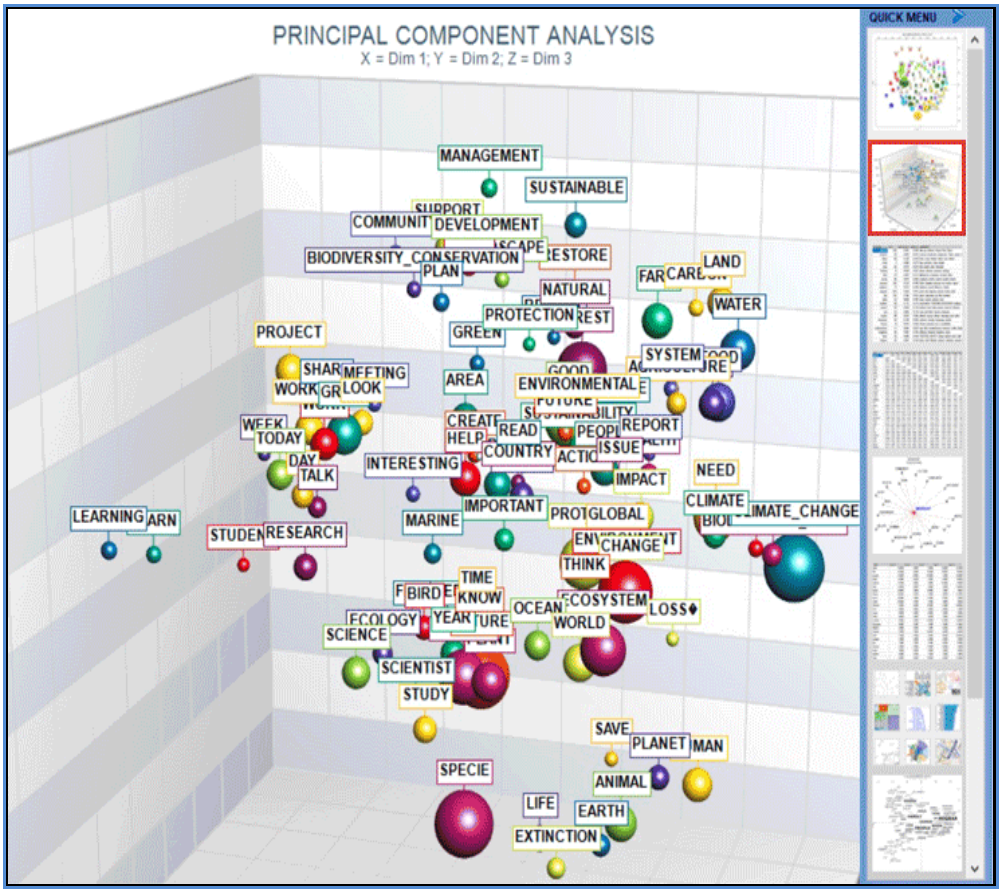




- Co-Word Analysis

L'uso di questa funzione **T-LAB** consente di analizzare le relazioni di **co-occorrenza** all'interno di gruppi di parole chiave.





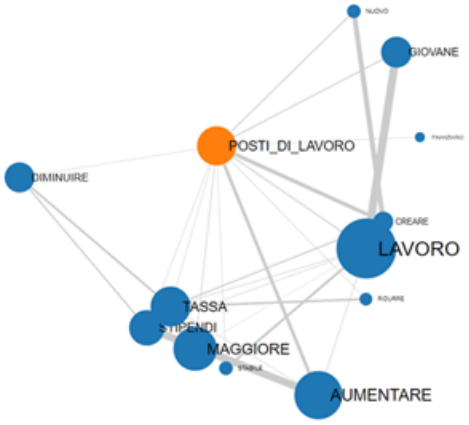
- Analisi delle Sequenze e Network Analysis

Questo strumento **T-LAB** tiene conto delle **posizioni** delle varie unità lessicali all'interno delle frasi e ci permette di rappresentare ed esplorare qualsiasi testo come una **rete** di relazioni.

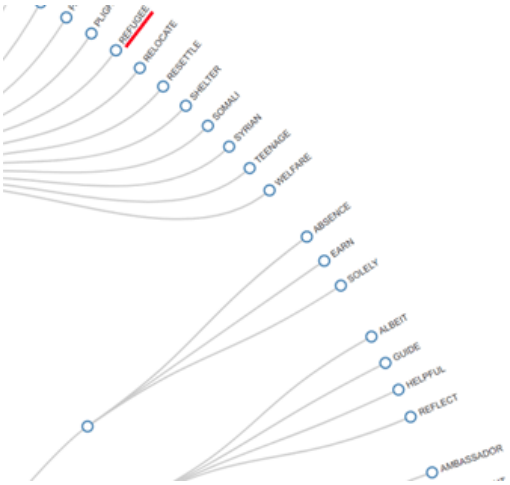
Ciò significa, dopo aver eseguito questo tipo di analisi, l'utilizzatore può verificare le relazioni tra i nodi della rete (cioè le parole chiave) a diversi livelli: a) in relazioni del tipo uno-a-uno; b) all'interno di 'ego network'; c) all'interno delle 'comunità' a cui appartengono; d) all'interno dell'intera rete costituita dal testo in analisi.

RELAZIONI DEL TIPO UNO-AD-UNO

EGO-NETWORK

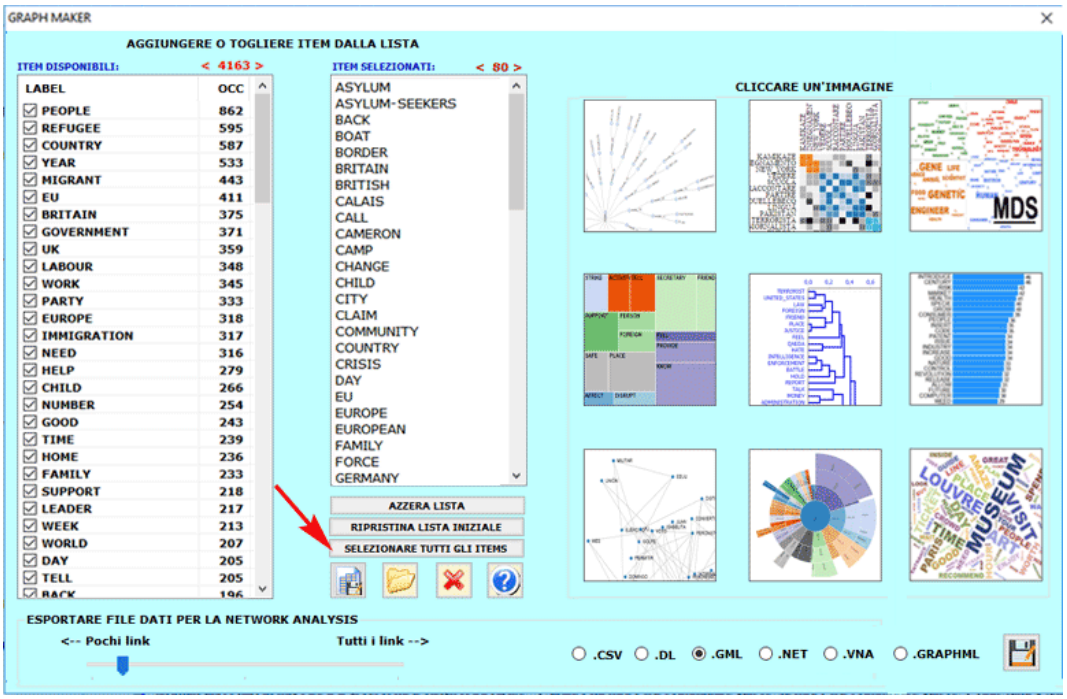


COMUNITA'



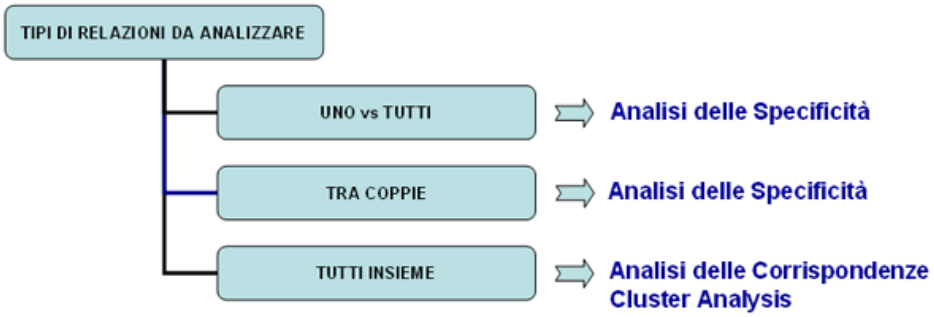
INTERA RETE

Inoltre, facendo clic sull'opzione **GRAPH MAKER**, l'utente può creare diversi tipi di grafici utilizzando elenchi personalizzati di parole chiave (vedi sotto).

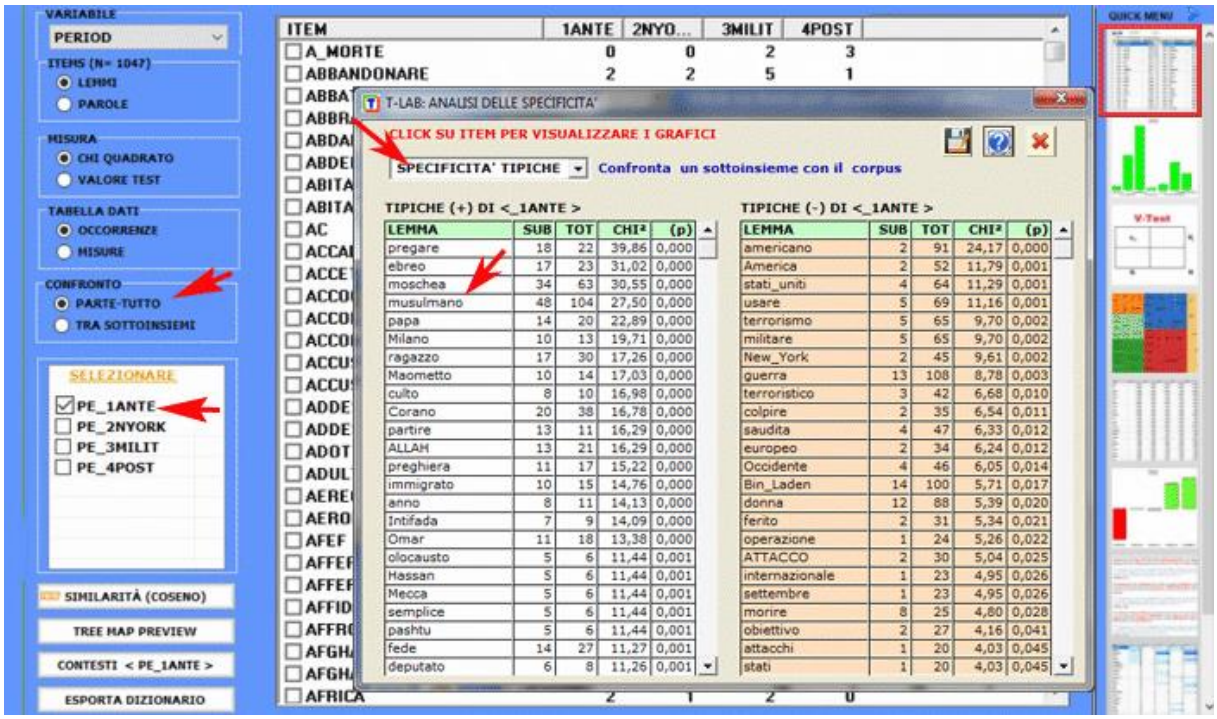


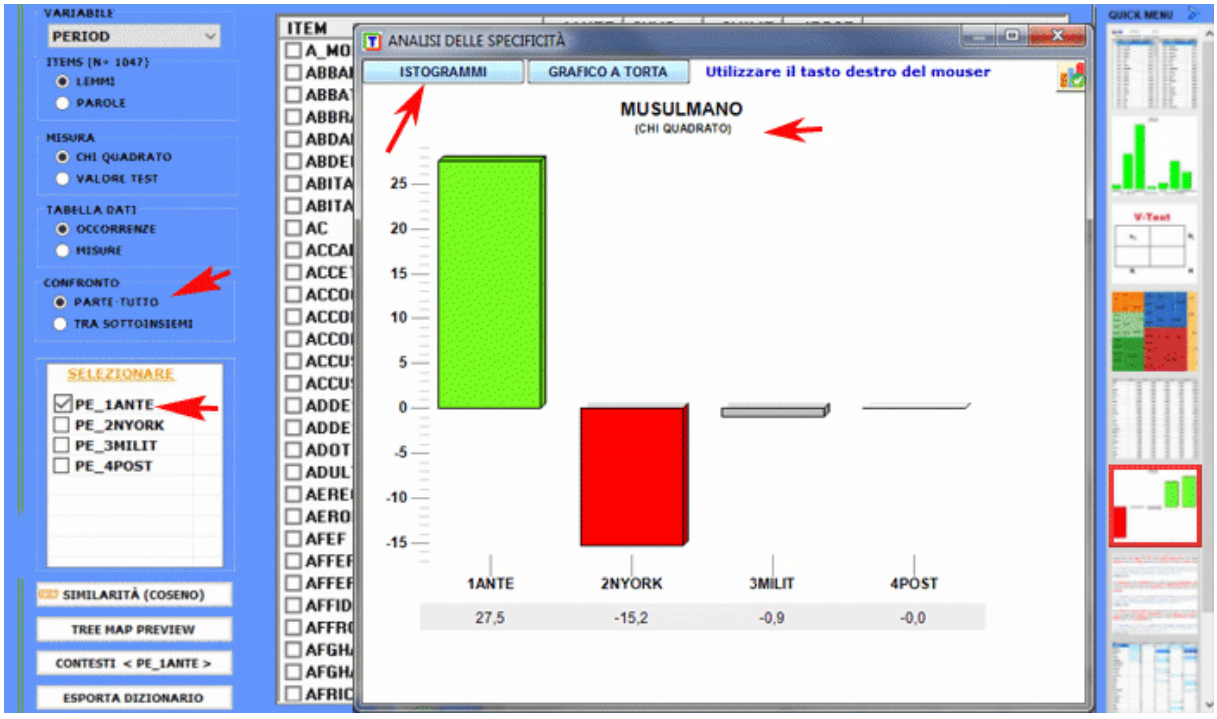
B : STRUMENTI PER ANALISI COMPARATIVE

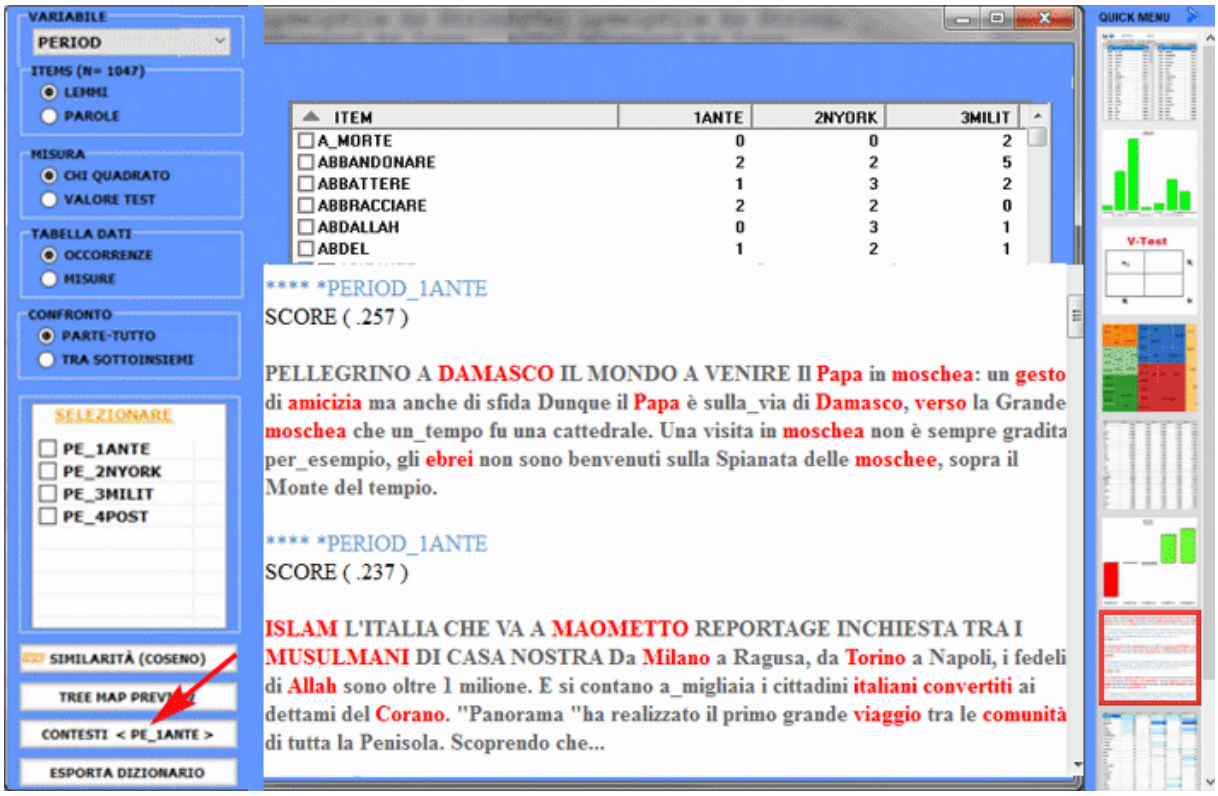
Questi strumenti consentono di analizzare vari tipi di relazioni tra le unità di contesto.



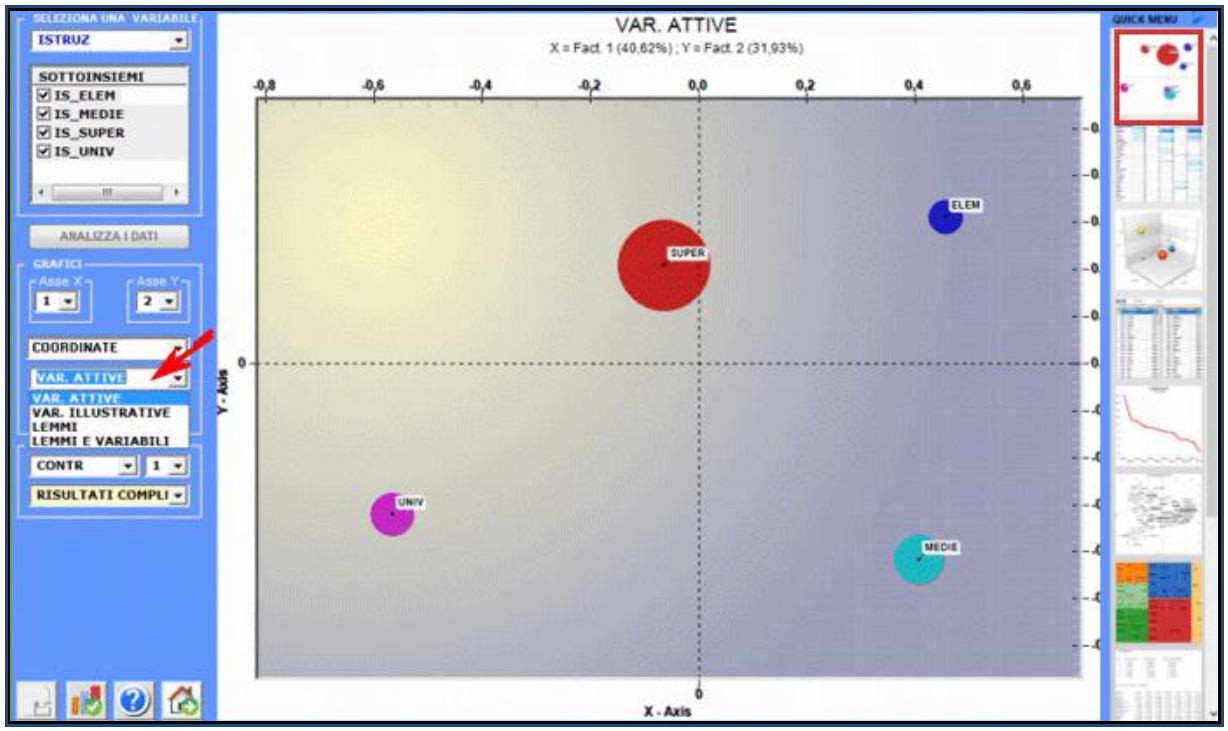
L'**Analisi delle Specificità** consente di verificare quali **parole** sono **tipiche** o **esclusive** di ogni specifico sottoinsieme del corpus. Inoltre permette di estrarre i **contesti tipici**, cioè i contesti elementari caratteristici, di ciascuno dei sottoinsiemi analizzati (ad esempio, le 'tipiche' frasi usate da specifiche leader politici).

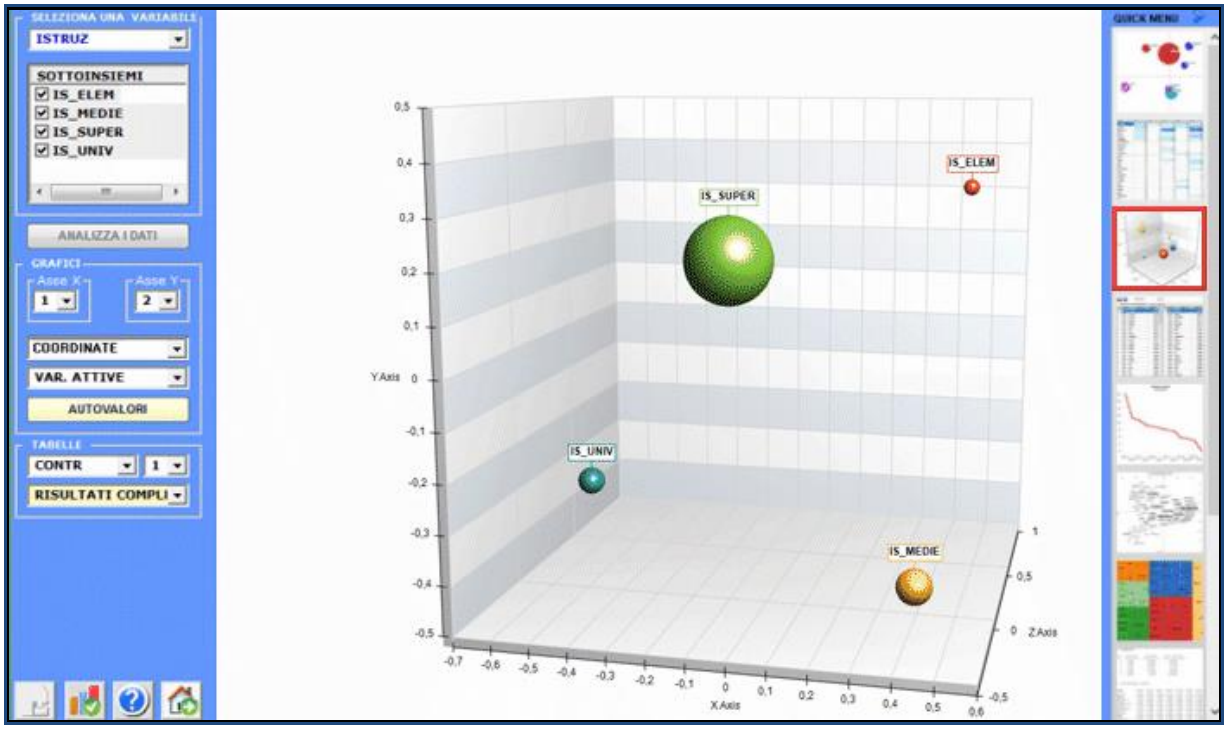

VARIABILE
PERIOD
ITEMS (N= 1047)
LEMMI
PAROLE
MISURA
CHI QUADRATO
VALORE TEST
TABELLA DATI
OCCORRENZE
MISURE
CONFRONTO
PARTE-TUTTO
TRA SOTTOINSIEMI
SELEZIONARE
PE_1ANTE
PE_2NYORK
PE_3MILIT
PE_4POST
SIMILARITÀ (COSENO)
TREE MAP PREVIEW
CONTESTI < PE_1ANTE >
ESPORTA DIZIONARIO
ITEM
1ANTE
2NYO...
3MILIT
4POST
A_MORTE 0 0 2 3
ABBANDONARE 2 2 5 1
T-LAB: ANALISI DELLE SPECIFICITA'
CLICK SU ITEM PER VISUALIZZARE I GRAFICI
SPECIFICITA' TIPICHE
Confronta un sottoinsieme con il corpus
TIPICHE (+) DI <_1ANTE >
LEMMA SUB TOT CHI² (p)
pregare 18 22 39,86 0,000
ebreo 17 23 31,02 0,000
moschea 34 63 30,55 0,000
musulmano 48 104 27,50 0,000
papa 14 20 22,89 0,000
Milano 10 13 19,71 0,000
ragazzo 17 30 17,26 0,000
Maometto 10 14 17,03 0,000
culto 8 10 16,98 0,000
Corano 20 38 16,78 0,000
partire 13 11 16,29 0,000
ALLAH 13 21 16,29 0,000
preghiera 11 17 15,22 0,000
immigrato 10 15 14,76 0,000
anno 8 11 14,13 0,000
Intifada 7 9 14,09 0,000
Omar 11 18 13,38 0,000
olocausto 5 6 11,44 0,001
Hassan 5 6 11,44 0,001
Mecca 5 6 11,44 0,001
semplice 5 6 11,44 0,001
pashtu 5 6 11,44 0,001
fede 14 27 11,27 0,001
deputato 6 8 11,26 0,001
TIPICHE (-) DI <_1ANTE >
LEMMA SUB TOT CHI² (p)
americano 2 91 24,17 0,000
America 2 52 11,79 0,001
stati_uniti 4 64 11,29 0,001
usare 5 69 11,16 0,001
terrorismo 5 65 9,70 0,002
militare 5 65 9,70 0,002
New_York 2 45 9,61 0,002
guerra 13 108 8,78 0,003
terroristico 3 42 6,68 0,010
colpire 2 35 6,54 0,011
saudita 4 47 6,33 0,012
europeo 2 34 6,24 0,012
Occidente 4 46 6,05 0,014
Bin_Laden 14 100 5,71 0,017
donna 12 88 5,39 0,020
ferito 2 31 5,34 0,021
operazione 1 24 5,26 0,022
ATTACCO 2 30 5,04 0,025
internazionale 1 23 4,95 0,026
settembre 1 23 4,95 0,026
morire 8 25 4,80 0,028
obiettivo 2 27 4,16 0,041
attacchi 1 20 4,03 0,045
stati 1 20 4,03 0,045
AFRICA 2 1 2 0
QUICK MENU


VARIABILE
PERIOD
ITEMS (N= 1047)
LEMMI
PAROLE
MISURA
CHI QUADRATO
VALORE TEST
TABELLA DATI
OCCORRENZE
MISURE
CONFRONTO
PARTE-TUTTO
TRA SOTTOINSIEMI
SELEZIONARE
PE_1ANTE
PE_2NYORK
PE_3MILIT
PE_4POST
SIMILARITÀ (COSENO)
TREE MAP PREVIEW
CONTESTI < PE_1ANTE >
ESPORTA DIZIONARIO
ITEM
ANALISI DELLE SPECIFICITÀ
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouser
MUSULMANO
(CHI QUADRATO)
25
20
15
10
5
0
-5
-10
-15
1ANTE
2NYORK
3MILIT
4POST
27,5
-15,2
-0,9
-0,0
QUICK MENU

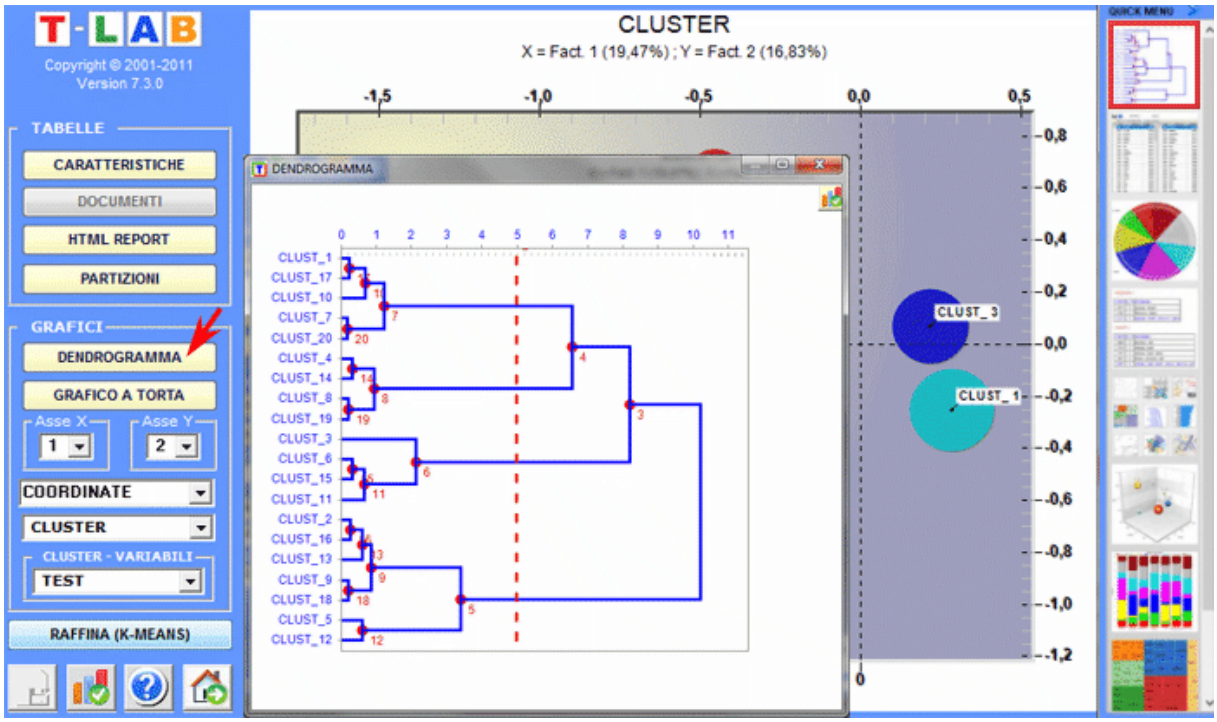


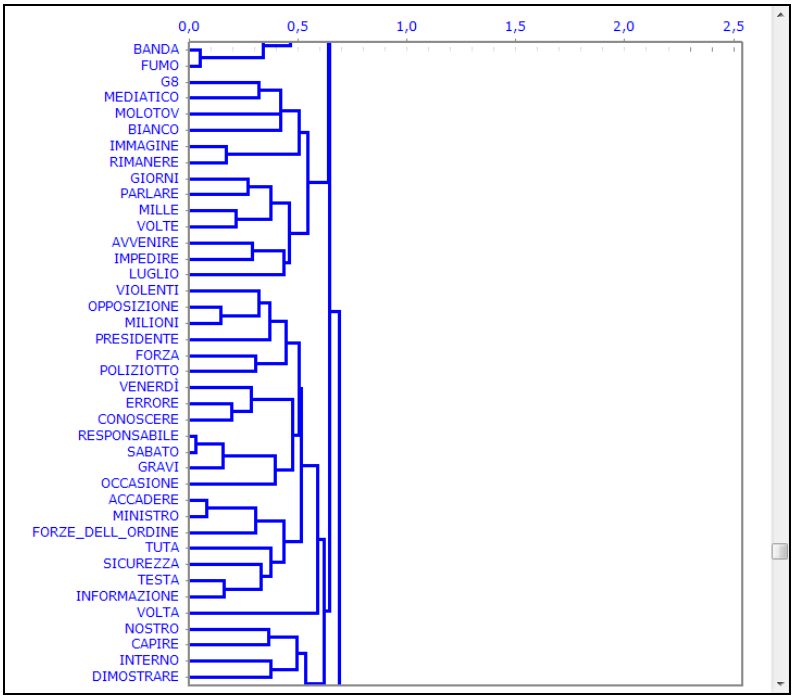
L'**Analisi delle Corrispondenze** consente di esplorare vari tipi di relazioni (somiglianze e differenze) tra gruppi di unità di contesto.





La **Cluster Analysis**, che può essere effettuata con varie tecniche, consente di individuare gruppi di unità di analisi che abbiano due caratteristiche complementari: massima omogeneità al loro interno e massima eterogeneità tra ciascuno di essi e gli altri.





C : STRUMENTI PER ANALISI TEMATICHE

Questi strumenti consentono di individuare, esaminare e mappare i “temi” presenti nei testi analizzati.

Poiché **tema** è una parola polisemica, in questo caso è utile far riferimento ad alcune definizione operative. Infatti, in questi strumenti **T-LAB**, “tema” è una label usata per indicare quattro diverse entità:

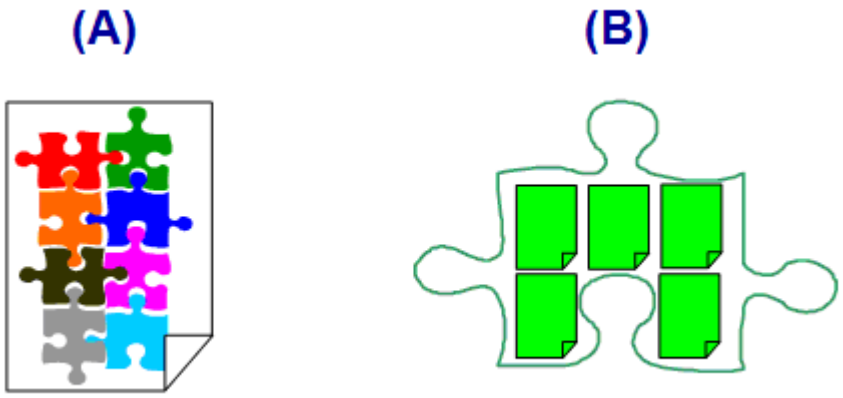
1 - un **cluster tematico di unità di contesto** caratterizzate dagli stessi pattern di parole chiave (vedi gli strumenti Analisi Tematica dei Contesti Elementari, Classificazione Tematica dei Documenti e Classificazione basata su Dizionari);

2 - un **gruppo tematico di parole-chiave** classificate come appartenenti alla stessa categoria (vedi lo strumento Classificazione Basata su Dizionari);

3 - una **componente di un modello probabilistico** che rappresenta ogni unità di contesto (sia essa un contesto elementare o un documento) come generato da una mistura di "temi" o "topics" (vedi gli strumenti Modellizzazione dei Temi Emergenti e Testi e Discorsi come Sistemi Dinamici);

4 - una **specifica parola chiave** usata per estrarre un insieme di contesti elementari in cui essa è associata con uno specifico gruppo di parole preselezionate dall'utilizzatore (vedi lo strumento Contesti Chiave di Parole Tematiche);

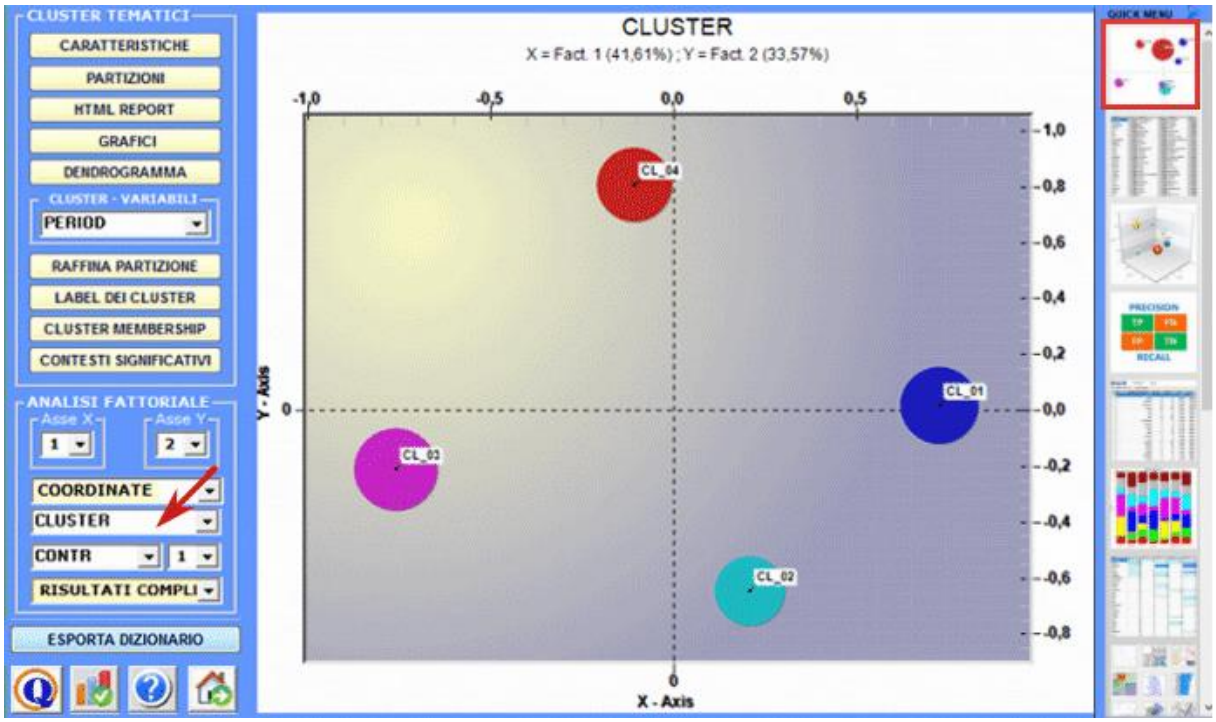
Per esempio, a seconda del tipo di strumento che stiamo usando, uno specifico documento può essere analizzato come composto da vari 'temi' (vedi 'A' sotto) o come appartenente a un insieme di documenti concernenti lo stesso 'tema' (vedi 'B' sotto). Infatti, nel caso 'A' ogni tema può corrispondere ad una parola o a una frase, mentre nel caso 'B' un tema può essere un'etichetta assegnata a un gruppo di documenti caratterizzati da gli stessi pattern di parole chiave.

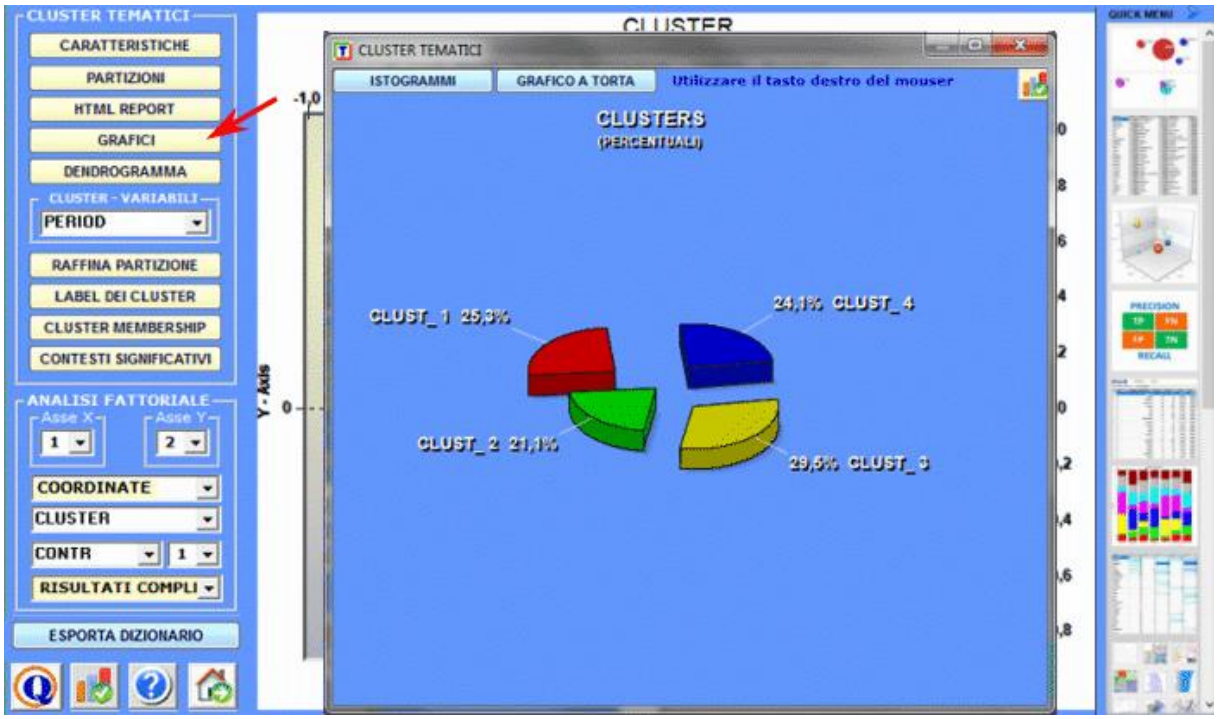


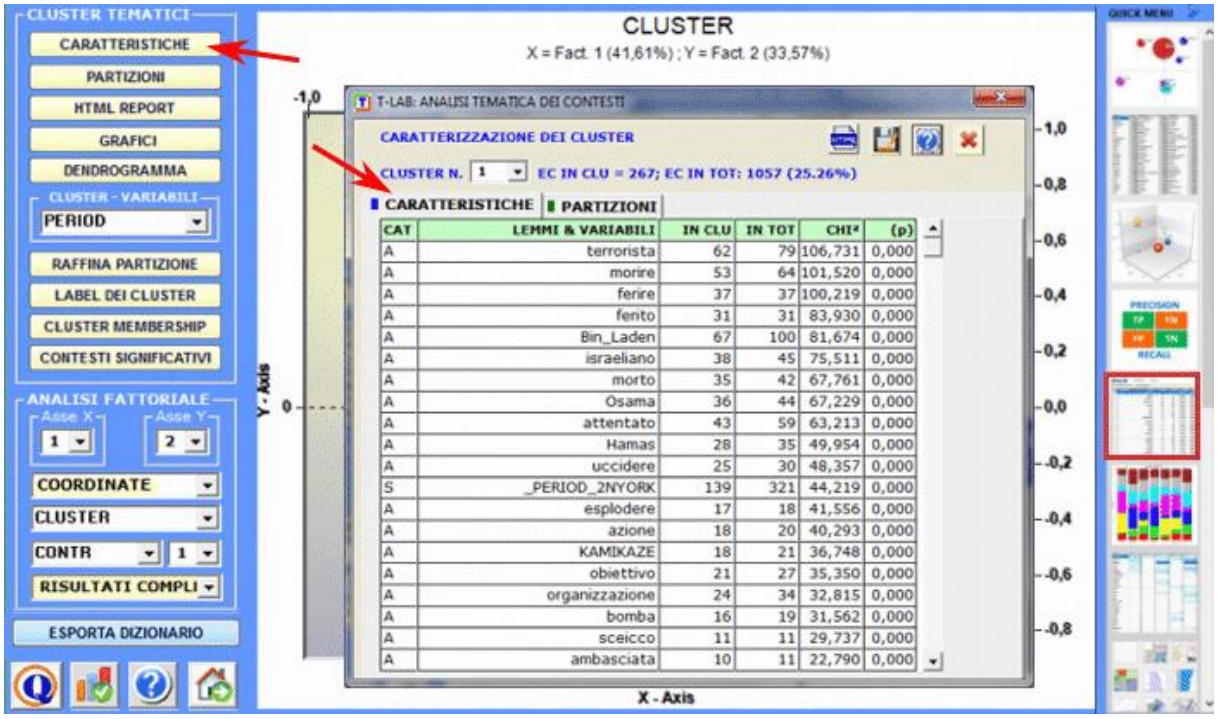
Nel dettaglio, i modi in cui **T-LAB** 'estrae' temi sono i seguenti:

1 - sia l' **Analisi Tematica dei Contesti Elementari** che la **Classificazione Tematica dei Documenti** funzionano nel modo seguente:

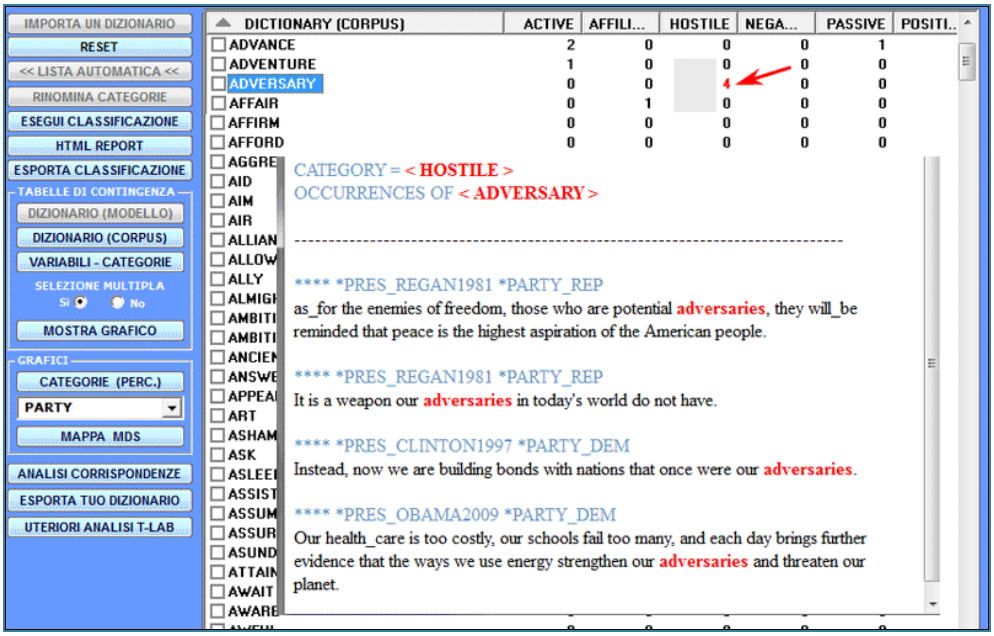
a- realizzano un'**analisi delle co-occorrenze** per individuare cluster tematici di unità di contesto;
b- realizzano un'**analisi comparativa** per confrontare i profili dei vari cluster;
c- producono vari tipi di grafici e tabelle (vedi sotto);
d- consentono di archiviare le **nuove variabili** ottenute (cluster tematici) e di utilizzarle in ulteriori analisi.

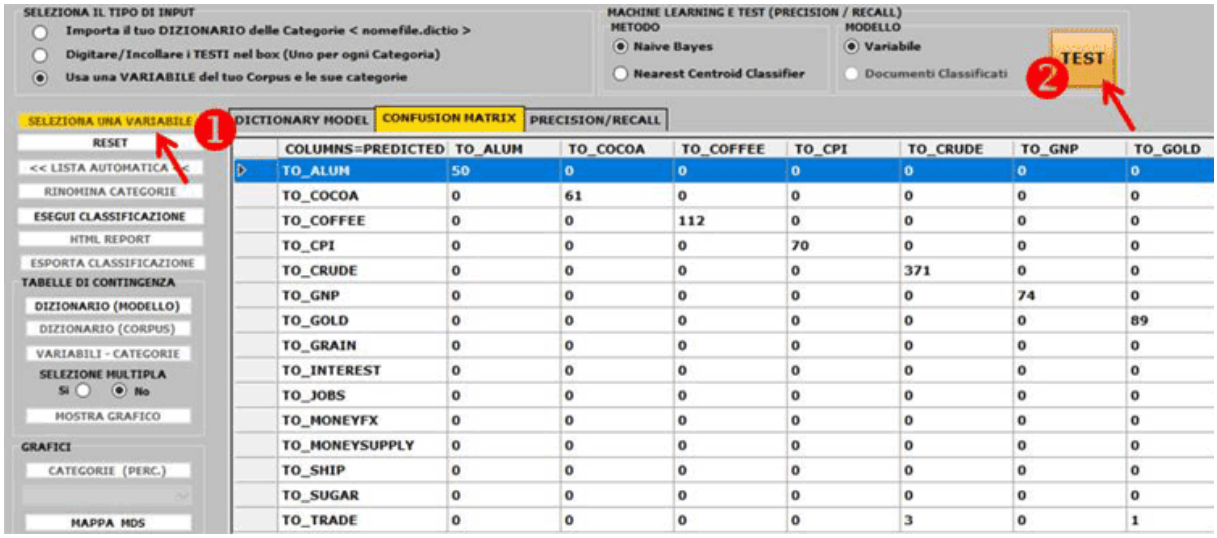




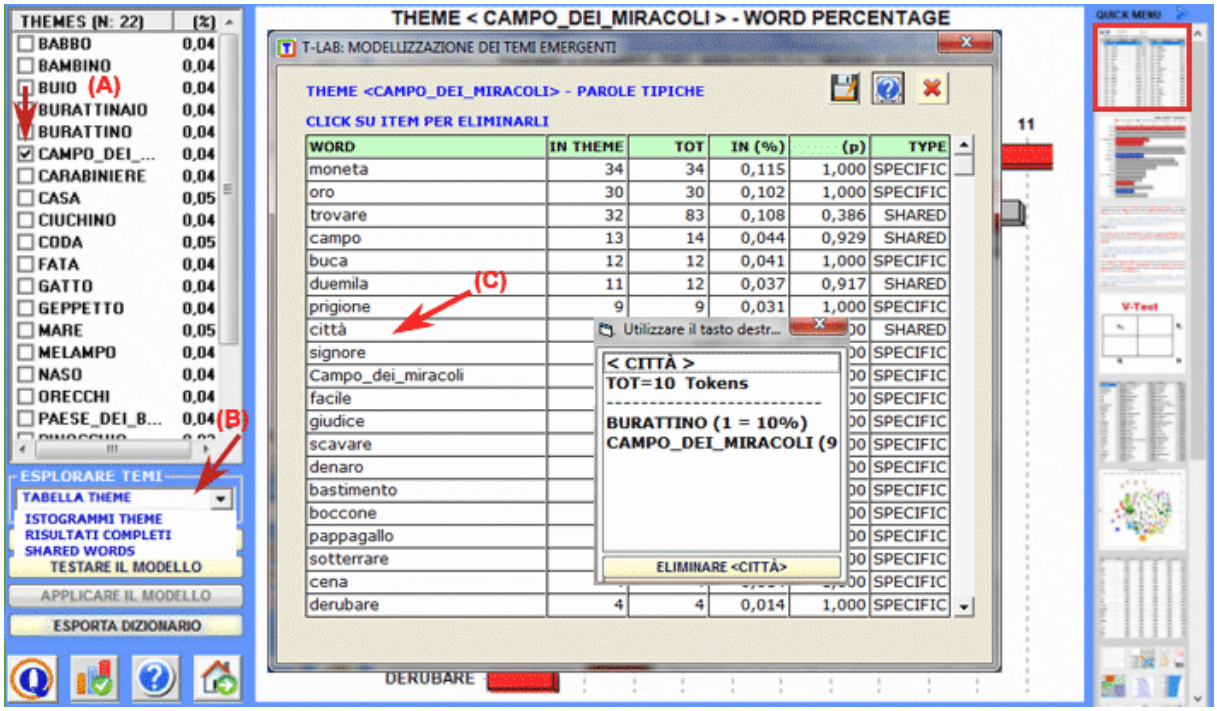


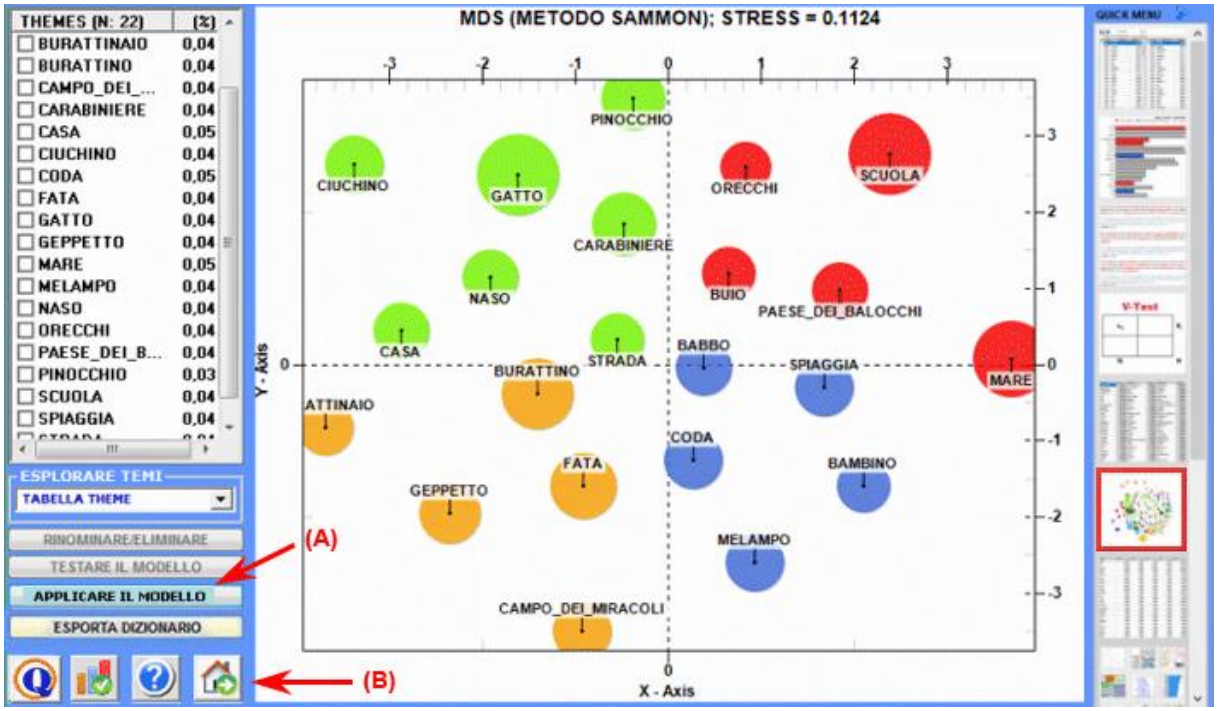
2 - tramite lo strumento **Classificazione Basata su Dizionari** possiamo facilmente costruire / testare / applicare modelli (ad esempio dizionari di categorie) sia per la classica analisi di contenuto che per la sentiment analysis. Infatti questo strumento ci permette di eseguire una classificazione automatica di tipo top-down sia delle unità lessicali (cioè parole e lemmi) che delle unità di contesto (cioè frasi, paragrafi e documenti brevi).


IMPORTA UN DIZIONARIO
RESET
<< LISTA AUTOMATICA <<
RINOMINA CATEGORIE
ESEGUI CLASSIFICAZIONE
HTML REPORT
ESPORTA CLASSIFICAZIONE
TABELLE DI CONTINGENZA
DIZIONARIO (MODELLO)
DIZIONARIO (CORPUS)
VARIABILI - CATEGORIE
SELEZIONE MULTIPLA
Si No
MOSTRA GRAFICO
GRAFICI
CATEGORIE (PERC.)
PARTY
MAPPA MDS
ANALISI CORRISPONDENZE
ESPORTA TUO DIZIONARIO
UTERIORI ANALISI T-LAB
DICTIONARY (CORPUS)
ACTIVE
AFFILI...
HOSTILE
NEGA...
PASSIVE
POSITI..
ADVANCE 2 0 0 0 1
ADVENTURE 1 0 0 0 0
ADVERSARY 0 0 4 0 0
AFFAIR 0 1 0 0 0
AFFIRM 0 0 0 0 0
AFFORD 0 0 0 0 0
CATEGORY = < HOSTILE >
OCCURRENCES OF < ADVERSARY >
**** *PRES_REGAN1981 *PARTY_REP
as_for the enemies of freedom, those who are potential adversaries, they will_be reminded that peace is the highest aspiration of the American people.
**** *PRES_REGAN1981 *PARTY_REP
It is a weapon our adversaries in today's world do not have.
**** *PRES_CLINTON1997 *PARTY_DEM
Instead, now we are building bonds with nations that once were our adversaries.
**** *PRES_OBAMA2009 *PARTY_DEM
Our health_care is too costly, our schools fail too many, and each day brings further evidence that the ways we use energy strengthen our adversaries and threaten our planet.

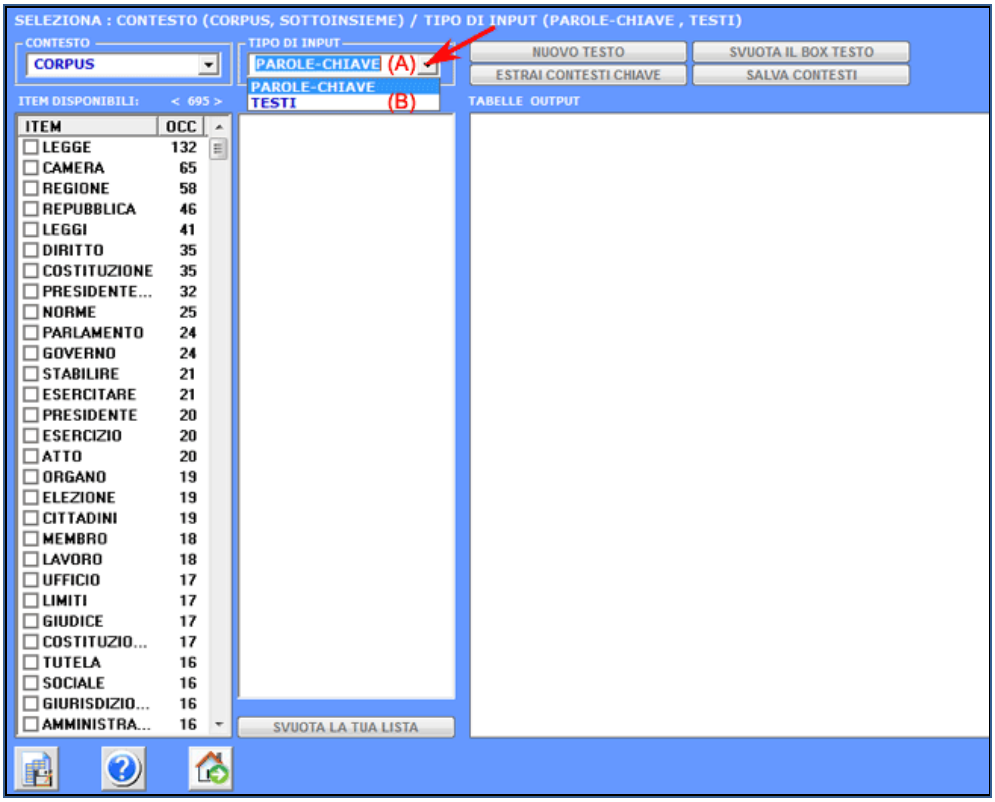

SELEZIONA IL TIPO DI INPUT
Importa il tuo DIZIONARIO delle Categorie < nomefile.dictio >
Digitare/Incollare i TESTI nel box (Uno per ogni Categoria)
Usa una VARIABILE del tuo Corpus e le sue categorie
MACHINE LEARNING E TEST (PRECISION / RECALL)
METODO
Naive Bayes
Nearest Centroid Classifier
MODELLO
Variabile
Documenti Classificati
TEST
SELEZIONA UNA VARIABILE
DICTIONARY MODEL
CONFUSION MATRIX
PRECISION/RECALL
RESET
<< LISTA AUTOMATICA <<
RINOMINA CATEGORIE
ESEGUI CLASSIFICAZIONE
HTML REPORT
ESPORTA CLASSIFICAZIONE
TABELLE DI CONTINGENZA
DIZIONARIO (MODELLO)
DIZIONARIO (CORPUS)
VARIABILI - CATEGORIE
SELEZIONE MULTIPLA
Si No
MOSTRA GRAFICO
GRAFICI
CATEGORIE (PERC.)
MAPPA MDS
COLUMNS=PREDICTED TO_ALUM TO_COCOA TO_COFFEE TO_CPI TO_CRUDE TO_GNP TO_GOLD
TO_ALUM 50 0 0 0 0 0 0
TO_COCOA 0 61 0 0 0 0 0
TO_COFFEE 0 0 112 0 0 0 0
TO_CPI 0 0 0 70 0 0 0
TO_CRUDE 0 0 0 0 371 0 0
TO_GNP 0 0 0 0 0 74 0
TO_GOLD 0 0 0 0 0 0 89
TO_GRAIN 0 0 0 0 0 0 0
TO_INTEREST 0 0 0 0 0 0 0
TO_JOBS 0 0 0 0 0 0 0
TO_MONEYFX 0 0 0 0 0 0 0
TO_MONEYSUPPLY 0 0 0 0 0 0 0
TO_SHIP 0 0 0 0 0 0 0
TO_SUGAR 0 0 0 0 0 0 0
TO_TRADE 0 0 0 0 3 0 1

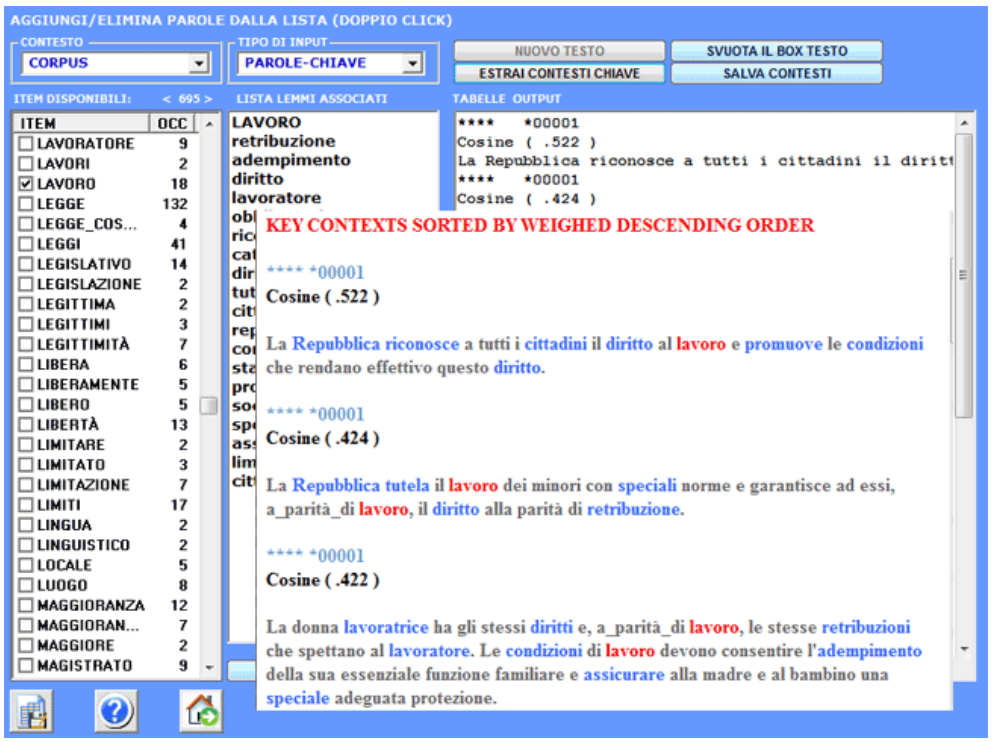
3 - tramite lo strumento **Modellizzazione dei Temi Emergenti** (vedi sotto) i componenti della 'mistura' tematica possono essere descritti attraverso il loro vocabolario caratteristico e possono essere utilizzati per la costruzione di griglie per l'analisi qualitativa e / o per la classificazione automatica delle unità di contesto (cioè contesti elementari o documenti).





4 - lo strumento **Contesti Chiave di Parole Tematiche** (vedi sotto) può essere utilizzato per due diversi scopi: (a) estrarre elenchi di unità di contesto (cioè contesti elementari) che permettono di approfondire il valore tematico di specifiche **parole chiave**; (b) estrarre gruppi di unità di contesto che risultano simili a una qualche **testo** 'esempio' scelto dall'utilizzatore.

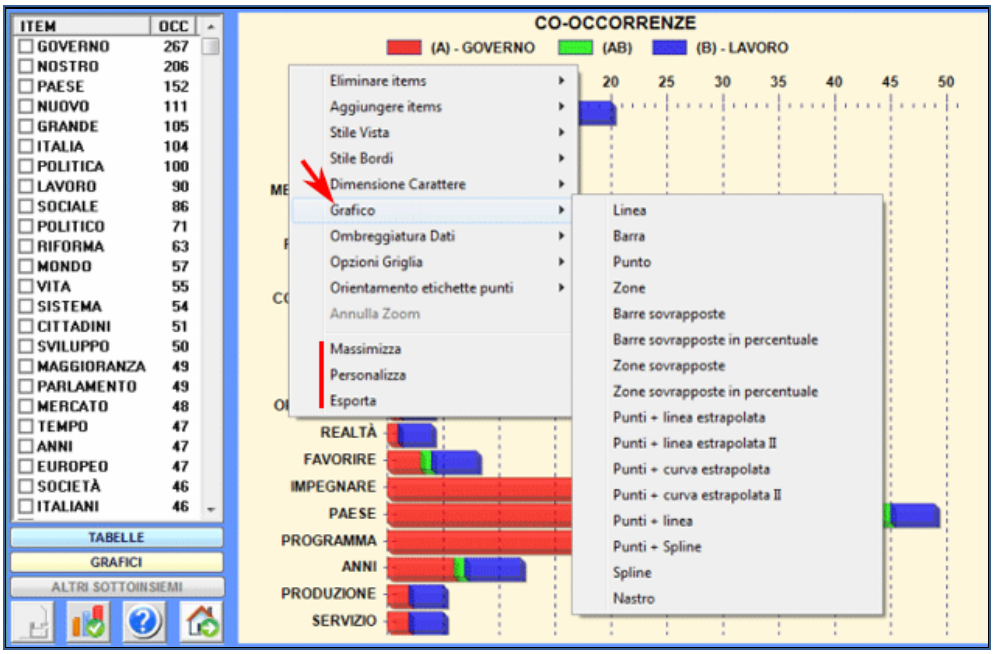


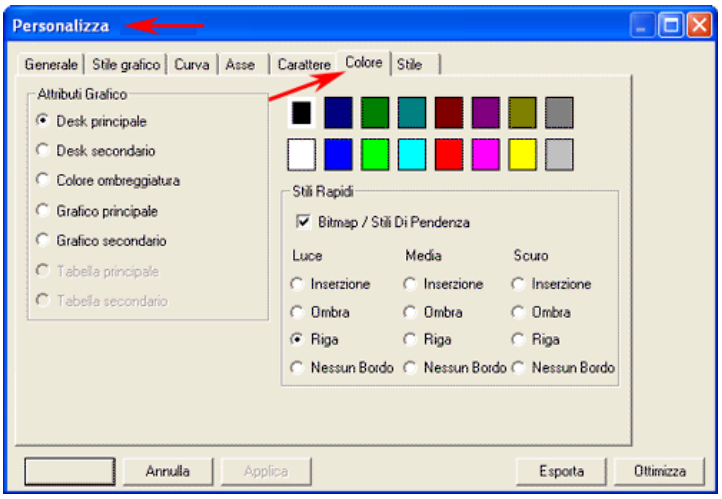


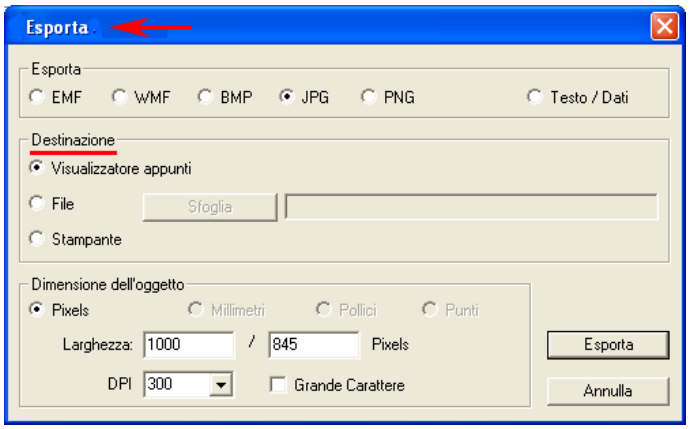
6 - L' INTERPRETAZIONE DEGLI OUTPUT consiste nella consultazione delle tabelle e dei grafici prodotti da **T-LAB**, nell'eventuale personalizzazione del loro formato e nel fare inferenze sul significato delle relazioni in essi rappresentate.

Nel caso delle **tabelle**, a seconda dei casi, **T-LAB** consente di esportarle in file con le seguenti estensioni: **.DAT**, **.TXT**, **.CSV**, **.XLXS**, **.HTML**. Ciò significa che, servendosi di qualunque editore di testi e/o di un qualche applicativo della suite Microsoft Office, l'utilizzatore può facilmente importarli e rielaborarli.

Nel caso dei **grafici**, appositi sub-menu attivati con il tasto destro del mouse consentono vari tipi di operazioni: zoom (clic con il tasto sinistro e selezionare un rettangolo), massimizzazione, personalizzazione ed esportazione degli output in diversi formati (vedi sotto, uso del tasto destro).







Alcuni criteri generali per l'interpretazione degli output **T-LAB** sono illustrati in un paper citato in Bibliografia e disponibile nel sito https://www.tlab.it (Lancia F.: 2007). In questo viene proposta l'ipotesi che gli output delle elaborazioni statistiche (tabelle e grafici) sono un tipo particolare di testi, cioè degli oggetti multi-semiotici caratterizzati dal fatto che le relazioni tra segni e simboli sono ordinate da misure che rinviano a specifici **codici**.

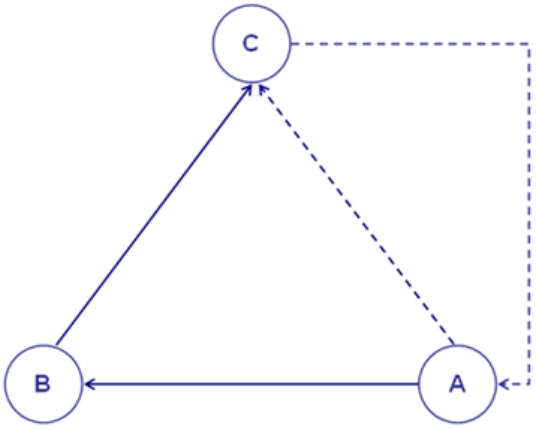
In altri termini, sia nel caso dei testi scritti in linguaggio naturale che in quelli scritti nel linguaggio della statistica, la possibilità di fare inferenze sulle relazioni che organizzano le **forme del contenuto** è fondata sul fatto che le relazioni tra le **forme dell'espressione** non sono casuali (random); infatti, nel primo caso (linguaggio naturale) le unità significanti si susseguono ordinate in modo lineare (una dopo l'altra nella catena del discorso), mentre nel secondo caso (tabelle e grafici) i principi di ordinamento sono costituiti dalle misure che determinano l'organizzazione degli **spazi semantici** multidimensionali.

Anche se gli spazi semantici rappresentati nelle mappe **T-LAB** sono molto vari, e ciascuno di essi richiede specifiche procedure interpretative, possiamo fare l'ipotesi che - in generale - la logica del processo inferenziale è la seguente:

A - rilevare una qualche relazione significativa tra le unità "presenti" sul piano dell'espressione (ad es. tra "dati" di tabelle e/o tra "label" di grafici);

B - esplorare e confrontare i tratti semantici delle stesse unità e i contesti a cui esse sono mentalmente e culturalmente associate (piano del contenuto);

C - costruire qualche ipotesi o qualche categoria di analisi che, nel contesto definito dal corpus, renda ragione delle relazioni tra forme dell'espressione e forme del contenuto.



Infine qualche informazione sui **vincoli attuali** delle opzioni **T-LAB**:

- dimensioni del corpus: max 90 Mb, pari a circa 55.000 pagine in formato .txt;
- documenti primari: max 30.000 (N.B.: quando nessuno dei testi supera i 2.000 caratteri, il limite è esteso a 99.999);
- variabili categoriali: max 50, ciascuna delle quali con max 150 modalità;
- modellizzazione dei temi emergenti: max 5.000 unità lessicali (*) per max 5.000.000 occorrenze;
- analisi tematica dei contesti elementari: max 300.000 righe (unità di contesto) x 5.000 colonne (unità lessicali);
- classificazione tematica dei documenti: max 99.999 righe (documenti) x 5.000 colonne (unità lessicali);
- analisi delle specificità (unità lessicali x categorie di una variabile): max 10.000 righe per 150 colonne;
- analisi delle corrispondenze (unità lessicali x categorie di una variabile): max 10.000 righe per 150 colonne;
- analisi delle corrispondenze (unità di contesto x unità lessicali): max 10.000 righe per 5.000 colonne;
- analisi delle corrispondenze multiple (contesti elementari x categorie di due più variabili): max 150.000 righe per 250 colonne;
- scomposizione in valori singolari (SVD): max 300.000 righe x 5.000 colonne;
- cluster analysis che utilizza i risultati di una precedente analisi delle corrispondenze (o SVD): max 10.000 righe (unità lessicali o contesti elementari);
- associazioni di parole, confronti tra coppie e co-word analysis: liste di max 5.000 unità lessicali;
- analisi delle sequenze: max 5.000 unità lessicali (o categorie) con testi di max 3.000.000 occorrenze.

(*) In **T-LAB**, 'unità lessicali' sono parole, multi-words, lemmi e categorie semantiche. Quindi, quando viene applicata la lemmatizzazione automatica, 5.000 unità lessicali corrispondono a circa 12.000 parole.

IMPOSTAZIONI DI ANALISI

Impostazioni automatiche e personalizzate

La scelta delle impostazioni **automatiche** (A) o **personalizzate** (B) riguarda la lista delle **Parole Chiave** utilizzate in tutte analisi effettuate con **T-LAB**. Tale scelta è reversibile fino a quando l'utilizzatore non effettua interventi che modificano il **Dizionario** del corpus.

A) IMPOSTAZIONI AUTOMATICHE

La scelta delle **impostazioni automatiche** comporta che la lista delle parole chiave includa **fino a un massimo di 5000 unità lessicali** selezionate automaticamente da **T-LAB** e appartenenti alle categorie grammaticali che sono più dense di significato: nomi, verbi, aggettivi e avverbi.

Il criterio di selezione utilizzato da T-LAB varia in funzione del tipo di corpus in analisi.

Se il corpus è costituito da un unico testo, le unità lessicali selezionate sono semplicemente quelle con i più elevati valori di **occorrenza**.

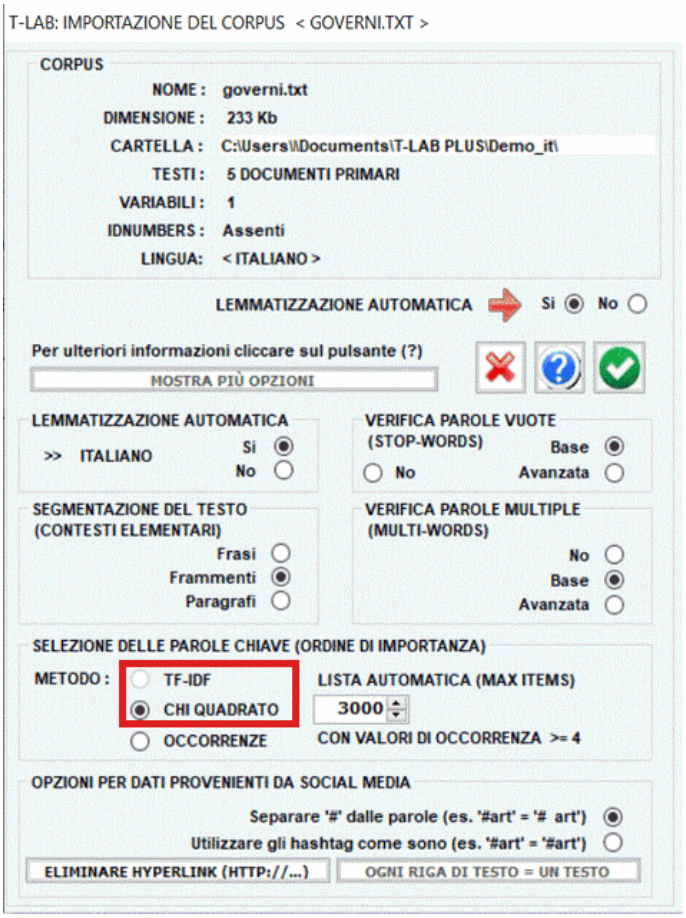
Se il corpus è costituito da due o più testi **T-LAB** applica il seguente algoritmo:

a) *seleziona le parole con valori di occorrenza superiori alla soglia minima;*

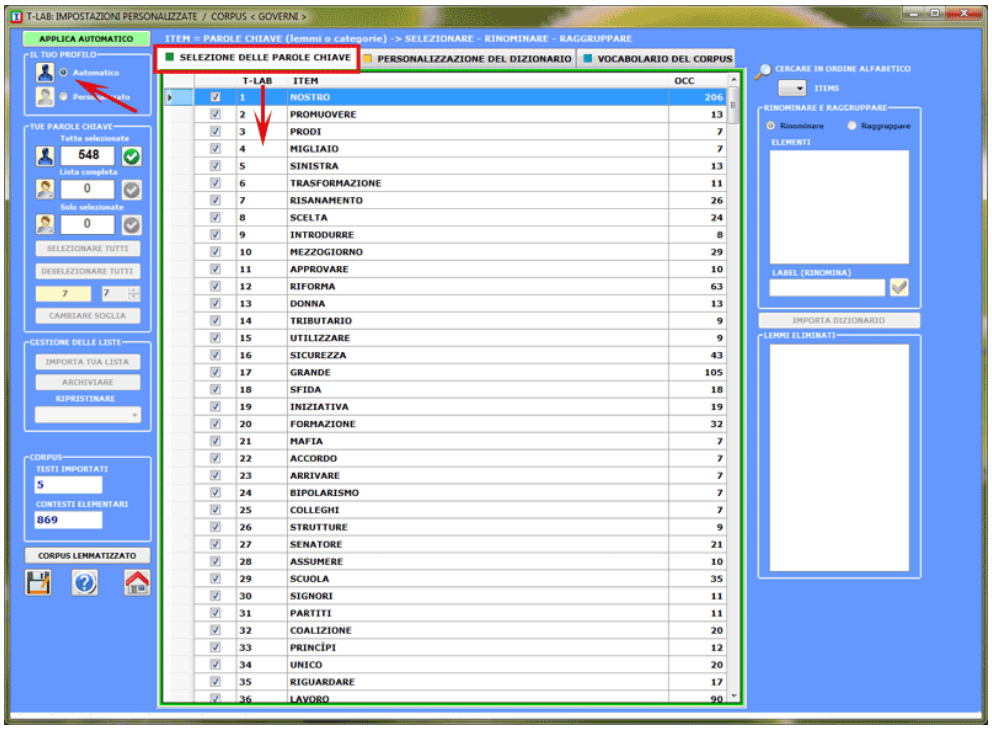
b) *calcola i valori del TF-IDF o applica il test del CHI quadro a tutti gli incroci di ogni parola selezionata per tutti i testi in analisi (N.B.: Nel caso del CHI quadro, il numero massimo dei testi è 500);*

c) *seleziona le parole con i valori maggiori nel metodo utilizzato (TF-IDF o CHI quadro), ovvero seleziona quelle parole che, nel corpus, fanno la differenza .*

N.B.:
- Nel caso il corpus sia costituito da due o più testi, l'utilizzatore può scegliere il criterio di selezione (CHI quadro o TF-IDF) nella fase di importazione (vedi immagine seguente).



- Quando è attiva l'opzione di impostazioni automatiche la tabella con la lista delle **Parole Chiave** include una colonna 'T-LAB' che indica la rilevanza di ogni item secondo il criterio selezionato (vedi immagine seguente).



B) IMPOSTAZIONI PERSONALIZZATE

La scelta delle **impostazioni personalizzate** consente all'utilizzatore di selezionare, ridenominare e raggruppare le unità lessicali (parole, lemmi o categorie) da includere nelle successive analisi **T-LAB**.

Nella tabella di riferimento è riportata la lista (lista 1) delle unità lessicali con valori di occorrenza uguali o superiori alla **soglia** prefissata. Alcune di queste, quelle indicate con "☑", fanno parte di una sub-lista (lista 2) creata da **T-LAB** (vedi impostazioni automatiche).

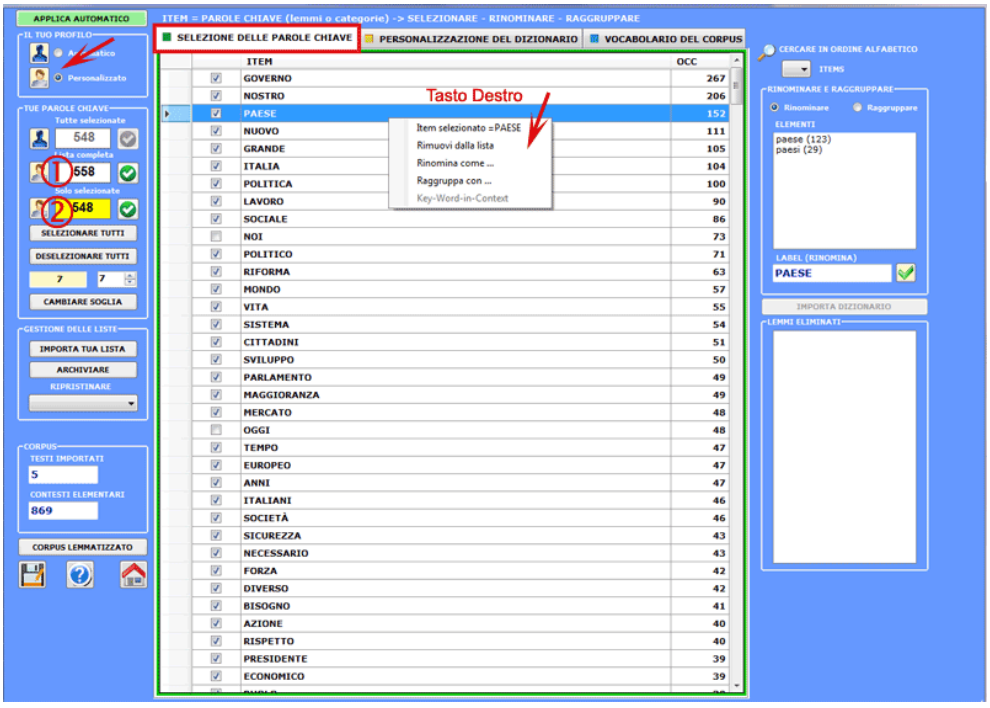
In funzione delle analisi che intende effettuare, l'utilizzatore può decidere se usare/modificare la lista (1) o la lista (2).

In entrambi i casi le operazioni possibili sono le seguenti:

- **modificare** il valore di soglia;
- **selezionare** i lemmi da mettere fuori analisi;
- **ripristinare** l'uso di uno o più lemmi;
- **selezionare/deselezionare** gli item da utilizzare.

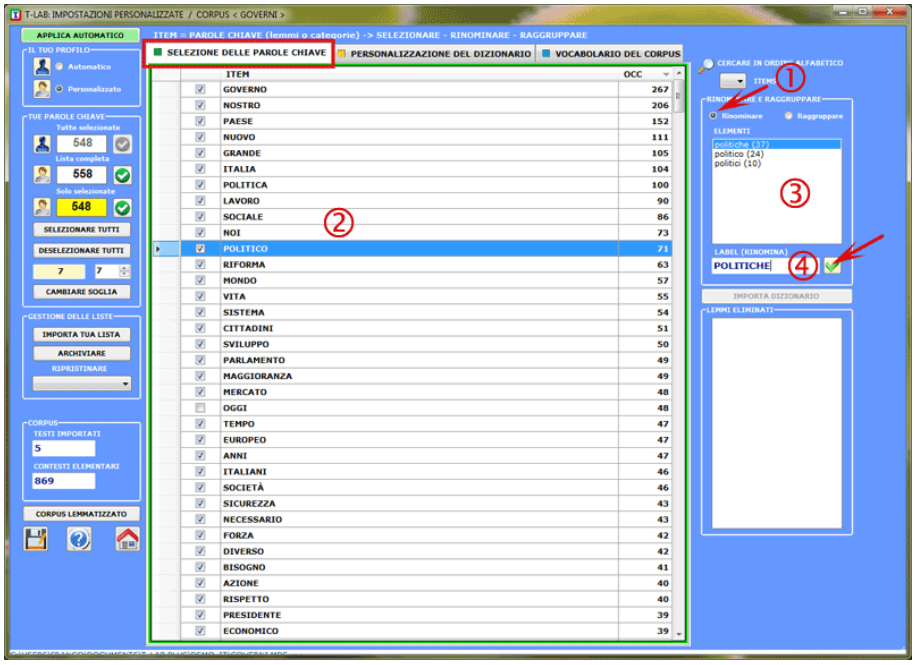
Un click sul pulsante "(1)" o sul pulsante "(2)" abilita l'opzione "personalizzata" delle impostazioni di analisi (vedi immagine seguente).

Le opzioni che riguardano gli **interventi sui singoli** lemmi sono accessibili tramite il tasto destro del mouse selezionando un qualunque item della tabella.



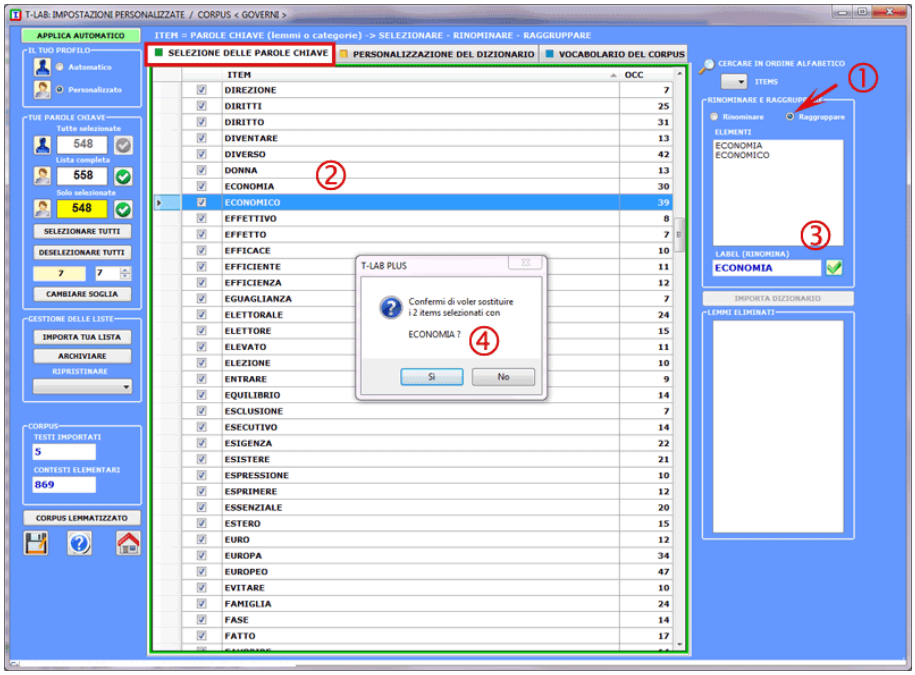
La **ridenominazione** di un singolo lemma va effettuata nel modo seguente:

1- accertarsi che sia attiva l'opzione "RINOMINARE";
2- cliccare su un item della lista;
3- scegliere una delle parole o digitare una label a propria scelta;
4- cliccare su "RINOMINA".

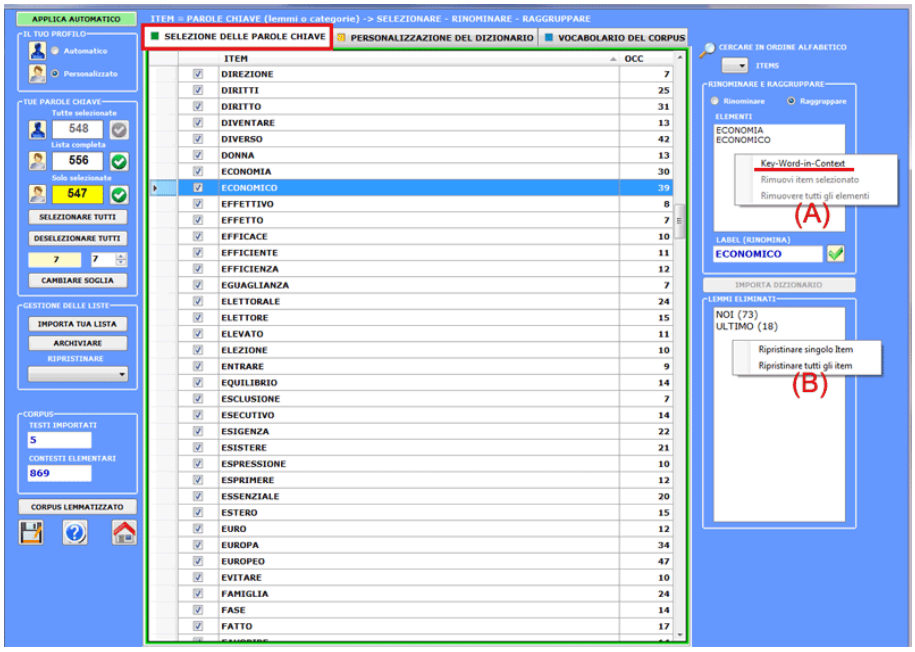


I **raggruppamenti** di due o più lemmi vanno effettuate nel modo seguente:

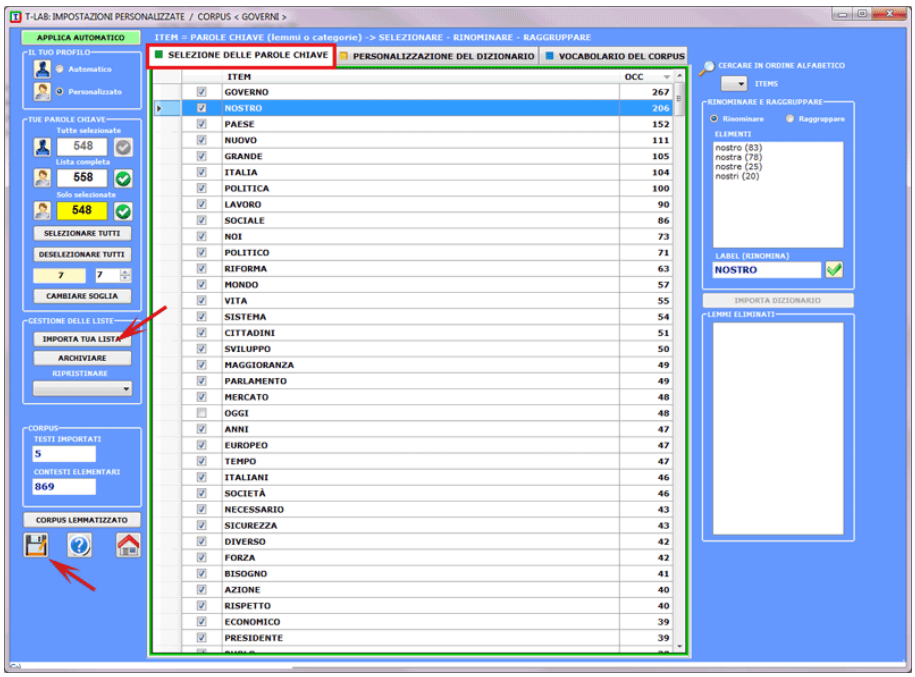
1- selezionare la modalità "RAGGRUPPARE";
2- cliccare su due o più item della lista;
3- scegliere uno dei lemmi o digitare una label a propria scelta;
4- cliccare su "RINOMINA".



Ulteriori opzioni sono attivabili usando il **tasto destro** nel box con gli item da rinominare/raggruppare (A) o nel box con i 'lemmi eliminati' (B).
In particolare, quando – nel caso (A) – viene selezionata l'opzione 'Key-Word-in-Context', è possibile accedere automaticamente allo strumento **Concordanze** e verificare i contesti di occorrenza dei vari item (vedi immagine seguente).



Uno specifico pulsante (vedi sotto) consente di **importare liste personalizzate di parole chiave**.
Ogni lista da importare (file MyList.diz) può includere fino a 10.000 records (min 20).
Ogni record (o riga) della lista deve essere costituito da una parola (Max 50 caratteri) e non deve contenere né spazi vuoti né segni di punteggiatura.
Un modello di file MyList.diz viene creato automaticamente da **T-LAB** ogni volta che viene salvata una lista delle parole chiave (si veda l'apposito pulsante in basso a sinistra).



Le **impostazioni di analisi**, ovvero sia la selezione dei lemmi (o categorie) che il **dizionario** in uso, possono essere salvate e ripristinate usando gli appositi pulsanti (Archiviare o Ripristinare). Ciò significa che lo stesso corpus - senza bisogno di ulteriori importazioni - può essere analizzato con vari dizionari e con diverse selezioni delle parole chiave. Ciò fino a un massimo di 10 diverse impostazioni.

In **T-LAB**, inoltre, è previsto che le impostazioni possono essere riviste in più sessioni, anche con vari utilizzi della funzione **Personalizzazione del Dizionario**.

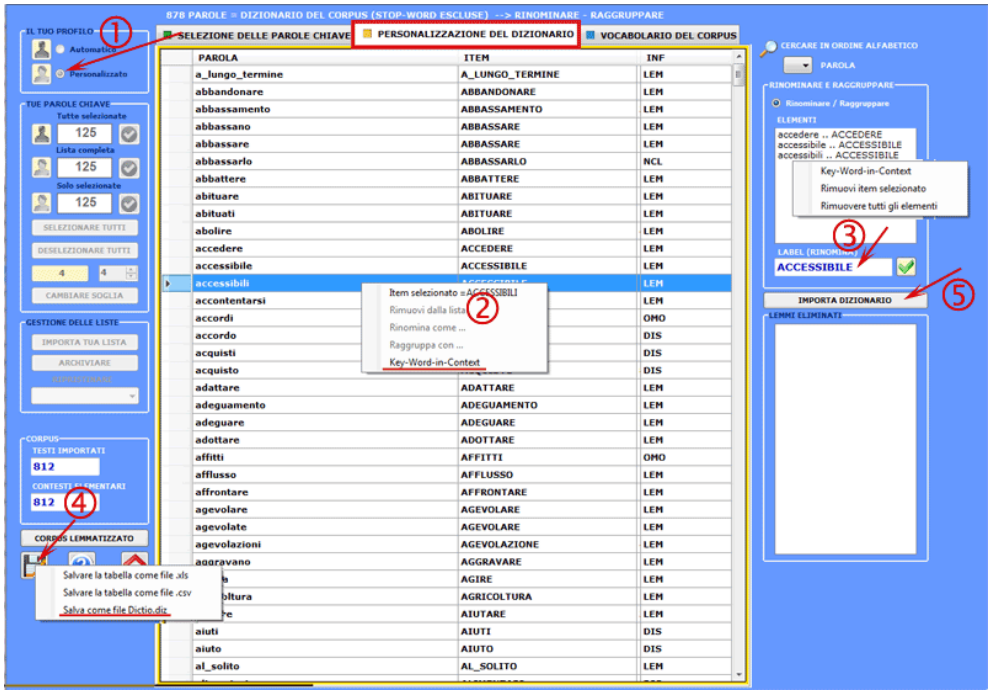
Personalizzazione del Dizionario

L'opzione **Personalizzazione del Dizionario** apre una finestra per eventuali interventi sul dizionario del corpus.

L'utilizzatore può ridenominare o raggruppare i **lemmi** disponibili (vedi sotto opzione '3'); inoltre può esportare il dizionario costruito (vedi sotto opzione '4') o importare un **dizionario personalizzato** (vedi sotto opzione '5').

Il punto di partenza è costituito da una tabella (il **Dizionario del Corpus**) che riporta le seguenti informazioni.

- corrispondenze forma/lemma;
- occorrenze di ogni forma nel corpus;
- alcune etichette che si riferiscono alla **lemmatizzazione automatica** (colonna "INF").



Prima di ogni intervento, selezionando una specifica forma e usando il tasto destro del mouse, è possibile verificare le **concordanze** (Key-Word-in-Context) che interessano (vedi sopra opzione'2').
In ogni caso, prima di ogni intervento, dopo aver cliccato il tab "selezione delle parole chiave", devono essere attivate le impostazioni personalizzate (vedi sopra opzione '1').

Gli **interventi possibili**, pur se diversi nelle loro intenzionalità (revisione delle lemmatizzazioni e/o applicazioni di griglie per l'analisi del contenuto), si traducono tutti in una riorganizzazione del database **T-LAB** e quindi in diverse tabelle per l'analisi dei dati. Ne deriva che tutti gli interventi vanno effettuati sulle forme (lemmi o categorie) ritenute interessanti ai fini delle analisi successive. **T-LAB**, infatti, rende disponibile un'ulteriore funzione - **Impostazioni Personalizzate** (vedi 'Selezione delle Parole Chiave') - attraverso la quale gli utilizzatori possono decidere quali lemmi "tenere" e quali "mettere fuori".

Le due funzioni (Personalizzazione del Dizionario e Impostazioni Personalizzate) sono molto integrate tra loro e l'utente può agevolmente muoversi dall'una all'altra, anche per cambiare le proprie scelte.
In **Personalizzazione del Dizionario**, per cambiare le label (o 'lemmi') attribuite alle parole, sono previste **due modalità di intervento**:

- una che consente di spostare le selezioni sul box a destra (click su un qualunque item) e, successivamente, di ridenominarle attraverso l'uso dell'opzione "rinomina" (In questo caso, la nuova label può essere definita utilizzando uno dei lemmi selezionati o digitando nel box "label"; vedi sopra opzione '3').

- una che prevede l'importazione di un dizionario personalizzato, riservata agli utilizzatori esperti che dispongono di loro liste per classificare le parole presenti in uno o più corpus (vedi sopra opzione '5').

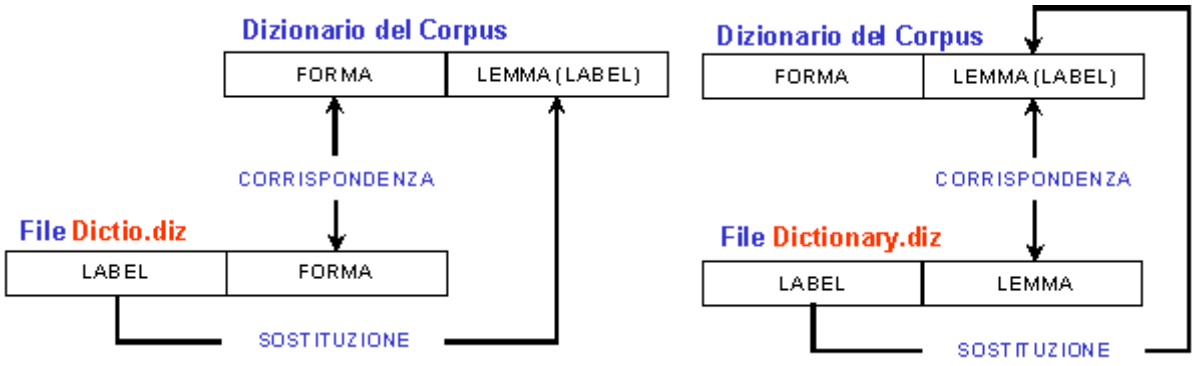
N.B.: L'uso del **tasto destro** nel box Rinominare/Raggruppare abilita un menu contestuale che consente tre operazioni: a) verificare le concordanze (KeyWord-in-Context) dell'item selezionato; b) rimuovere l'item selezionato dal box; c) rimuovere tutti gli item selezionati dal box.

Per **l'importazione di un dizionario personalizzato** si richiede che l'utilizzatore abbia predisposto un file **Dictio.diz** o un file **Dizionary.diz**, i quali possono essere costituiti da "n" righe, ciascuna con una coppia di stringhe separate dal carattere ";".
In entrambi i casi, la lunghezza massima di ogni stringa (parola, lemma o categoria) è di 50 caratteri e, al suo interno, non devono essere presenti né spazi vuoti (blank) né apostrofi.
Per ogni coppia, la prima stringa - quella a sinistra - indica la label (lemma o categoria) definita dall'utilizzatore, la seconda la parola (caso **Dictio.diz**) o il lemma (caso **Dictionary.diz**) corrispondente già presente nel dizionario **T-LAB.**

Ecco qualche esempio:

(File **Dictio.diz**)	(File **Dictionary.diz**)
ACCOGLIERE;accogliamo ACCOGLIERE;accogliate ACCOGLIERE;accoglie ACCOGLIERE;accoglie ------ PREPARARE;preparerà PREPARARE;preparerai PREPARARE;prepareranno PREPARARE;preparerebbe	ACCOGLIENZA;accoglienza ACCOGLIENZA;accogliere ACCOGLIENZA;accogliente ------ PENSIERO_ASTRATTO;concettualizzare PENSIERO_ASTRATTO;analisi PENSIERO_ASTRATTO;analizzare PENSIERO_ASTRATTO;interpretare

A seconda del tipo di file importato, i cambiamenti nel dizionario del corpus seguiranno una logica diversa (vedi sotto).



N.B. :

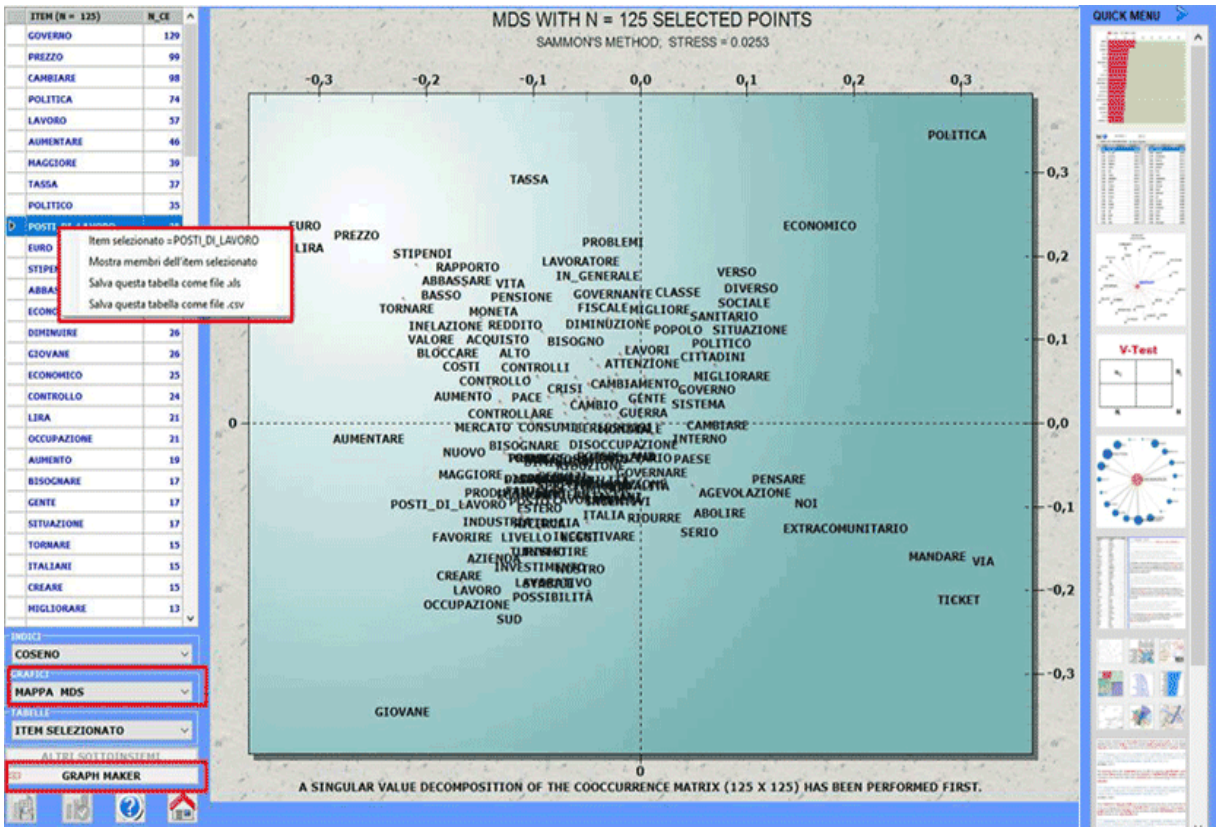
- Utilizzando l'opzione **corpus lemmatizzato** è possibile esportare una copia del corpus (file .txt) in cui ogni forma è sostituita con il corrispondente lemma.

ANALISI DELLE CO-OCCORRENZE

Associazioni di parole

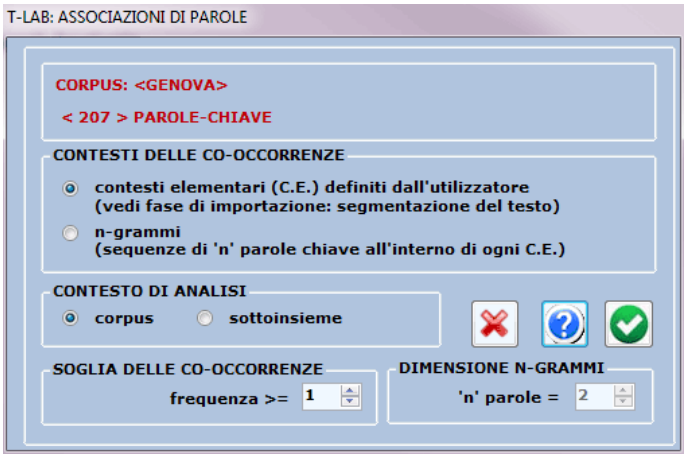
N.B.: Le immagini di questa sezione fanno riferimento a una precedente versione di T-LAB. In **T-LAB 10** l'aspetto è leggermente diverso. Inoltre: a) una nuova opzione permette all'utilizzatore di visualizzare una **Map Overview** con le parole più rilevanti; b) un nuovo strumento (**GRAPH MAKER**) consente di creare ed esportare vari tipi di grafici dinamici in formato HTML; c) il **tasto destro** sulle tabelle con le parole chiave rende disponibili opzioni supplementari; d) una galleria di immagini funziona come un menu aggiuntivo e consente di passare da un output all'altro con un solo clic.

Alcune di queste nuove funzionalità sono evidenziate nell'immagine seguente.



Questo strumento **T-LAB** consente di verificare le relazioni di **co-occorrenza** e di **similarità** che, all'interno del corpus o di un suo sottoinsieme, determinano il significato locale delle **parole chiave** selezionate dall'utilizzatore.

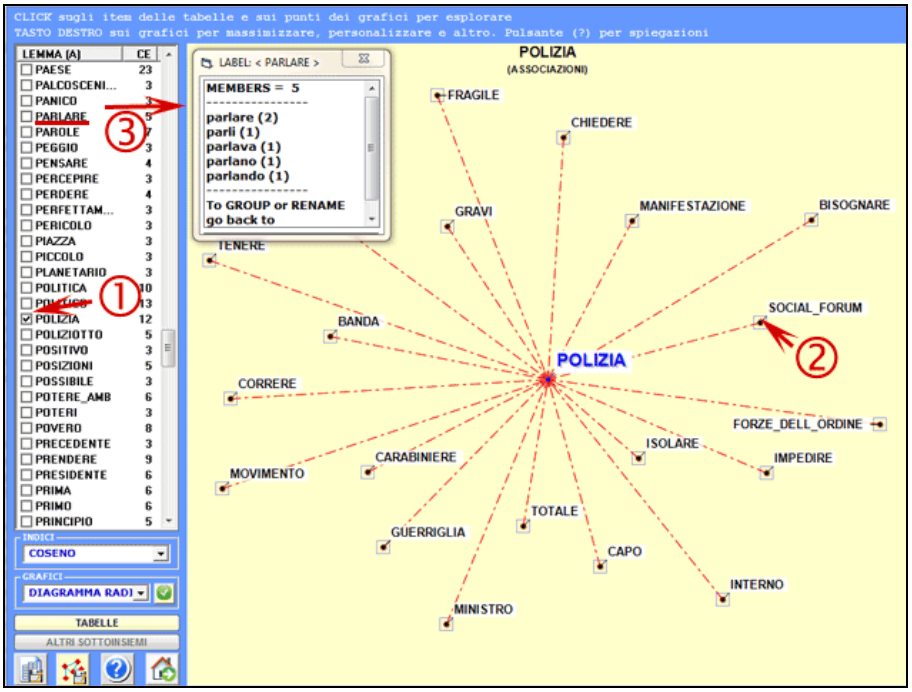
Tale verifica può essere effettuata tramite opzioni predefinite (**A**) o tramite opzioni selezionate dall'utilizzatore (**B**).



Nel primo caso (**A**: opzioni predefinite) le **co-occorrenze** delle parole sono calcolate all'interno dei **contesti elementari** selezionati in fase di importazione del corpus (es., frasi, frammenti, paragrafi, etc.); diversamente, nel secondo caso (**B**: opzioni selezionate dall'utilizzatore) le co-occorrenze possono essere anche calcolate all'interno di sequenze di parole di lunghezza variabile (cioè **n-grammi**, vedi corrispondente sezione del glossario) ed è anche possibile decidere la soglia minima (cioè la frequenza) delle co-occorrenze da considerare.

La finestra di lavoro (vedi sotto) è resa disponibile subito dopo aver effettuato il calcolo delle co-occorrenze tra tutte le parole incluse nella lista selezionata dall'utilizzatore.

Sulla sinistra di questa finestra è riportata una tabella con la lista delle parole chiave e i valori numerici che indicano la quantità di contesti elementari o n-grammi in cui ciascuna parola risulta presente.



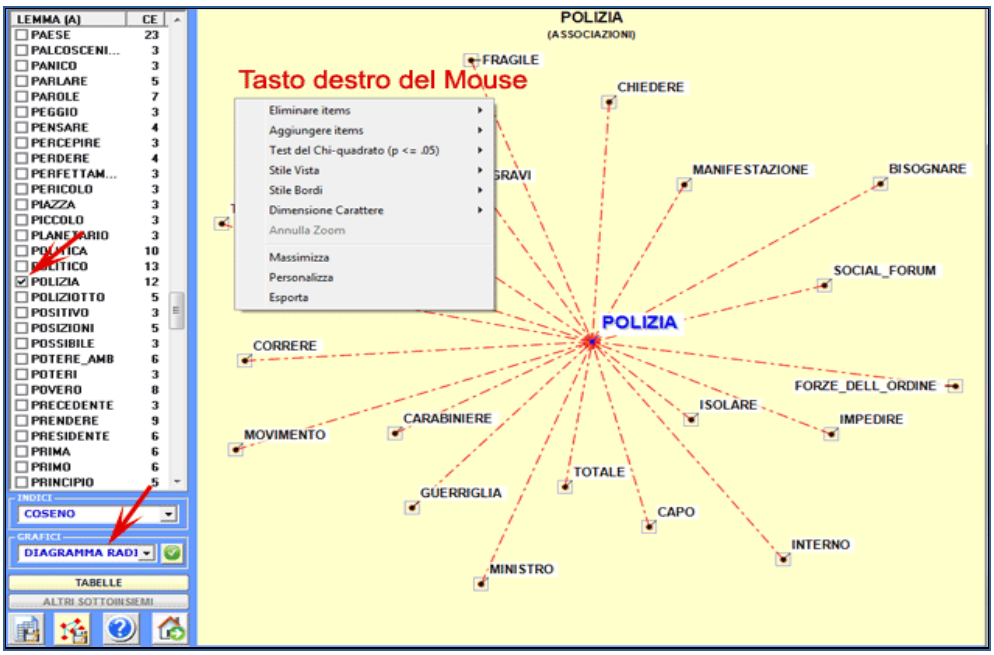
Un semplice click sugli item della tabella (opzione '1') o sui punti dei grafici (opzione '2') consente di verificare le associazioni relative a ciascuna parola target. Diversamente un click sulle label incluse nella tabella (opzione '3') consente di verificare gli item inclusi in ogni lemma.

Di volta in volta, la selezione delle parole associate è effettuata tramite il calcolo di un **Indice di Associazione** (vedi corrispondente sezione del glossario) o tramite un indice di somiglianza del **secondo ordine** (vedi spiegazione al termine di questa sezione). Nel primo caso gli indici disponibili sono sei (**Coseno, Dice, Jaccard, Equivalenza, Inclusione e Informazione Mutua**) e il loro calcolo è piuttosto rapido; diversamente, nel caso degli indici del secondo ordine - e soprattutto quando il corpus è di notevoli dimensioni - l'analisi dei dati può richiedere minuti. Inoltre va tenuto conto del fatto che, nel caso degli indici del **secondo ordine**, i risultati sono tanto più affidabili quanto più numerose le parole incluse nella lista.

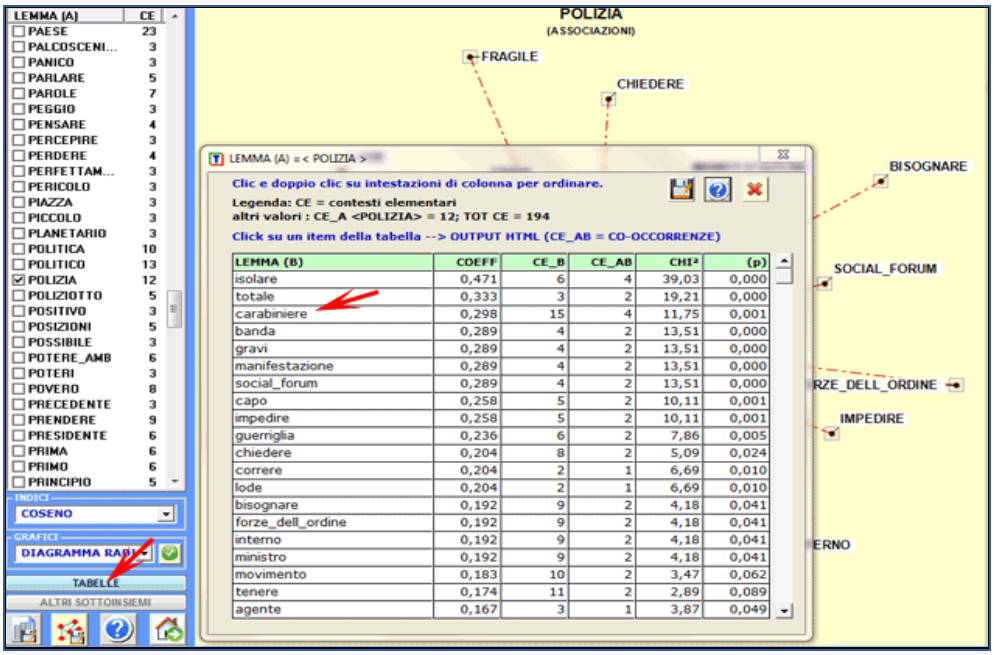
Ad ogni interrogazione, **T-LAB** produce grafici e tabelle.

Sia le tabelle che i grafici possono essere esportate tramite l'uso di appositi pulsanti.

Nei **diagrammi radiali**, il lemma selezionato è posto al centro. Gli altri sono distribuiti intorno ad esso, ciascuno a una distanza proporzionale al suo grado di associazione. Le relazioni significative sono quindi del tipo uno-ad-uno, tra il lemma centrale e ciascuno degli altri.
Ogni click su un item produce un nuovo grafico e, tramite l'uso del tasto destro del mouse, è possibile aprire una finestra di dialogo che consente vari tipi di personalizzazione (vedi sotto).



Le **tabelle** contengono dati che consentono di verificare le relazioni tra occorrenze e co-occorrenze delle parole (Max. 50) che risultano più associate a quella selezionata.

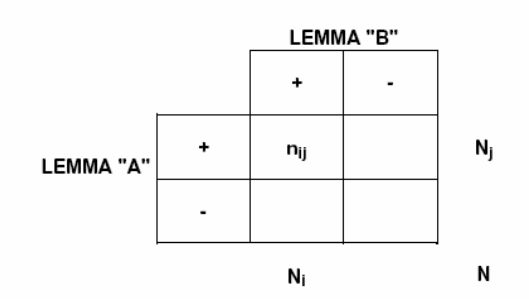


LEMMA (B)	COEFF	CE_B	CE_AB	CHI²	(p)
isolare	0,471	6	4	39,03	0,000
totale	0,333	3	2	19,21	0,000
carabiniere	0,298	15	4	11,75	0,001
banda	0,289	4	2	13,51	0,000
gravi	0,289	4	2	13,51	0,000
manifestazione	0,289	4	2	13,51	0,000
social_forum	0,289	4	2	13,51	0,000
capo	0,258	5	2	10,11	0,001
impedire	0,258	5	2	10,11	0,001
guerriglia	0,236	6	2	7,86	0,005
chiedere	0,204	8	2	5,09	0,024
correre	0,204	2	1	6,69	0,010
lode	0,204	2	1	6,69	0,010
bisognare	0,192	9	2	4,18	0,041
forze_dell_ordine	0,192	9	2	4,18	0,041
interno	0,192	9	2	4,18	0,041
ministro	0,192	9	2	4,18	0,041
movimento	0,183	10	2	3,47	0,062
tenere	0,174	11	2	2,89	0,089
agente	0,167	3	1	3,87	0,049

Le chiavi di lettura sono le seguenti:

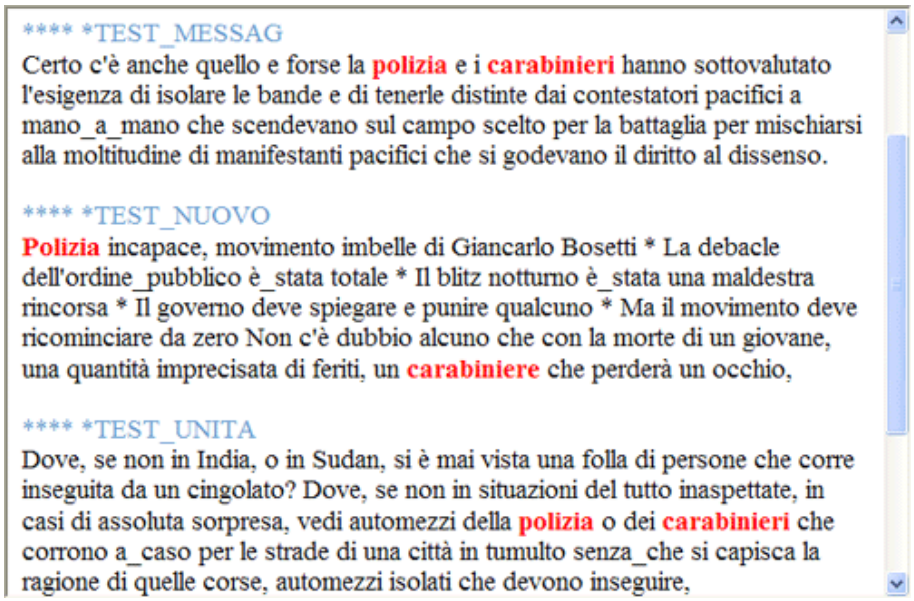
- **LEMMA (A)** = lemma selezionato (quello al centro del grafico);
- **LEMMA (B)** = lemmi associati a quello selezionato;
- **COEFF** = valore dell'indice selezionato;
- **TOT CE** = totale dei contesti elementari (CE) o degli n-grammi analizzati;
- **CE_A** = totale dei CE in cui è presente il lemma selezionato (A);
- **CE_B** = totale dei CE in cui è presente ogni lemma associato (B);
- **CE_AB** = totale dei CE in cui i lemmi "A" e "B" sono associati (co-occorrenze);
- **CHI2** = valore del chi quadro che esprime la significatività delle co-occorrenze;
- **(p)** = probabilità associata al valore del chi quadro (def=1).

Nel caso del **chi quadro**, per ogni coppia di lemmi ("A" e B") la struttura della tabella analizzata è la seguente

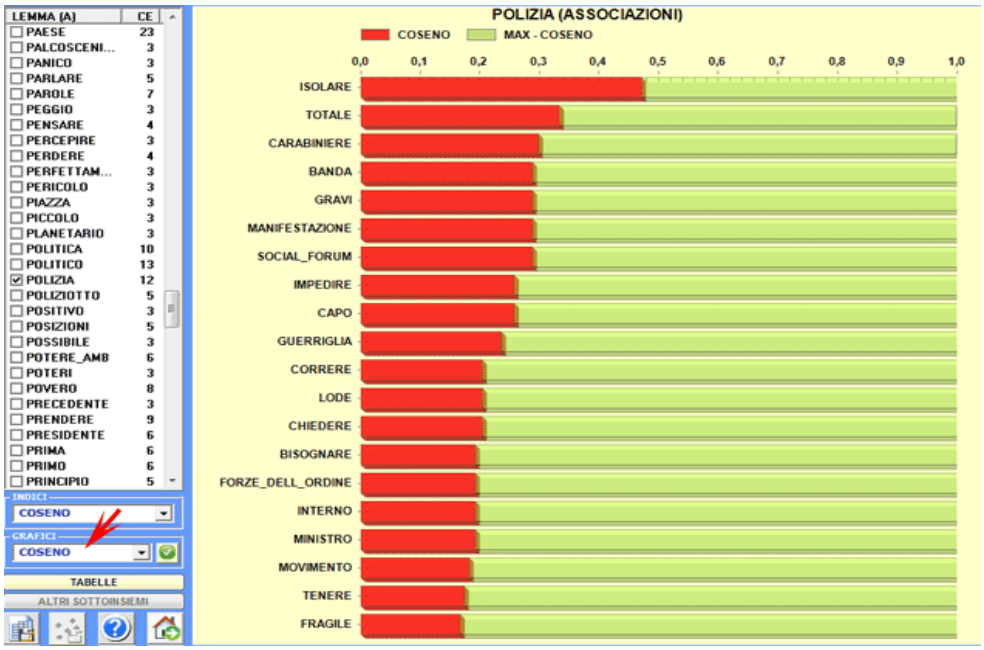


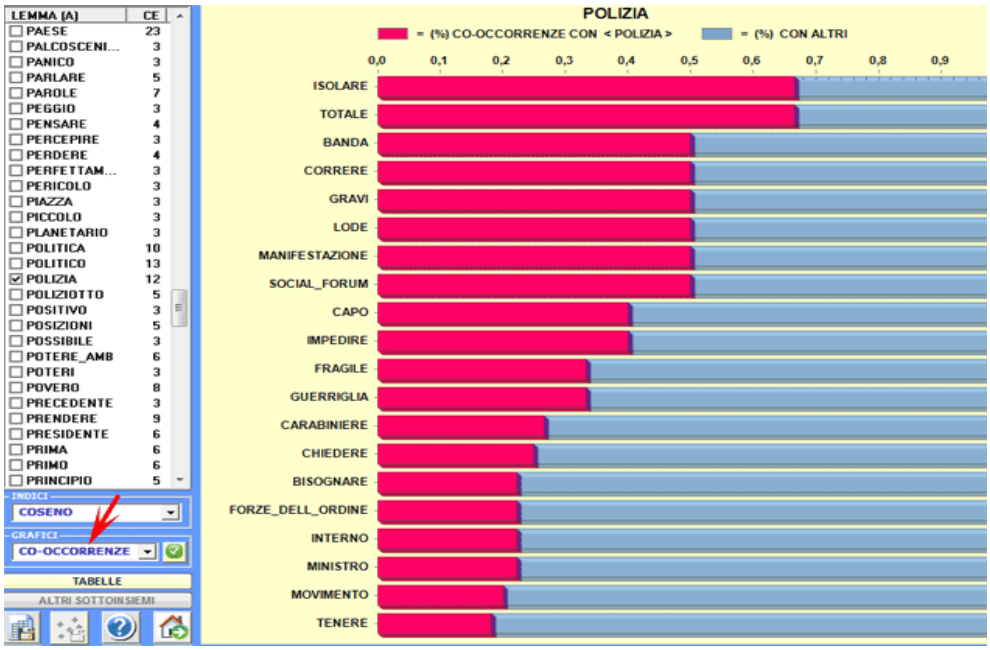
Dove : nij = CE_AB, Nj = CE_A, Ni = CE_B, N = TOT CE.

Un click su ogni item della tabella (es. "carabiniere") consente di visualizzare e di salvare un file con tutti i contesti elementari in cui esso è presente insieme alla parola centrale (es. co-occorrenze di ‘carabiniere’ e ‘polizia’).

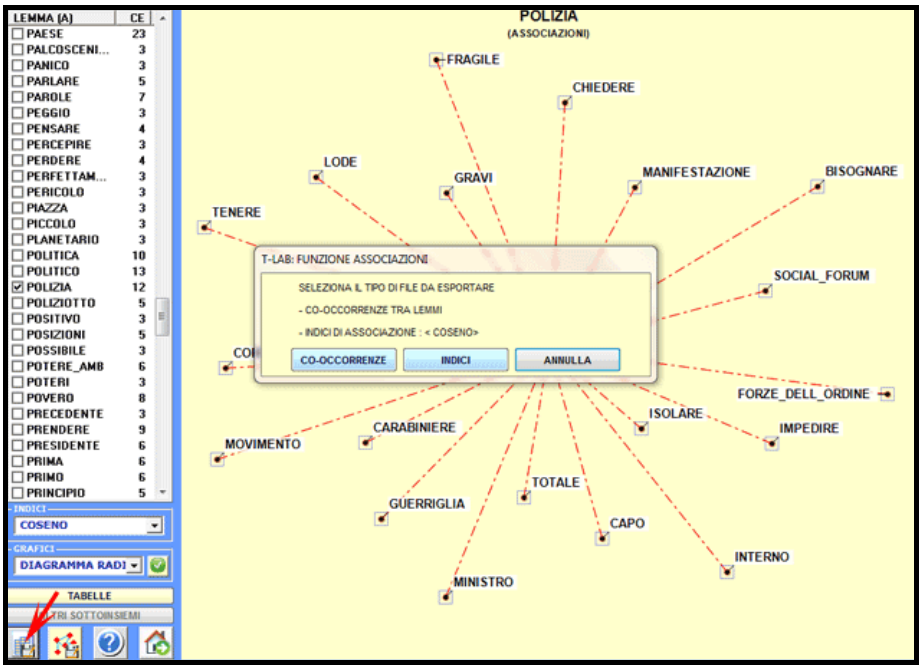


Ulteriori grafici (Istogrammi) consentono di apprezzare le differenze tra i valori del **coefficiente** utilizzato e tra le **percentuali** delle co-occorrenze (vedi sotto).



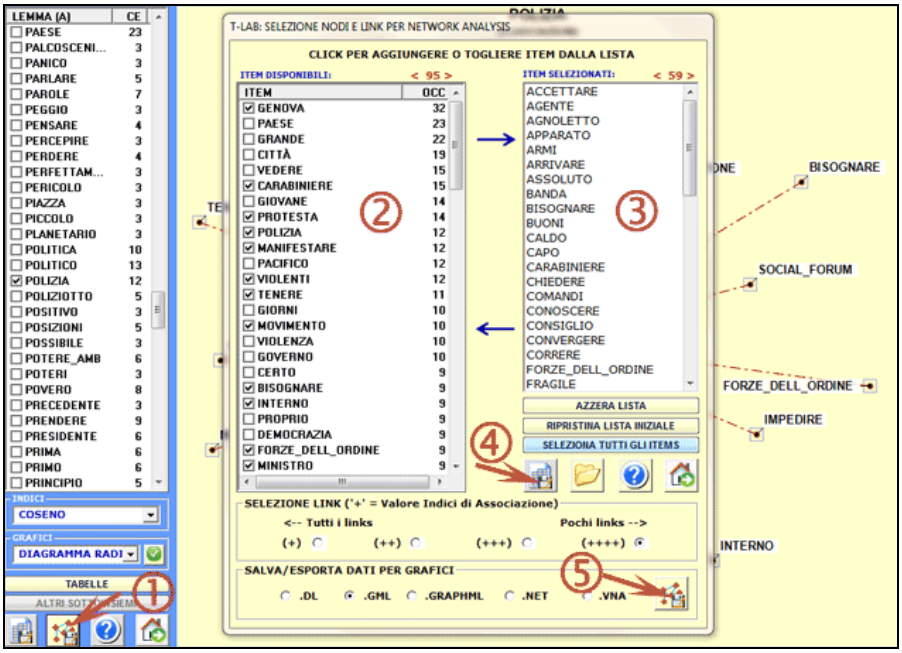


Cliccando sul pulsante in basso a sinistra l'utilizzatore può esportare **vari tipi di tabelle** (vedi immagine seguente).



Una ulteriore finestra **T-LAB** (vedi immagine seguente, step 1) consente di creare file grafici che possono essere editati con software per la network analysis quali Gephi, Pajek, Ucinet, yEd ed altri. In questo caso, i **nodi** della rete sono costituiti dalle parole associate con la parola target. Le opzioni disponibili sono le seguenti: selezionare gli item (cioè i 'nodi') da inserire nei grafici (vedi sotto, step 2 e 3), esportare la corrispondente matrice di adiacenza (vedi sotto, step 4), esportare il tipo di file grafico prescelto (vedi sotto, step 5).

N.B.: In **T-LAB 10** la finestra seguente è stata sostituita con lo strumento GRAPH MAKER.

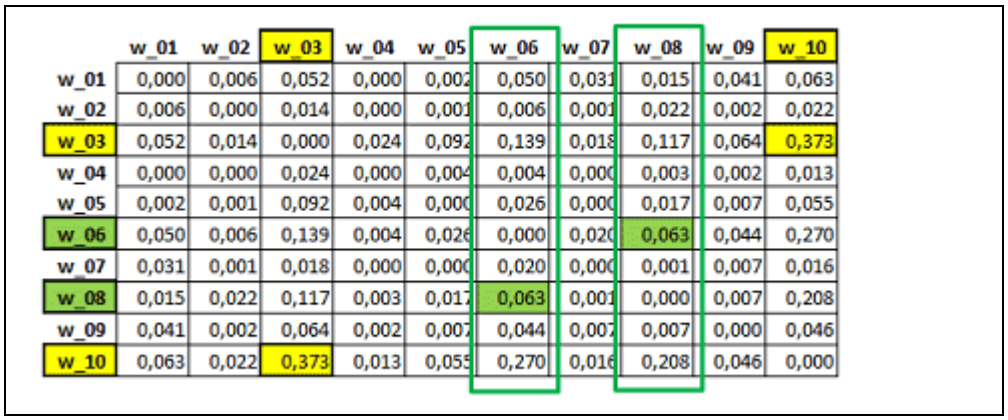


Ad esempio, file .gml esportati da **T-LAB** possono consentire di realizzare grafici come quelli riportati alla fine di questa sezione.

Le modalità di calcolo dei vari indici di 'associazione' (o prossimità) sono illustrate nella corrispondente sezione del Manuale/Help (vedi glossario). Come si potrà verificare, tutti questi indici sono ottenuti attraverso una normalizzazione dei valori di co-occorrenza concernenti coppie di parole; quindi - nei calcoli del **primo ordine** - due parole mai co-occorrenti hanno un indice di associazione pari a '0'. Diversamente, gli indici del **secondo ordine** evidenziano fenomeni di **similarità** concernenti l'uso (e quindi il significato) delle parole che non dipendono direttamente dalle loro co-occorrenze; infatti, in questo caso, due parole mai co-occorrenti possono avere un indice di associazione anche molto elevato.

Utilizzando alcuni concetti della linguistica strutturale, possiamo affermare che mentre gli indici del primo ordine rilevano fenomeni concernenti l'asse sintagmatico (combinazione e prossimità 'in praesentia', cioè parole 'l'una accanto all'altra' in una specifica frase), gli indici del secondo ordine rivelano fenomeni concernenti l'asse paradigmatico (associazione e similarità 'in absentia', cioè relazioni di quasi-sinonimia tra due o più termini usati dallo stesso autore). In effetti, all'interno dei testi analizzati, le parole con una elevata similarità del secondo ordine sono spesso quasi-sinonimi.

Per capire il modo in cui **T-LAB** calcola gli indici del 'secondo ordine', è utile ricordare che gli indici del 'primo ordine' possono essere utilizzati per costruire matrici di prossimità come la seguente (A).

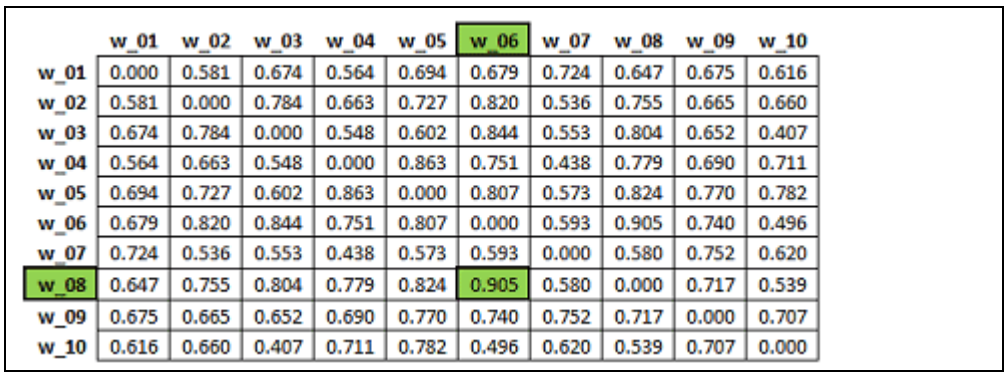


	w_01	w_02	w_03	w_04	w_05	w_06	w_07	w_08	w_09	w_10
w_01	0,000	0,006	0,052	0,000	0,002	0,050	0,031	0,015	0,041	0,063
w_02	0,006	0,000	0,014	0,000	0,001	0,006	0,001	0,022	0,002	0,022
w_03	0,052	0,014	0,000	0,024	0,092	0,139	0,018	0,117	0,064	0,373
w_04	0,000	0,000	0,024	0,000	0,004	0,004	0,000	0,003	0,002	0,013
w_05	0,002	0,001	0,092	0,004	0,000	0,026	0,000	0,017	0,007	0,055
w_06	0,050	0,006	0,139	0,004	0,026	0,000	0,020	0,063	0,044	0,270
w_07	0,031	0,001	0,018	0,000	0,000	0,020	0,000	0,001	0,007	0,016
w_08	0,015	0,022	0,117	0,003	0,017	0,063	0,001	0,000	0,007	0,208
w_09	0,041	0,002	0,064	0,002	0,007	0,044	0,007	0,007	0,000	0,046
w_10	0,063	0,022	0,373	0,013	0,055	0,270	0,016	0,208	0,046	0,000

Matrice (A) – Similarità del Primo Ordine

In questa matrice simmetrica (A), il valore 0.373 (in giallo) corrisponde al più elevato indice del 'primo ordine' ed indica l'associazione tra le parole 'w_03' e 'w_10'. Più specificamente, si tratta di un indice di equivalenza ottenuto dividendo il quadrato delle loro co-occorrenze per il prodotto delle loro occorrenze (360^2/627*553).

A partire dalla matrice di cui sopra (A), **T-LAB** costruisce una seconda matrice (B) ottenuta calcolando i coseni risultanti dal confronto di tutte le colonne contenenti gli indici del primo ordine (vedi matrice 'A'). Come di può verificare, nella seguente tabella 'B' il valore di similarità più elevato riguarda la relazione tra le parole 'w_06' e 'w_08'. Ciò significa che i rispettivi vettori (vedi le due colonne evidenziate in verde nella matrice 'A'), dopo essere stati opportunamente normalizzati, risultano essere tra loro molto simili (coseno = 0.905), anche se l'associazione del 'primo ordine' tra le parole due parole in questione risulta piuttosto bassa (0.063).



	w_01	w_02	w_03	w_04	w_05	w_06	w_07	w_08	w_09	w_10
w_01	0.000	0.581	0.674	0.564	0.694	0.679	0.724	0.647	0.675	0.616
w_02	0.581	0.000	0.784	0.663	0.727	0.820	0.536	0.755	0.665	0.660
w_03	0.674	0.784	0.000	0.548	0.602	0.844	0.553	0.804	0.652	0.407
w_04	0.564	0.663	0.548	0.000	0.863	0.751	0.438	0.779	0.690	0.711
w_05	0.694	0.727	0.602	0.863	0.000	0.807	0.573	0.824	0.770	0.782
w_06	0.679	0.820	0.844	0.751	0.807	0.000	0.593	0.905	0.740	0.496
w_07	0.724	0.536	0.553	0.438	0.573	0.593	0.000	0.580	0.752	0.620
w_08	0.647	0.755	0.804	0.779	0.824	0.905	0.580	0.000	0.717	0.539
w_09	0.675	0.665	0.652	0.690	0.770	0.740	0.752	0.717	0.000	0.707
w_10	0.616	0.660	0.407	0.711	0.782	0.496	0.620	0.539	0.707	0.000

Matrice (B) – Similarità del Secondo Ordine

Detto in altri termini, un indice del 'primo ordine' è ottenuto applicando una formula che include valori di co-occorrenza e occorrenza, mentre un indice del 'secondo ordine' è ottenuto moltiplicando due vettori normalizzati.

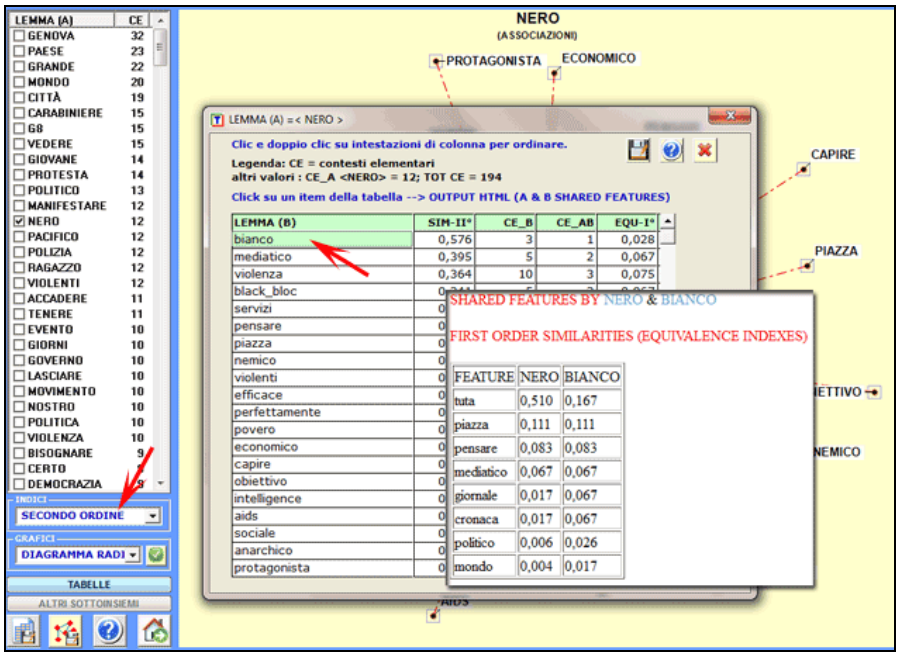
Al di là delle modalità di calcolo, va sottolineato il fatto che nei due casi ('A' e 'B') vengono rilevati due diversi fenomeni. Nel primo caso ('A'), infatti, il focus è sulle co-occorrenze; diversamente, nel secondo caso ('B') - e indipendentemente dalle loro co-occorrenze - il focus è sulle somiglianze tra 'profili' i cui dati fanno riferimento all'uso delle parole da parte degli

autori dei testi analizzati.

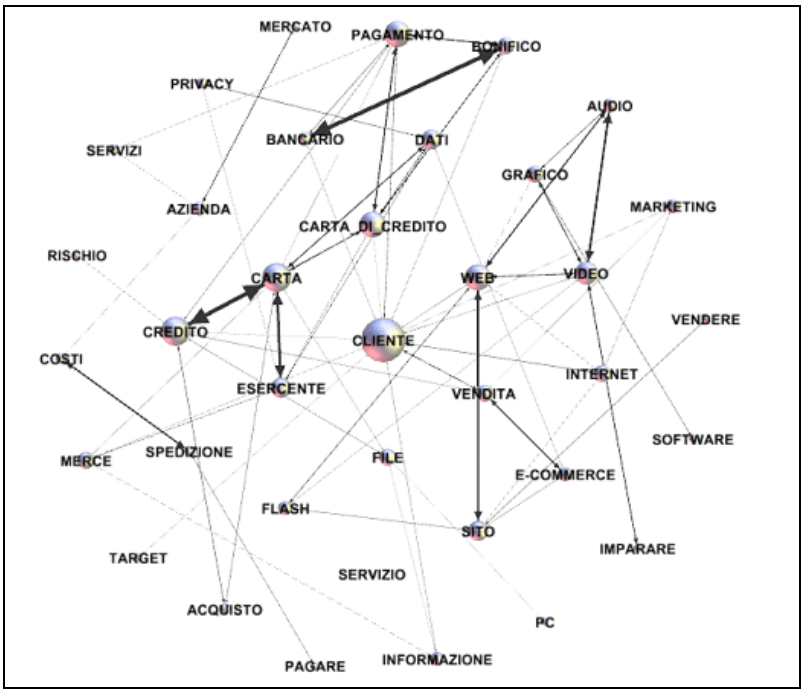
Tanto per fare un esempio, nell'analisi di **Pinocchio** del primo ordine il termine 'fata' risulta prevalentemente associato (vedi co-occorrenze) con 'buona' e 'capelli turchini'; diversamente, nell'analisi del secondo ordine, il termine che risulta più simile a 'fata' è 'mamma', anche se le co-occorrenze tra questi due termini ('fata' e 'mamma') sono - all'interno della fiaba di Collodi - pressoché irrilevanti (cioè solo 3).

Le tabelle visualizzate da **T-LAB** consentono di verificare sia le similarità del secondo ordine (vedi sotto colonna SIM-II°) che gli indici del primo ordine (EQU-I°, cioè indice di equivalenza).

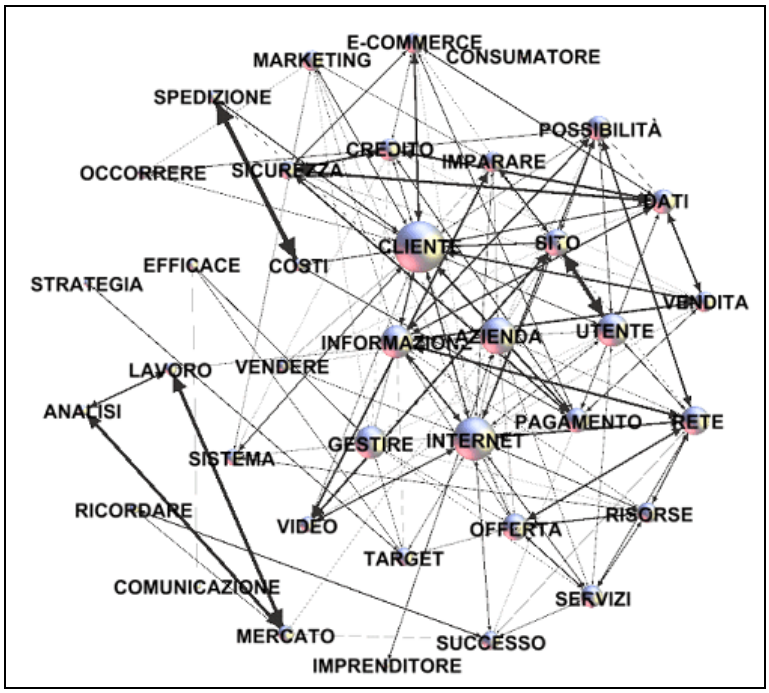
Inoltre, cliccando su ogni item di questa tabella, è possibile visualizzare file HTML che consentono di verificare quali 'caratteristiche' ('features') determinano la somiglianza tra ogni coppia di parole. Ad esempio, la tabella seguente mostra che la somiglianza del secondo ordine tra ‘nero’ e ‘bianco’ è in primo luogo determinata da caratteristiche condivise quali ‘tuta’, ‘piazza’, etc..



In alcuni casi e per specifici obiettivi può risultare particolarmente interessante confrontare le reti semantiche di specifiche parole target ottenute esportando le rispettive prossimità del primo ordine e similarità del secondo ordine. Ad esempio, i grafici seguenti sono stati ottenuti esportando file .gml prodotti da **T-LAB** e successivamente importati nel software Gephi (vedi https://gephi.org/). In entrambi i casi, la parola target è 'cliente', all'interno di un corpus costituito da una mailing list concernente il commercio elettronico. Come si può rilevare, nel primo caso ('A' - associazioni del primo ordine) le relazioni tra i nodi rinviano a specifici sintagmi quali 'pagamento con bonifico bancario', 'sito web', 'costi di spedizione', 'tramite carta di credito', etc.; diversamente, nel secondo caso ('B' - associazioni del secondo ordine), le relazioni tra i nodi con più elevato scambio sembrano rinviare a un paradigma di tipo gestionale.



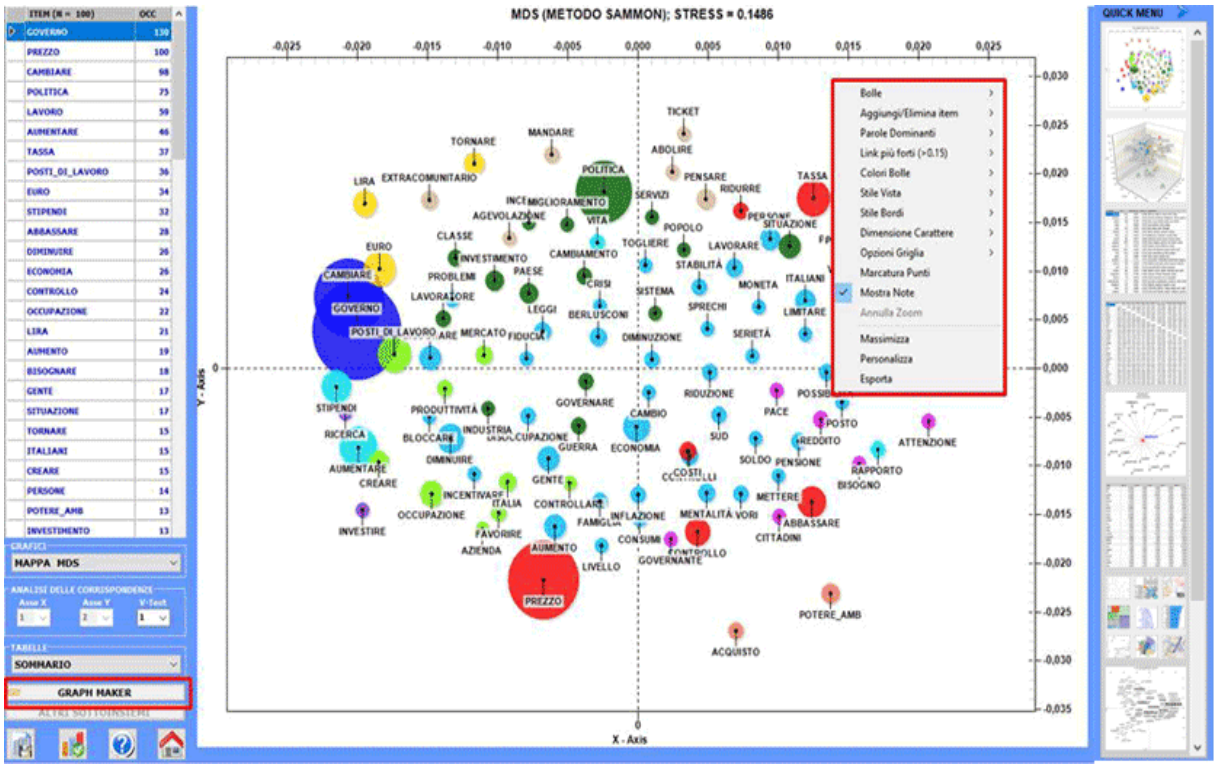
(A) - Associazioni del Primo Ordine



(B) - Associazioni del Secondo Ordine

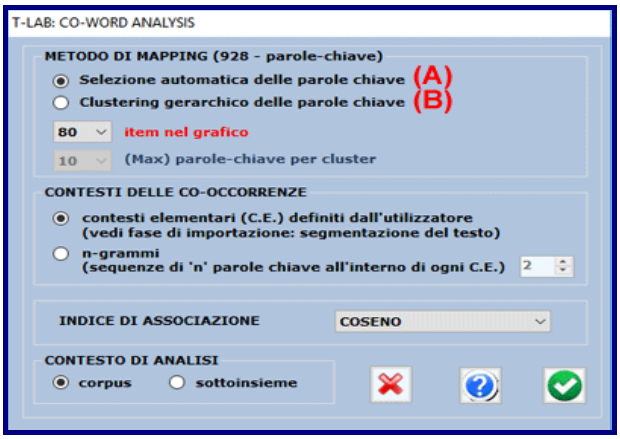
Co-Word Analysis e Mappe Concettuali

N.B.: Le immagini di questa sezione fanno riferimento all'interfaccia una precedente versione di T-LAB. In **T-LAB 10** l'aspetto è leggermente diverso. Inoltre: a) quando è attiva l'opzione 'selezione automatica delle parole chiave', nella mappa **MDS** vengono usati differenti colori per distinguere i vari gruppi (cluster) di elementi; b) è stata aggiunta una tecnica di visualizzazione denominata t-SNE (t-Distributed Stochastic Neighbor Embedding); c) un nuovo strumento (**GRAPH MAKER**) consente di creare ed esportare vari tipi di grafici dinamici in formato HTML; d) il **tasto destro** sulle tabelle con le parole chiave rende disponibili opzioni supplementari; e) una galleria di immagini funziona come un menu aggiuntivo e consente di passare da un output all'altro con un solo clic.
Alcune di queste nuove funzionalità sono evidenziate nell'immagine seguente.



L'uso di questa funzione **T-LAB** consente di realizzare due tipi di analisi concernenti le **co-occorrenze** delle parole:

- **A -** tra **singole parole-chiave** (lemmi o categorie), se la loro quantità non supera 500 elementi (min 10);
- **B -** tra (ed entro) piccoli **cluster**, denominati **Nuclei Tematici**, se la quantità delle parole-chiave selezionate supera 100 elementi (max 3000).

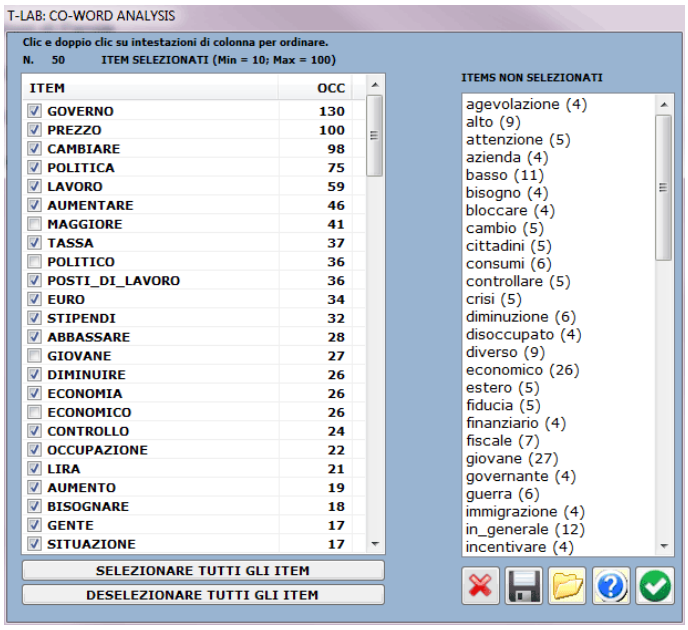


L'utilizzatore può selezionare l'indice di associazione da utilizzare e, solo nel caso dell'opzione B, sia la quantità massima di cluster da ottenere (da 50 a 100) che la quantità massima di parole-chiave per cluster.

La procedura di calcolo utilizzata prevede i seguenti passi:

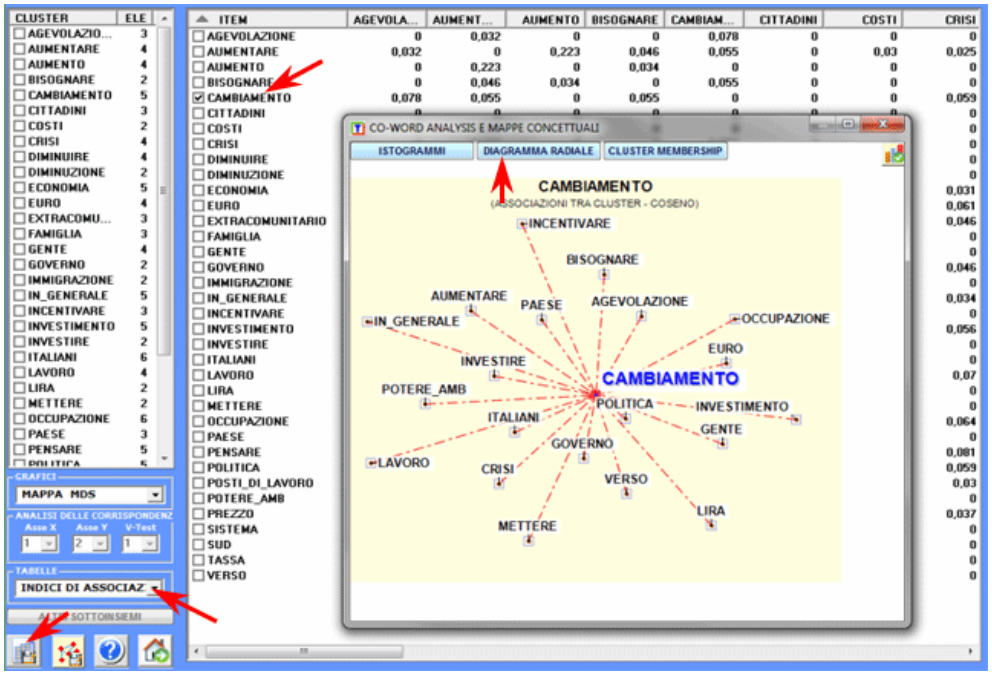
1. *costruzione di una matrice delle co-occorrenze (relazioni tra parole);*
2. *calcolo degli indici di associazione selezionati (Coseno, Dice, Jaccard, Equivalenza, Inclusione, Informazione Mutua);*
3. *clusterizzazione gerarchica applicata alla matrice delle dissimilarità;*
4. *costruzione di una seconda matrice delle dissimilarità (relazioni tra cluster);*
5. *rappresentazione grafica mediante multidimensional scaling e analisi delle corrispondenze.*

N.B.:
- nei casi "A" (vedi sopra), l'utilizzatore può rivedere e personalizzare la selezione delle parole-chiave (vedi immagine seguente) e **T-LAB** non effettua i passaggi 3 e 4;



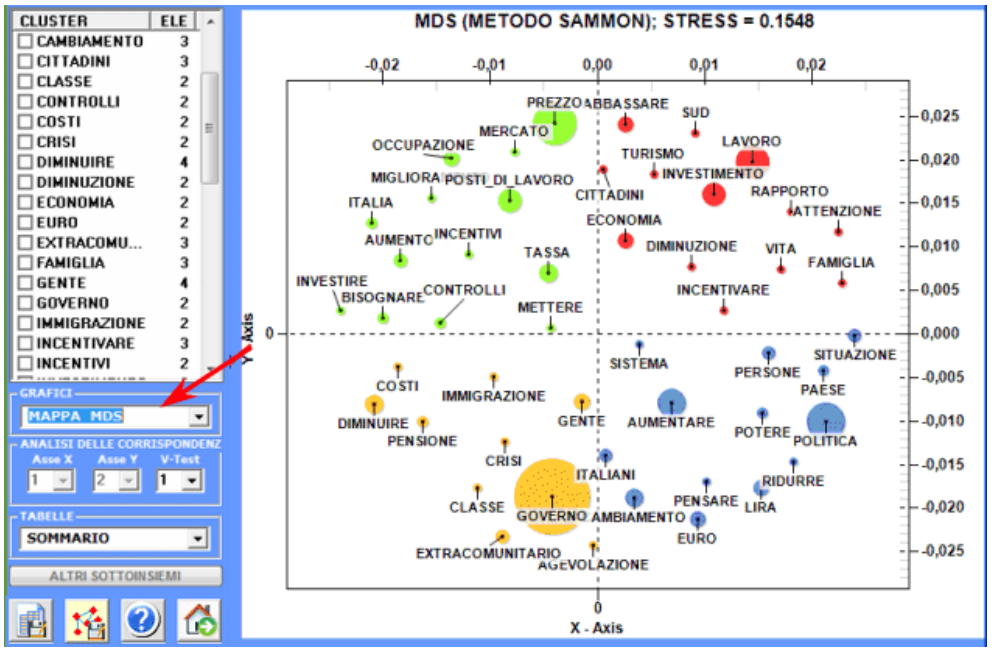
- un'accurata **selezione delle parole-chiave**, garantisce una migliore qualità dei risultati;
- poiché le **parole multiple** non classificate da **T-LAB** sono casi specifici di co-occorrenza e l'opzione "B" li individua come cluster (ad esempio, "Torri" + "Gemelle"), si consiglia di

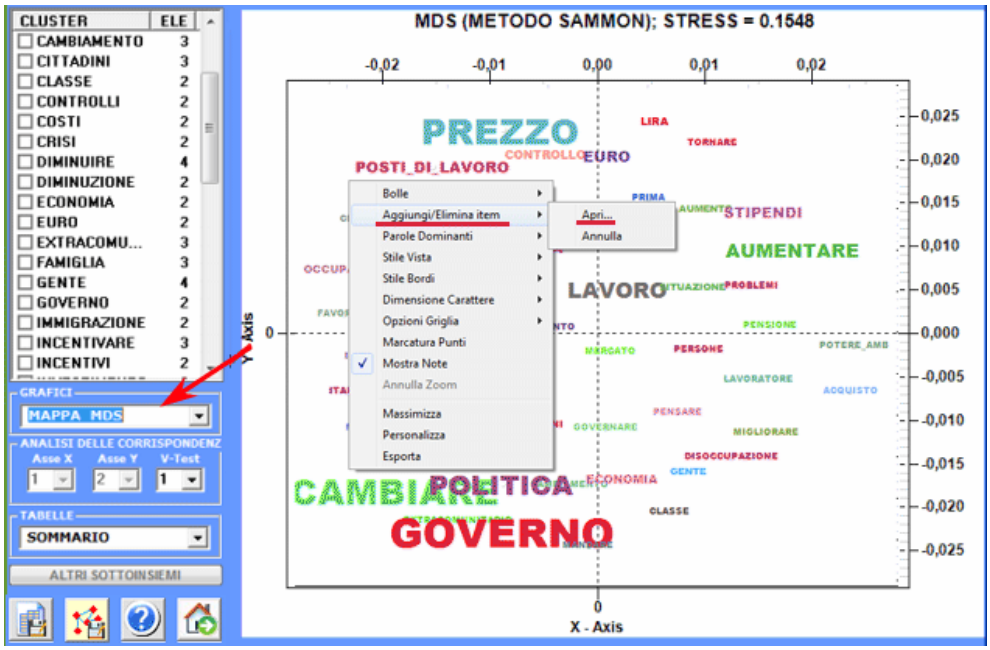
risolvere questi casi durante la fase di importazione del corpus. In ogni caso, senza ripetere il processo di importazione, è possibile effettuare gli opportuni interventi tramite la funzione **Personalizzazione del Dizionario** (ad esempio, attribuendo la label "Torri_Gemelle" ai due distinti item "Torri" e "Gemelle");
- le tabelle utilizzate per le rappresentazioni grafiche possono essere esplorate e salvate con qualche click (vedi immagine seguente).



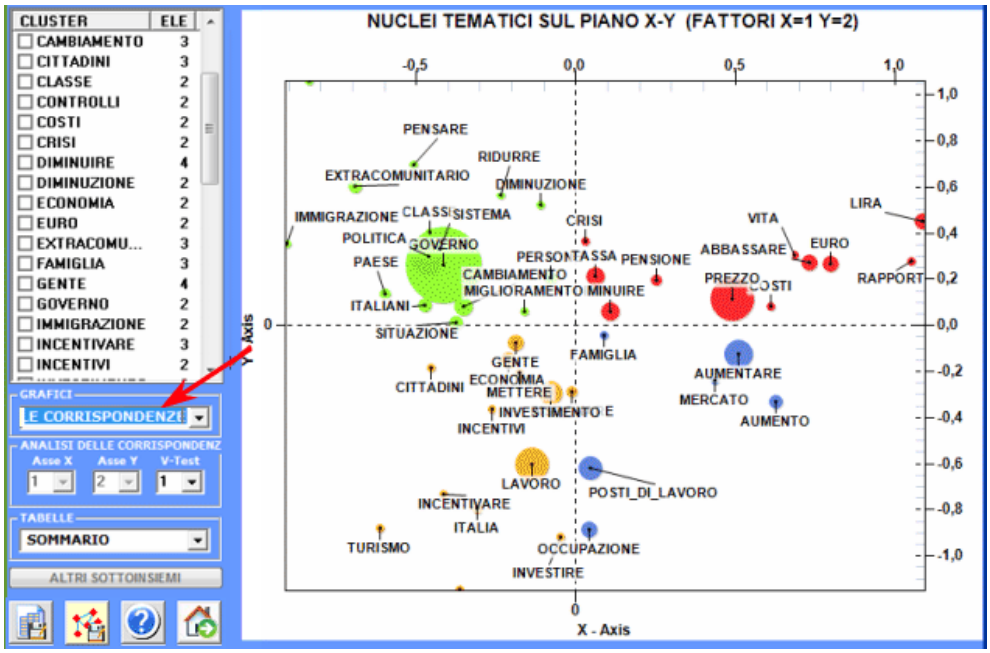
Al termine dell'analisi automatica, i grafici prodotti da **T-LAB** sono di quattro tipi (vedi sotto), e ciascuno di essi può essere personalizzato tramite la corrispondente finestra di dialogo (tasto destro del mouse).

1 – Mappa MDS

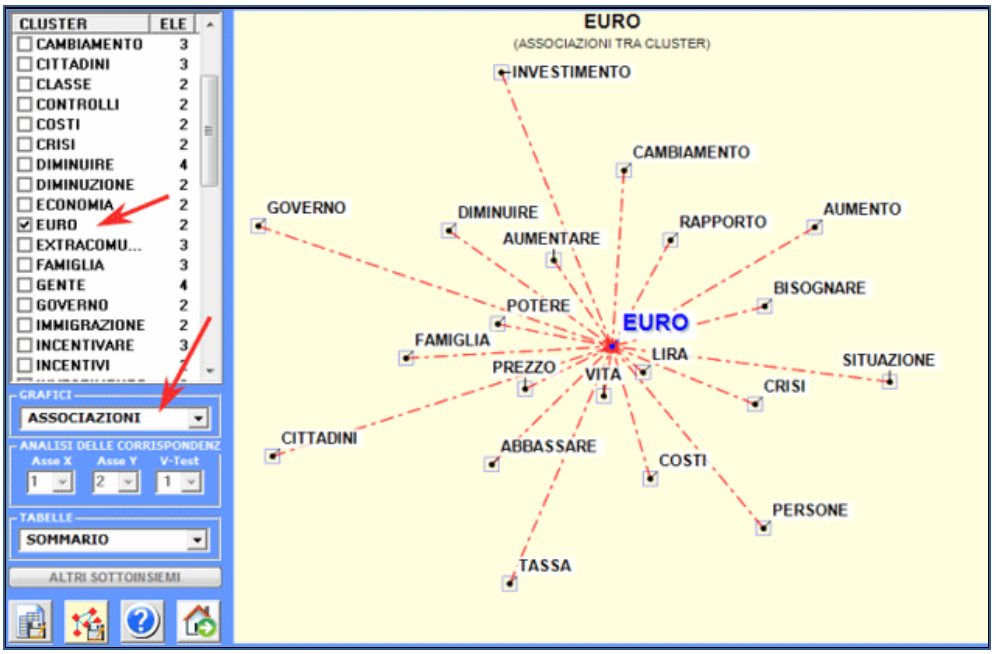




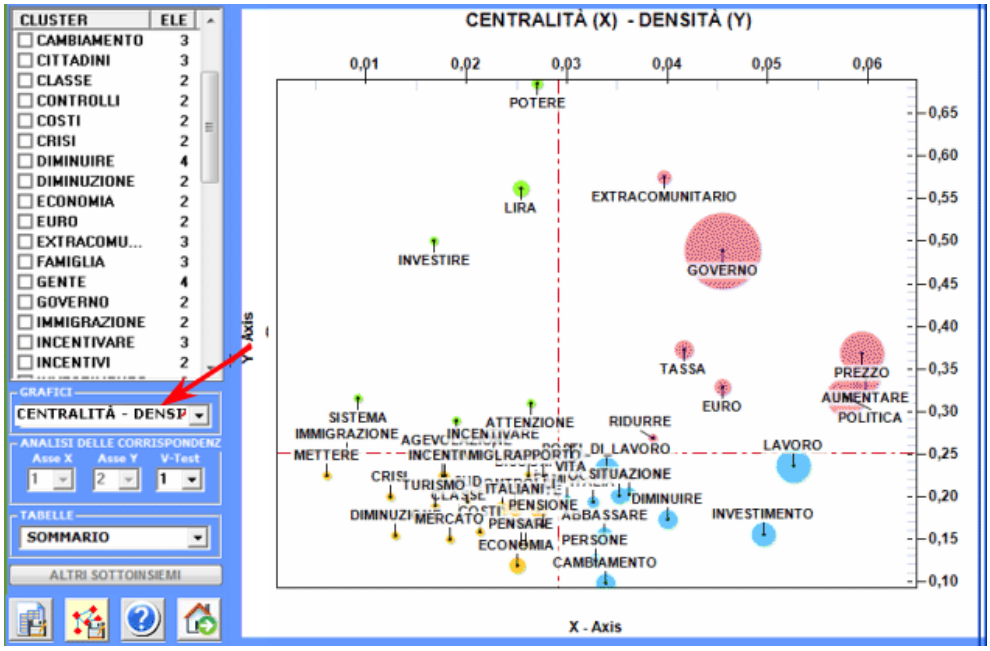
2 – Analisi Fattoriale delle Corrispondenze



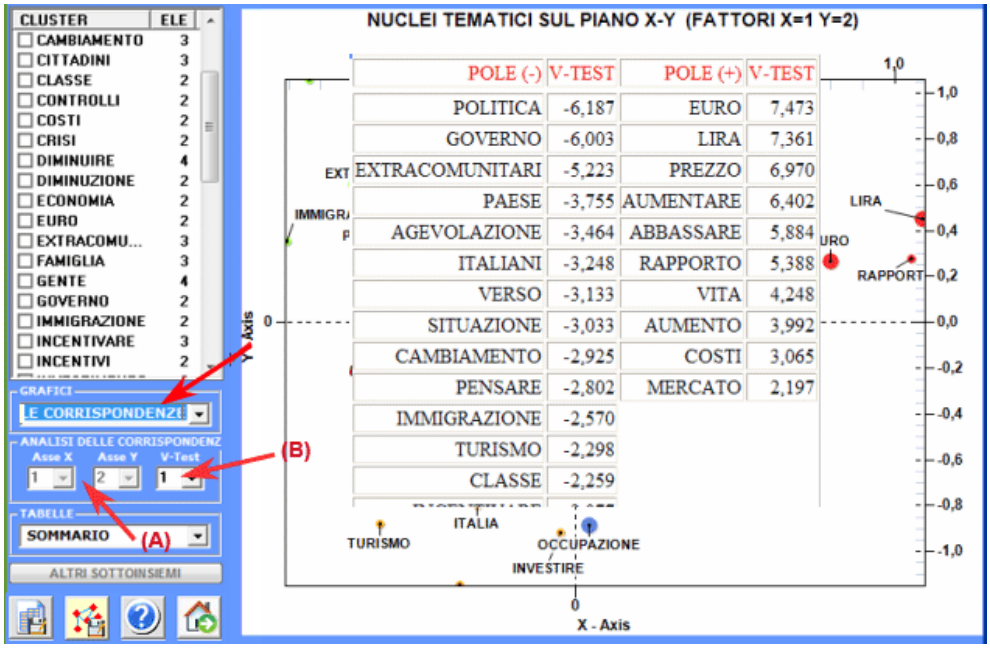
3 – Diagramma delle Associazioni



4 – Diagramma con le misure di Centralità e Densità (solo nel caso della cluster analisi)

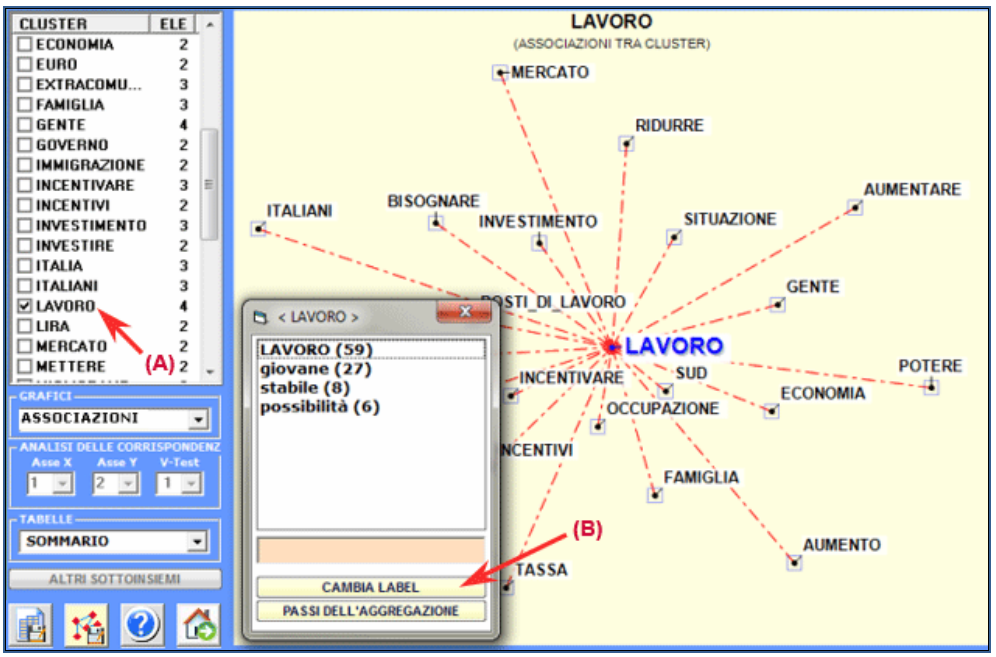


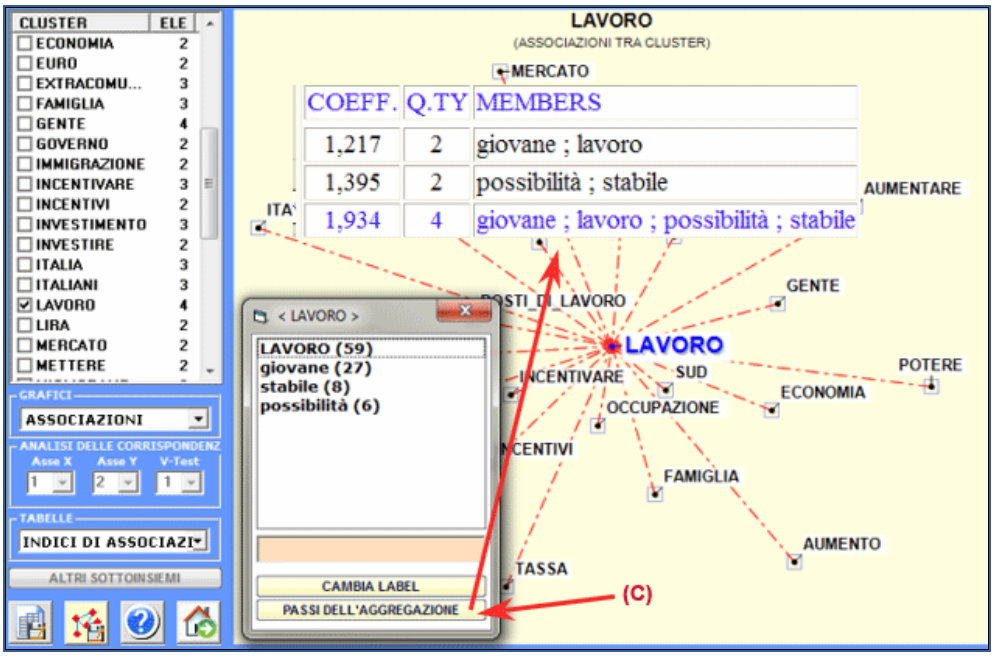
In particolare, i risultati ottenuti mediante **Analisi Fattoriale delle Corrispondenze** possono essere esplorati attraverso grafici che utilizzano le coordinate sui primi dieci assi fattori (vedi sotto "A").
Poiché **T-LAB** consente di verificare i **Valori Test** di ogni fattore (vedi sotto "B"), questo tipo di output può essere utilizzato per un'accurata interpretazione delle relazioni tra cluster e/o tra parole-chiave.

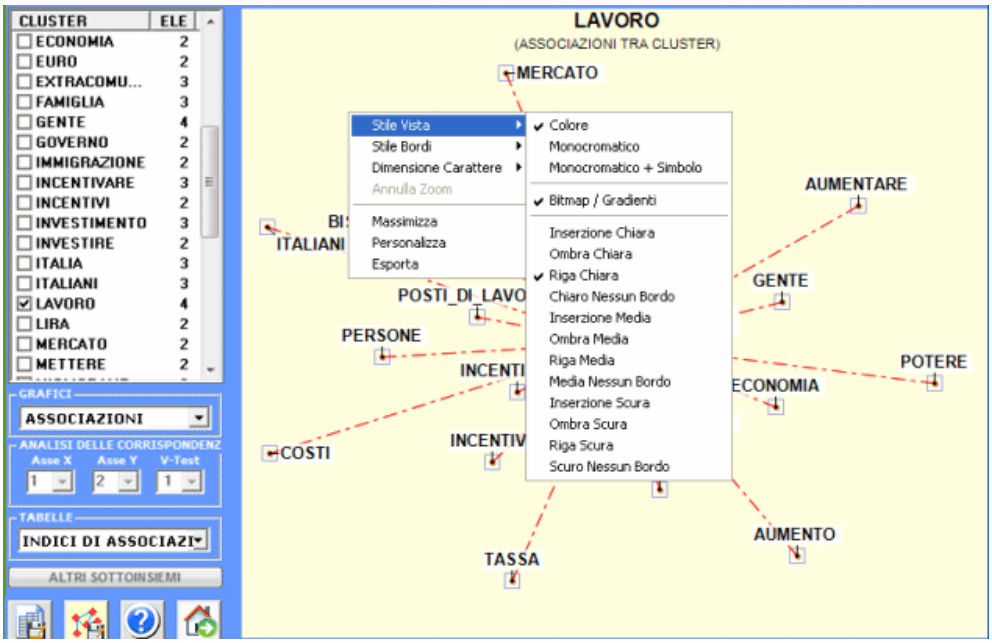


L'esplorazione e/o la personalizzazione dei vari grafici è consentita nei modi seguenti:

AZIONE	RISULTATO
check un item della tabella click su un punto del grafico	grafico delle relative associazioni
click su una label della colonna "CLUSTER" (vedi sotto "A")	lista con gli elementi del cluster
click sul pulsante "cambia label" (vedi sotto "B")	nuova label attribuita al cluster
click sul pulsante "passi dell'aggregazione" (vedi sotto "C")	file HTML con le aggregazioni nel cluster selezionato
tasto destro del mouse	personalizzazione dei grafici

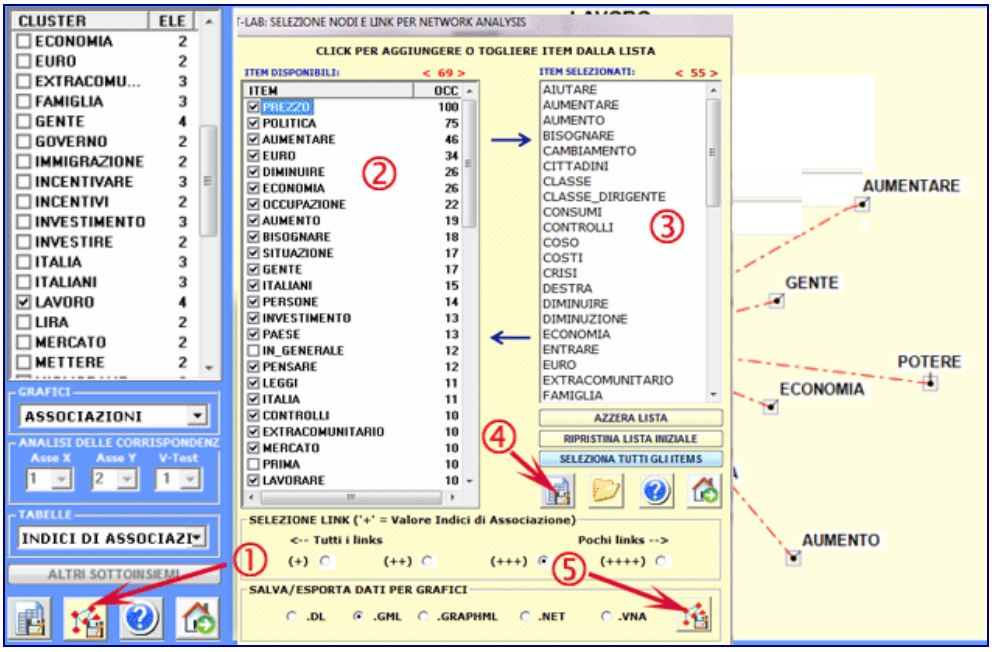




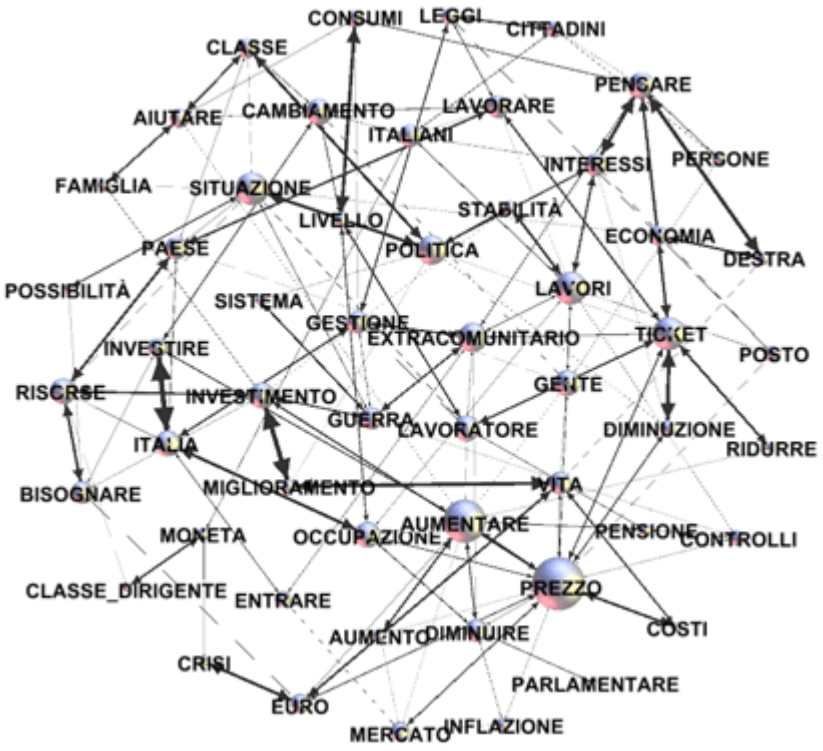


Una ulteriore finestra **T-LAB** (vedi immagine seguente, step 1) consente di creare file grafici che possono essere editati con software per la **network analysis** quali Gephi, Pajek, Ucinet, yEd ed altri. In questo caso, le opzioni disponibili sono le seguenti: selezionare gli item (cioè i nodi) da inserire nei grafici (vedi sotto, step 2 e 3), esportare la corrispondente matrice di adiacenza (vedi sotto, step 4), esportare il tipo di file prescelto (vedi sotto, step 5).

N.B.: In **T-LAB 10** la finestra seguente è stata sostituita con lo strumento **GRAPH MAKER**.

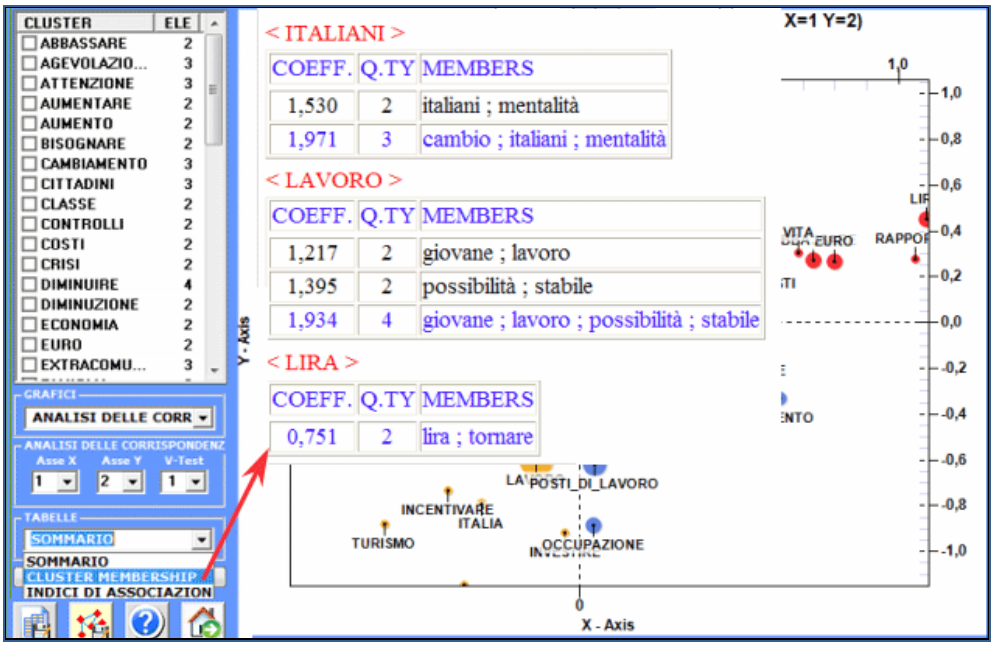


Ad esempio, un file .gml esportato da T-LAB può consentire di realizzare un grafico come il seguente.



Le tabelle esportabili con questo strumento **T-LAB** sono di tre tipi:

1 – La tabella "**Cluster Membership**" (vedi sotto) riporta la storia dell'aggregazione gerarchica interna a ciascun cluster;



2 – La tabella "**Sommario**" (vedi sotto) include le seguenti misure:

- **ECQ** = quantità dei contesti elementari in cui due o più elementi del cluster sono co-occorrenti;
- **Centrality** = media degli indici di associazione concernenti le relazioni tra cluster;
- **Density** = media degli indici di associazioni interni a ciascun cluster.

CLUSTER	ECQ	CENTRALITY	DENSITY	MEMBERS
ABBASSARE	2	0,034	0,154	ABBASSARE; INFLAZIONE
AGEVOLAZIONE	2	0,025	0,144	AGEVOLAZIONE; FINANZIARIO; PENSARE
ATTENZIONE	5	0,028	0,310	ATTENZIONE; SOCIALE; VERSO
AUMENTARE	21	0,061	0,182	AUMENTARE; BLOCCARE; RAPPORTO; STIPENDI
AUMENTO	2	0,025	0,187	AUMENTO; PRODUTTIVITÀ
BISOGNARE	2	0,028	0,217	BISOGNARE; FIDUCIA
CAMBIAMENTO	4	0,034	0,094	CAMBIAMENTO; POLITICO; SERIETÀ; TOGLIERE
CITTADINI	2	0,019	0,224	CITTADINI; GOVERNANTE; PACE
CONTROLLI	6	0,029	0,143	ALTO; CONTROLLI; LAVORATORE; PENSIONE
COSTI	1	0,023	0,158	COSTI; MONETA
CRISI	1	0,013	0,200	CRISI; MONDIALE
DIMINUIRE	7	0,039	0,173	DIMINUIRE; DISOCCUPAZIONE; PROBLEMI; SANITARIO
DIMINUZIONE	1	0,014	0,154	DIMINUZIONE; FISCALE
ECONOMIA	2	0,026	0,118	ECONOMIA; LEGGI
EURO	5	0,047	0,329	EURO; VALORE
EXTRACOMUNITARIO	4	0,033	0,478	EXTRACOMUNITARIO; MANDARE

3- La tabella "**Indici di Associazione**" (vedi sotto) riporta le misure delle relazioni di somiglianza tra i cluster (between) e tra le parole interne a ciascun cluster (within).

BETWEEN

< ECONOMIA >

CLUSTER	INDEX
lavoro	0,117
GOVERNO	0,087
investimento	0,086
sud	0,085
abbassare	0,085
ridurre	0,082
situazione	0,060
miglioramento	0,055
crisi	0,052
turismo	0,050
cittadini	0,046
diminuzione	0,046
prezzo	0,044
posti_di_lavoro	0,043
AGEVOLAZIONE	0,037
bisognare	0,035
paese	0,034
italiani	0,030
persone	0,029
gente	0,028
euro	0,026
TASSE	0,025
aumentare	0,018
politica	0,016

WITHIN

< DIMINUZIONE >

LEMMA_A	LEMMA_B	INDEX
diminuzione	fiscale	0,154

< ECONOMIA >

LEMMA_A	LEMMA_B	INDEX
economia	leggi	0,118

< EURO >

LEMMA_A	LEMMA_B	INDEX
euro	valore	0,329

< EXTRACOMUNITARIO >

LEMMA_A	LEMMA_B	INDEX
extracomunitario	mandare	0,478

< FAMIGLIA >

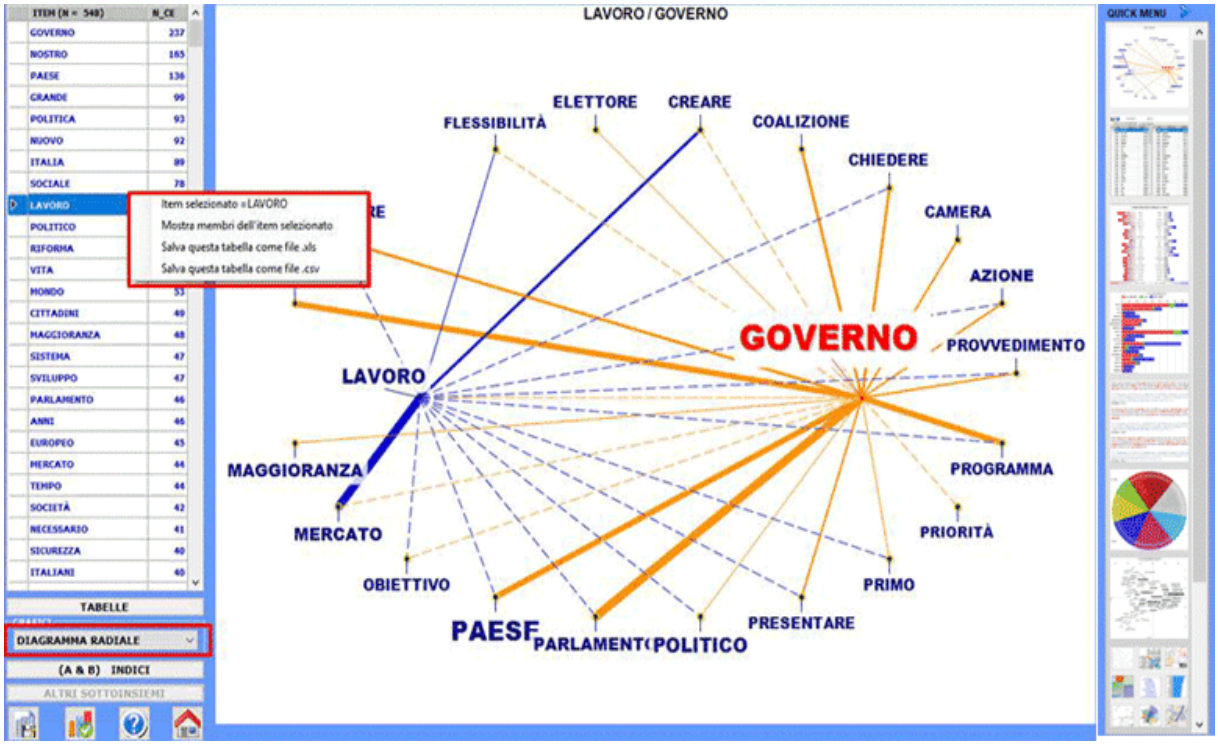
LEMMA_A	LEMMA_B	INDEX
bisogno	reddito	0,224
famiglia	reddito	0,169

N.B.:
- nel caso in cui non sia stata effettuata una Cluster Analysis, la tabella "Cluster Membership" non è ovviamente disponibile, la tabella "Sommario" è semplificata e quella con gli "Indici di Associazione" si riferisce alle relazioni di co-occorrenza tra le parole;
- all'uscita da questa analisi, il **dizionario** dei Nuclei Tematici (cioè la lista delle label attribuite ai cluster di parole) può essere esportato e, dopo un'attenta revisione, può essere importato tramite la funzione **Personalizzazione del Dizionario**. In questo modo, l'utilizzatore potrà realizzare analisi di secondo livello e passare dalle parole (primo livello) ai **temi** o ai **concetti.**

Confronto tra coppie di parole chiave

N.B.: Le immagini di questa sezione fanno riferimento all'interfaccia di una precedente versione di **T-LAB**. In **T-LAB 10** l'aspetto è leggermente diverso: a) il **tasto destro** sulle tabelle con le parole chiave rende disponibili opzioni supplementari; b) disponibile un nuovo **diagramma radiale** che consente di verificare rapidamente le differenze tra le associazioni di parole; c) una galleria di immagini funziona come un menu aggiuntivo e consente di passare da un output all'altro con un solo clic.

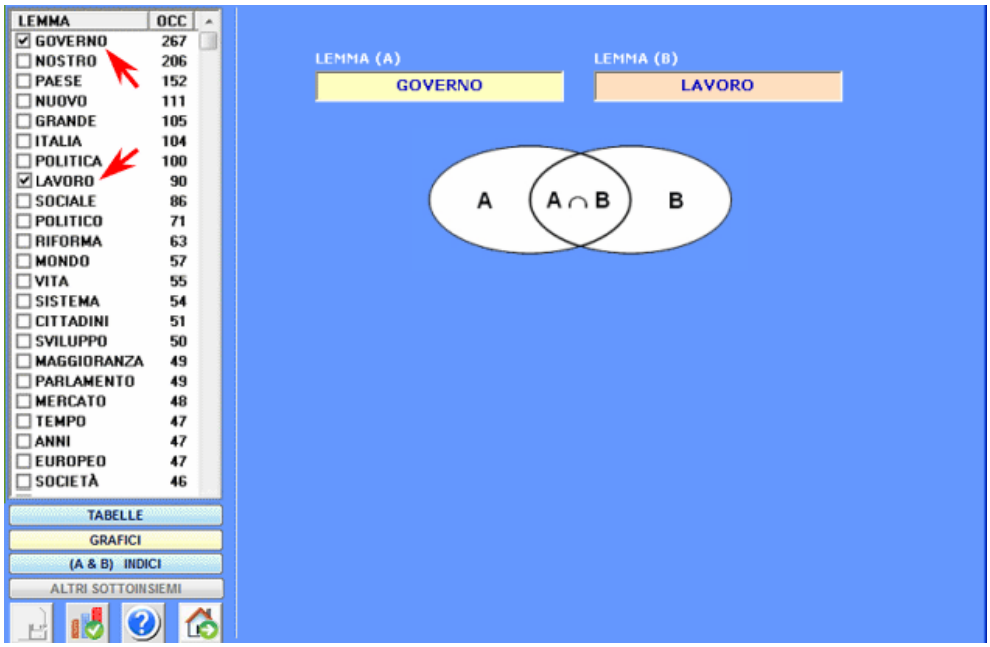
Alcune di queste nuove funzionalità sono evidenziate nell'immagine seguente.



Questo strumento **T-LAB** consente di confrontare insiemi di **contesti elementari** (cioè contesti di co-occorrenza) in cui sono presenti gli elementi di una coppia di **parole chiave**.

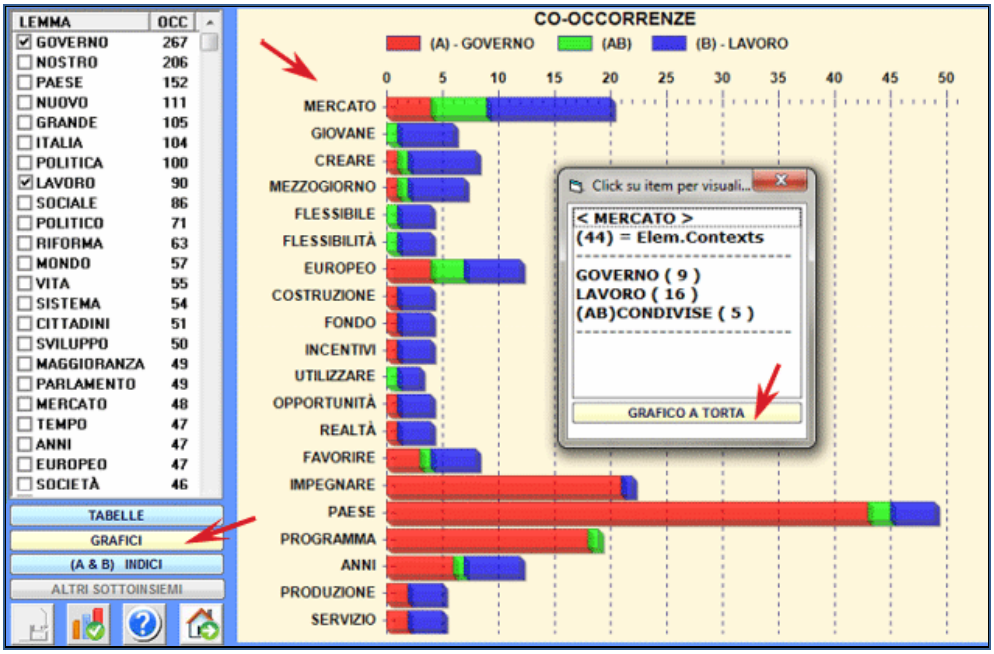
Sulla sinistra è riportata la tabella con la lista dei **lemmi** selezionati e i corrispondenti valori di **occorrenza** nel **corpus** o in un suo **sottoinsieme**.

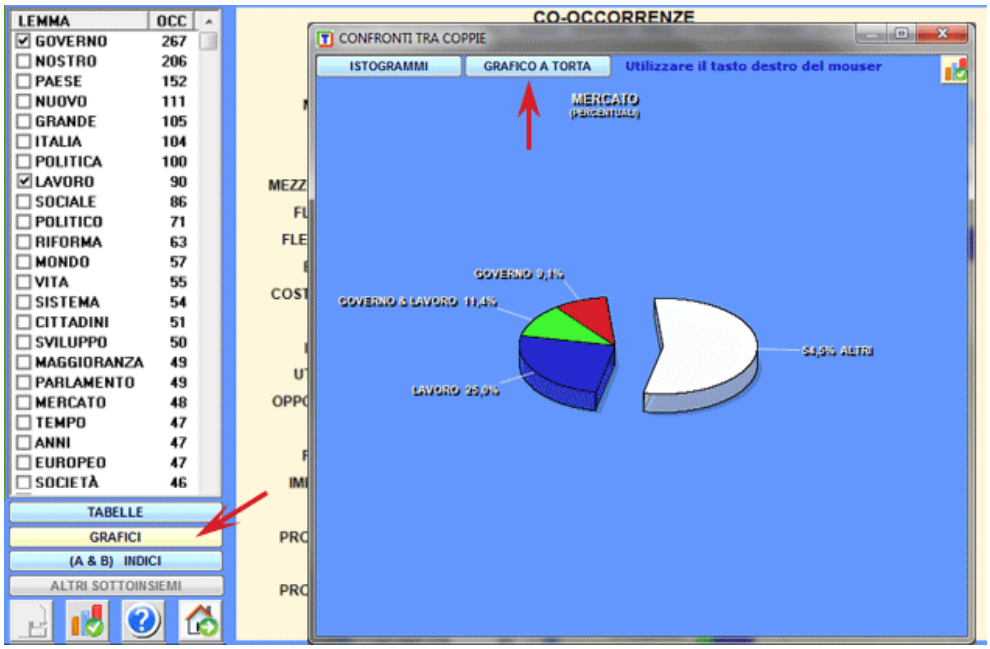
L'utilizzatore, con un semplice click, è invitato a selezionare - uno dopo l'altro - due di essi (una "coppia").

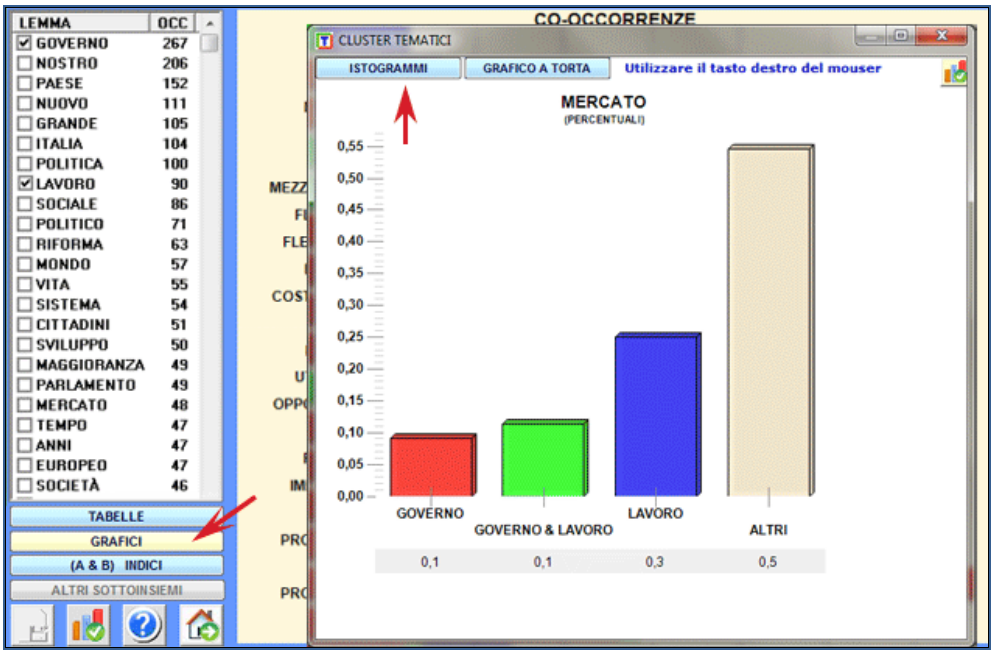


Degli istogrammi (vedi sotto) consentono di meglio apprezzare la quantità dei contesti elementari in cui ogni lemma è in relazione di co-occorrenza con la parola chiave "A" (colore rosso), con la parola chiave "B" (colore blu) o con entrambe (AB: colore verde).

Con un doppio click su ogni label del grafico è possibile verificare i rispettivi valori di co-occorrenza e ottenere grafici a torta e istogrammi (vedi sotto).





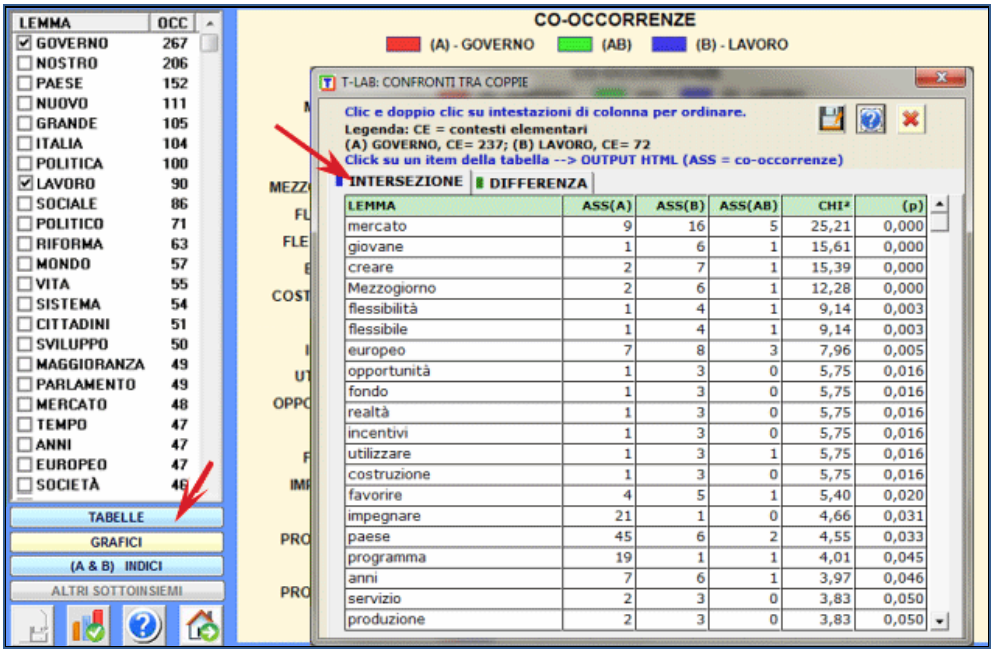


I confronti proposti da **T-LAB** riguardano le relazioni tra gli elementi della "coppia" e ciascuna delle parole contenute nella tabella (vedi sotto).

Siano:
A = insieme dei **contesti elementari** (TOT. C.E. = 237) in cui è presente la prima parola della coppia (es. "governo");
B = insieme dei contesti elementari (TOT. C.E. = 72) in cui è presente la seconda parola della coppia (es. "lavoro").

Il primo tipo di confronto concerne le **associazioni condivise** (vedi pulsante "**intersezione**"), cioè un confronto che prende in considerazione le parole che sono presenti sia in "A" che in "B".

Nella tabella di output ogni riga riporta i valori corrispondenti ai confronti di ciascun lemma.



LEMMA	ASS(A)	ASS(B)	ASS(AB)	CHI²	(p)
mercato	9	16	5	25,21	0,000
giovane	1	6	1	15,61	0,000
creare	2	7	1	15,39	0,000
Mezzogiorno	2	6	1	12,28	0,000
flessibilità	1	4	1	9,14	0,003
flessibile	1	4	1	9,14	0,003
europeo	7	8	3	7,96	0,005
opportunità	1	3	0	5,75	0,016
fondo	1	3	0	5,75	0,016
realtà	1	3	0	5,75	0,016
incentivi	1	3	0	5,75	0,016
utilizzare	1	3	1	5,75	0,016
costruzione	1	3	0	5,75	0,016
favorire	4	5	1	5,40	0,020
impegnare	21	1	0	4,66	0,031
paese	45	6	2	4,55	0,033
programma	19	1	1	4,01	0,045
anni	7	6	1	3,97	0,046
servizio	2	3	0	3,83	0,050
produzione	2	3	0	3,83	0,050

Le chiavi di lettura sono le seguenti:

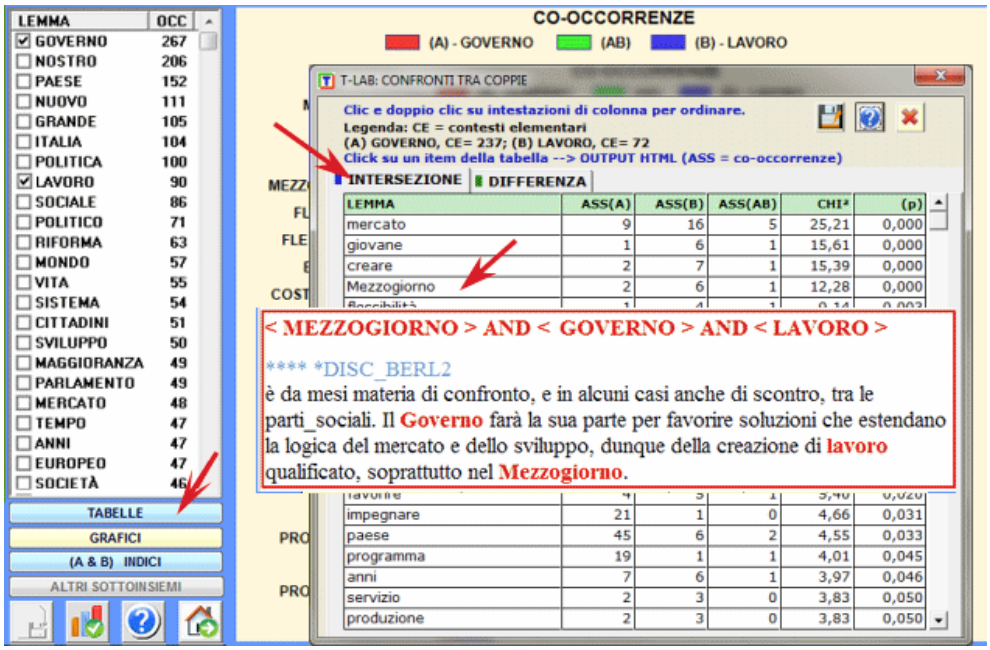
- **ASS (A)** = quantità di contesti elementari in cui ogni lemma è presente (co-occorrenze) in (A);
- **ASS (B)** = quantità di contesti elementari in cui ogni lemma è presente in (B);
- **ASS (AB)** = quantità di contesti elementari in cui ogni lemma è presente sia in (A) che in (B);
- **CHI2** = valori del chi-quadrato;
- (p) = probabilità associata al valore del chi quadrato.

In questo caso, per ogni parola chiave (es. "mercato") **T-LAB** costruisce una tabella come la seguente e applica ad essa il test del **CHI quadro**:

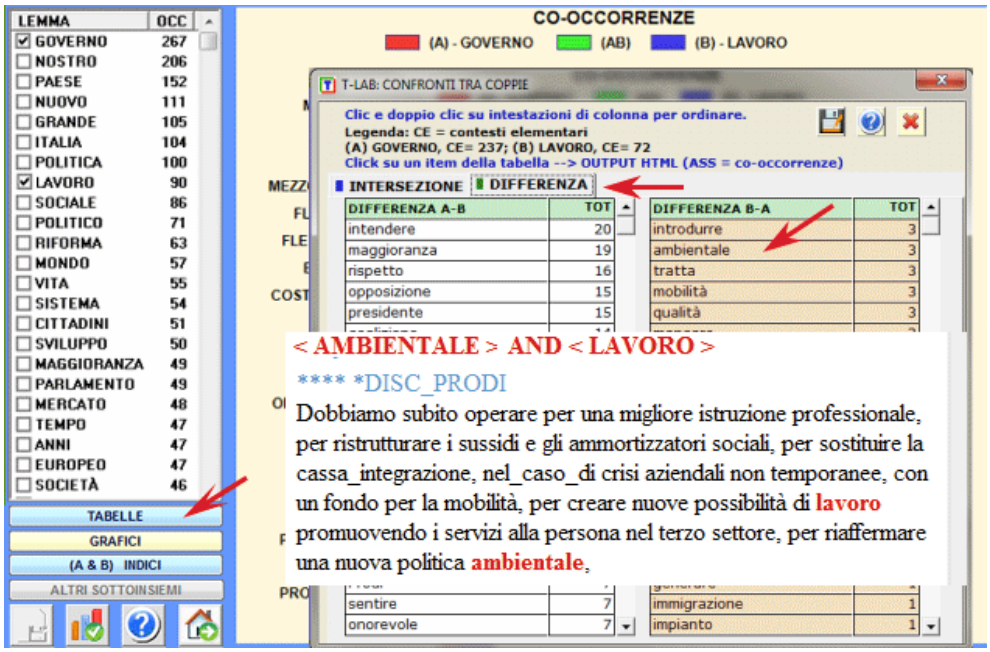
	ASSOC.	NON ASSOC.	TOT.
A	9	228	237
B	16	56	72
	25	284	309

Nella riga (A) sono indicate le quantità di contesti elementari in cui "mercato" è presente (9) o assente (228) nell'insieme dei contesti (237) propri della prima parola della coppia ("Governo").
Nella riga (B) sono indicate le quantità di contesti elementari in cui "mercato" è presente (16) o assente (56) nell'insieme dei contesti (72) propri della seconda parola della coppi ("Lavoro"). N.B.: In questo caso, il valore del CHI quadro è pari a 25.21.

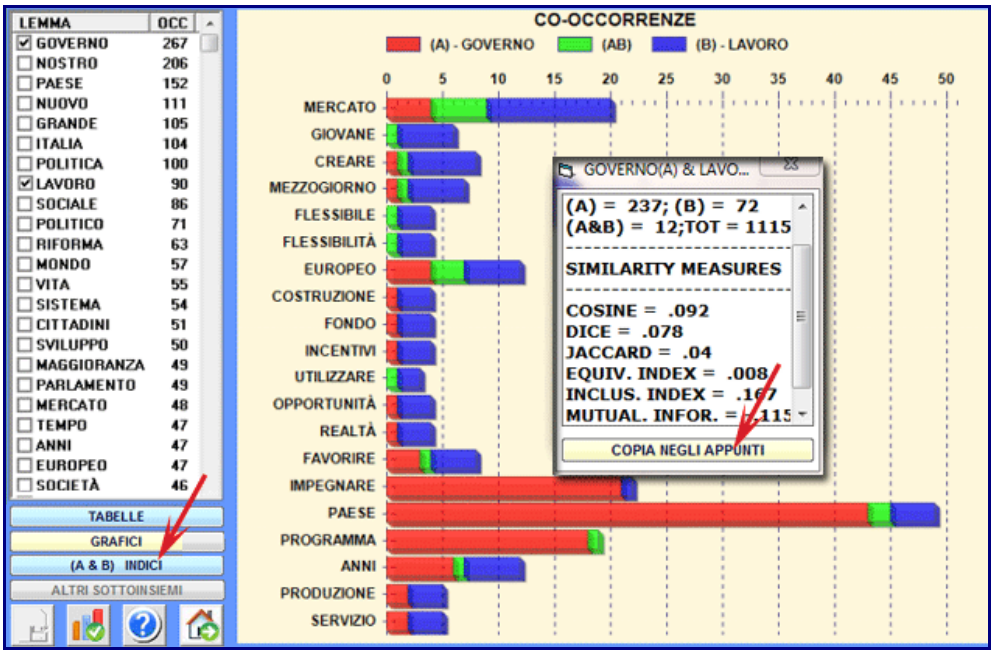
Inoltre, con un doppio click sugli item della tabella output (vedi sotto), è possibile visualizzare un file HTML con gli elementi delle colonne ASS(A), ASS(B) e ASS(AB).



Il secondo tipo di confronto concerne le **associazioni esclusive** (vedi pulsante "**differenza**"), cioè un confronto tra le parole che sono presenti solo in "A" o solo in "B".
In questo caso **T-LAB** propone due tabelle con le parole chiave che in modo esclusivo sono associate al primo "o" al secondo termine della coppia. In entrambe le tabelle, la colonna "TOT" indica la quantità di contesti elementari in cui ogni lemma è associato solo con uno dei due termini della coppia.



Infine, cliccando i pulsanti appropriati (vedi immagine seguente) è possibile verificare ed esportare tutti gli indici di similarità concernenti ogni coppia di parole analizzata.


LEMMA
OCC
GOVERNO 267
NOSTRO 206
PAESE 152
NUOVO 111
GRANDE 105
ITALIA 104
POLITICA 100
LAVORO 90
SOCIALE 86
POLITICO 71
RIFORMA 63
MONDO 57
VITA 55
SISTEMA 54
CITTADINI 51
SVILUPPO 50
MAGGIORANZA 49
PARLAMENTO 49
MERCATO 48
TEMPO 47
ANNI 47
EUROPEO 47
SOCIETÀ 46
TABELLE
GRAFICI
(A & B) INDICI
ALTRI SOTTOINSIEMI
CO-OCCORRENZE
(A) - GOVERNO
(AB)
(B) - LAVORO
0 5 10 15 20 25 30 35 40 45 50
MERCATO
GIOVANE
CREARE
MEZZOGIORNO
FLESSIBILE
FLESSIBILITÀ
EUROPEO
COSTRUZIONE
FONDO
INCENTIVI
UTILIZZARE
OPPORTUNITÀ
REALTÀ
FAVORIRE
IMPEGNARE
PAESE
PROGRAMMA
ANNI
PRODUZIONE
SERVIZIO
GOVERNO(A) & LAVO...
(A) = 237; (B) = 72
(A&B) = 12;TOT = 1115
SIMILARITY MEASURES
COSINE = .092
DICE = .078
JACCARD = .04
EQUIV. INDEX = .008
INCLUS. INDEX = .167
MUTUAL. INFOR. =
COPIA NEGLI APPUNTI

Analisi delle Sequenze e Network Analysis

Questo strumento **T-LAB** tiene conto delle **posizioni** delle varie unità lessicali all'interno delle frasi e ci permette di rappresentare ed esplorare qualsiasi testo come una **rete** di relazioni.

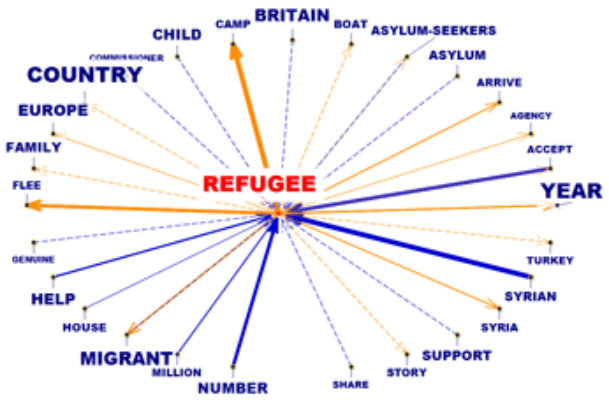
Le varie opzioni disponibili possono essere utilizzate per scopi quali Co-Word Analysis, Analisi Tematiche e Disambiguazioni.

Infatti, dopo aver costruito duce matrici in cui sono registrate tutte le coppie di predecessori e successori, **T-LAB** calcola le **probabilità di transizione** (catene di Markov) e fornisce vari output concernenti le parole target.

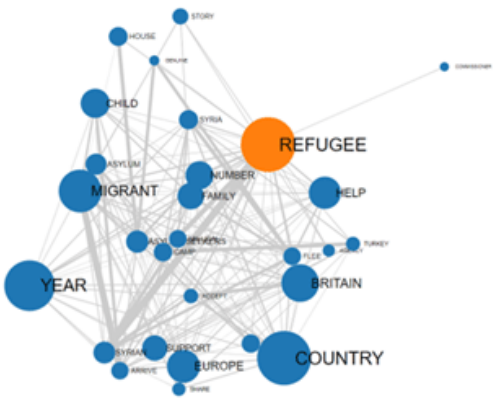
Inoltre, è possibile eseguire una **cluster analysis** ed esplorare le relazioni semantiche tra le parole sia all'interno dell'intera rete che all'interno di 'cluster tematici' (N.B.: In questo caso, l'algoritmo di clusterizzazione è costituito dal 'Louvain method' sviluppato da Blondel V.D., Guillame J.-L , Lambiotte R., Lefebre E., 2008. E, in T-LAB, la tabella input è costituita da links ‘directed’ e ‘weighted’).

Ciò significa, dopo aver eseguito questo tipo di analisi, l'utilizzatore può verificare le relazioni tra i nodi della rete (cioè le parole chiave) a diversi livelli: a) in relazioni del tipo uno-a-uno; b) all'interno di 'ego network'; c) all'interno delle 'comunità' a cui appartengono; d) all'interno dell'intera rete costituita dal testo in analisi.

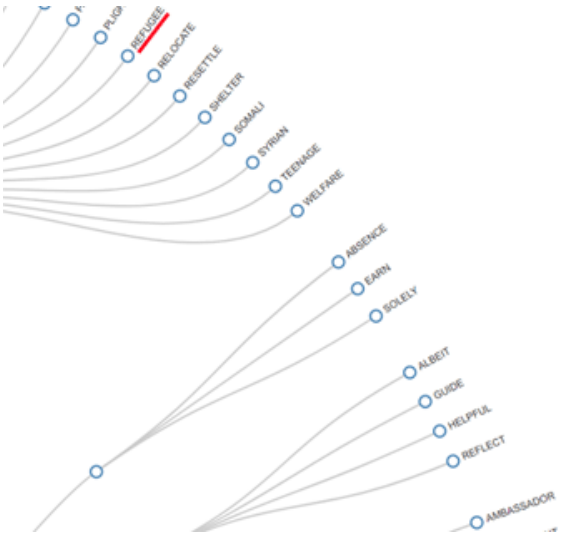
RELAZIONI DEL TIPO UNO-AD-UNO



EGO-NETWORK



COMUNITA' **INTERA RETE**



Le informazioni su come utilizzare le varie opzioni di analisi sono organizzate in tre sezioni:

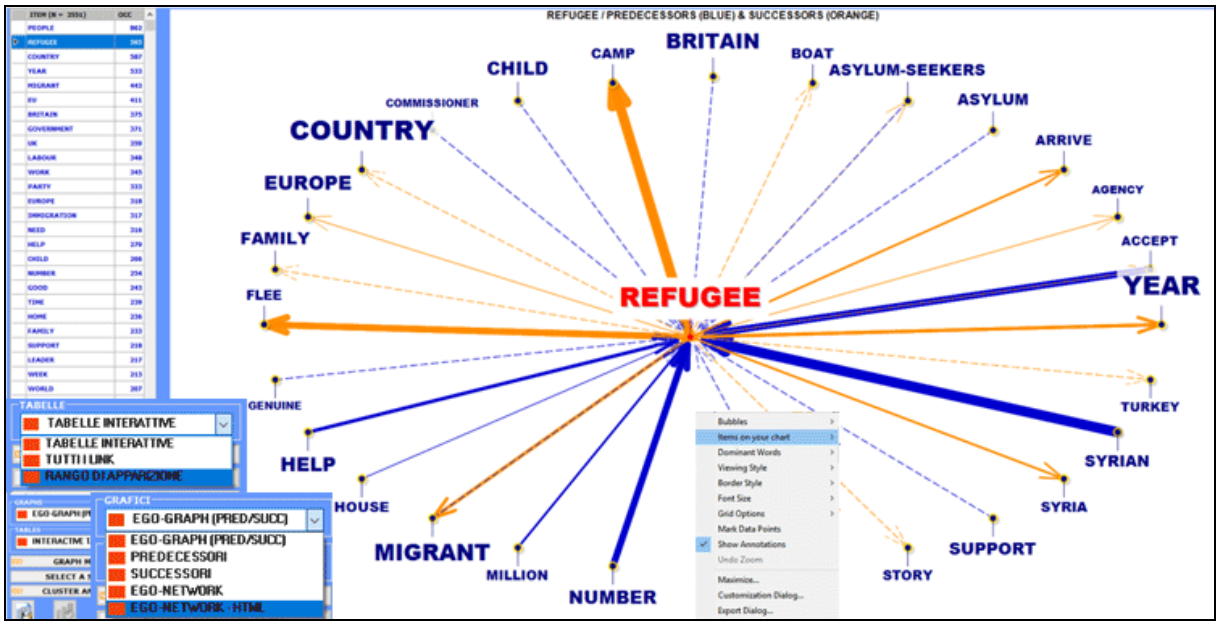
A - Esplorare le connessioni del tipo uno-a-uno e le "ego network";
B - Esplorare le 'comunità' (cioè i cluster tematici) e l'intera rete;
C - Alcuni dettagli tecnici.

N.B.: Per motivi editoriali, questa pagina include esempi di analisi tratti da un corpus i cui testi sono in lingua Inglese.

A - ESPLORARE LE CONNESSIONI DEL TIPO UNO-A-UNO E LE "EGO NETWORK"

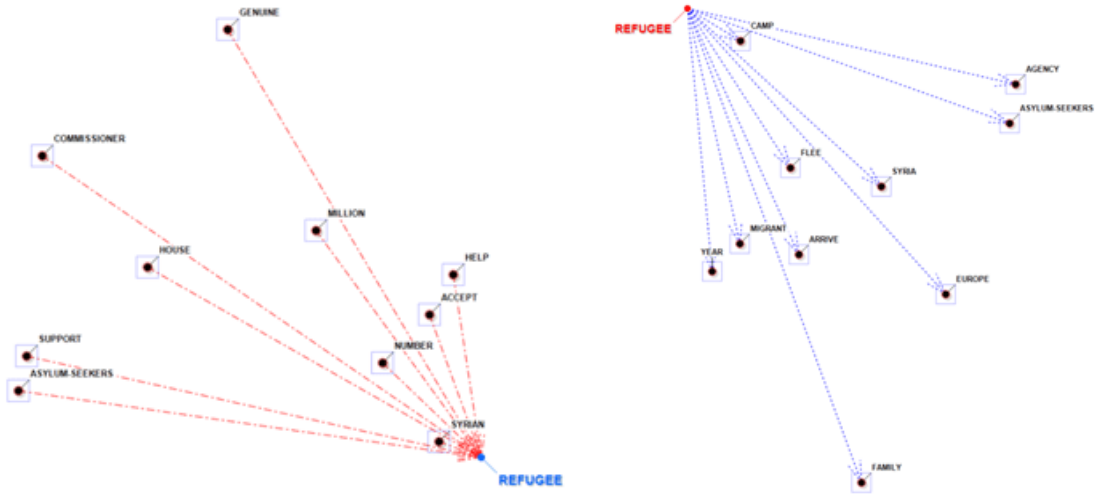
Quando l'analisi automatica è terminata, sono disponibili diversi grafici e tabelle che consentono di verificare le relazioni e i dati concernenti le parole chiave selezionate (N.B.: A questo scopo è sufficiente un clic su un item delle tabelle o su un qualsiasi punto mostrato nei grafici).

Tutti i **grafici** possono essere personalizzati ed esportati in diversi formati (usare il pulsante destro del mouse).

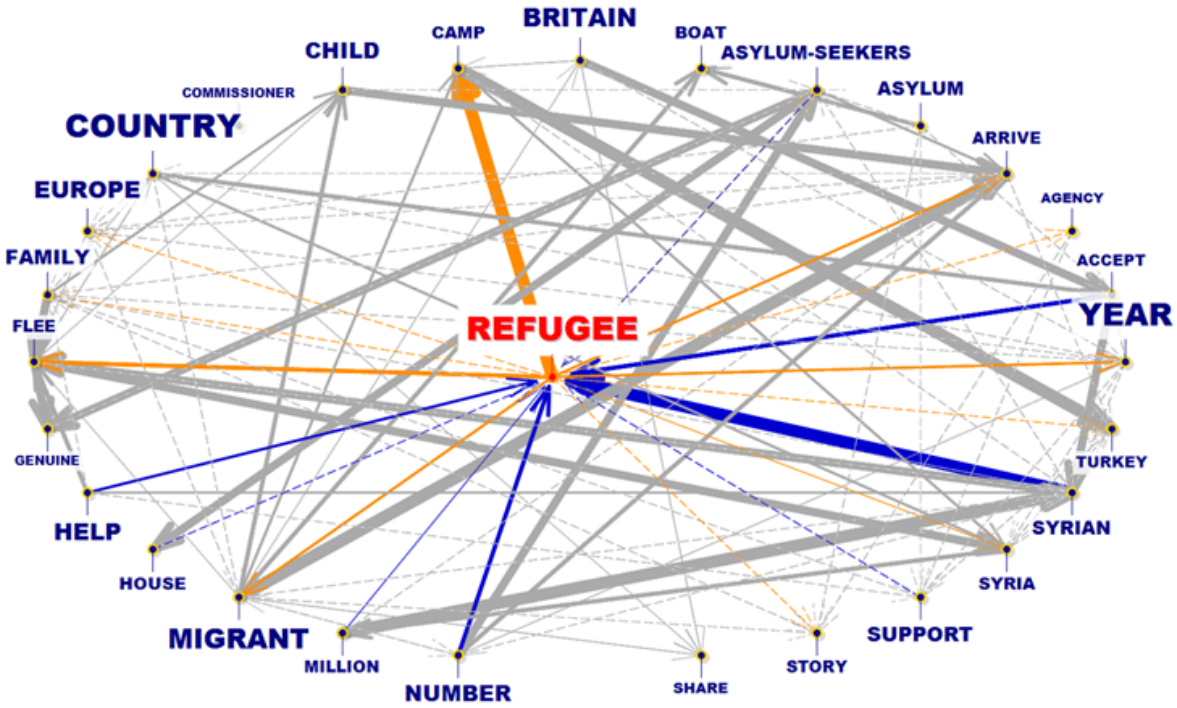


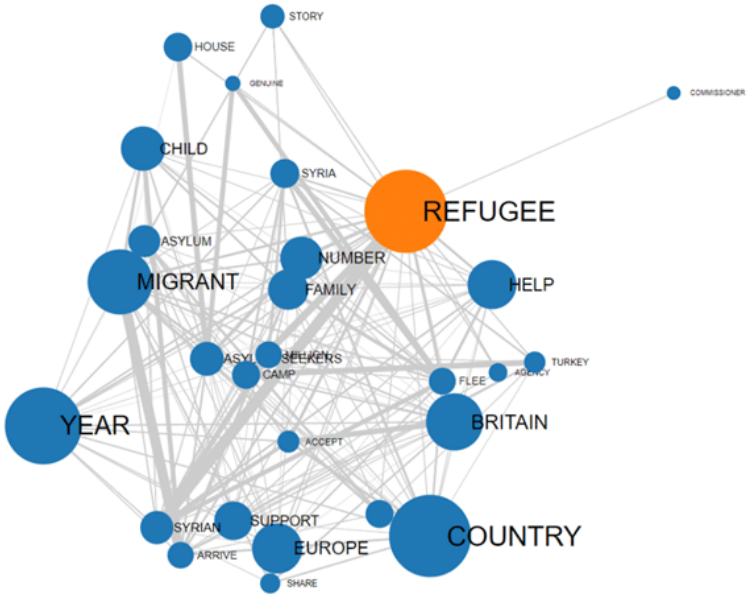
In due dei grafici gli elementi più vicini a quelli selezionati sono quelli che hanno le probabilità più elevate di venire prima (predecessori) o dopo (successori) di essi.

PREDECESSORI **SUCCESSORI**



Negli altri casi, la vicinanza tra i termini-chiave è rappresentata mediante i diversi spessori delle frecce che li connettono (vedi immagini seguenti).



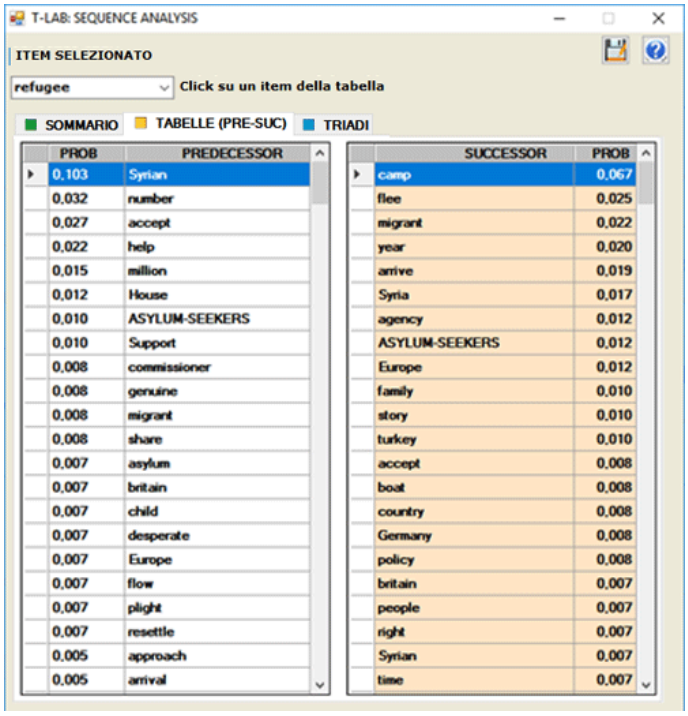


Tutti i dati possono essere verificati tramite vari tipi di **tabelle**.

Nel dettaglio:

Le **TABELLE INTERATTIVE** mostrano le liste dei predecessori e dei successori associati con le parole chiave selezionate.

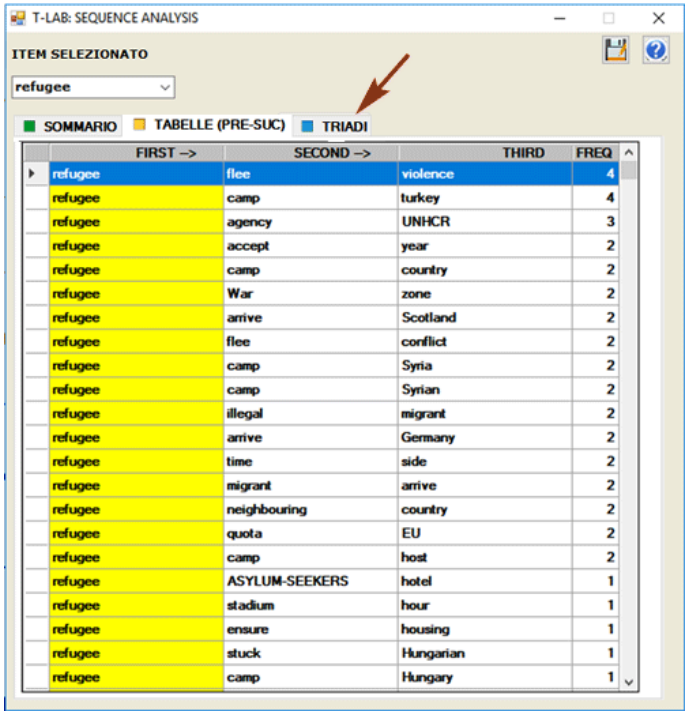
L'ordinamento è di tipo decrescente sui valori di probabilità ("PROB"). Ad esempio, nella tabella seguente, la probabilità che "camp" segua "refugee" è 0.067, ovvero pari al 6.7%.



PROB	PREDECESSOR
0,103	Syrian
0,032	number
0,027	accept
0,022	help
0,015	million
0,012	House
0,010	ASYLUM-SEEKERS
0,010	Support
0,008	commissioner
0,008	genuine
0,008	migrant
0,008	share
0,007	asylum
0,007	britain
0,007	child
0,007	desperate
0,007	Europe
0,007	flow
0,007	plight
0,007	resettle
0,005	approach
0,005	arrival

SUCCESSOR	PROB
camp	0,067
flee	0,025
migrant	0,022
year	0,020
arrive	0,019
Syria	0,017
agency	0,012
ASYLUM-SEEKERS	0,012
Europe	0,012
family	0,010
story	0,010
turkey	0,010
accept	0,008
boat	0,008
country	0,008
Germany	0,008
policy	0,008
britain	0,007
people	0,007
right	0,007
Syrian	0,007
time	0,007

L'opzione **TRIADI** consente di visualizzare alcune tabelle con sequenze di tre elementi in cui il lemma selezionato è in prima, seconda o terza posizione. Per ciascuna triade **T-LAB** riporta le corrispondenti occorrenze. (N.B.: All'interno delle triadi, le parole vuote non sono incluse).

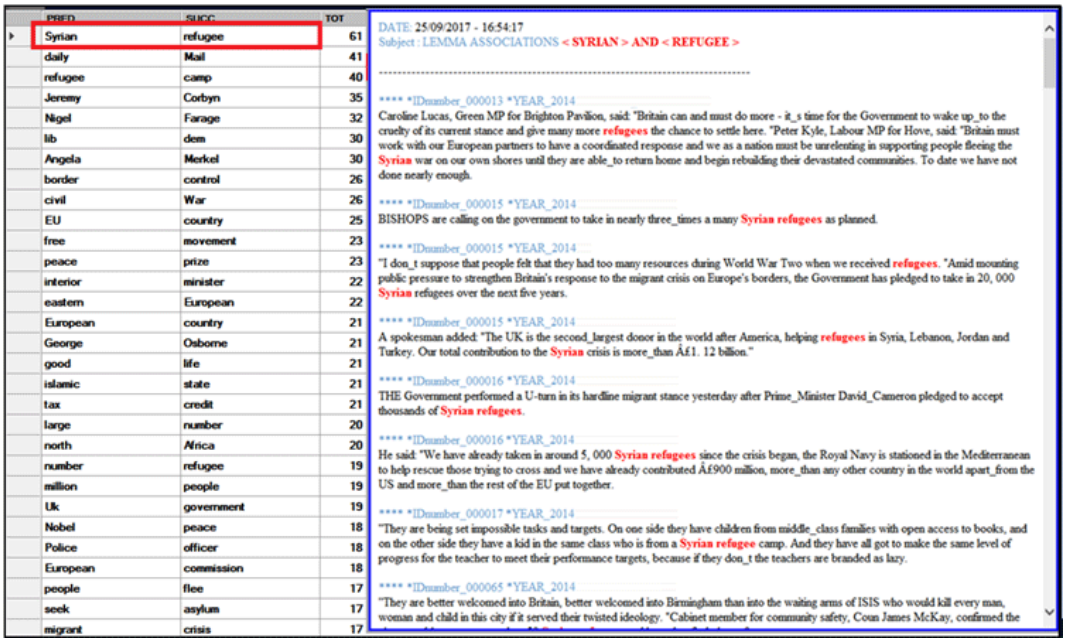


FIRST -->	SECOND -->	THIRD	FREQ
refugee	flee	violence	4
refugee	camp	turkey	4
refugee	agency	UNHCR	3
refugee	accept	year	2
refugee	camp	country	2
refugee	War	zone	2
refugee	arrive	Scotland	2
refugee	flee	conflict	2
refugee	camp	Syria	2
refugee	camp	Syrian	2
refugee	illegal	migrant	2
refugee	arrive	Germany	2
refugee	time	side	2
refugee	migrant	arrive	2
refugee	neighbouring	country	2
refugee	quota	EU	2
refugee	camp	host	2
refugee	ASYLUM-SEEKERS	hotel	1
refugee	stadium	hour	1
refugee	ensure	housing	1
refugee	stuck	Hungarian	1
refugee	camp	Hungary	1

La tabella **TUTTI I LINK** (vedi sotto), che è particolarmente utile per disambiguare i significati delle parole, contiene tutte le coppie di predecessori e successori, e anche le rispettive occorrenze.
Facendo clic su una riga di questa tabella, tutti i segmenti di testo (cioè contesti elementari) in

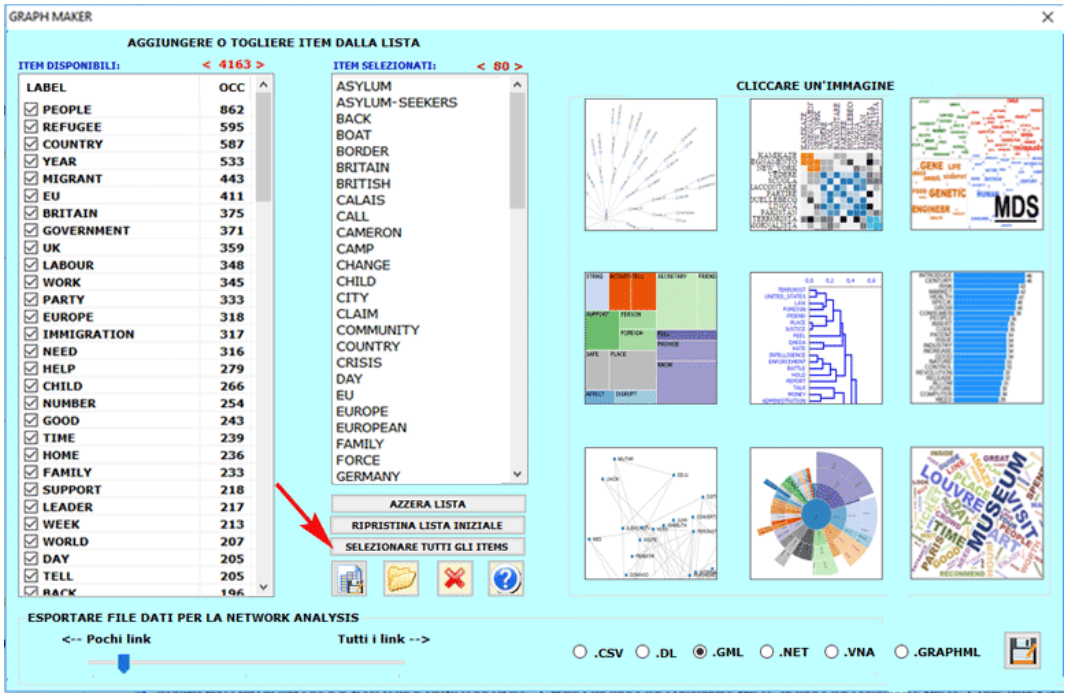
cui i due membri di ciascuna coppia sono presenti allo stesso tempo (cioè co-occorrenze) verranno visualizzati in formato HTML sul lato destro della tabella.

.



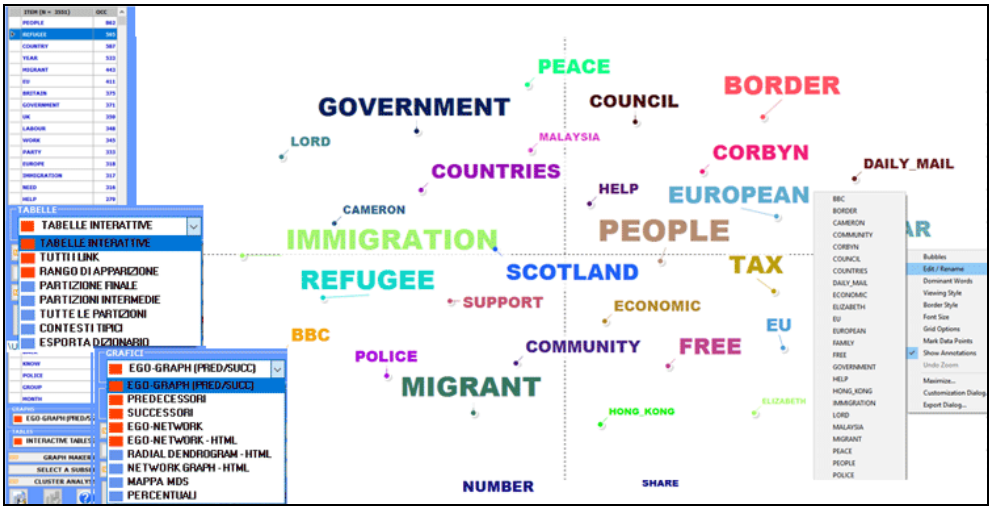
La tabella **RANGO DI APPARIZIONE**, con la frequenza e l'ordine medio di apparizione (o evocazione) di ogni termine all'interno dei segmenti di testo, viene mostrata solo quando il corpus è costituito da brevi testi, ad esempio risposte a domande aperte.

In qualsiasi momento, facendo clic sull'opzione **GRAPH MAKER**, l'utente può creare diversi tipi di grafici utilizzando elenchi personalizzati di parole chiave (vedi sotto).
N.B.: Gli utenti esperti che sono interessati ad esportare file in diversi formati (e.g., dl .gml .vna .graphml) con i dati relativi a tutti i link, possono fare clic sul pulsante 'SELEZIONARE TUTTI GLI ITEMS'.



B - ESPLORARE LE 'COMUNITÀ' (CIOÈ I CLUSTER TEMATICI) E L'INTERA RETE

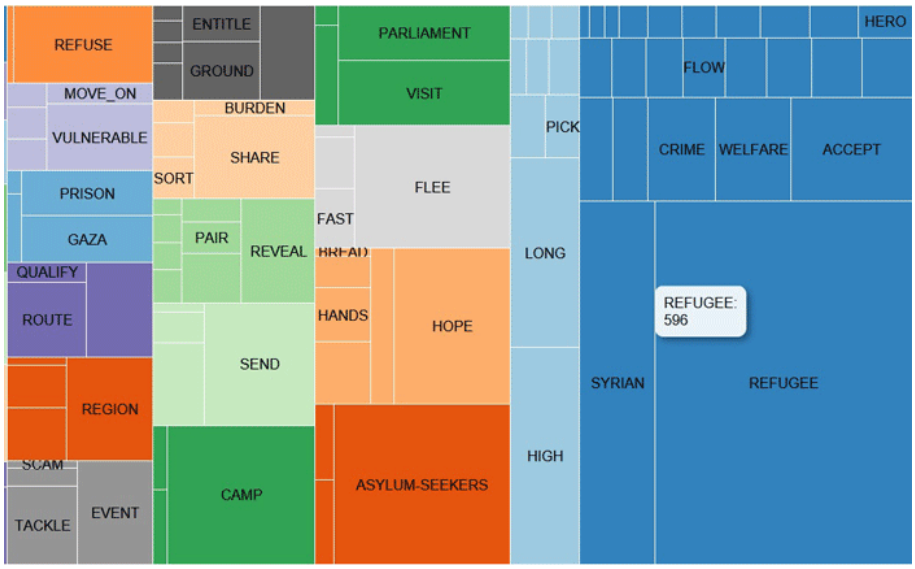
Quando si effettua un'analisi cluster, vengono resi disponibili **ulteriori grafici e tabelle** che consentono di esplorare tutti i livelli interni alla rete analizzata (vedi sotto gli item contrassegnati con un piccolo rettangolo in colore blu) .



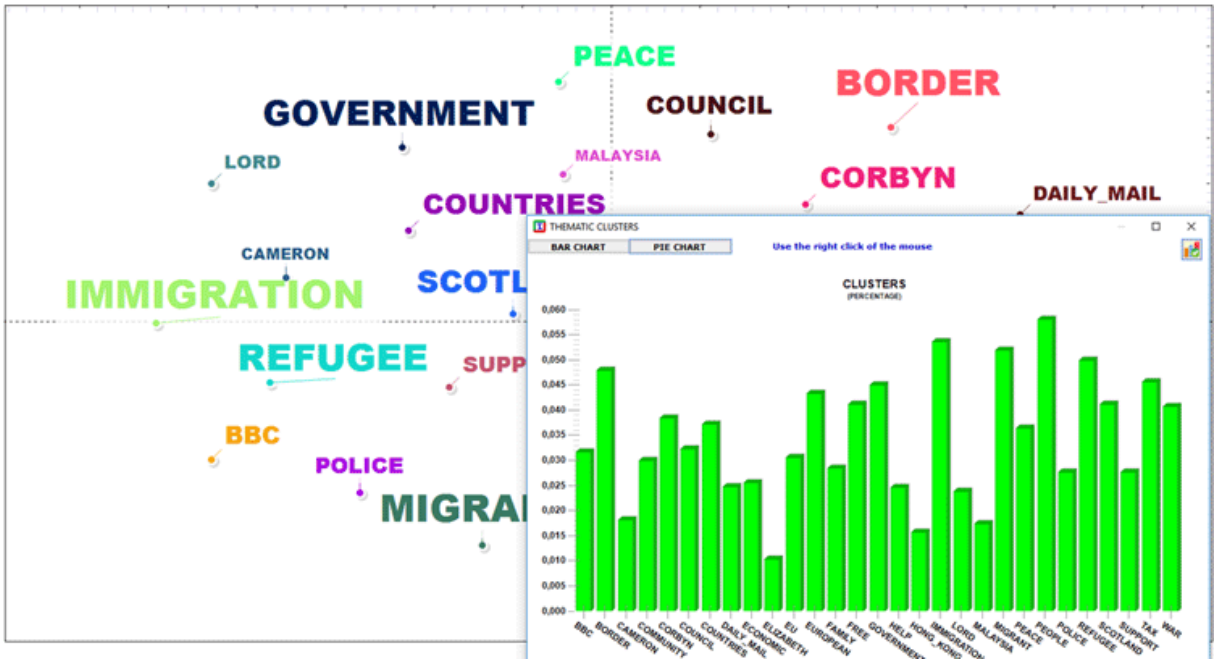
Una prima tabella riassume le caratteristiche (cioè i termini chiave) della **PARTIZIONE FINALE** ottenuta dall'algoritmo di clusterizzazione.
In tale tabella, le caratteristiche di ciascun cluster tematico sono ordinate mediante il relativo valore **TF-IDF** (vedi sotto).
N.B.: Quando un cluster della partizione finale include solo due parole, di solito questo significa che un caso di multiword non è stato risolto durante la fase di pre-trattamento dei dati).

10_REFUGEE	TF-IDF_10	11_NICK	TF-IDF_11	12_KONG	TF-IDF_12	14_MIGRANT	TF-IDF_14
REFUGEE	692,605	NICK	50,809	KONG	45,461	MIGRANT	203,235
SYRIAN	288,808	CLEGG	45,461	HONG	42,786	MINISTER	112,314
CAMP	187,190	CAMERON	34,764	CHARGE	37,438	BOAT	101,618
ASYLUM-SEEKERS	101,618	FOOTBALL	26,741	NETWORK	34,764	CHANGE	90,921
FLEE	96,269	CAR	24,067	TRAFFIC	29,416	CLAIM	90,921
ACCEPT	90,921	CASE	21,393	VIOLENCE	29,416	RESCUE	74,876
SCHEME	61,505	LEGACY	21,393	FARM	29,416	INTERIOR	66,854
HIGH	53,483	PRIME_MINISTER	21,393	FOOD	29,416	SMALL	64,180
SHARE	45,461	THOUGHT	18,719	INDUSTRY	26,741	BUSINESS	64,180
REFUSE	45,461	UNHAPPY	16,045	VICTIM	26,741	BENEFIT	58,831
RESETTLEMENT	42,786	RECALL	16,045	DOMESTIC	24,067	ROMANIAN	58,831
VULNERABLE	42,786	SERIOUS	16,045	INFRASTRUCTURE	21,393	ITALIAN	56,157
RESETTLE	40,112	HIT	13,371	ABUSE	21,393	MILLION	50,809
COMMISSIONER	37,438	MATCH	13,371	SMUGGLE	21,393	WORKER	50,809
HOPE	34,764	BELIEVE	13,371	SPENCER	21,393	NAVY	48,135
PERIOD	34,764	FAN	13,371	TERMS	18,719	BULGARIAN	48,135
RELOCATION	32,090	DELIGHT	13,371	MARKS	18,719	FISH	48,135
SEND	32,090	DEVOTE	10,697	PRODUCTION	18,719	SHIP	45,461
HOST	32,090	FEDERATION	10,697	SEXUAL	18,719	VESSEL	42,786
CURRENTLY	32,090	BLAIR	10,697	BRISTOL	18,719	CLIMATE	40,112
EVENT	32,090	BOMBER	10,697	BOOST	16,045	LIFE	37,438
UNHCR	32,090	ABSOLUTELY	10,697	CAMPAIGN	16,045	LAUNCH	37,438
CRIME	29,416	ACQUIRE	10,697	HOSPITALITY	16,045	ROYAL	34,764
LONG	26,741	MOUTH	10,697	WAIT	16,045	EXAMPLE	34,764
VISIT	26,741	HONOUR	10,697	PREPARATION	13,371	CONTRIBUTE	32,090
MAIN	24,067	ROBINSON	10,697	BREATH	13,371	PORT	32,090
REGION	24,067	THREAT	10,697	COAT	13,371	TAXPAYER	32,090
REFLECT	21,393	SUICIDE	10,697	AGRICULTURAL	13,371	SKILLED	29,416
PRISON	21,393	SURE	10,697	BRAVE	10,697	AFRICAN	29,416
PROGRAM	21,393	POOR	10,697	COMPETITION	10,697	APPLY	26,741
PAIR	21,393	WIFE	10,697	CHARACTER	10,697	AUSTRALIA	26,741
TRAFFICKER	21,393	TERRIFY	8,022	GLASS	10,697	CABINET	26,741
SHELTER	21,393	TRANSPORT	8,022	EXPORT	10,697	COASTGUARD	26,741

Facendo clic su una qualunque parola della tabella PARTIZIONE FINALE (così come della tabella TUTTE LE PARTITIONI), un grafico dinamico del tipo TreeMap ci consente di verificare le 'comunità' a cui essa risulta appartenere (vedi sotto).

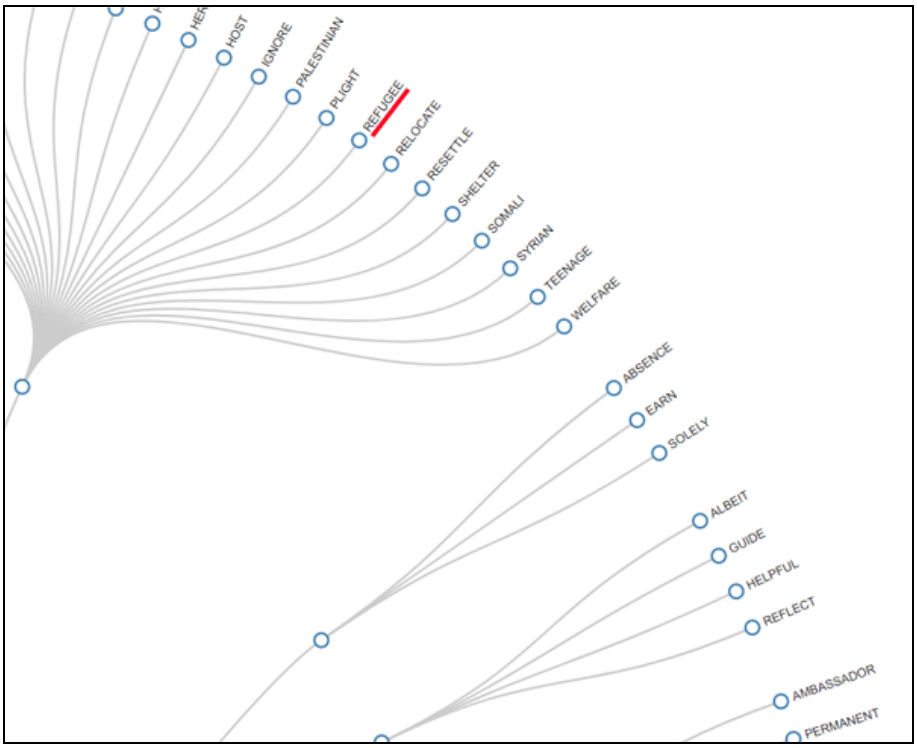


La **MAPPA MDS** e il grafico **PERCENTUALI** (vedi sotto) ci permettono di verificare il 'peso' di ciascun cluster, così come le relazioni tra i vari cluster all'interno della partizione finale (vedi sotto).

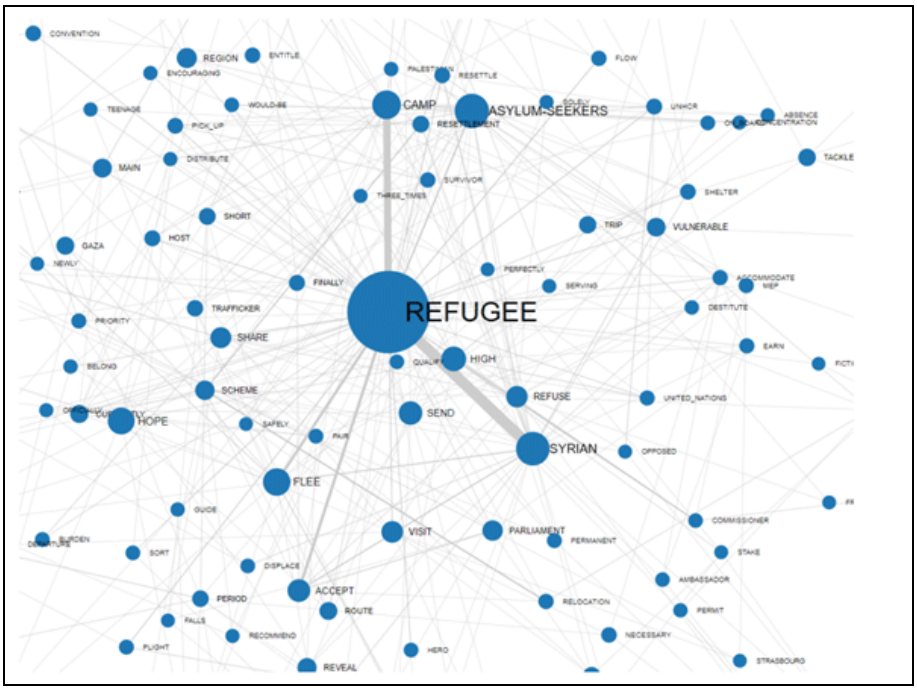


A seconda del numero di parole chiave, due grafici in formato HTML ci permettono di verificare le loro relazioni sia all'interno dell'intera rete che all'interno del cluster a cui appartengono (vedi sotto).

DENDROGRAMMA RADIALE



NETWORK (FORCE-DIRECTED GRAPH)

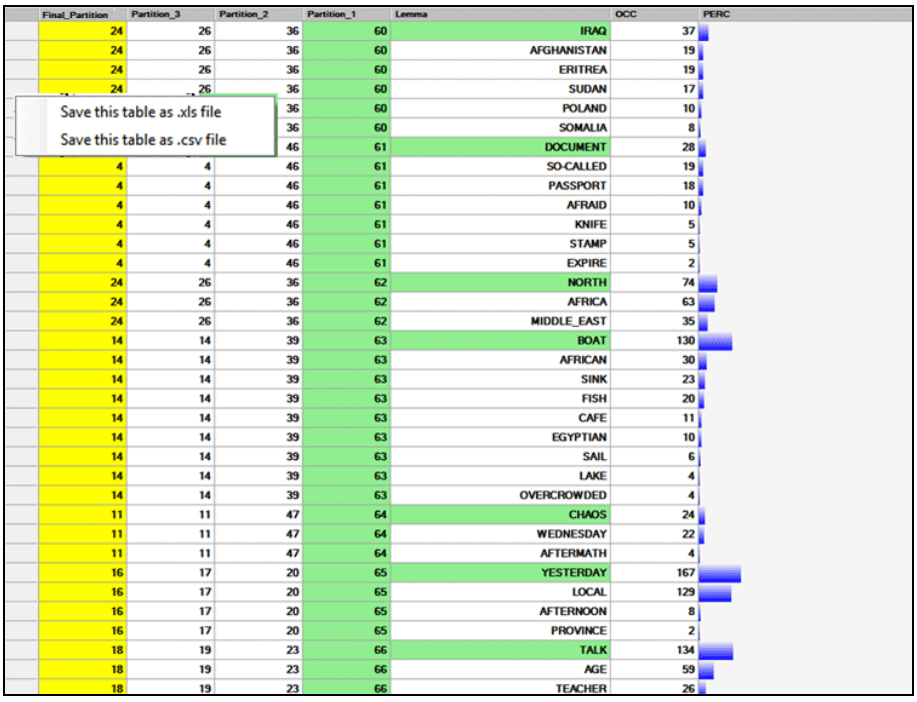


Altre tre tabelle ci forniscono ulteriori informazioni ottenute tramite la cluster analysis.

In particolare:

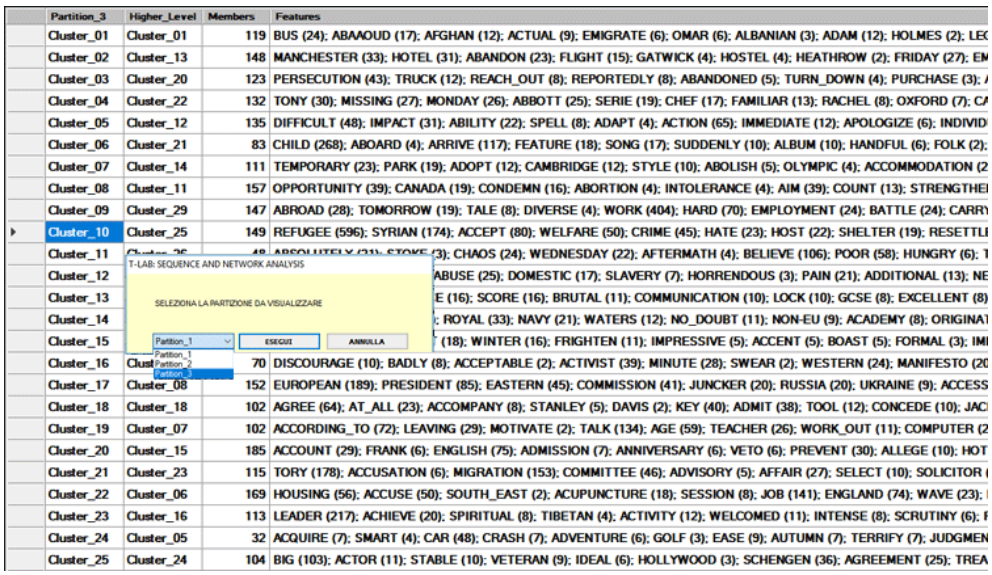
La tabella **TUTTE LE PARTIZIONI** consente di verificare come le parole chiave sono state raggruppate in ciascuna partizione della cluster analisi (vedi immagine a seguire, dove i

numeri nelle colonne partitioni si riferiscono ai vari cluster).
N.B.: Per impostazione predefinita, questa tabella viene presentata ordinata sulla prima partizione (cioè quella con il maggior numero di cluster), e ogni passaggio da un piccolo cluster all'altro è marcato evidenziando in verde la prima parola che ad esso appartiene.



Final_Partition	Partition_3	Partition_2	Partition_1	Lemma	OCC	PERC
24	26	36	60	IRAQ	37	
24	26	36	60	AFGHANISTAN	19	
24	26	36	60	ERITREA	19	
24	26	36	60	SUDAN	17	
		36	60	POLAND	10	
		36	60	SOMALIA	8	
		46	61	DOCUMENT	28	
4	4	46	61	SO-CALLED	19	
4	4	46	61	PASSPORT	18	
4	4	46	61	AFRAID	10	
4	4	46	61	KNIFE	5	
4	4	46	61	STAMP	5	
4	4	46	61	EXPIRE	2	
24	26	36	62	NORTH	74	
24	26	36	62	AFRICA	63	
24	26	36	62	MIDDLE_EAST	35	
14	14	39	63	BOAT	130	
14	14	39	63	AFRICAN	30	
14	14	39	63	SINK	23	
14	14	39	63	FISH	20	
14	14	39	63	CAFE	11	
14	14	39	63	EGYPTIAN	10	
14	14	39	63	SAIL	6	
14	14	39	63	LAKE	4	
14	14	39	63	OVERCROWDED	4	
11	11	47	64	CHAOS	24	
11	11	47	64	WEDNESDAY	22	
11	11	47	64	AFTERMATH	4	
16	17	20	65	YESTERDAY	167	
16	17	20	65	LOCAL	129	
16	17	20	65	AFTERNOON	8	
16	17	20	65	PROVINCE	2	
18	19	23	66	TALK	134	
18	19	23	66	AGE	59	
18	19	23	66	TEACHER	26	

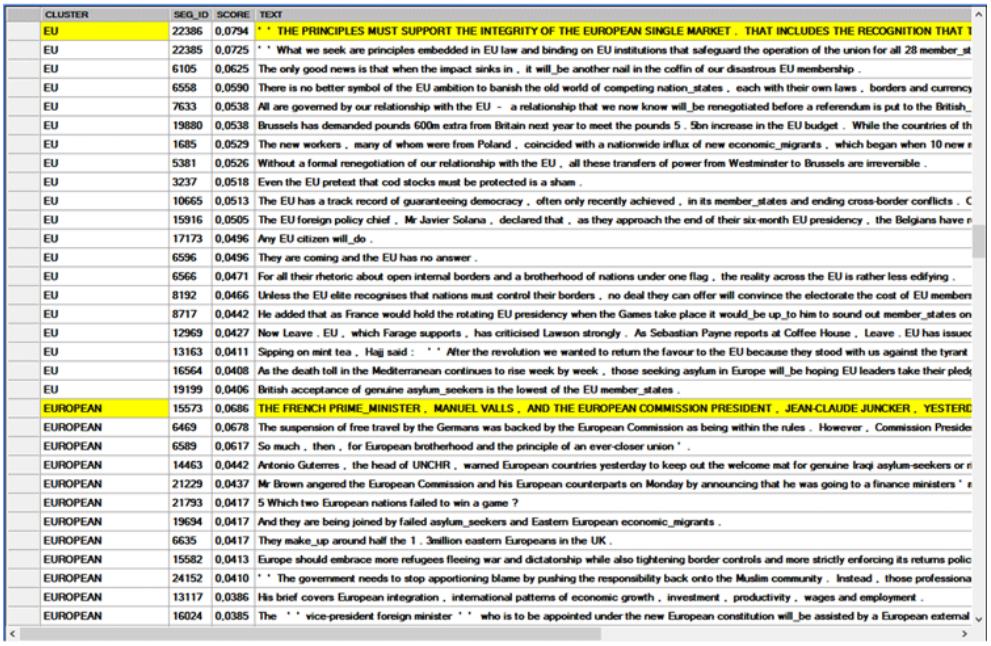
La tabella **PARTIZIONI INTERMEDIE** consente di verificare come le parole-chiave sono state raggruppate all'interno di ogni partizione selezionata. E, di volta in volta, le parole caratteristiche di ogni cluster tematico sono ordinate per i valori decrescenti delle loro occorrenze (vedi sotto).



Partition_3	Higher_Level	Members	Features
Cluster_01	Cluster_01	119	BUS (24); ABAAOUD (17); AFGHAN (12); ACTUAL (9); EMIGRATE (6); OMAR (6); ALBANIAN (3); ADAM (12); HOLMES (2); LE
Cluster_02	Cluster_13	148	MANCHESTER (33); HOTEL (31); ABANDON (23); FLIGHT (15); GATWICK (4); HOSTEL (4); HEATHROW (2); FRIDAY (27); EM
Cluster_03	Cluster_20	123	PERSECUTION (43); TRUCK (12); REACH_OUT (8); REPORTEDLY (8); ABANDONED (5); TURN_DOWN (4); PURCHASE (3);
Cluster_04	Cluster_22	132	TONY (30); MISSING (27); MONDAY (26); ABBOTT (25); SERIE (19); CHEF (17); FAMILIAR (13); RACHEL (8); OXFORD (7); CA
Cluster_05	Cluster_12	135	DIFFICULT (48); IMPACT (31); ABILITY (22); SPELL (8); ADAPT (4); ACTION (65); IMMEDIATE (12); APOLOGIZE (6); INDIVID
Cluster_06	Cluster_21	83	CHILD (268); ABOARD (4); ARRIVE (117); FEATURE (18); SONG (17); SUDDENLY (10); ALBUM (10); HANDFUL (6); FOLK (2);
Cluster_07	Cluster_14	111	TEMPORARY (23); PARK (19); ADOPT (12); CAMBRIDGE (12); STYLE (10); ABOLISH (5); OLYMPIC (4); ACCOMMODATION (2
Cluster_08	Cluster_11	157	OPPORTUNITY (39); CANADA (19); CONDEMN (16); ABORTION (4); INTOLERANCE (4); AIM (39); COUNT (13); STRENGTHE
Cluster_09	Cluster_29	147	ABROAD (28); TOMORROW (19); TALE (8); DIVERSE (4); WORK (404); HARD (70); EMPLOYMENT (24); BATTLE (24); CARR
Cluster_10	Cluster_25	149	REFUGEE (596); SYRIAN (174); ACCEPT (80); WELFARE (50); CRIME (45); HATE (23); HOST (22); SHELTER (19); RESETTL
Cluster_11			(3); CHAOS (24); WEDNESDAY (22); AFTERMATH (4); BELIEVE (106); POOR (58); HUNGRY (6);
Cluster_12			ABUSE (25); DOMESTIC (17); SLAVERY (7); HORRENDOUS (3); PAIN (21); ADDITIONAL (13); NE
Cluster_13			E (16); SCORE (16); BRUTAL (11); COMMUNICATION (10); LOCK (10); GCSE (8); EXCELLENT (8)
Cluster_14			ROYAL (33); NAVY (21); WATERS (12); NO_DOUBT (11); NON-EU (9); ACADEMY (8); ORIGINA
Cluster_15			(18); WINTER (16); FRIGHTEN (11); IMPRESSIVE (5); ACCENT (5); BOAST (5); FORMAL (3); IM
Cluster_16	Clust	70	DISCOURAGE (10); BADLY (8); ACCEPTABLE (2); ACTIVIST (39); MINUTE (28); SWEAR (2); WESTERN (24); MANIFESTO (20
Cluster_17	Cluster_08	152	EUROPEAN (189); PRESIDENT (85); EASTERN (45); COMMISSION (41); JUNCKER (20); RUSSIA (20); UKRAINE (9); ACCESS
Cluster_18	Cluster_18	102	AGREE (64); AT_ALL (23); ACCOMPANY (8); STANLEY (5); DAVIS (2); KEY (40); ADMIT (38); TOOL (12); CONCEDE (10); JAC
Cluster_19	Cluster_07	102	ACCORDING_TO (72); LEAVING (29); MOTIVATE (2); TALK (134); AGE (59); TEACHER (26); WORK_OUT (11); COMPUTER (2
Cluster_20	Cluster_15	185	ACCOUNT (29); FRANK (6); ENGLISH (75); ADMISSION (7); ANNIVERSARY (6); VETO (6); PREVENT (30); ALLEGE (10); HOT
Cluster_21	Cluster_23	115	TORY (178); ACCUSATION (6); MIGRATION (153); COMMITTEE (46); ADVISORY (5); AFFAIR (27); SELECT (10); SOLICITOR
Cluster_22	Cluster_06	169	HOUSING (56); ACCUSE (50); SOUTH_EAST (2); ACUPUNCTURE (18); SESSION (8); JOB (141); ENGLAND (74); WAVE (23);
Cluster_23	Cluster_16	113	LEADER (217); ACHIEVE (20); SPIRITUAL (8); TIBETAN (4); ACTIVITY (12); WELCOMED (11); INTENSE (8); SCRUTINY (6);
Cluster_24	Cluster_05	32	ACQUIRE (7); SMART (4); CAR (48); CRASH (7); ADVENTURE (6); GOLF (3); EASE (9); AUTUMN (7); TERRIFY (7); JUDGMEN
Cluster_25	Cluster_24	104	BIG (103); ACTOR (11); STABLE (10); VETERAN (9); IDEAL (6); HOLLYWOOD (3); SCHENGEN (36); AGREEMENT (25); TREA

La tabella **CONTESTI TIPICI** consente di controllare i segmenti di testo che hanno il più alto punteggio di associazione con i vari cluster della partizione finale. In questa tabella il "punteggio" si riferisce alla somiglianza (indice coseno) tra il vettore delle caratteristiche di ciascun cluster e il vettore in cui viene rappresentato ogni segmento di testo.

N.B. Il segmento di testo più significativo di ciascun cluster è evidenziato in giallo.



	CLUSTER	SEG_ID	SCORE	TEXT
	EU	22386	0,0794	' ' THE PRINCIPLES MUST SUPPORT THE INTEGRITY OF THE EUROPEAN SINGLE MARKET . THAT INCLUDES THE RECOGNITION THAT T
	EU	22385	0,0725	' ' What we seek are principles embedded in EU law and binding on EU institutions that safeguard the operation of the union for all 28 member_st
	EU	6105	0,0625	The only good news is that when the impact sinks in , it will_be another nail in the coffin of our disastrous EU membership .
	EU	6558	0,0590	There is no better symbol of the EU ambition to banish the old world of competing nation_states , each with their own laws , borders and currency
	EU	7633	0,0538	All are governed by our relationship with the EU - a relationship that we now know will_be renegotiated before a referendum is put to the British_
	EU	19880	0,0538	Brussels has demanded pounds 600m extra from Britain next year to meet the pounds 5 . 5bn increase in the EU budget . While the countries of th
	EU	1685	0,0529	The new workers , many of whom were from Poland , coincided with a nationwide influx of new economic_migrants , which began when 10 new s
	EU	5381	0,0526	Without a formal renegotiation of our relationship with the EU , all these transfers of power from Westminster to Brussels are irreversible .
	EU	3237	0,0518	Even the EU pretext that cod stocks must be protected is a sham .
	EU	10665	0,0513	The EU has a track record of guaranteeing democracy , often only recently achieved , in its member_states and ending cross-border conflicts . C
	EU	15916	0,0505	The EU foreign policy chief , Mr Javier Solana , declared that , as they approach the end of their six-month EU presidency , the Belgians have n
	EU	17173	0,0496	Any EU citizen will_do .
	EU	6596	0,0496	They are coming and the EU has no answer .
	EU	6566	0,0471	For all their rhetoric about open internal borders and a brotherhood of nations under one flag , the reality across the EU is rather less edifying .
	EU	8192	0,0466	Unless the EU elite recognises that nations must control their borders , no deal they can offer will convince the electorate the cost of EU members
	EU	8717	0,0442	He added that as France would hold the rotating EU presidency when the Games take place it would_be up_to him to sound out member_states on
	EU	12969	0,0427	Now Leave . EU , which Farage supports , has criticised Lawson strongly . As Sebastian Payne reports at Coffee House , Leave . EU has issuec
	EU	13163	0,0411	Sipping on mint tea , Hajj said : ' ' After the revolution we wanted to return the favour to the EU because they stood with us against the tyrant
	EU	16564	0,0408	As the death toll in the Mediterranean continues to rise week by week , those seeking asylum in Europe will_be hoping EU leaders take their pledg
	EU	19199	0,0406	British acceptance of genuine asylum_seekers is the lowest of the EU member_states .
	EUROPEAN	15573	0,0686	THE FRENCH PRIME_MINISTER , MANUEL VALLS , AND THE EUROPEAN COMMISSION PRESIDENT , JEAN-CLAUDE JUNCKER , YESTERD
	EUROPEAN	6469	0,0678	The suspension of free travel by the Germans was backed by the European Commission as being within the rules . However , Commission Preside
	EUROPEAN	6589	0,0617	So much , then , for European brotherhood and the principle of an ever-closer union ' .
	EUROPEAN	14463	0,0442	Antonio Guterres , the head of UNCHR , warned European countries yesterday to keep out the welcome mat for genuine Iraqi asylum-seekers or r
	EUROPEAN	21229	0,0437	Mr Brown angered the European Commission and his European counterparts on Monday by announcing that he was going to a finance ministers ' m
	EUROPEAN	21793	0,0417	5 Which two European nations failed to win a game ?
	EUROPEAN	19694	0,0417	And they are being joined by failed asylum_seekers and Eastern European economic_migrants .
	EUROPEAN	6635	0,0417	They make_up around half the 1 . 3million eastern Europeans in the UK .
	EUROPEAN	15582	0,0413	Europe should embrace more refugees fleeing war and dictatorship while also tightening border controls and more strictly enforcing its returns polic
	EUROPEAN	24152	0,0410	' ' The government needs to stop apportioning blame by pushing the responsibility back onto the Muslim community . Instead , those professiona
	EUROPEAN	13117	0,0386	His brief covers European integration , international patterns of economic growth , investment , productivity , wages and employment .
	EUROPEAN	16024	0,0385	The ' ' vice-president foreign minister ' ' who is to be appointed under the new European constitution will_be assisted by a European external

Come altri casi di analisi tematica, **T-LAB** permette di **esportare il dizionario** della partizione finale che può essere utilizzato per ulteriori analisi.

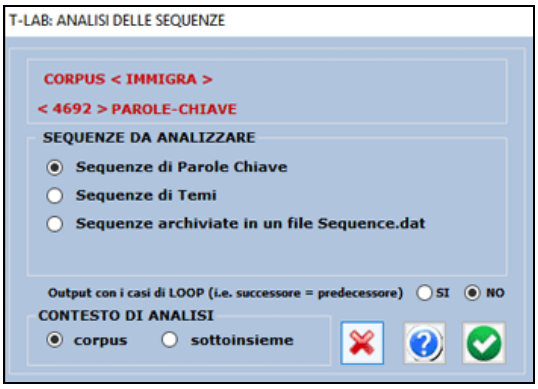
C - ALCUNI DETTAGLI TECNICI

I tipi di sequenze che questo strumento T-LAB ci consente di analizzare sono le seguenti:

1- **Sequenze di Parole-Chiave**, i cui elementi sono unità lessicali (vale a dire parole o lemmi) presenti nel corpus o in un sottoinsieme di esso. In questo caso il numero massimo di 'nodi' (vale a dire i 'tipi' di unità lessicali) è 5.000;
N.B.: Quando viene applicata la lemmatizzazione automatica, 5.000 unità lessicali corrispondono a circa 12,000 parole.

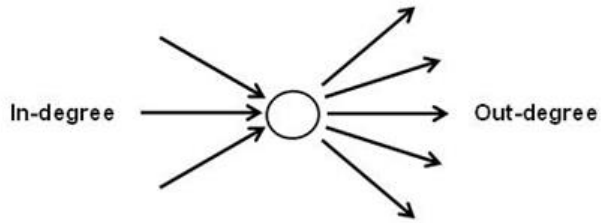
2- **Sequenze di Temi**, i cui elementi sono unità di contesto (cioè contesti elementari) classificate da uno strumento **T-LAB** per l'analisi tematica.
N.B.: In questo caso, poiché la sequenza dei contesti elementari (frasi o paragrafi) caratterizza l'intera 'catena' (predecessori e successori) del corpus, **T-LAB** realizza una forma specifica di **analisi del discorso**, i cui nodi (vale a dire i 'temi') possono variare da un minimo 5 a un massimo di 5.

3 - **Sequenze archiviate in un file Sequence.dat** predisposto dall'utilizzatore (vedi relative spiegazioni alla fine di questa sezione). In questo caso il numero massimo di record è 50.000 e il numero di 'tipi' (ossia nodi) non deve superare 5.000.

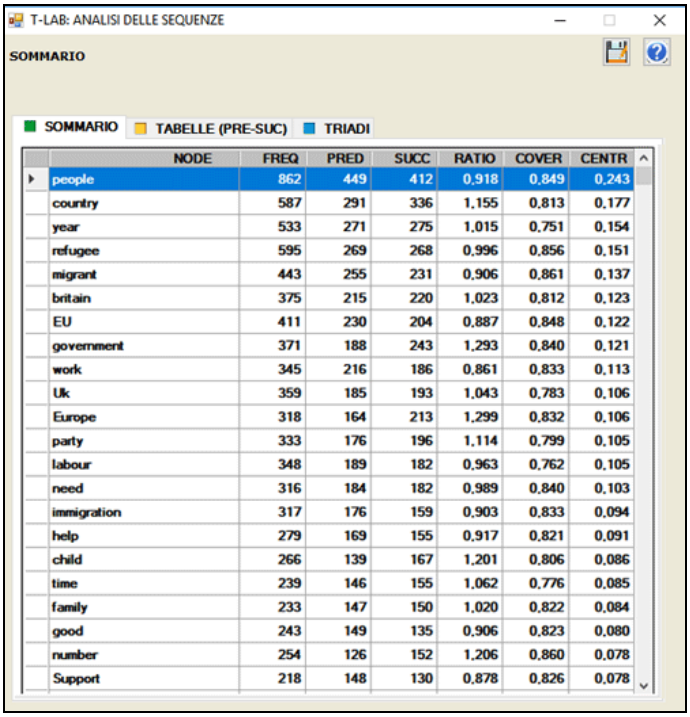


Le informazioni seguenti sono fornite per aiutare l'utente a comprendere meglio i dati riportati nella tabella **SOMMARIO**.

Secondo la teoria dei grafi, i predecessori e i successori di ogni nodo (nel nostro caso, unità lessicali o temi) possono essere rappresentati con delle frecce (archi) in ingresso (in-degree = tipi di predecessori), o in uscita (out-degree = tipi di successori).



Ad esempio, nella tabella seguente "people" ha 412 tipi di successori e 449 tipi of predecessori.
E il valore centrality degree è pari a 0.243.



NODE	FREQ	PRED	SUCC	RATIO	COVER	CENTR
people	862	449	412	0,918	0,849	0,243
country	587	291	336	1,155	0,813	0,177
year	533	271	275	1,015	0,751	0,154
refugee	595	269	268	0,996	0,856	0,151
migrant	443	255	231	0,906	0,861	0,137
britain	375	215	220	1,023	0,812	0,123
EU	411	230	204	0,887	0,848	0,122
government	371	188	243	1,293	0,840	0,121
work	345	216	186	0,861	0,833	0,113
Uk	359	185	193	1,043	0,783	0,106
Europe	318	164	213	1,299	0,832	0,106
party	333	176	196	1,114	0,799	0,105
labour	348	189	182	0,963	0,762	0,105
need	316	184	182	0,989	0,840	0,103
immigration	317	176	159	0,903	0,833	0,094
help	279	169	155	0,917	0,821	0,091
child	266	139	167	1,201	0,806	0,086
time	239	146	155	1,062	0,776	0,085
family	233	147	150	1,020	0,822	0,084
good	243	149	135	0,906	0,823	0,080
number	254	126	152	1,206	0,860	0,078
Support	218	148	130	0,878	0,826	0,078

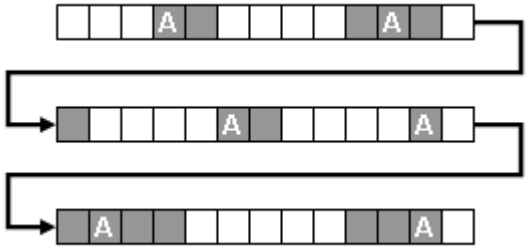
In base al loro rapporto (successori/predecessori), è possibile verificare la varietà semantica generata dal nodo in questione:

- se è maggiore di quanta ne riceve (ratio > 1), il nodo è definito "sorgente";
- se è tanta quanta ne riceve (ratio = 1), il nodo è definito "relais";
- se è minore di quanta ne riceve (ratio < 1), il nodo è definito "assorbente".

Nella stessa tabella, per ogni unità lessicale, la colonna "cover" (coverage) indica in che misura (percentuale) le sue occorrenze sono precedute o seguite da unità lessicali incluse nella lista definita dall'utilizzatore.

Quando le unità analizzate "coprono" la totalità di quelle presenti nel corpus, il valore di "cover" è uguale a 1; diversamente, è un valore inferiore.
Inoltre: quando il valore di "cover" è uguale a 1, anche la sommatoria delle probabilità (sia per i predecessori che per i successori) è uguale a 1; diversamente, è un valore inferiore.
In entrambi i casi, le percentuali "residue" sono determinate dal fatto che vi sono predecessori e successori non inclusi nell'analisi.

Si consideri ad esempio la sequenza rappresentata nell'immagine seguente. Essa è costituita da 39 eventi: di questi, solo 16 (le ipotetiche unità in analisi) sono "coperti" (quadratini in grigio). Ciò a causa del fatto che alcuni di essi, ad esempio quelli corrispondenti alle occorrenze dell'unità lessicale "A", hanno come predecessori e successori anche unità lessicali non incluse nell'analisi (quadratini in bianco).



Differentemente, quando l'utilizzatore analizza una sequenza di temi o un file esterno tutti gli eventi sono "coperti".

N.B.: Per analizzare un file esterno, l'utente deve preparare il corrispondente file 'Sequence.dat'; quindi, dopo aver aperto un progetto esistente, deve selezionare l' opzione "Sequenze registrate in un file Sequence.dat".

Il metodo di calcolo e gli output (grafici e tabelle) sono analoghi a quelli già descritti (vedi sopra).

Il file Sequence.dat, che può contenere ogni tipo di sequenze (ad es. nomi degli interlocutori di una conversazione, categorie ottenute mediante analisi di contenuto, nomi di eventi, etc.), deve essere costituito da "N" record (min. 50 max 50.000), ciascuno costituito da una label di max 50 caratteri, senza spazi bianchi e senza segni di punteggiatura.

I tipi di eventi (tags) non devono essere più di 5.000.

La struttura del file Sequence.dat è quindi quella di un semplice elenco (vedi esempi seguenti):

EXAMPLE_01	**EXAMPLE_02**	**EXAMPLE_03**
Hamlet	activist	event_01
King	food	event_03
Hamlet	genetic	event_02
Queen	conservative	event_03
Hamlet	activist	event_03
Queen	genetic	event_01
Hamlet	conservative	event_05
King	activist	event_02
Queen	commerce	event_05
Hamlet	conservative	event_01
King	activist	event_02
... ...	... ...	... ...

Sia nel caso delle sequenze concernenti le unità lessicali (o temi) del corpus che nel caso delle sequenze registrate in un file esterno (Sequence.dat), **T-LAB** produce alcune tabelle collocate all'interno della cartella MY-OUTPUT.

Concordanze

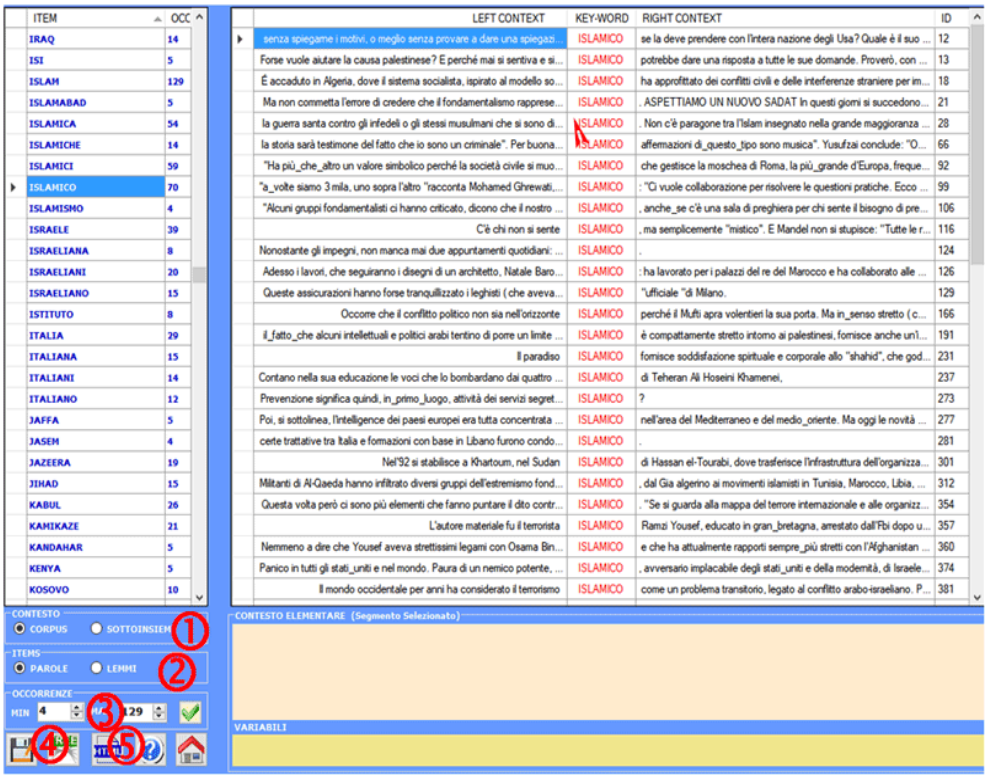
Questo strumento **T-LAB** ci consente di verificare i contesti di occorrenza di ogni unità lessicale.

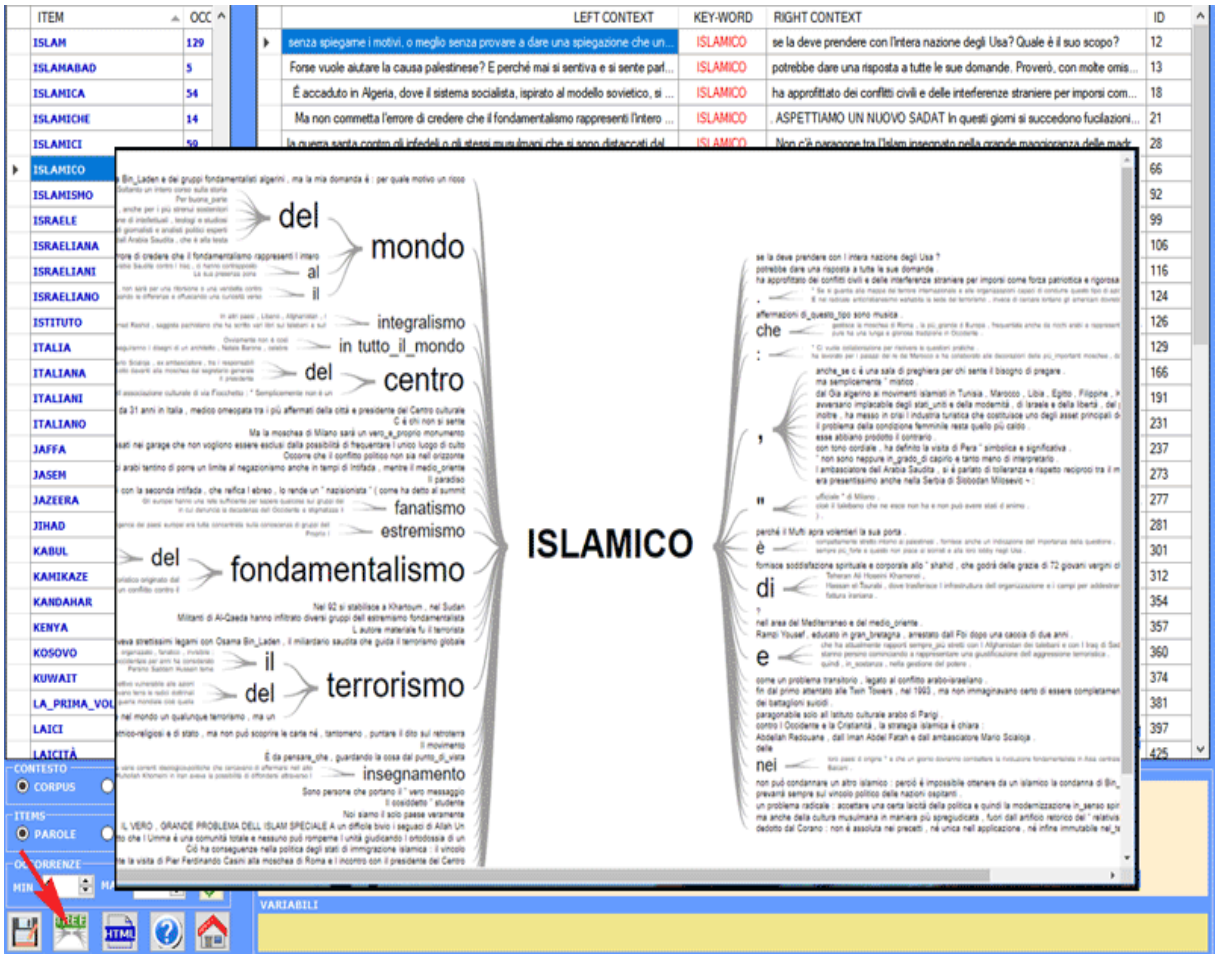
Le ricerche del tipo KWIC (Key-Word-in-Context) possono essere effettuate per **forme** e per **lemmi** (vedi sotto '2'), sia all'interno dell'intero **corpus** che all'interno di un suo **sottoinsieme** (vedi sotto '1').

E' inoltre possibile definire il range delle occorrenze (min e max; vedi sotto '3').

Per ogni unità lessicale, con un semplice click sulla riga corrispondente, è possibile verificare quali sono i suoi contesti di occorrenza (i **contesti elementari**).

Appositi pulsanti consentono di salvare un file con tutti i contesti selezionati, sia in formato 'Word Tree' (vedi sotto '4') che i formato HTML (vedi sotto '5').





DATE: 13/08/2018 - 18:24:17
CONCORDANCES **ISLAMICO**

**** *PERIOD_1ANTE *DC_PAPA
senza spiegarne i motivi, o meglio senza provare a dare una spiegazione che un cittadino non ferrato di politica riesca a capire. Infatti ripetutamente sento parlare del miliardario Osama Bin_Laden e dei gruppi fondamentalisti algerini, ma la mia domanda è: per quale motivo un ricco **ISLAMICO** se la deve prendere con l'intera nazione degli Usa? Quale è il suo scopo?

**** *PERIOD_1ANTE *DC_PAPA
Forse vuole aiutare la causa palestinese? E perché mai si sentiva e si sente parlare di eccidi di massa in Algeria? A che cosa mirano i fondamentalisti in quella regione? Soltanto un intero corso sulla storia del mondo **ISLAMICO** potrebbe dare una risposta a tutte le sue domande. Proverò, con molte omissioni e semplificazioni.

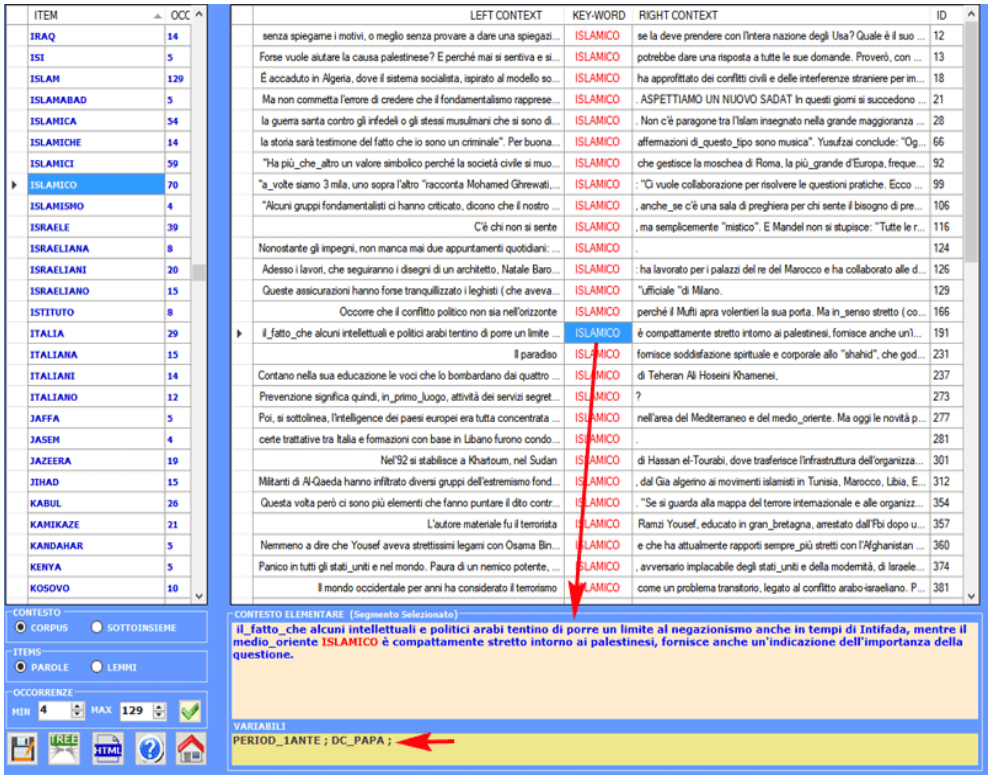
**** *PERIOD_1ANTE *DC_PAPA
É accaduto in Algeria, dove il sistema socialista, ispirato al modello sovietico, si era trasformato in una oligarchia militare, inefficiente e venale. In altri paesi, Libano, Afghanistan, l'integralismo **ISLAMICO** ha approfittato dei conflitti civili e delle interferenze straniere per imporsi come forza patriottica e rigorosamente morale.

**** *PERIOD_1ANTE *DC_PAPA
Ma non commetta l'errore di credere che il fondamentalismo rappresenti l'intero mondo **ISLAMICO**. ASPETTIAMO UN NUOVO SADAT In questi giorni si succedono fucilazioni di palestinesi eseguite dagli uomini dai volti coperti della milizia palestinese. I processi sono sommari, senza difesa, durano poche ore e hanno come epilogo scontato la pena di morte.

**** *PERIOD_1ANTE *DC_SERVIZI
la guerra santa contro gli infedeli o gli stessi musulmani che si sono distaccati dall'insegnamento del Profeta. Ovviamente non è così in tutto_il_mondo **ISLAMICO**. Non c'è paragone tra l'Islam insegnato nella grande maggioranza delle madrassa egiziane, tunisine, o turche o malesi, un Islam tollerante, aperto alla modernità, e quello insegnato qui,

Inoltre, cliccando al centro dei segmenti mostrati è possibile visualizzare il loro contenuto e

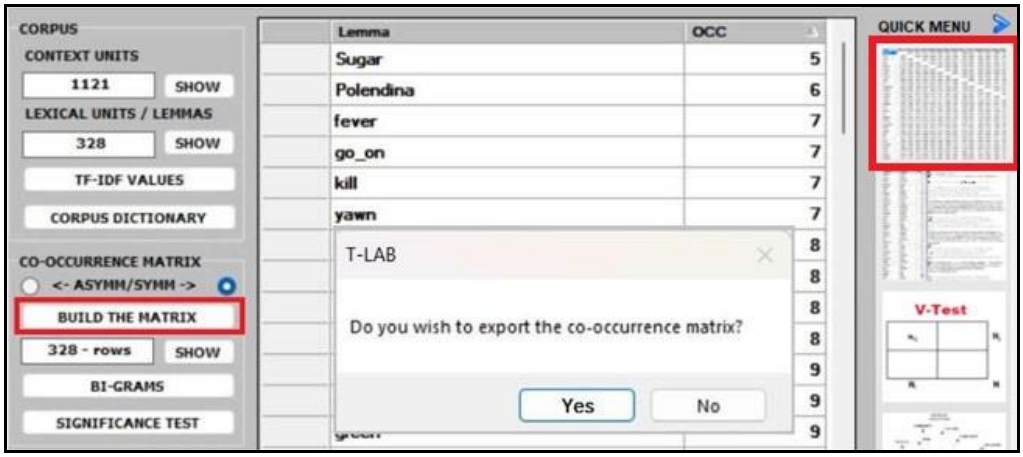
verificare le categorie utilizzate nelle rispettive linee di codifica (vedi sotto)



Co-occurrence Toolkit

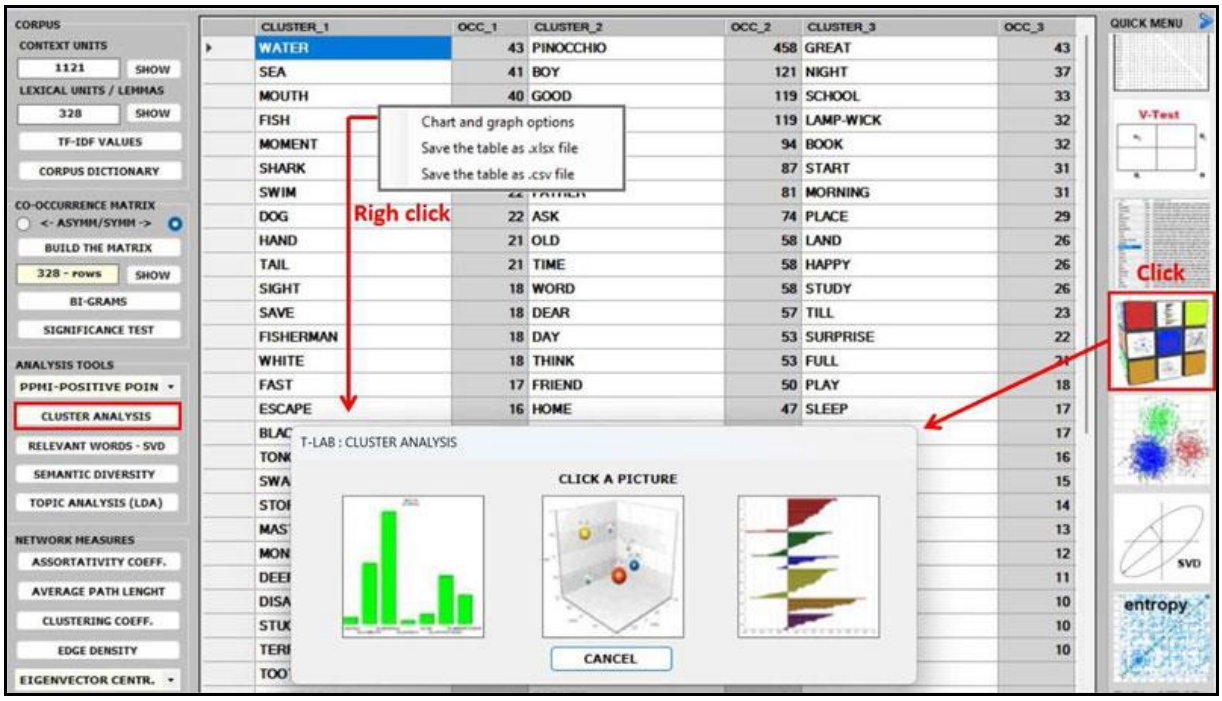
N.B .: Questa sezione dell'help è disponibile solo in inglese.

This tool, which can be used for a variety of tasks, offers a set of techniques for building and analysing **word co-occurrence matrices** with up to 5,000 columns.
The matrices to be built can be both **symmetric** and **asymmetric**, and they can represent the co-occurrences of the words either within the whole **corpus** or within a **subset** of it.



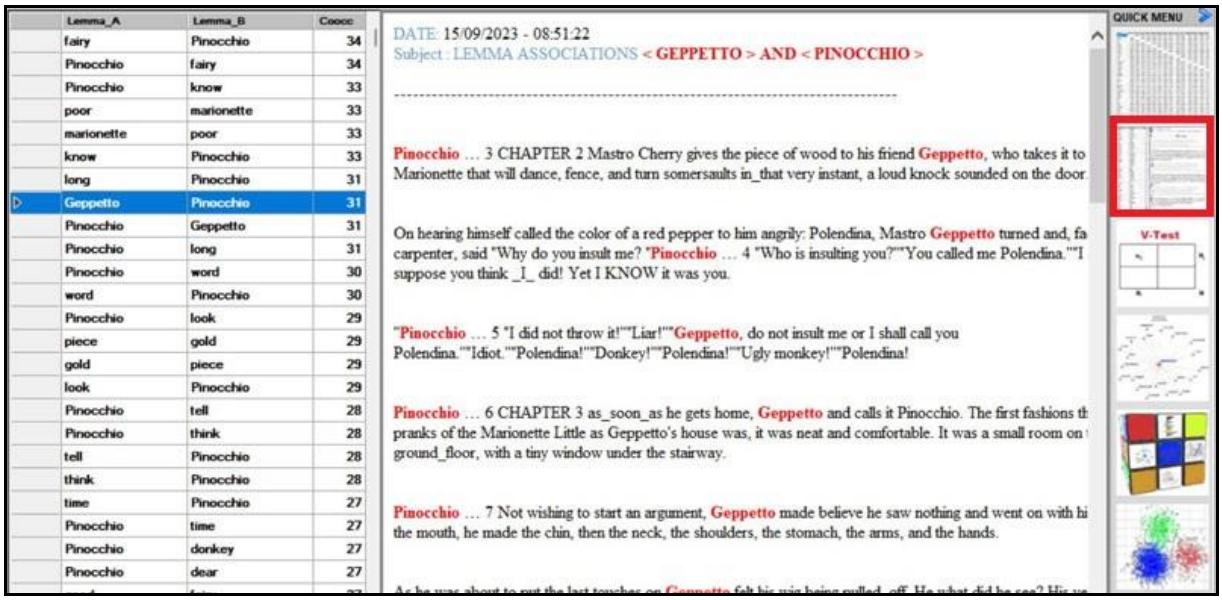
N.B.: In the case of word co-occurrences, the difference between symmetric and asymmetric matrices is that symmetric matrices assume that the order of words does not matter (i.e., they are represented as undirected graphs where the values in a row and a column are the same), while asymmetric matrices take into account the direction of co-occurrence and, for this reason, are represented as directed graph where the values in a row (i.e., successor) and a column (i.e., predecessor) are not necessarily the same.

Whichever tool you are using, the way to export tables and graphs is very simple (see picture below).



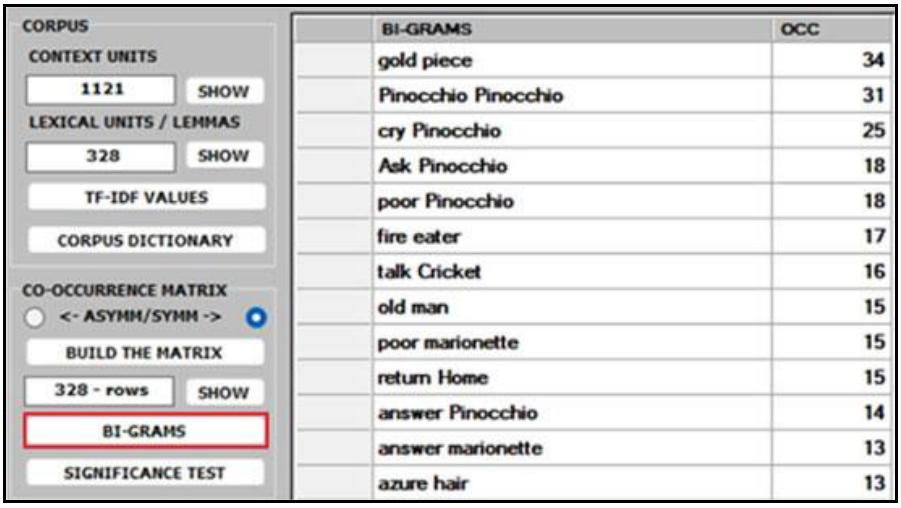
After building any co-occurrence matrix, the user is allowed to extract the relevant information by using about fifteen options listed on the left menu (see the above picture).

N.B.:
- all the below pictures have been obtained by analysing the English version of "The Adventures of Pinocchio" (by Carlo Collodi) and its symmetric word co-occurrence matrix.
- all items in the tables are 'lemmas' because a **T-LAB** lemmatization has been performed on the Pinocchio corpus first.
- whatever matrix you are analysing, it is always possible to check the text segments in which pairs of words co-occur (see picture below).

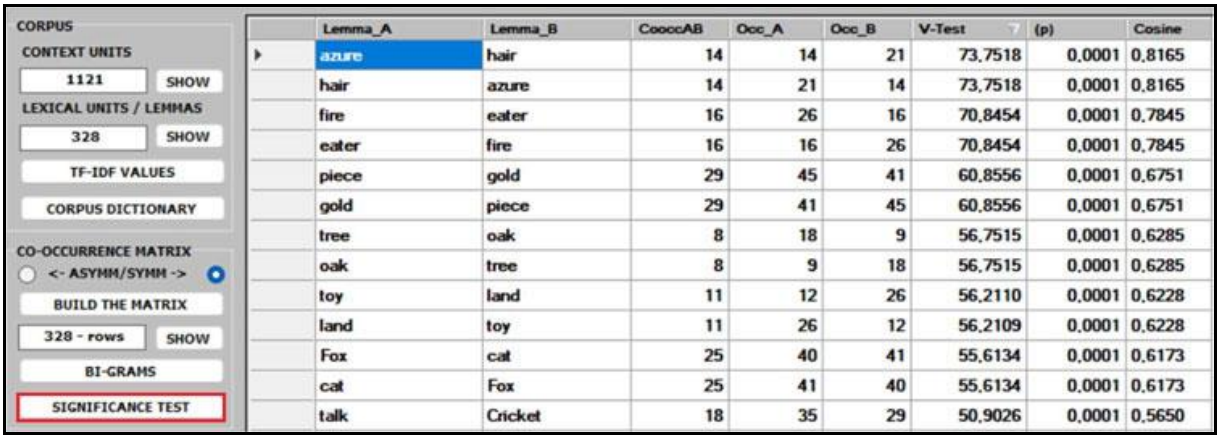


Below are the descriptions of the various analysis options:

- both the **BI-GRAMS** and the **SIGNIFICANCE TEST** extract pairs of words (e.g., collocations) which can be relevant for customizing the corpus dictionary an also for detecting small groups of related words which can affect any cluster analysis (see pictures below).

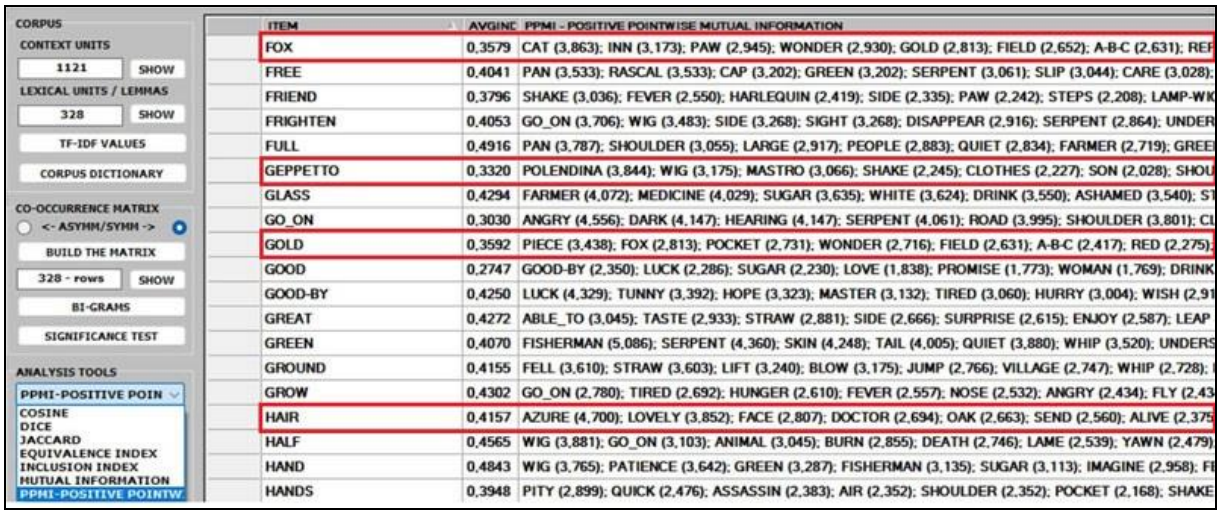


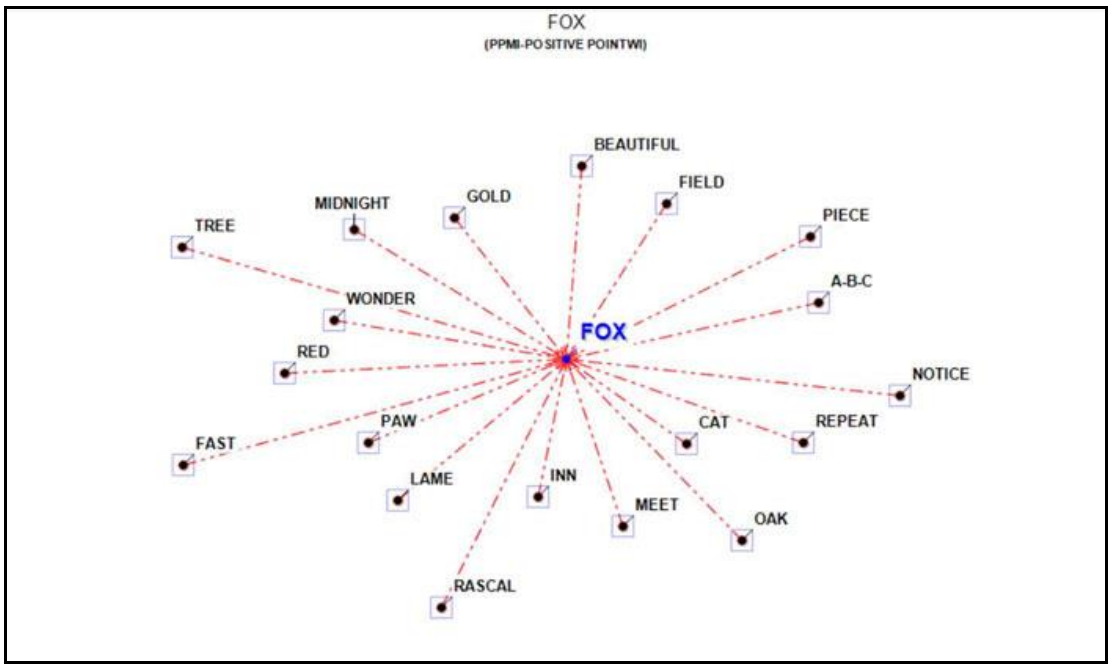
BI-GRAMS	OCC
gold piece	34
Pinocchio Pinocchio	31
cry Pinocchio	25
Ask Pinocchio	18
poor Pinocchio	18
fire eater	17
talk Cricket	16
old man	15
poor marionette	15
return Home	15
answer Pinocchio	14
answer marionette	13
azure hair	13



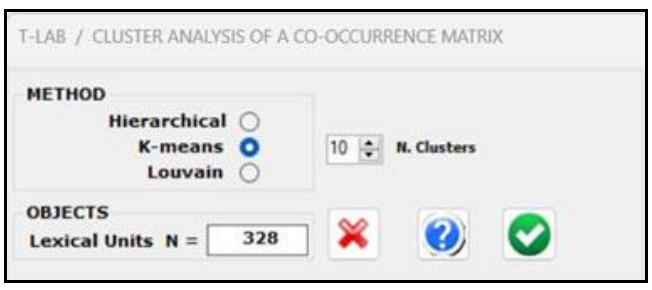
Lemma_A	Lemma_B	CooccAB	Occ_A	Occ_B	V-Test	(p)	Cosine
azure	hair	14	14	21	73,7518	0,0001	0,8165
hair	azure	14	21	14	73,7518	0,0001	0,8165
fire	eater	16	26	16	70,8454	0,0001	0,7845
eater	fire	16	16	26	70,8454	0,0001	0,7845
piece	gold	29	45	41	60,8556	0,0001	0,6751
gold	piece	29	41	45	60,8556	0,0001	0,6751
tree	oak	8	18	9	56,7515	0,0001	0,6285
oak	tree	8	9	18	56,7515	0,0001	0,6285
toy	land	11	12	26	56,2110	0,0001	0,6228
land	toy	11	26	12	56,2109	0,0001	0,6228
Fox	cat	25	40	41	55,6134	0,0001	0,6173
cat	Fox	25	41	40	55,6134	0,0001	0,6173
talk	Cricket	18	35	29	50,9026	0,0001	0,5650

- the **ASSOCIATIONS** option, in addition to the indexes used by other **T-LAB** tools (see Word Associations and Co-Word Analysis), includes the **PPMI** (i.e., Positive Pointwise Mutual Information), which is a measure of how much more likely two words are to co-occur than by chance, based on their probabilities in a text corpus. It can be used to distinguish between words that are simply co-occurring by chance and words that are semantically related. It can also reduce the effect of high-frequency words that co-occur with many other words by chance. Moreover, unlike other indexes (e.g., Cosine, Dice, Jaccard etc.) its maximum value is not '1' and its upper bound can vary.



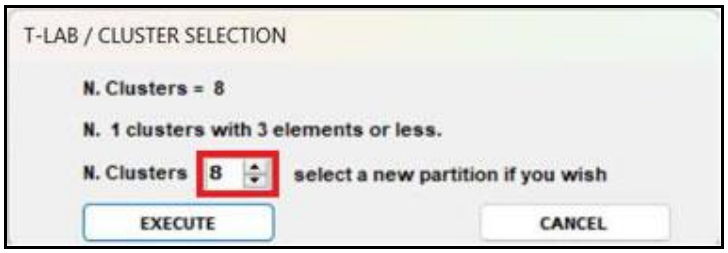


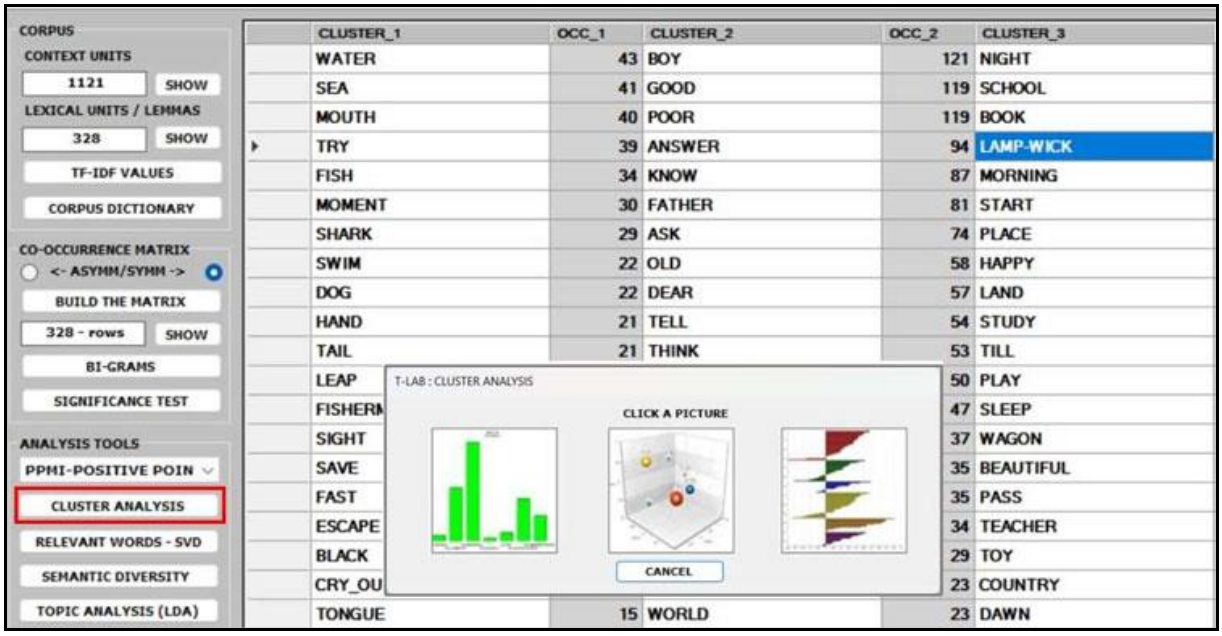
- the **CLUSTER ANALYSIS** offers three methods for analysing a word co-occurrence matrix: **Hierarchical**, **K-means** and **Louvain**.



All the above three methods use vectors which are normalized by the cosine coefficient, and one of them (i.e., the K-means) performs the clustering on the first 10 dimensions obtained by a SVD (i.e., Singular Value Decomposition) of the normalized word co-occurrence matrix.
To evaluate the quality of clustering results, **T-LAB** provides the **Silhouette** scores for each data point. Moreover, when clicking the '**Q**' button located at the bottom left corner of the screen, the user is allowed to obtain three different quality indices (i.e.: Calinski-Harabasz, Dunn and ICC-rho).

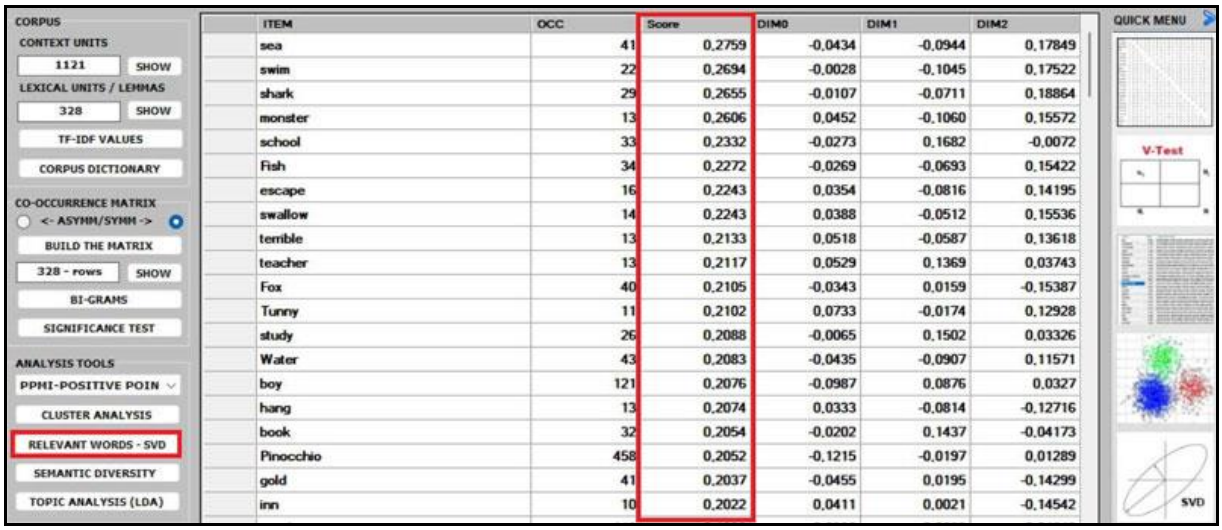
N.B.:
- Depending on the clustering method, the **relationships between words within each cluster** can be visualized through different types of charts and graphs.
- When performing a hierarchical clustering, the user is allowed to change the number of clusters (i.e., the cluster partition) within a range from 3 to 20.



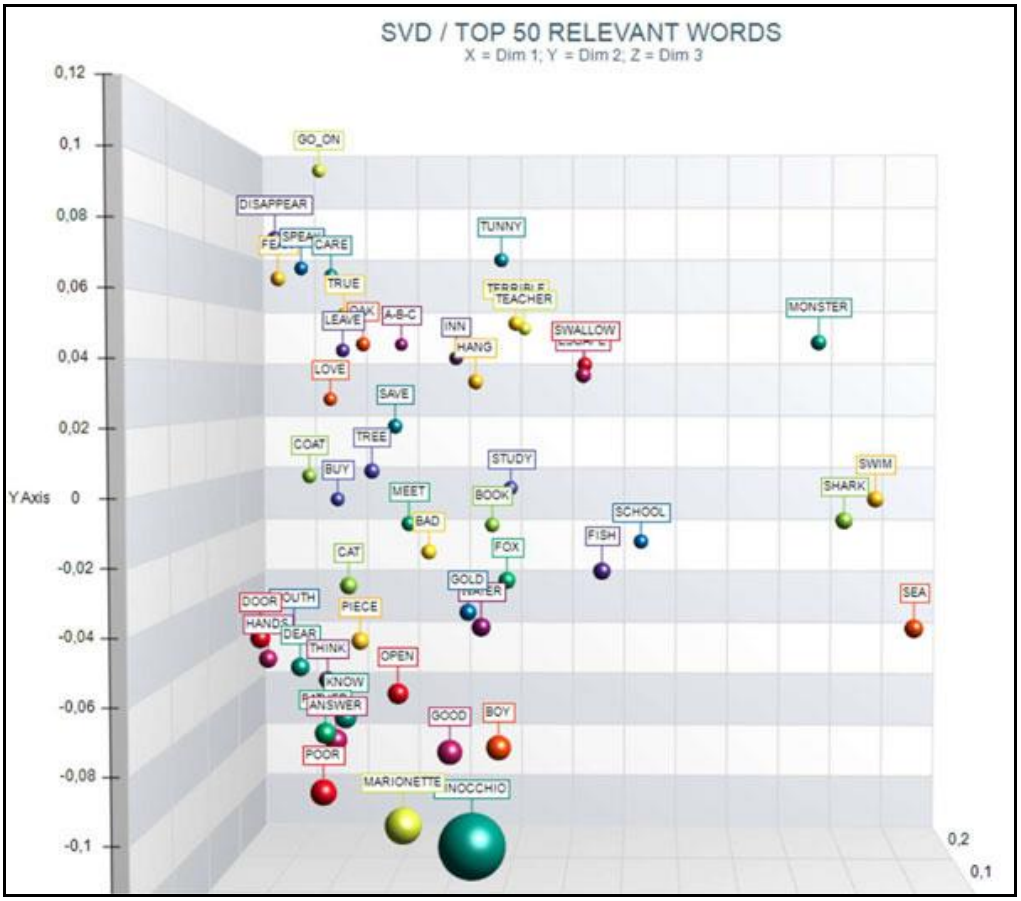


- the **RELEVANT WORDS - SVD** provides a relevance score for each word, which is computed by summing the square of its first 3 dimensions (i.e., the eigenvectors), each one multiplied by its corresponding singular value, and then by computing the square root of that sum.

This means that the words with the higher scores are the farthest from the point of origin, which is the point where the horizontal axis (x-axis) and the vertical axis (y-axis) intersect. And, for this reason, they are the words that most contribute to organizing semantic polarizations, which can also have emotional connotations.
N.B.: In this case, the SVD is performed on a centered matrix and therefore it is equivalent to PCA.

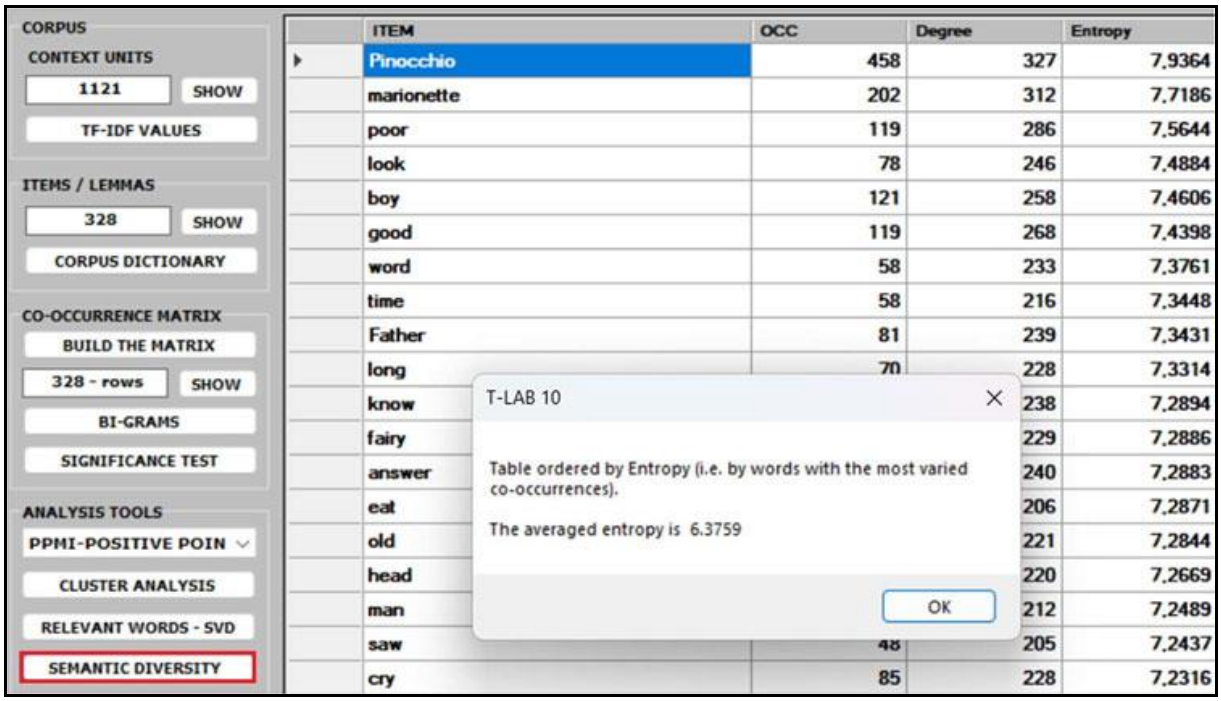


ITEM	OCC	Score	DIM0	DIM1	DIM2
sea	41	0,2759	-0,0434	-0,0944	0,17849
swim	22	0,2694	-0,0028	-0,1045	0,17522
shark	29	0,2655	-0,0107	-0,0711	0,18864
monster	13	0,2606	0,0452	-0,1060	0,15572
school	33	0,2332	-0,0273	0,1682	-0,0072
Fish	34	0,2272	-0,0269	-0,0693	0,15422
escape	16	0,2243	0,0354	-0,0816	0,14195
swallow	14	0,2243	0,0388	-0,0512	0,15536
terrible	13	0,2133	0,0518	-0,0587	0,13618
teacher	13	0,2117	0,0529	0,1369	0,03743
Fox	40	0,2105	-0,0343	0,0159	-0,15387
Tunny	11	0,2102	0,0733	-0,0174	0,12928
study	26	0,2088	-0,0065	0,1502	0,03326
Water	43	0,2083	-0,0435	-0,0907	0,11571
boy	121	0,2076	-0,0987	0,0876	0,0327
hang	13	0,2074	0,0333	-0,0814	-0,12716
book	32	0,2054	-0,0202	0,1437	-0,04173
Pinocchio	458	0,2052	-0,1215	-0,0197	0,01289
gold	41	0,2037	-0,0455	0,0195	-0,14299
inn	10	0,2022	0,0411	0,0021	-0,14542



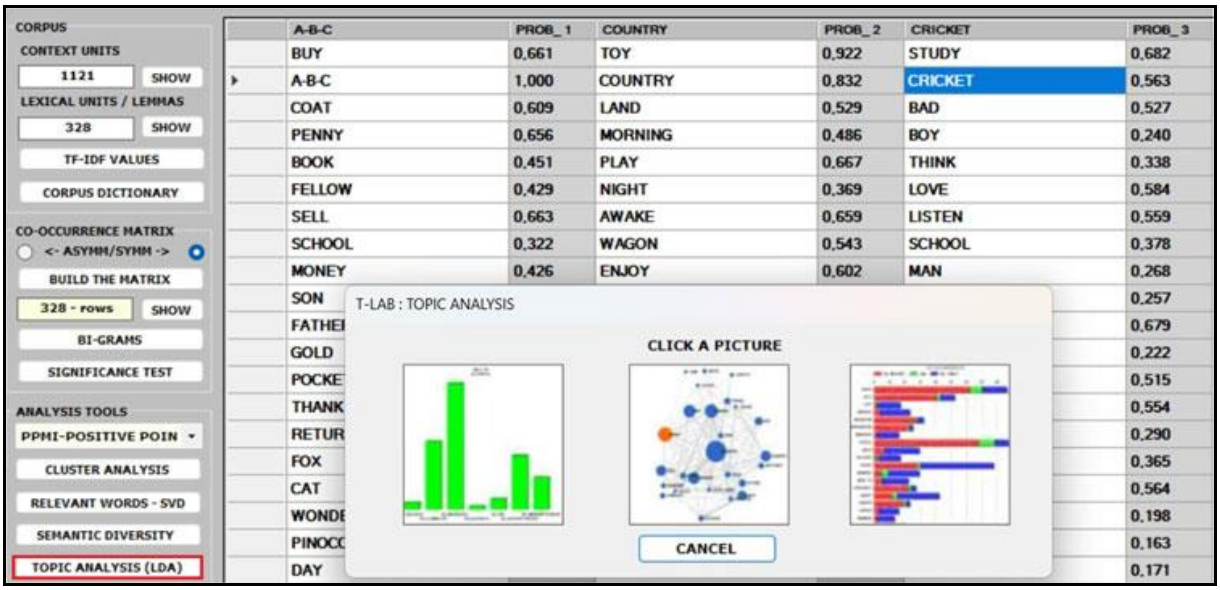
- the **SEMANTIC DIVERSITY** of each word (i.e., its ability to have links with many other words) is measured by means of the **entropy** index.

N.B.: The average entropy of the word co-occurrence matrix can be used to quantify the 'complexity' of a text, since more complex texts (i.e., texts in which many words cooccur with a variety of other words) tend to have higher entropy than simpler texts (i.e., texts in which many words cooccur with only a few other words and – for that reason – are more predictable). And, since high entropy corresponds to low predictability, it may be also interesting to check which words in a text have higher predictability values (i.e., low entropy).



- the **TOPIC ANALYSIS** of the word co-occurrence matrix uses the same algorithm of the **T-LAB** Modeling of Emerging Themes tool (i.e., Latent Dirichlet Allocation and the Gibbs Sampling); however, in this case, both the indexes of the matrix (i.e., the 'i' and the 'j') refer to the same words and the values correspond to their co-occurrences. As can be verified, the results of this approach are quite interesting and consistent.

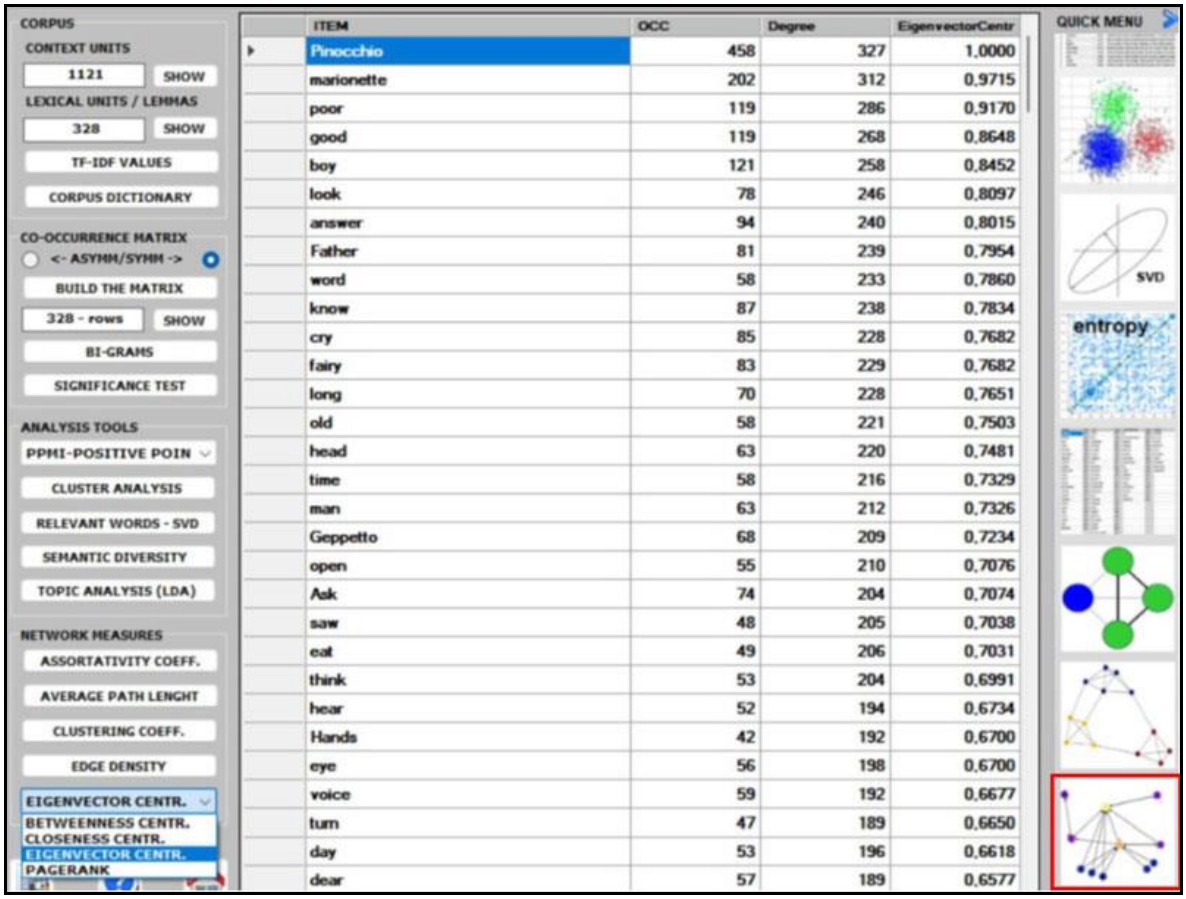
  N.B.: In the table below, the words are ordered by their frequency within each topic.



- Regarding the five **CENTRALITY MEASURES** (i.e., Betweenness centrality, Closeness centrality, Eigenvector centrality, Katz centrality and PageRank centrality) we observe

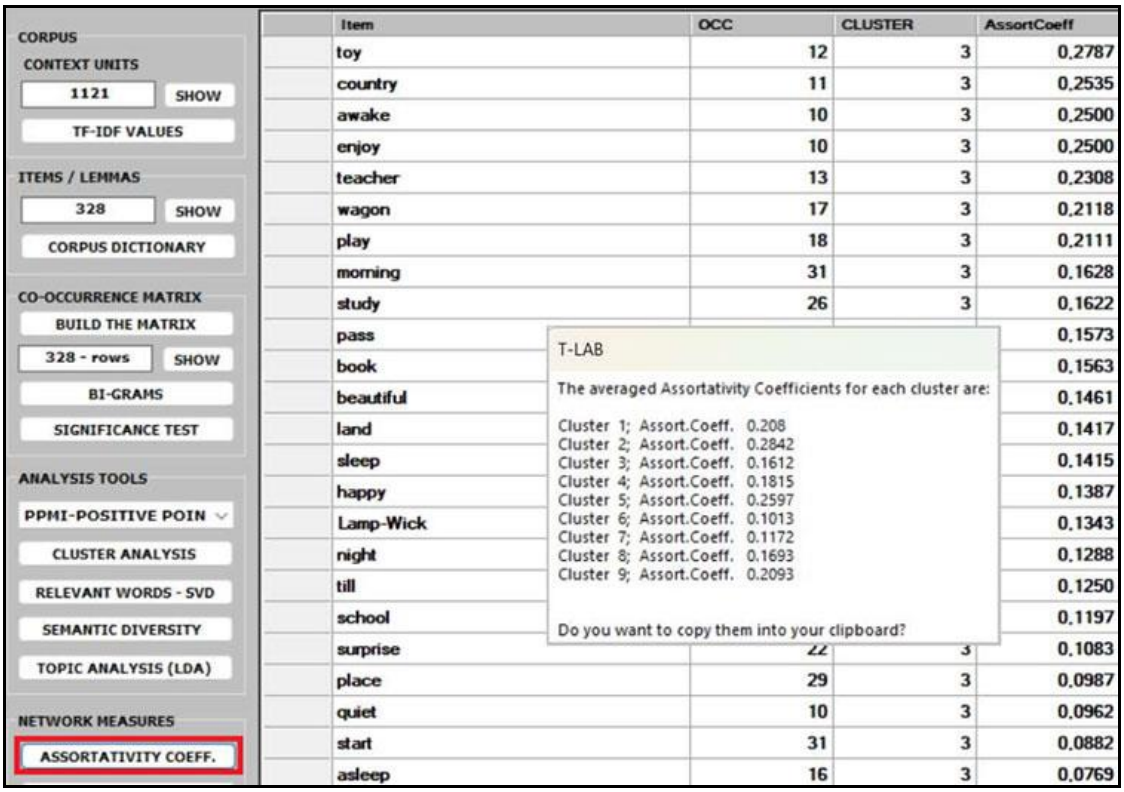
that, especially in the case of a symmetric word co-occurrence matrix, they are closely related to each other. Moreover, they usually rank more highly the words with higher occurrence values. The only exception seems to be the Betweenness centrality. In fact, it is possible for a vertex to have high betweenness centrality (i.e., to be able to connect important parts of the network) without having high indegree or high outdegree.

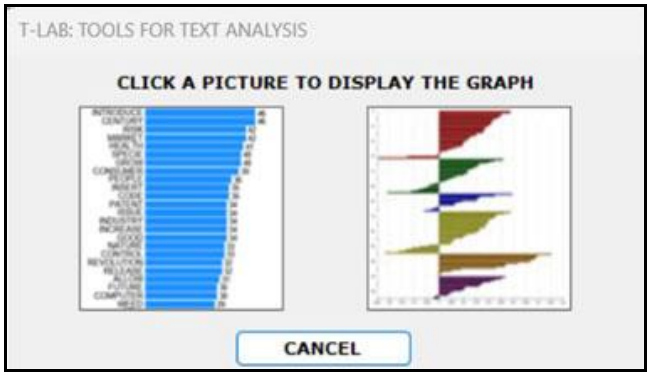
N.B.:
- All definitions of centrality measures, as well as their algorithms, can be easily checked on Wikipedia.
- In **T-LAB**, all the results of centrality measures are normalized to the maximum value. This means that all the results are between 0 and 1, which makes them easier to compare.



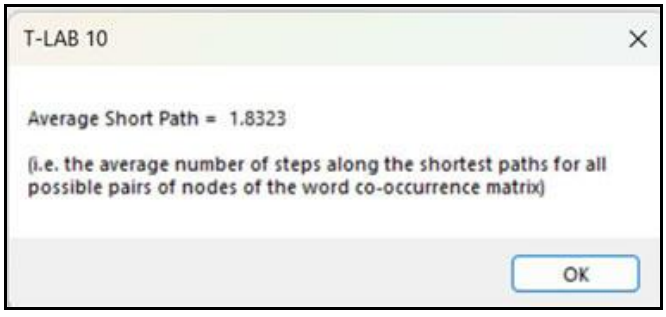
ITEM	OCC	Degree	EigenvectorCentr
Pinocchio	458	327	1,0000
marionette	202	312	0,9715
poor	119	286	0,9170
good	119	268	0,8648
boy	121	258	0,8452
look	78	246	0,8097
answer	94	240	0,8015
Father	81	239	0,7954
word	58	233	0,7860
know	87	238	0,7834
cry	85	228	0,7682
fairy	83	229	0,7682
long	70	228	0,7651
old	58	221	0,7503
head	63	220	0,7481
time	58	216	0,7329
man	63	212	0,7326
Geppetto	68	209	0,7234
open	55	210	0,7076
Ask	74	204	0,7074
saw	48	205	0,7038
eat	49	206	0,7031
think	53	204	0,6991
hear	52	194	0,6734
Hands	42	192	0,6700
eye	56	198	0,6700
voice	59	192	0,6677
turn	47	189	0,6650
day	53	196	0,6618
dear	57	189	0,6577

- the **ASSORTATIVITY COEFFICIENT** is a measure of how likely nodes of a certain type are to be connected to other nodes of the same type (i.e., 'similar' in some respects). In the case of **T-LAB**, the types refer to the results of a previous cluster analysis. Therefore, (a) if– for any 'i' node – the assortativity coefficient is positive and high, then it indicates that the node is strongly connected with other nodes of the same cluster; (b) if – for any 'k' cluster - the average assortativity coefficient is positive and high, then it indicates that the nodes which belong to the cluster are strongly connected with each other; (c) a global average high positive assortativity coefficient indicates that the clustering algorithm has successfully grouped nodes based on their links within the cluster they belong to. This means that nodes within the same cluster are more likely to be connected to each other than nodes from different clusters.



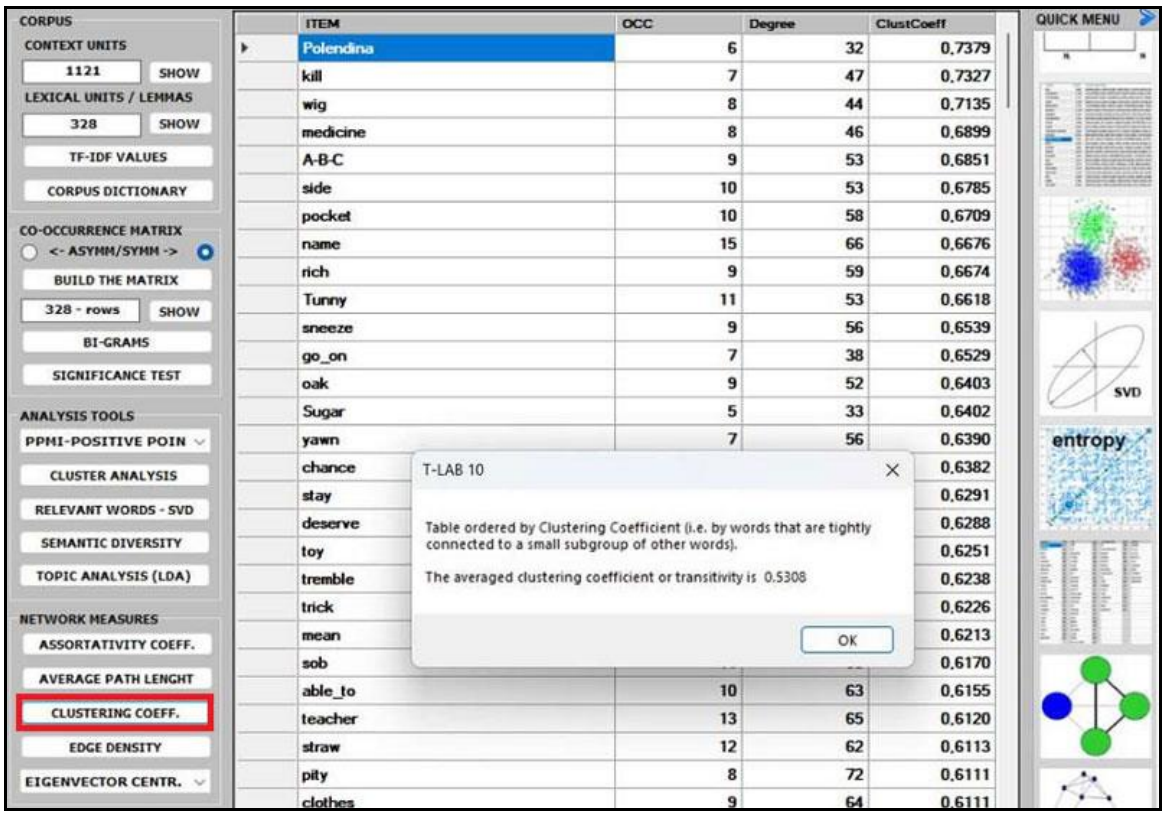


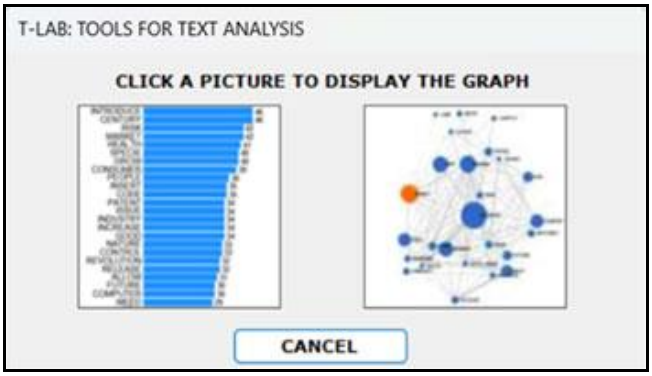
- the **AVERAGE PATH LENGTH** (or average short path), in this case, is defined as the average number of steps along the shortest paths for all possible pairs of nodes of the word co-occurrence matrix.



- the **CLUSTERING COEFFICIENT** deserves special attention. In fact, the 'local' clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together and to pair up with each other (i.e., something like 'The friend of my friend is my friend.'). In other words, the clustering coefficient of a node (i.e., word) quantifies how close its neighbours (i.e., other words) are to being a tightly connected subgroup (i.e., a clique). It is computed as the proportion of the 'actual' connections among its neighbours compared with the number of all its 'possible' connections. Its maximum value is '1', and the average clustering coefficient of all nodes it is also known as 'transitivity' of the network.

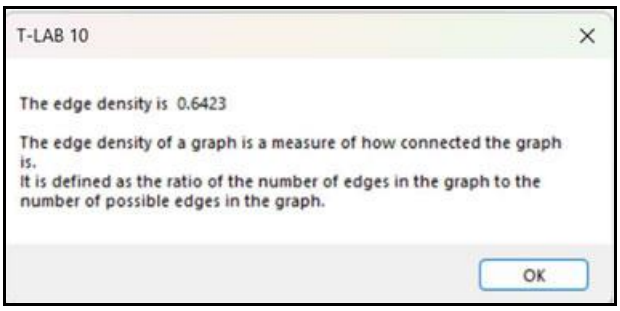
  N.B.:
  - When a network has a large clustering coefficient and a small average path length it can be considered a 'small world' (see Wikipedia).





- the **EDGE DENSITY** is a measure of how connected the graph is. It is defined as the ratio of the actual number of edges in the graph to the possible number of edges in the graph. A high edge density indicates that the nodes in the graph are more likely to be connected to each other. This means that there are many paths between any two nodes in the graph. A low edge density indicates that the nodes in the graph are more likely to be disconnected from each other. This means that there are few paths between any two nodes in the graph.

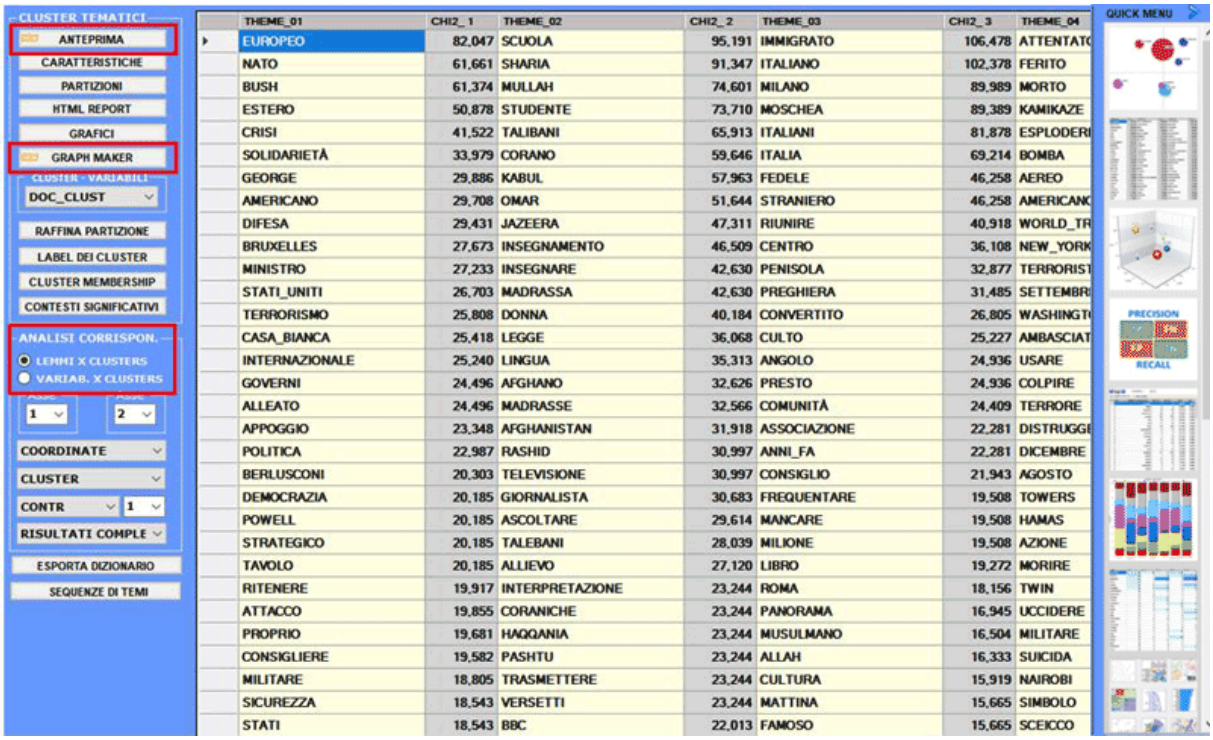
N.B.: It appears that there is a positive correlation between edge density and clustering coefficient. In fact, both measures refer to the connectivity of a graph and can be used to compare the properties of different graphs (i.e., in this case, the properties of different co-occurrence matrices).



ANALISI TEMATICHE

Analisi Tematica dei Contesti Elementari

N.B.: Le immagini di questa sezione fanno riferimento a una versione precedente di **T-LAB**. In **T-LAB 10** l'aspetto è leggermente diverso. Inoltre: a) un nuovo pulsante (TREE MAP PREVIEW) consente di creare grafici dinamici in formato HTML; b) il pulsante DENDROGRAMMA è stato sostituito con lo strumento GRAPH MAKER; c) è disponibile una ulteriore tabella che mostra in varie colonne le parole tipiche di ogni cluster; d) è possibile effettuare ulteriori analisi delle corrispondenze tra i cluster tematici e ciascuna delle variabili disponibili; e) una galleria di immagini funziona come un menu aggiuntivo e consente di passare da un output all'altro con un solo clic.
Alcune di queste nuove funzionalità sono evidenziate nell'immagine seguente.



	THEME_01	CHI2_1	THEME_02	CHI2_2	THEME_03	CHI2_3	THEME_04
▸	EUROPEO	82,047	SCUOLA	95,191	IMMIGRATO	106,478	ATTENTATO
	NATO	61,661	SHARIA	91,347	ITALIANO	102,378	FERITO
	BUSH	61,374	MULLAH	74,601	MILANO	89,989	MORTO
	ESTERO	50,878	STUDENTE	73,710	MOSCHEA	89,389	KAMIKAZE
	CRISI	41,522	TALIBANI	65,913	ITALIANI	81,878	ESPLODERI
	SOLIDARIETÀ	33,979	CORANO	59,646	ITALIA	69,214	BOMBA
	GEORGE	29,886	KABUL	57,963	FEDELE	46,258	AEREO
	AMERICANO	29,708	OMAR	51,644	STRANIERO	46,258	AMERICANO
	DIFESA	29,431	JAZEERA	47,311	RIUNIRE	40,918	WORLD_TR
	BRUXELLES	27,673	INSEGNAMENTO	46,509	CENTRO	36,108	NEW_YORK
	MINISTRO	27,233	INSEGNARE	42,630	PENISOLA	32,877	TERRORIST
	STATI_UNITI	26,703	MADRASSA	42,630	PREGHIERA	31,485	SETTEMBRI
	TERRORISMO	25,808	DONNA	40,184	CONVERTITO	26,805	WASHINGTO
	CASA_BIANCA	25,418	LEGGE	36,068	CULTO	25,227	AMBASCIAT
	INTERNAZIONALE	25,240	LINGUA	35,313	ANGOLO	24,936	USARE
	GOVERNI	24,496	AFGHANO	32,626	PRESTO	24,936	COLPIRE
	ALLEATO	24,496	MADRASSE	32,566	COMUNITÀ	24,409	TERRORE
	APPOGGIO	23,348	AFGHANISTAN	31,918	ASSOCIAZIONE	22,281	DISTRUGGE
	POLITICA	22,987	RASHID	30,997	ANNI_FA	22,281	DICEMBRE
	BERLUSCONI	20,303	TELEVISIONE	30,997	CONSIGLIO	21,943	AGOSTO
	DEMOCRAZIA	20,185	GIORNALISTA	30,683	FREQUENTARE	19,508	TOWERS
	POWELL	20,185	ASCOLTARE	29,614	MANCARE	19,508	HAMAS
	STRATEGICO	20,185	TALEBANI	28,039	MILIONE	19,508	AZIONE
	TAVOLO	20,185	ALLIEVO	27,120	LIBRO	19,272	MORIRE
	RITENERE	19,917	INTERPRETAZIONE	23,244	ROMA	18,156	TWIN
	ATTACCO	19,855	CORANICHE	23,244	PANORAMA	16,945	UCCIDERE
	PROPRIO	19,681	HAQQANIA	23,244	MUSULMANO	16,504	MILITARE
	CONSIGLIERE	19,582	PASHTU	23,244	ALLAH	16,333	SUICIDA
	MILITARE	18,805	TRASMETTERE	23,244	CULTURA	15,919	NAIROBI
	SICUREZZA	18,543	VERSETTI	23,244	MATTINA	15,665	SIMBOLO
	STATI	18,543	BBC	22,013	FAMOSO	15,665	SCEICCO

Questo strumento **T-LAB** consente di costruire ed esplorare una **rappresentazione dei contenuti** del corpus attraverso pochi e significativi **cluster tematici** (minimo 3, massimo 50), ciascuno dei quali:

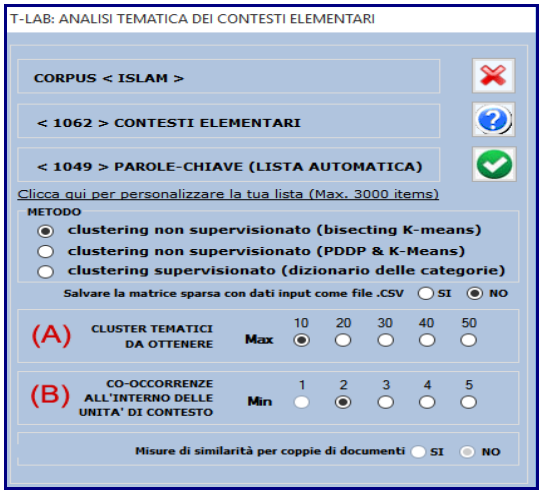
a) risulta costituito da un insieme contesti elementari (frasi, paragrafi o testi brevi quali risposte a domande aperte) caratterizzati dagli stessi pattern di parole chiave;

b) è descritto attraverso le unità lessicali (parole, lemmi o categorie) e le variabili (se presenti) che più caratterizzano i contesti elementari da cui è composto.

Per molti versi, si può affermare che il risultato dell'analisi propone una mappatura delle isotopie (iso = uguale; topoi = luoghi) intese come temi "generali" o "specifici" (Rastier, 2002: 204) caratterizzati dalla co-occorrenza di tratti semantici. In effetti ogni cluster,

caratterizzato da insiemi di unità lessicali che condividono gli stessi contesti di riferimento, consente di ricostruire "un filo" del discorso all'interno della trama complessiva costituita dal corpus o da un suo sottoinsieme.

Il processo di analisi può essere effettuato tramite un metodo di clustering 'non supervisionato' (nel caso specifico, un algoritmo bisecting K-Means) o tramite una classificazione 'supervisionata' (vale a dire approccio top-down). Quando si sceglie il secondo (cioè classificazione supervisionata), viene richiesto di importare un dizionario delle categorie, sia esso creato tramite una precedente analisi **T-LAB** che costruito dall'utilizzatore.



Una finestra di dialogo (vedi sopra) consente di scegliere alcuni parametri dell'analisi.

In particolare:

- il parametro (A) permette di fissare il numero massimo di partizioni da includere negli output **T-LAB**;
- il parametro (B) permette di escludere dall'analisi le unità di contesto che non contengono un numero minimo di parole chiave (co-occorrenze) incluse nella lista predisposta dall'utilizzatore.

N.B.:
Quando si seleziona l'opzione 'classificazione supervisionata', poiché il numero di cluster che devono essere ottenuti coincide con il numero di categorie presenti nel dizionario, il parametro 'A' non è disponibile;
I suddetti parametri producono cambiamenti significativi nei risultati dell'analisi solo quando il numero di unità di contesto è molto grande e/o quando esse sono costituite da testi corti.

Nel caso di **clustering non supervisionato** (opzione di default), la **procedura di analisi** è costituito dai seguenti passaggi step:

a - costruzione di una tabella dati unità di contesto x unità lessicali (max 300.000 righe x 3.000 colonne), con valori del tipo presenza/assenza;
b - pretrattamento dei dati tramite TF-IDF e trasformazione di ogni vettore riga a lunghezza 1 (norma euclidea);
c - uso della misura del coseno e clusterizzazione delle unità di contesto tramite il metodo bisecting K-means (riferimenti: Steinbach, Karypis, & Kumar, 2000; Savaresi, Booley, 2001);
d - archiviazione delle varie partizioni ottenute e, per ciascuna di esse:
e - costruzione di una tabella di contingenza unità lessicali x cluster (n x k);

f - test del chi quadro applicato a tutti gli incroci cluster x unità lessicali;
g - analisi delle corrispondenze della tabella di contingenza unità lessicali x cluster (riferimenti: Benzécri, 1984; Greenacre, 1984; Lebart, Salem, 1994).

N.B.: A partire da T-LAB 2016, la clusterizzazione delle unità di contesto (vedi sopra step 'c') può essere ottenuta sia usando l'algoritmo bisecting K-means algorithm (1) che usando una versione 'not centered' dell'algoritmo PDDP(*Principal* Direction Divisive Partitioning) proposto da D. Booley (1998) per selezionare i centroidi delle varie bisezioni K-means.

La principale differenza tra i due algoritmi sta nel metodo attraverso il quale i due centroidi vengono ottenuti; infatti, nel primo caso (1) essi sono il risultato di una reiterazione, mentre nel secondo caso (2) sono ottenuti tramite SVD (i.e. Singular Value Decomposition), cioè tramite un algoritmo 'one-shot'(see Savaresi, S.M., & Boley, D.L. , 2004).

Quindi, questa procedura realizza un tipo di **analisi delle co-occorrenze** (step a-b-c) e, a seguire, un tipo di **analisi comparativa** (e-f-g). In particolare, l'analisi comparativa usa come colonne delle tabelle di contingenza le modalità della "nuova variabile" derivata dall'analisi delle co-occorrenze (modalità della nuova variabile = cluster tematici).

N.B.: Quando l'utilizzatore decidere di ripetere/applicare i risultati di una precedente analisi tematica (sia **Analisi Tematica dei Contesti Elementari** che **Modellizzazione dei Temi Emergenti**), **T-LAB** realizza soltanto un'analisi comparativa dei cluster già ottenuti (passi e-f-g).

Nel caso di **classificazione supervisionata**, le fasi dell'analisi comparativa sono le stesse (vedi sopra e-f-g), mentre l'analisi delle co-occorrenze viene eseguita come segue:

a) normalizzazione dei seed vectors (vale a dire i profili delle co-occorrenze) corrispondenti alle 'k' categorie del dizionario importato;

b) calcolo degli indici del coseno e delle distanze euclidee tra ogni 'i' unità di contesto e ogni 'k' vettore 'seme';

c) assegnazione di ogni 'i' unità di contesto alla 'k' classe o categoria per la quale il seme corrispondente risulta il più simile (in questo caso, la massima somiglianza del coseno e la minima distanza euclidea devono coincidere, altrimenti **T-LAB** considerare la 'i' unità di contesto come non classificata).

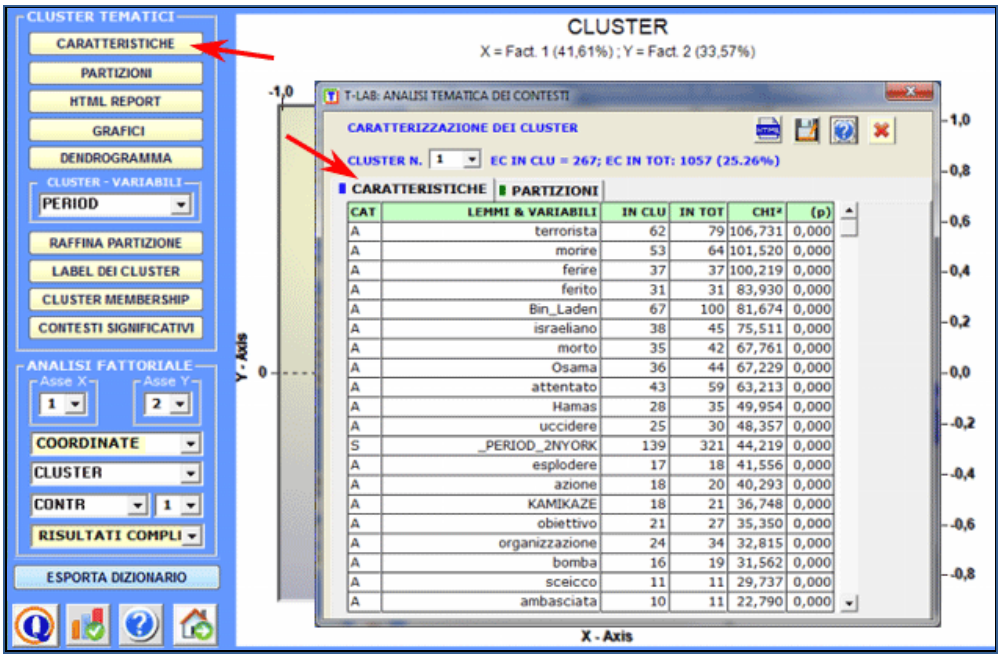
Al termine dell'analisi l'utilizzatore può agevolmente effettuare le seguenti operazioni:

1 - esplorare le caratteristiche dei cluster;
2 - esplorare le relazioni tra cluster;
3 - esplorare le relazioni tra cluster e variabili;
4 - esplorare le diverse partizioni dei cluster;
5 - raffinare i risultati della partizione prescelta e, se necessario, ripetere alcuni dei passi sopra descritti (1,2,3);
6 - assegnare label ai cluster;
7 - verificare quali contesti elementari appartengono a ciascun cluster;
8 - verificare il "peso" di ciascun contesto elementare entro il cluster a cui appartiene;
9- esportare una classificazione tematica dei documenti (solo nel caso in cui il corpus è costituito da almeno 2 documenti primari e questi non sono testi corti trattati come contesti elementari);

10- archiviare la partizione selezionata per esplorarla con altri strumenti **T-LAB**;
11- esportare un dizionario delle categorie;
12 – verificare la qualità della partizione scelta e la coerenza semantica dei vari temi;
13 - inoltre, quando il corpus è strutturato come un discorso o una conversazione, cioè quando le unità di contesto si succedono secondo un preciso ordine temporale, è possibile esplorare in modo dinamico le **sequenze di temi** (vedi sotto, parte finale di questa sezione).

Nel dettaglio:

1 - Esplorare le caratteristiche dei cluster



CAT	LEMMI & VARIABILI	IN CLU	IN TOT	CHI²	(p)
A	terrorista	62	79	106,731	0,000
A	morire	53	64	101,520	0,000
A	ferire	37	37	100,219	0,000
A	ferito	31	31	83,930	0,000
A	Bin_Laden	67	100	81,674	0,000
A	israeliano	38	45	75,511	0,000
A	morto	35	42	67,761	0,000
A	Osama	36	44	67,229	0,000
A	attentato	43	59	63,213	0,000
A	Hamas	28	35	49,954	0,000
A	uccidere	25	30	48,357	0,000
S	_PERIOD_2NYORK	139	321	44,219	0,000
A	esplodere	17	18	41,556	0,000
A	azione	18	20	40,293	0,000
A	KAMIKAZE	18	21	36,748	0,000
A	obiettivo	21	27	35,350	0,000
A	organizzazione	24	34	32,815	0,000
A	bomba	16	19	31,562	0,000
A	sceicco	11	11	29,737	0,000
A	ambasciata	10	11	22,790	0,000

Cliccando il pulsante **CARATTERISTICHE**, per ogni cluster vengono mostrate le unità lessicali e le variabili che lo caratterizzano; e, per ciascuna di esse (unità lessicali o variabili), sono riportati: i valori del chi quadro e le sommatorie dei contesti elementari in cui risulta presente, sia all'interno del cluster selezionato ("IN CLUST") che all'interno dell'insieme analizzato ("IN TOT"). Inoltre, nella colonna "CAT", viene indicato se la caratteristica è stata selezionata dall'utilizzatore nella funzione Impostazioni di Analisi ("A") oppure se è stata proposta da **T-LAB** come descrizione "supplementare" ("S").

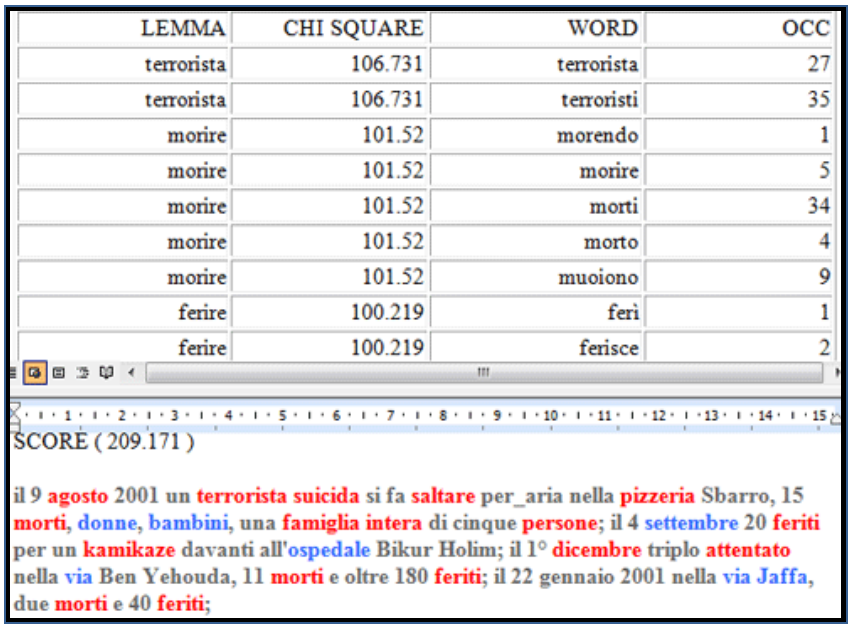
Nel caso del test del chi quadro la struttura della tabella analizzata è la seguente:

	Cluster "A"	Other Clusters	
Word "a"	n_{ij}		N_j
Other Words			
	N_i		N

Dove:

$\mathbf{n_{ij}}$ si riferisce alle occorrenze della parola (a) all'interno del cluster selezionato (A);
$\mathbf{N_j}$ si riferisce a tutte le occorrenze della parola (a) all'interno del corpus (o del sottoinsieme) in analisi;
$\mathbf{N_i}$ si riferisce a tutte le occorrenze all'interno del cluster selezionato (A);
N si riferisce a tutte le occorrenze della tabella di contingenza parole per cluster.

Un **report HTML** (vedi sotto) consente una dettagliata verifica delle caratteristiche dei cluster. In questo, oltre alla lista delle parole tipiche, vengono mostrati - ordinati in modo decrescente in base al rispettivo peso (score) - i contesti elementari che più caratterizzano il cluster in esame.

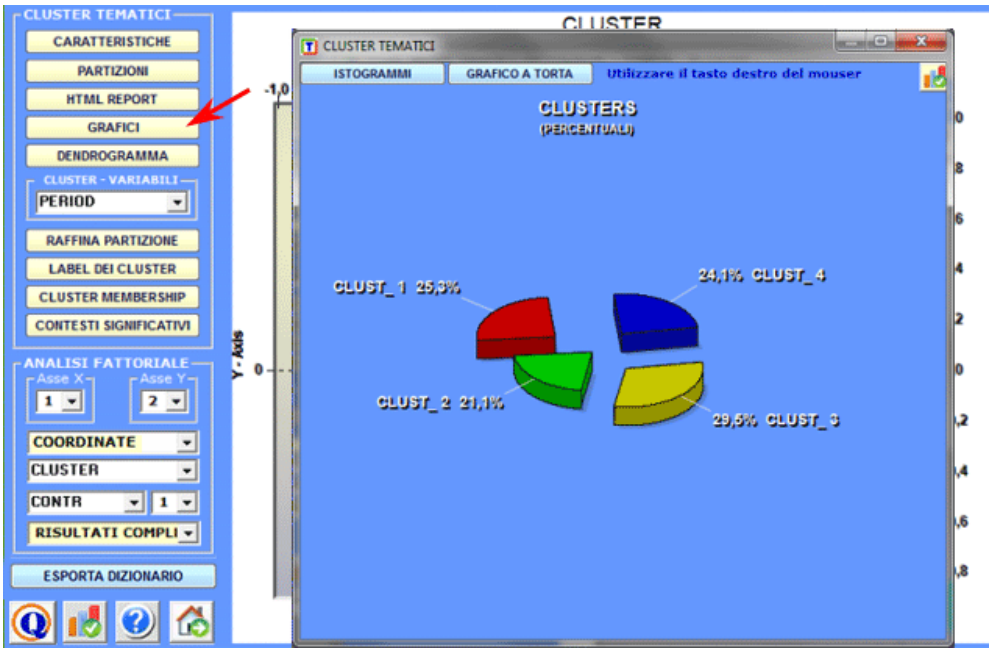


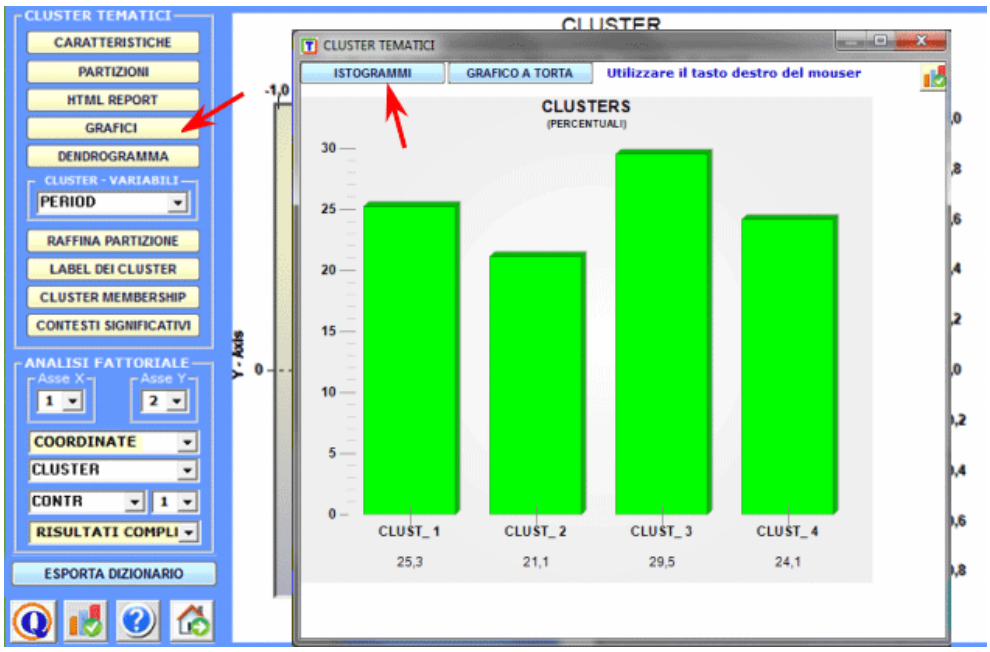
LEMMA	CHI SQUARE	WORD	OCC
terrorista	106.731	terrorista	27
terrorista	106.731	terroristi	35
morire	101.52	morendo	1
morire	101.52	morire	5
morire	101.52	morti	34
morire	101.52	morto	4
morire	101.52	muoiono	9
ferire	100.219	ferì	1
ferire	100.219	ferisce	2

SCORE (209.171)

il 9 agosto 2001 un terrorista suicida si fa saltare per_aria nella pizzeria Sbarro, 15 morti, donne, bambini, una famiglia intera di cinque persone; il 4 settembre 20 feriti per un kamikaze davanti all'ospedale Bikur Holim; il 1° dicembre triplo attentato nella via Ben Yehouda, 11 morti e oltre 180 feriti; il 22 gennaio 2001 nella via Jaffa, due morti e 40 feriti;

Grafici a torta e **Istogrammi** (vedi sotto) consentono di verificare la percentuale delle unità di contesto appartenenti ad ogni cluster.


CLUSTER TEMATICI
CARATTERISTICHE
PARTIZIONI
HTML REPORT
GRAFICI
DENDROGRAMMA
CLUSTER - VARIABILI
PERIOD
RAFFINA PARTIZIONE
LABEL DEI CLUSTER
CLUSTER MEMBERSHIP
CONTESTI SIGNIFICATIVI
ANALISI FATTORIALE
Asse X
Asse Y
1
2
COORDINATE
CLUSTER
CONTR
1
RISULTATI COMPLI
ESPORTA DIZIONARIO
CLUSTER
CLUSTER TEMATICI
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouser
CLUSTERS
(PERCENTUALI)
CLUST_1 25,3%
24,1% CLUST_4
CLUST_2 21,1%
29,5% CLUST_3
Y - Axis

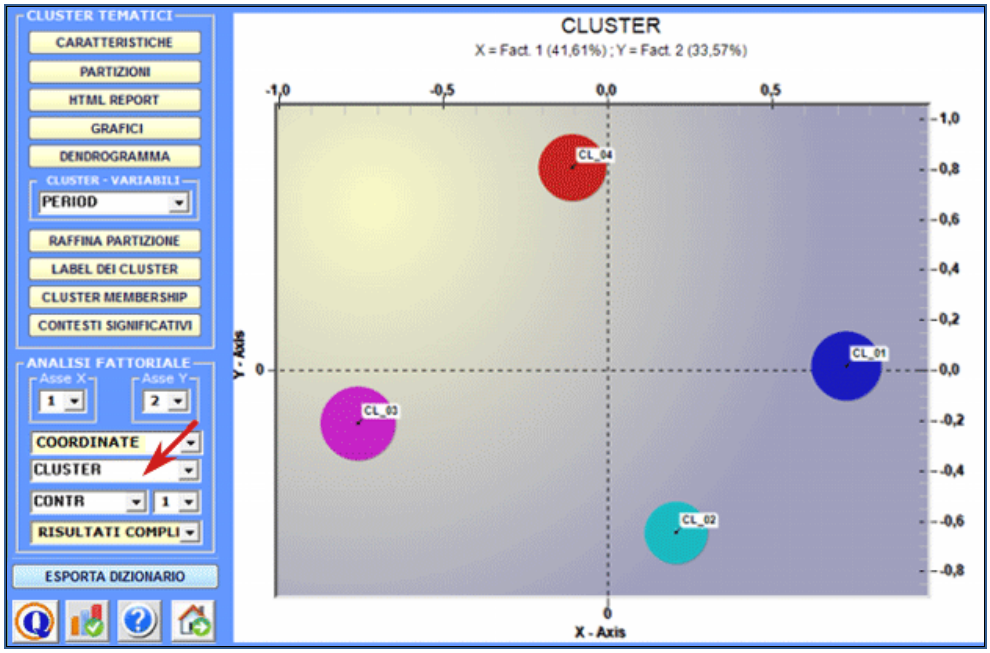

CLUSTER TEMATICI
CARATTERISTICHE
PARTIZIONI
HTML REPORT
GRAFICI
DENDROGRAMMA
CLUSTER - VARIABILI
PERIOD
RAFFINA PARTIZIONE
LABEL DEI CLUSTER
CLUSTER MEMBERSHIP
CONTESTI SIGNIFICATIVI
ANALISI FATTORIALE
Asse X
Asse Y
COORDINATE
CLUSTER
CONTR
RISULTATI COMPLI
ESPORTA DIZIONARIO
CLUSTER TEMATICI
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouser
CLUSTERS
(PERCENTUALI)
30
25
20
15
10
5
0
CLUST_1
CLUST_2
CLUST_3
CLUST_4
25,3
21,1
29,5
24,1
Y - Axis

2 - Esplorare le relazioni tra cluster

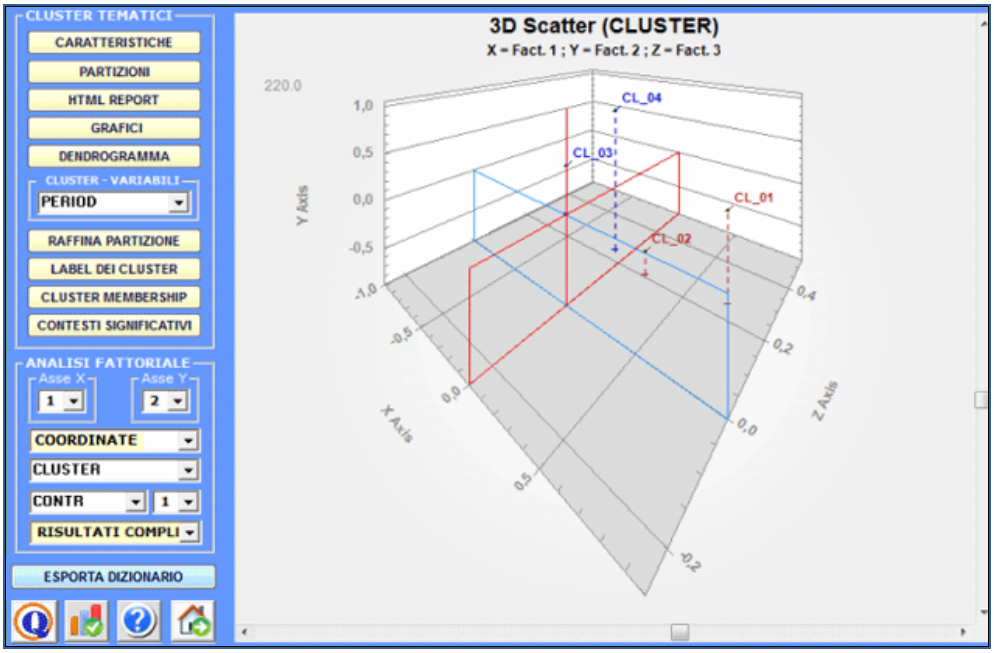
Alcuni grafici, ottenuti tramite **Analisi delle Corrispondenze** consentono di esplorare le relazioni tra i cluster all'interno di spazi bidimensionali.

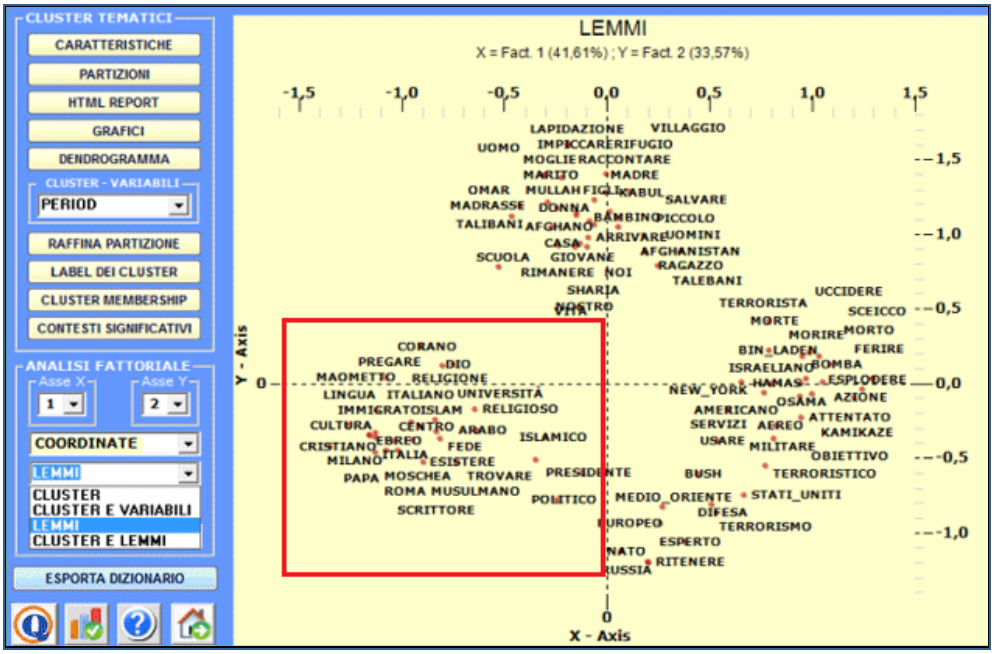
Più specificamente:

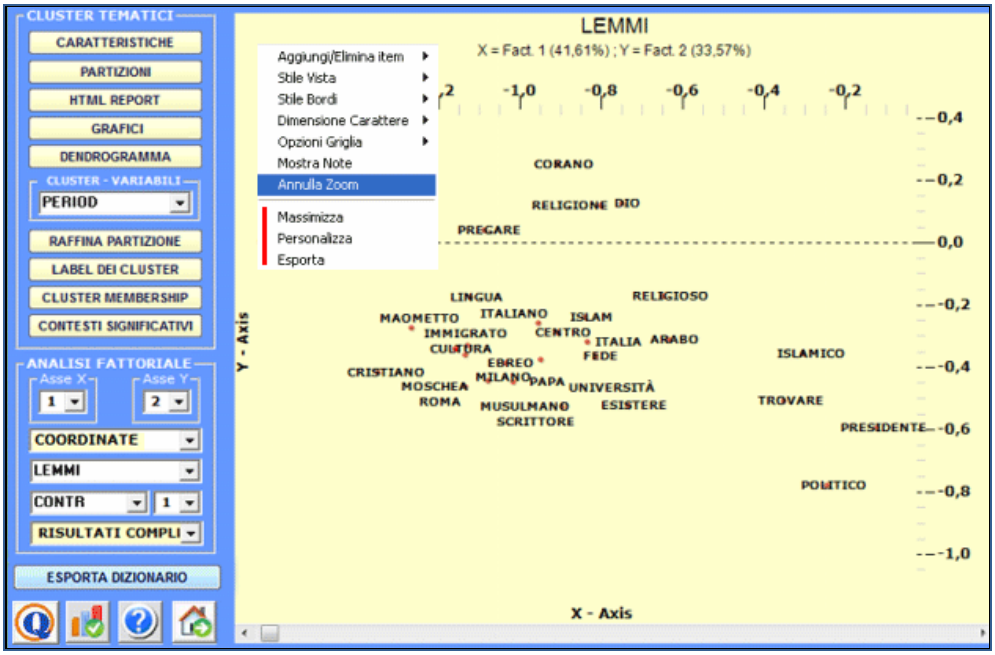
- Per esplorare le varie combinazioni degli assi fattoriali è sufficiente selezionarli negli appositi box ("Asse X", "Asse Y");- Per ciascuna delle combinazioni (X-Y), è possibile visualizzare vari tipi di elementi (cluster, lemmi e variabili).



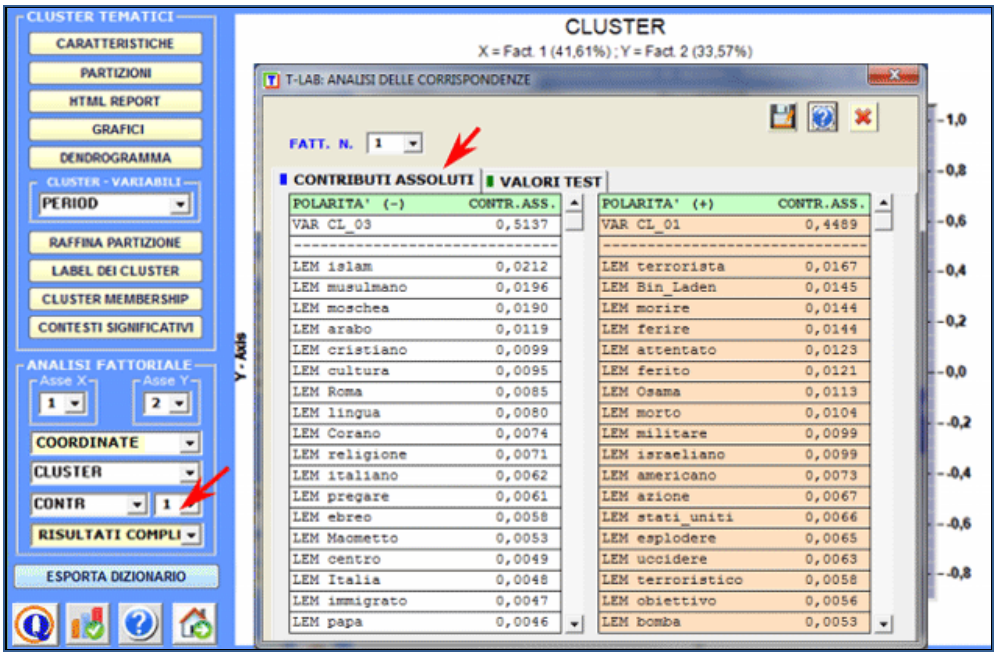
Tutti i grafici possono essere personalizzati tramite l'uso di apposite finestre di dialogo (uso del tasto destro del mouse). Inoltre quando i cluster tematici sono più di tre, le loro relazioni possono essere esplorate tramite grafici 3D (vedi sotto).



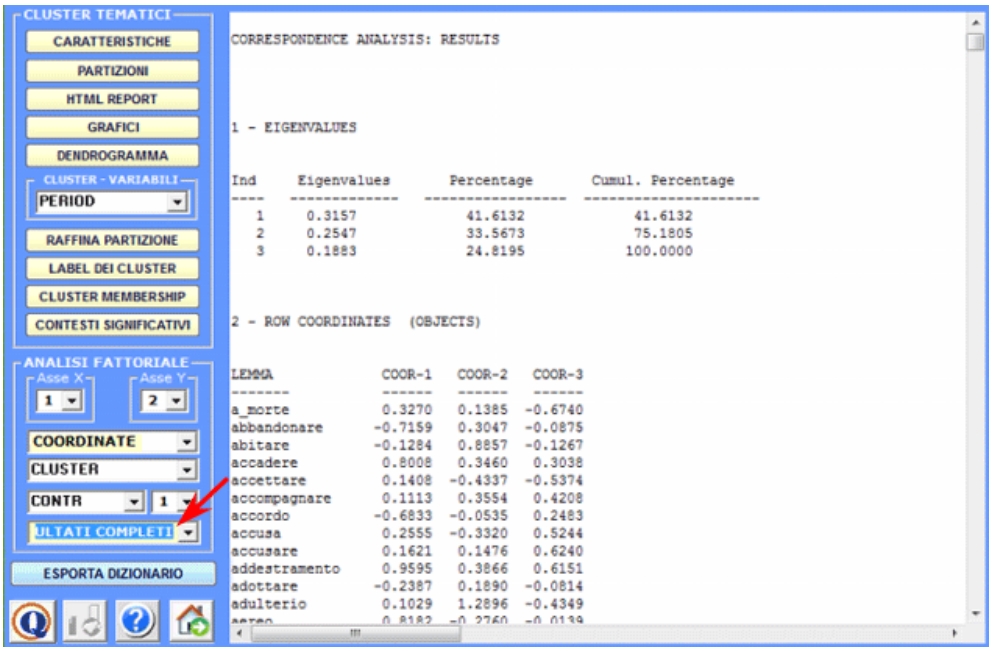




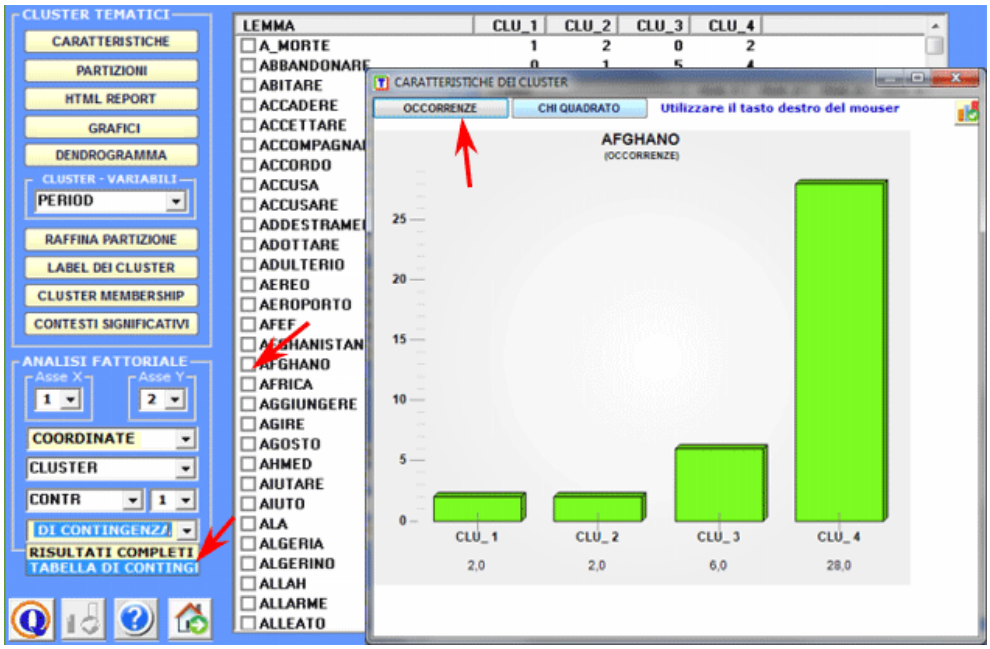
Come nella funzione Analisi delle Corrispondenze, selezionando gli appositi pulsanti (vedi sotto) compaiono delle tabelle che riportano le caratteristiche delle polarità fattoriali in esame (X-Y).

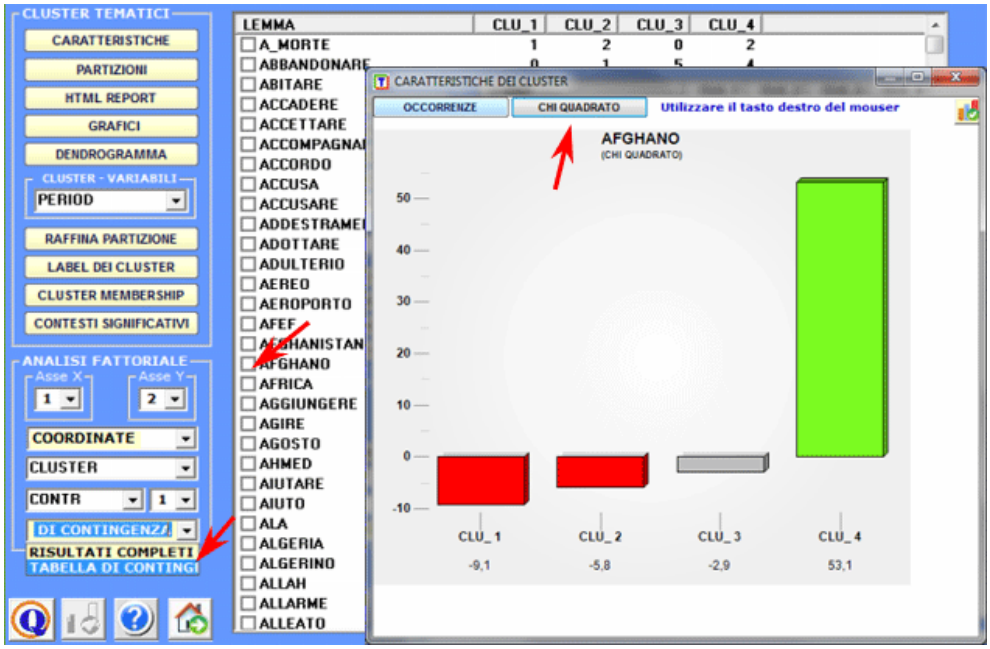


Una specifica opzione (vedi sotto) consente di visualizzare ed esportare i **Risultati Completi** dell'analisi delle corrispondenze unità lessicali x cluster.



Una ulteriore opzione (vedi sotto) consente di visualizzare/esportare la **Tabella di Contingenza** e di creare grafici che mostrano sia le distribuzioni delle singole parole all'interno dei cluster che i rispettivi valori del chi quadrato.



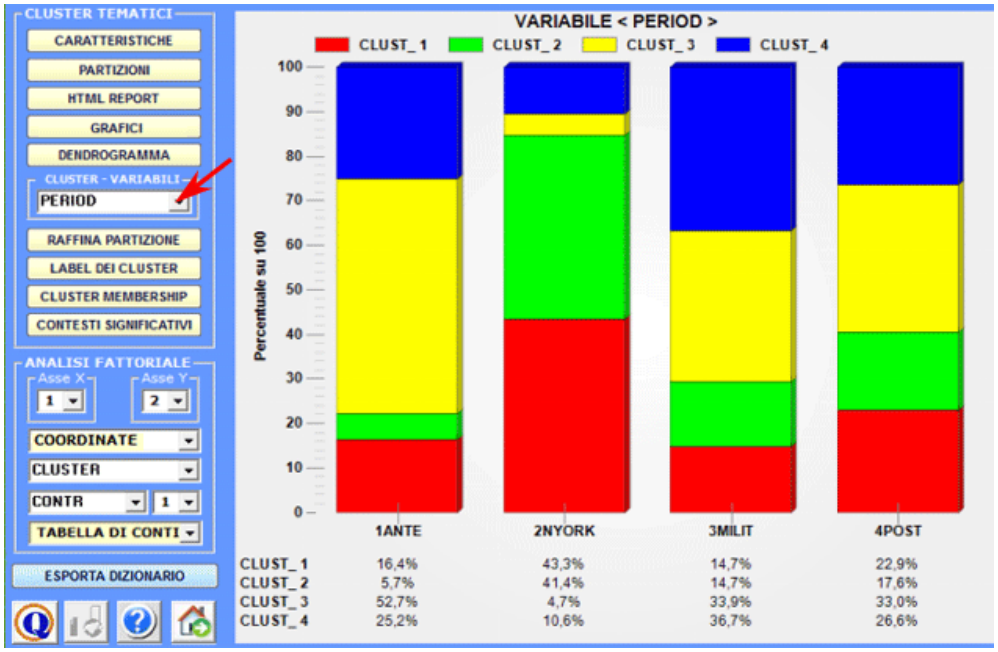


Inoltre, cliccando su specifiche celle della stessa tabella, è possibile creare file HTML con tutti i contesti elementari in cui la parola in riga è presente nel cluster in colonna.

N.B.: In questa tabella sono incluse sia le parole 'attive' ('A') che quelle 'supplementari' ('S').

3 - Esplorare le relazioni tra cluster e variabili

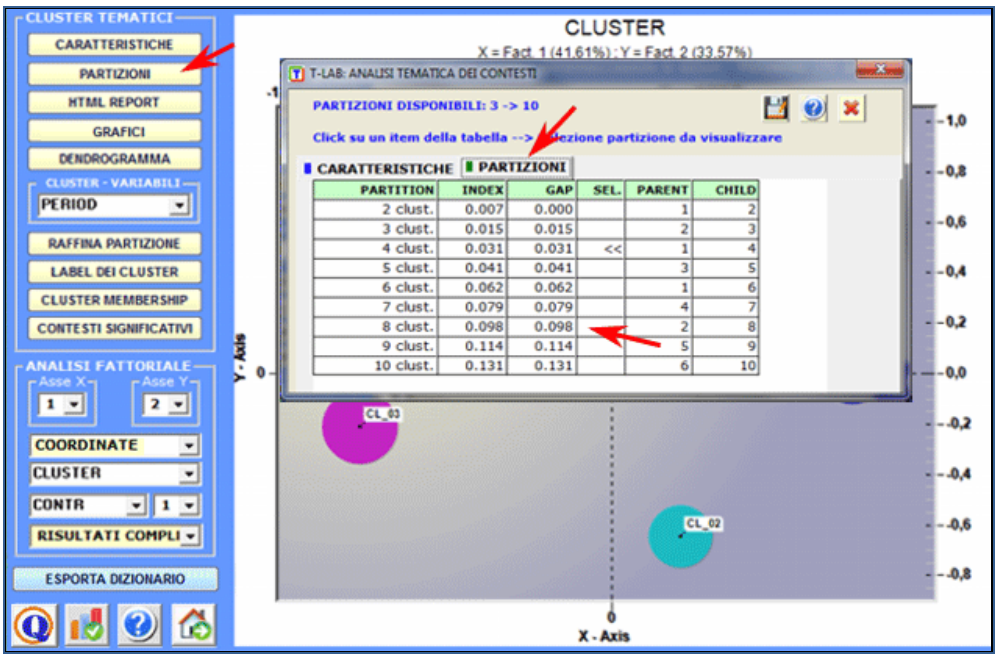
Alcuni istogrammi consentono di verificare le relazioni tra cluster e modalità delle variabili.



Ulteriori relazioni tra cluster e variabili possono essere esplorate con le opzioni disponibili nella sezione **Analisi Fattoriale** (vedi sopra)

4 - Esplorare le diverse partizioni

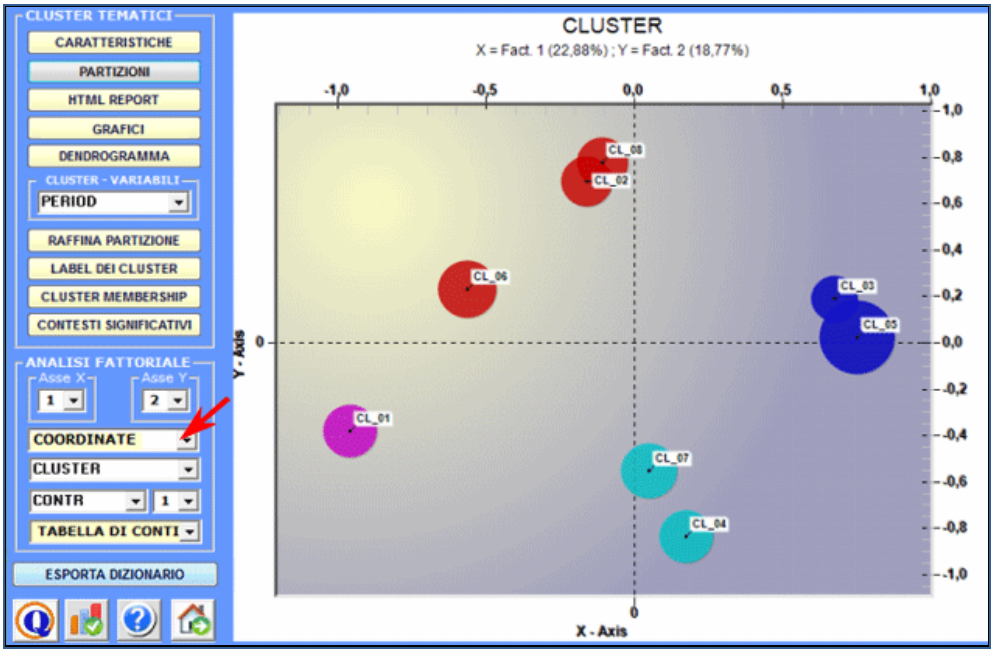
Poiché l'algoritmo usato da **T-LAB** (bisecting K-Means) produce una clusterizzazione gerarchica, l'utilizzatore può agevolmente esplorare diverse soluzioni dell'analisi: partizioni da 3 a 50 clusters.



Per ogni partizione ottenuta, un'apposita tabella (vedi sopra) riporta i seguenti valori:

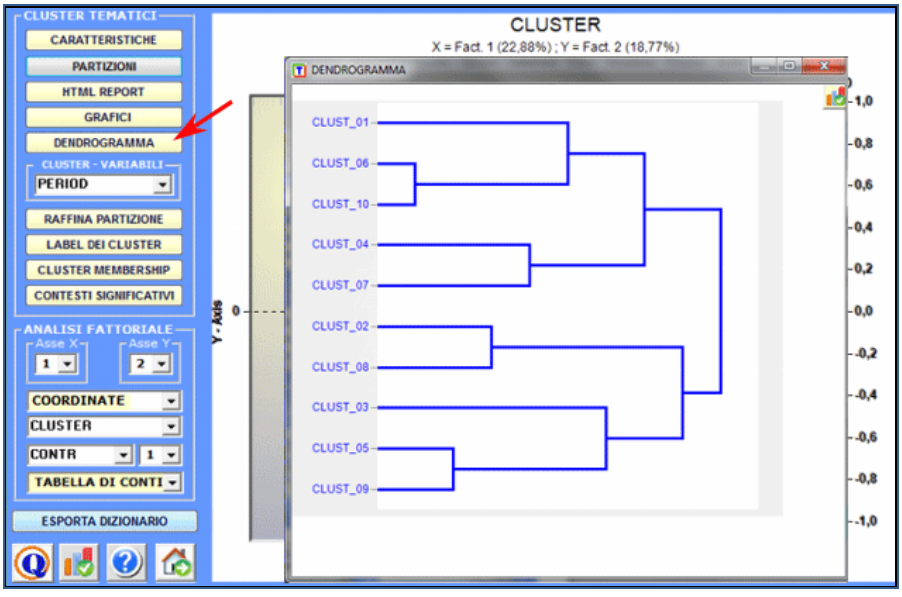
- "Index", che corrisponde al rapporto tra varianza intercluster e varianza totale;
- "Gap", che indica la differenza tra il valore dell'index e quello della partizione immediatamente precedente:
- Numero del cluster "figlio" (child) ottenuto attraverso dalla bi-sezione del corrispondente "genitore" ("parent").

L'opzione **partizioni** (vedi sopra) consente di esplorare agevolmente le caratteristiche delle varie soluzioni disponibili.

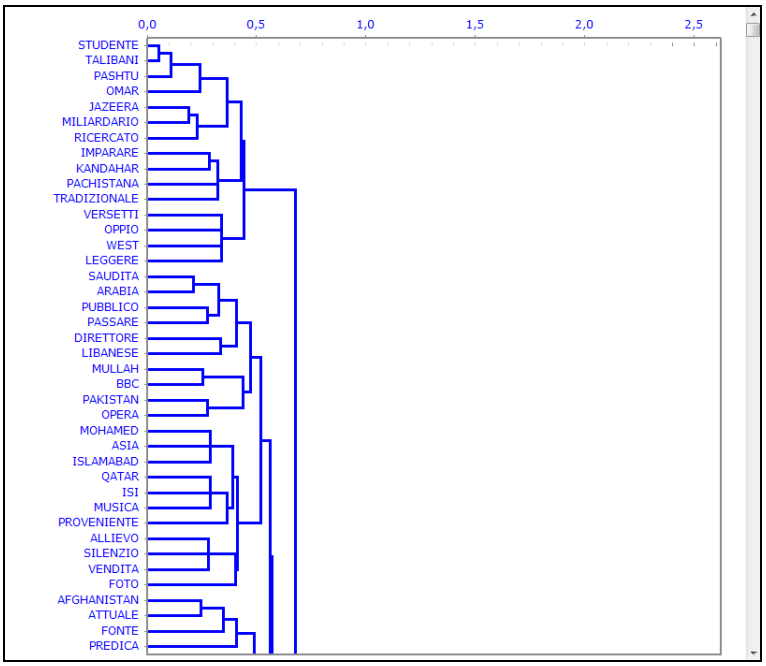


Inoltre l'opzione **dendrogramma** (vedi sotto) consente due possibilità:

A) verificare l'albero delle varie bi-sezioni dei cluster



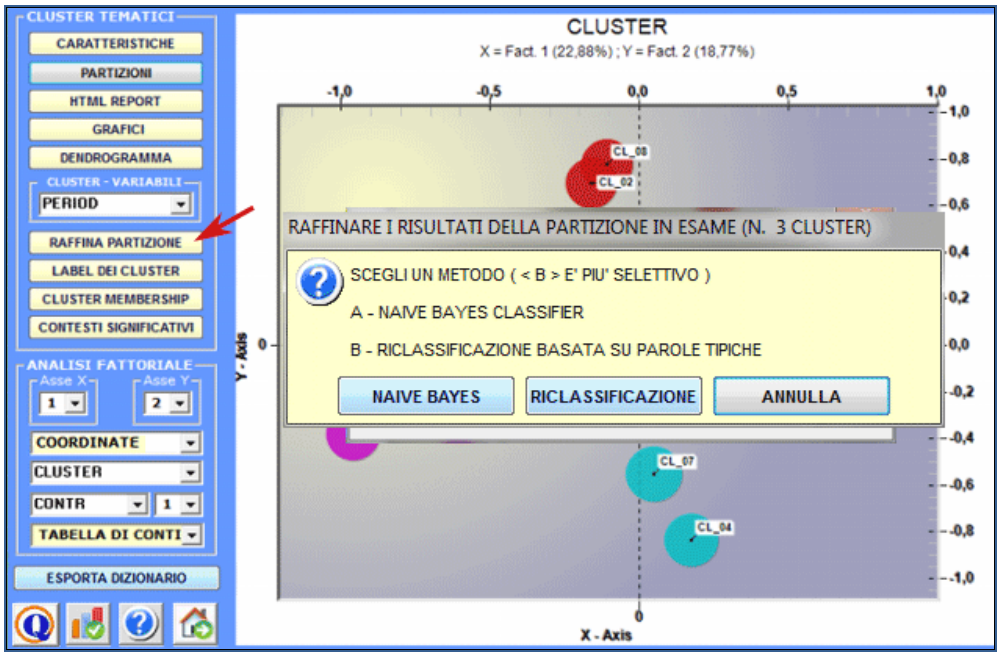
B) verificare l'albero delle parole caratteristiche a ciascun cluster.



5 - Raffinare i risultati della partizione prescelta

Dopo aver esplorato diverse soluzioni, l'utilizzatore può raffinare i risultati della partizione prescelta e, se necessario, ripetere alcuni dei passi sopra descritti (1,2,3).

A questo scopo sono disponibili due metodi (vedi immagine seguente).



Quando viene scelto il metodo '**A**' (cioè **Naïve Bayes Classifier**), questa funzione T-LAB consente di escludere dall'analisi tutte le unità di contesto la cui appartenenza a un cluster non soddisfa i seguenti criteri:
a) per ogni unità di contesto, il cluster di appartenenza determinato mediante l'algoritmo del bisecting K-Means (unsupervised clustering) e quello determinato mediante il Naïve Bayes Classifier (supervised clustering) deve essere il medesimo;
b) il massimo valore della probabilità a posteriori, corrispondente all'appartenenza della i-unità di contesto al k-cluster, deve essere – in termini percentuali – superiore di almeno il 50 % ai valori delle probabilità a posteriori computate per la stessa i-unità di contesto nei rimanenti cluster.

Diversamente, nel caso del metodo di '**B**' (cioè **Riclassificazione basata su base Parole Tipiche**) **T-LAB** considera le caratteristiche del cluster, cioè le parole con un significativo valore de Chi-Quadro, come item di un dizionario delle categorie ed esegue le tre fasi della 'classificazione supervisionata' descritte all'inizio di questa sezione. Quindi, quando l'utente è interessato a ri-applicare dizionari e a comparare i relativi risultati, si consiglia vivamente di usare questo metodo.

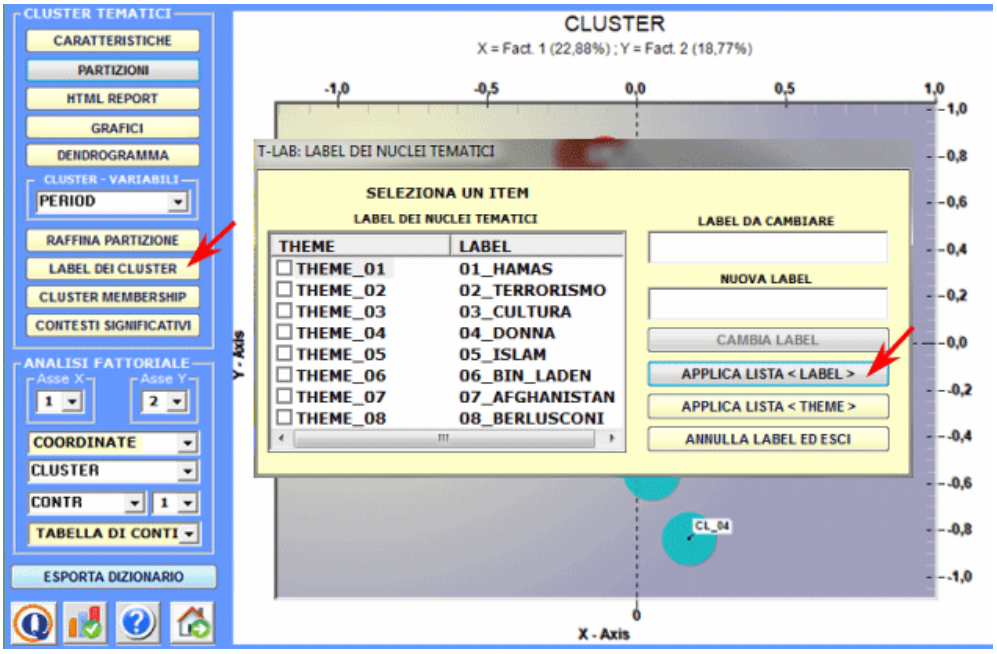
Tutti i risultati di questo calcolo sono in una tabella esportata da T-LAB (vedi sotto), che contiene i valori delle probabilità a posteriori espressi in termini percentuali.

Context_ID	OLD	NEW	MATC	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CONTEXT
'00001000001	3	3	YES	0	0	1	0	0	0	0	0	PARTITO DI ALLAH In pubblico il
'00001000002	5	5	YES	0	0	0	0	1	0	0	0	La disfatta è_stata evitata in extre
'00001000003	5	5	YES	0	0	0	0	1	0	0	0	Secondo e più grave motivo di pre
'00001000004	8	8	YES	0	0	0	0	0	0	0	1	Uno schiaffo al regime , che avev
'00001000005	1	1	YES	1	0	0	0	0	0	0	0	In molte circoscrizioni nel Delta e
'00001000006	8	8	YES	0	0	0	0	0	0	0	1	Il regime si trova a gestire una sit
'00001000007	5	5	YES	0	0	0	0	1	0	0	0	Alcuni dirigenti sono a favore dell
'00001000008	8	8	YES	0	0	0	0	0	0	0	1	' ' Sbagliano ' ' continua Salal
'00001000009	1	1	YES	1	0	0	0	0	0	0	0	L ' editorialista del quotidiano Al-
'00002000001	6	6	YES	0	0	0	0	0	1	0	0	I VERI BERSAGLI DI Bin_Laden
'00002000002	5	5	YES	0	0	0	0	1	0	0	0	Sono uno studente universitario d
'00002000003	6	6	YES	0	0	0	0	0	1	0	0	senza spiegarne i motivi , o meg
'00002000004	1	1	YES	1	0	0	0	0	0	0	0	Forse vuole aiutare la causa pale
'00002000005	5	5	YES	0	0	0	0	1	0	0	0	Per una parte dell ' Islam , come
'00002000006	2	2	YES	0	1	0	0	0	0	0	0	le fazioni radicali del mondo mus
'00002000007	5	5	YES	0	0	0	0	1	0	0	0	É bene ricordare che i maggiori b
'00002000008	6	6	YES	0	0	0	0	0	0,89	0,11	0	In alcune circostanze questi regir
'00002000009	7	7	YES	0	0	0	0	0	0	1	0	É accaduto in Algeria , dove il si
'00002000010	5	5	YES	0	0	0	0	1	0	0	0	In Palestina la nascita di uno stat
'00002000011	6	6	YES	0	0	0	0	0	1	0	0	É il simbolo del capitalismo , è l
'00002000012	1	1	YES	1	0	0	0	0	0	0	0	Ma non commetta l ' errore di cre
'00002000013	7	7	YES	0	0	0	0	0	0	1	0	Le esecuzioni avvengono davanti
'00002000014	1	1	YES	1	0	0	0	0	0	0	0	Ma oggi , per la_prima_volta , ar
'00002000015	1	1	YES	1	0	0	0	0	0	0	0	tant ' è che gli uccisi accusati di
'00003000001	7	7	YES	0	0	0	0	0	0	1	0	É QUI LA SCUOLA DEL TERRO
'00003000002	4	4	YES	0	0	0	1	0	0	0	0	E tutto gira intorno a un uomo mi
'00003000003	7	7	YES	0	0	0	0	0	0	1	0	É questa l ' età in cui i bambini e
'00003000004	5	5	YES	0	0	0	0	1	0	0	0	la guerra santa contro gli infedeli
'00003000005	7	7	YES	0	0	0	0	0	0	1	0	tra le montagne dell ' Afghanistar
'00003000006	7	7	YES	0	0	0	0	0	0	1	0	Descritto nei minimi dettagli delle

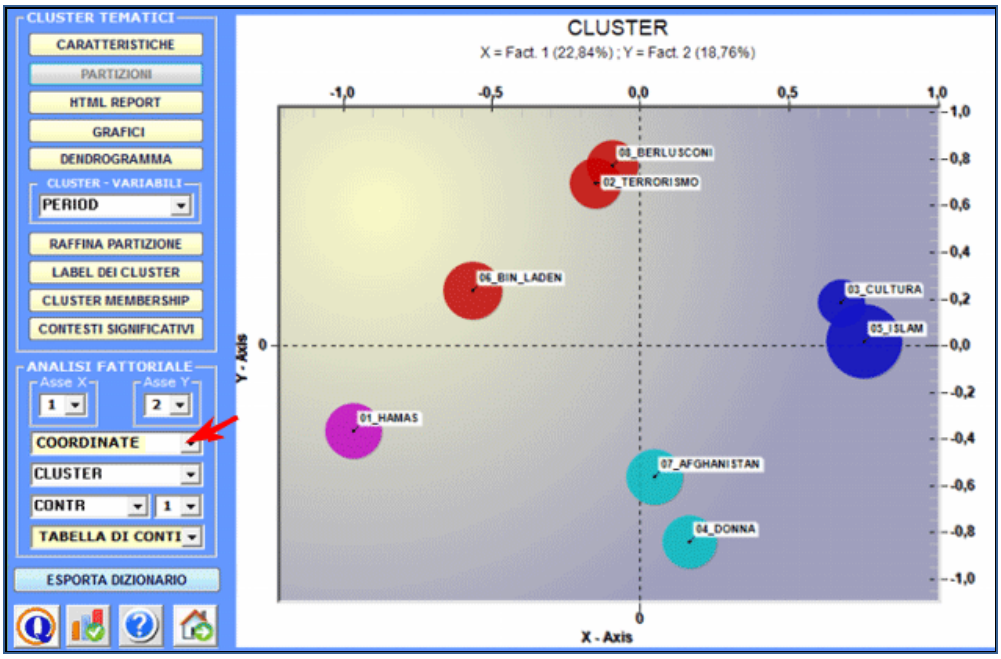
6- Assegnare label ai cluster

Un'apposita funzione **T-LAB** consente di attribuire label ai cluster.

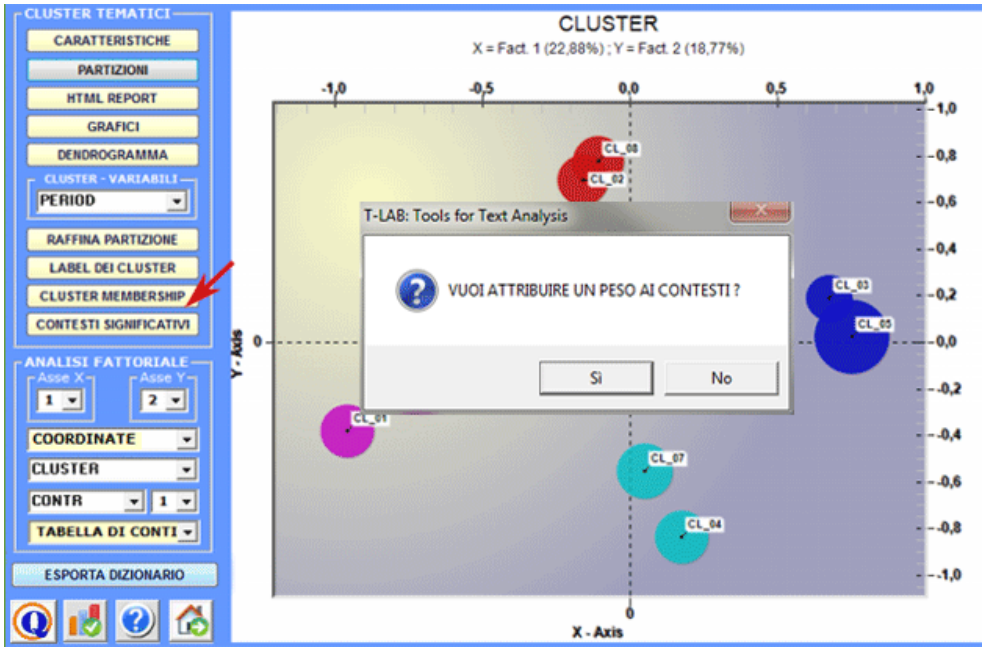
(N.B: Al primo uso alcune label sono proposte automaticamente dal software).



Le label attribuite ai vari cluster possono essere visualizzate nei vari grafici disponibili (vedi sotto).

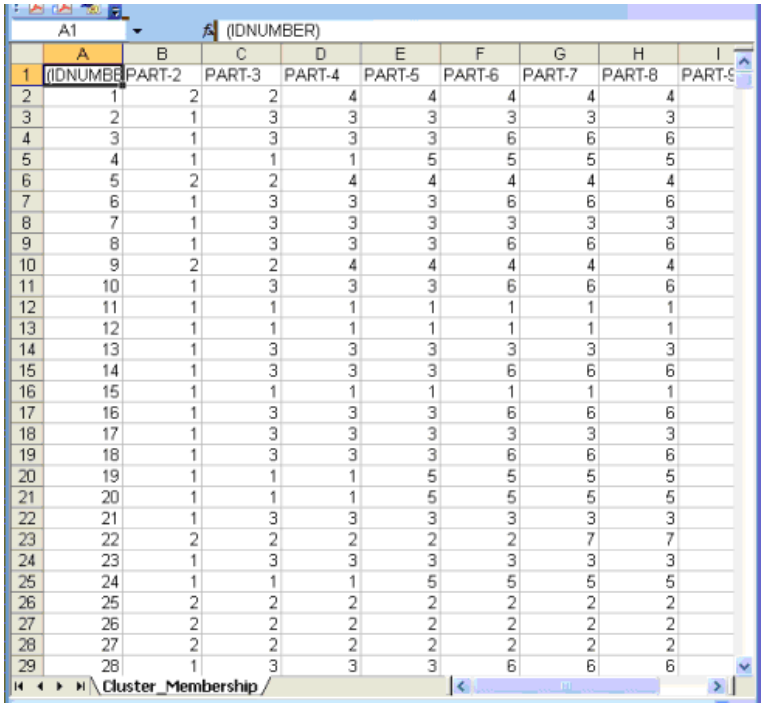


7 - Verificare quali contesti elementari appartengono a ciascun cluster; (8) Verificare il "peso" di ciascun contesto elementare entro il cluster a cui appartiene; (9) Esportare una classificazione tematica dei documenti.



Infatti il pulsante **Cluster Membership** consente di esportare tre tipi di tabelle (vedi sotto) in formato MS Excel:

a – "Cluster_Partitions.xls", con tutte le corrispondenze unità di contesto x cluster all'interno delle varie partizioni;



	A	B	C	D	E	F	G	H	I
1	(IDNUMBE	PART-2	PART-3	PART-4	PART-5	PART-6	PART-7	PART-8	PART-9
2	1	2	2	4	4	4	4	4	
3	2	1	3	3	3	3	3	3	
4	3	1	3	3	3	6	6	6	
5	4	1	1	1	5	5	5	5	
6	5	2	2	4	4	4	4	4	
7	6	1	3	3	3	6	6	6	
8	7	1	3	3	3	3	3	3	
9	8	1	3	3	3	6	6	6	
10	9	2	2	4	4	4	4	4	
11	10	1	3	3	3	6	6	6	
12	11	1	1	1	1	1	1	1	
13	12	1	1	1	1	1	1	1	
14	13	1	3	3	3	3	3	3	
15	14	1	3	3	3	6	6	6	
16	15	1	1	1	1	1	1	1	
17	16	1	3	3	3	6	6	6	
18	17	1	3	3	3	3	3	3	
19	18	1	3	3	3	6	6	6	
20	19	1	1	1	5	5	5	5	
21	20	1	1	1	5	5	5	5	
22	21	1	3	3	3	3	3	3	
23	22	2	2	2	2	2	7	7	
24	23	1	3	3	3	3	3	3	
25	24	1	1	1	5	5	5	5	
26	25	2	2	2	2	2	2	2	
27	26	2	2	2	2	2	2	2	
28	27	2	2	2	2	2	2	2	
29	28	1	3	3	3	6	6	6	

Cluster_Membership

b – "Themes-Contexts.xls" (vedi sotto) con le corrispondenze unità di contesto x cluster all'interno della partizione selezionata.

(IDNUMBER)	THEME	SCORE	CONTEXT
'00001000001	03_CULTURA	5,91	PARTITO DI ALLAH In pubblico il presidente Mubarak ribadisce di essere sod
'00001000002	05_ISLAM	0,9	La disfatta è_stata evitata in extremis grazie all ' appoggio degli indipendenti
'00001000003	05_ISLAM	4,96	Secondo e più grave motivo di preoccupazione : gli odiati Fratelli musulmani
'00001000004	08_BERLUSCONI	0,44	Uno schiaffo al regime , che aveva perseguitato gli islamici proprio in previsio
'00001000005	01_HAMAS	13,66	In molte circoscrizioni nel Delta e del Cairo , ritenute capisaldi della confrater
'00001000006	08_BERLUSCONI	0,99	Il regime si trova a gestire una situazione paradossale : insiste nel considera
'00001000007	05_ISLAM	0,3	Alcuni dirigenti sono a favore della riabilitazione della confraternita che , sost
'00001000008	08_BERLUSCONI	3,93	' ' Sbagliano ' ' continua Salah ' ' perché gli islamici hanno dimostrato d
'00001000009	01_HAMAS	13,97	L ' editorialista del quotidiano Al-Ahram Mohammed Said Ahmed è pessimist
'00002000001	06_BIN_LADEN	3,62	I VERI BERSAGLI DI Bin_Laden Riformatori arabi e musulmani : ecco i nemi
'00002000002	05_ISLAM	0,91	Sono uno studente universitario di 19 anni e , sebbene la riforma Berlinguer a
'00002000003	06_BIN_LADEN	14,81	senza spiegarne i motivi , o meglio senza provare a dare una spiegazione ch
'00002000004	01_HAMAS	2,93	Forse vuole aiutare la causa palestinese ? E perché mai si sentiva e si sente
'00002000005	05_ISLAM	16	Per una parte dell ' Islam , come del_resto per altre correnti delle grandi relig
'00002000006	02_TERRORISMO	0,4	le fazioni radicali del mondo musulmano cercarono di mobilitare le masse cor
'00002000007	05_ISLAM	10,02	É bene ricordare che i maggiori bersagli dell ' islamismo radicale furono anzit
'00002000008	06_BIN_LADEN	0	In alcune circostanze questi regimi secolari giustificarono la protesta popolare
'00002000009	07_AFGHANISTAN	4,69	É accaduto in Algeria , dove il sistema socialista , ispirato al modello sovieti
'00002000010	05_ISLAM	1,17	In Palestina la nascita di uno stato ebraico ha permesso a due grandi movime
'00002000011	06_BIN_LADEN	6,44	É il simbolo del capitalismo , è l ' espressione più compiuta della modernità
'00002000012	01_HAMAS	6,76	Ma non commetta l ' errore di credere che il fondamentalismo rappresenti l '
'00002000013	07_AFGHANISTAN	1,12	Le esecuzioni avvengono davanti alle porte delle case dei condannati , davant
'00002000014	01_HAMAS	7,17	Ma oggi , per la_prima_volta , arabi palestinesi hanno parlato alla tivù israeli
'00002000015	01_HAMAS	10,05	tant ' è che gli uccisi accusati di collusione con Israele sono persone che ha
'00003000001	07_AFGHANISTAN	14,63	É QUI LA SCUOLA DEL TERRORISMO AFGHANISTAN RAPPORTO DA UN
'00003000002	04_DONNA	13,17	E tutto gira intorno a un uomo misterioso , il mullah Omar . E a tonnellate di
'00003000003	07_AFGHANISTAN	3,38	É questa l ' età in cui i bambini entrano nelle madrassa , le scuole coraniche
'00003000004	05_ISLAM	31,6	la guerra santa contro gli infedeli o gli stessi musulmani che si sono distacca
'00003000005	07_AFGHANISTAN	11,17	tra le montagne dell ' Afghanistan e tra gli altipiani e le pianure semidesertich
'00003000006	07_AFGHANISTAN	7,32	Descritto nei minimi dettagli delle sue grotte e delle sue cime innevate da Ed

In particolare, il valore di rilevanza (score) assegnato ad ogni j-contesto elementare appartenente al k-cluster è calcolato nel modo seguente:

$$score_j = \sum_{j \in k} X_{i,j} \times \frac{n_j}{N}$$

Dove:

Scorej = valore di rilevanza attribuito al contesto elementare (j);

$\mathbf{\Sigma X_{ij}}$ = somma dei valori del chi-quadrato corrispondenti alle parole chiave (i) trovate nel contesto elementare in questione (j) e che sono risultate tipiche del cluster (k);

nj = totale delle parole chiave (parole distinte), tipiche del cluster (k), trovate nel contesto elementare (j);

N = totale delle parole chiave (parole distinte) tipiche del cluster (k).

c - " Ec_Document_Classification.xls" (file output fornito soltanto quando il corpus si compone almeno di almeno 2 documenti primari e questi non sono testi corti trattati come contesti elementari) che elenca le “appartenenze miste” di ogni documento (vedi sotto).

DOC_ID	VAR_01	BEST_CL	CLUST_1	CLUST_2	CLUST_3	CLUST_4	CLUST_5	CLUST_6	CLUST_7	CLUST_8
1	PE_1ANTE	1	0,613	0	0,131	0	0,137	0	0	0,119
2	PE_1ANTE	5	0,313	0,005	0	0	0,326	0,289	0,067	0
3	PE_1ANTE	7	0,028	0	0,042	0,224	0,261	0,055	0,381	0,009
4	PE_1ANTE	3	0	0	0,579	0	0,409	0	0,012	0
5	PE_1ANTE	5	0	0	0,101	0,083	0,796	0	0,016	0,004
6	PE_1ANTE	3	0	0	0,47	0,342	0,156	0	0,002	0,031
7	PE_1ANTE	3	0	0	0,745	0	0,255	0	0	0
8	PE_1ANTE	5	0	0	0	0	1	0	0	0
9	PE_1ANTE	3	0	0,027	0,902	0,029	0	0,008	0,019	0,015
10	PE_1ANTE	5	0,005	0,013	0,185	0	0,797	0	0	0
11	PE_1ANTE	1	0,858	0,037	0,077	0,006	0	0,009	0,013	0
12	PE_1ANTE	5	0	0	0	0	1	0	0	0
13	PE_1ANTE	3	0	0	0,665	0	0,335	0	0	0
14	PE_2NYORK	2	0	0,615	0	0	0	0,045	0	0,34
15	PE_2NYORK	1	0,881	0,003	0	0	0,008	0,088	0,014	0,008
16	PE_2NYORK	6	0,003	0,2	0	0,012	0	0,704	0	0,081
17	PE_2NYORK	1	0,489	0,217	0	0	0,018	0,24	0,002	0,034
18	PE_2NYORK	6	0,035	0,153	0	0	0	0,775	0	0,038
19	PE_2NYORK	8	0,132	0,073	0	0	0	0,206	0,007	0,582
20	PE_2NYORK	8	0	0,237	0	0	0,079	0,098	0	0,586
21	PE_2NYORK	2	0,085	0,534	0	0	0	0,264	0	0,117
22	PE_2NYORK	2	0,108	0,845	0	0	0	0,001	0,028	0,017
23	PE_2NYORK	2	0,27	0,286	0,034	0	0,082	0,136	0,025	0,166
24	PE_2NYORK	4	0,161	0,046	0,029	0,588	0,001	0,062	0,113	0
25	PE_2NYORK	1	0,47	0,141	0	0	0	0,182	0,005	0,202
26	PE_2NYORK	1	0,637	0	0	0,007	0,283	0,024	0,049	0
27	PE_3MILIT	4	0	0	0,01	0,853	0,119	0	0,014	0,004
28	PE_3MILIT	5	0	0,007	0,301	0	0,637	0	0	0,054
29	PE_3MILIT	3	0	0,03	0,403	0	0,205	0,183	0	0,179
30	PE_3MILIT	2	0	0,557	0,15	0,121	0,172	0	0	0

In questo caso i valori derivano dalla formula già illustrata (vedi punto “b"), sommando gli score dei contesti elementari appartenenti a ogni documento ed applicando un calcolo di percentuali.

10 - Archiviare la partizione selezionata per esplorarla con altri strumenti T-LAB

All'uscita dalla funzione Analisi Tematica dei Contesti Elementari, alcuni messaggi ricordano che è possibile esplorare i cluster ottenuti con altri strumenti **T-LAB**.

Scegliendo l'opzione "Salva", la variabile < **CONT_CLUST** > (cluster di contesti elementari) resta disponibile solo in alcuni tipi di analisi (es. Sequenze di Temi, Associazioni di Parole, Confronti tra Coppie e Co-Word Analysis) e fino a quando l'utilizzatore modifica la lista delle parole chiave.

11 - Esportare un dizionario delle categorie

Quando viene selezionata questa opzione, **T-LAB** crea due file:
a) un file dizionario con estensione **.dictio** pronto per essere importato tramite uno degli strumenti per l'analisi tematica. In tale dizionario ciascun cluster corrisponde a una categoria descritta tramite le sue parole caratteristiche, cioè da tutte le parole con un significativo valore del chi-quadro al suo interno;
b) un file **MyList.diz** pronto per essere importato tramite la funzione ‘Impostazioni Personalizzate’. Poiché tale file contiene l’elenco alfabetico di tutte le parole con un

significativo valore del chi-quadro, cioè di tutte le parole che determinano la differenze tra cluster tematici, il suo uso può consentire di ripetere alcune analisi con una modalità 'più selettiva' e discriminante.

12 – Verificare la qualità della partizione scelta e la coerenza semantica dei vari temi

Quando viene cliccato il pulsante 'Indici di Qualità', **T-LAB** crea un file HTML in cui sono riportate varie misure.
Le prime di queste si riferiscono alla qualità della partizione in 'k' cluster, cioè – ad esempio – al rapporto tra varianza interna ed esterna.
Le seconde si riferiscono alla 'coerenza semantica' di ciascuno dei cluster, e più specificatamente alle similarità tra prime 10 parole caratteristiche di ogni tema.
Nel dettaglio:
- le prime 10 parole sono quelle con il più alto valore del chi-quadro;
- le misure di similarità sono calcolate usando il coefficiente del coseno;
- come nel caso dello strumento **Associazioni di Parole**, il coefficiente del coseno è calcolato verificando le co-occorrenze di ogni coppia di parole all'interno dei segmenti di testo definiti come contesti elementari.

13 – Esplorare Sequenze di Temi

A differenza dello strumento 'Sequenze di Temi' incluso in un sottomenu T-LAB per l'analisi delle co-occorrenze, questa opzione è stata specificamente progettata per integrare l'analisi tematica dei contesti elementari. Più specificamente: il suo uso ha senso solo quando l'intero corpus può essere considerato come un discorso e/o quando le sue varie sezioni (ad esempio: capitoli di un libro, parti di una intervista, interventi di vari partecipanti a una conversazione o una discussione, etc.) si susseguono con un preciso ordine temporale.

In questo caso le relazioni analizzate sono quelle tra contesti elementari (fino a un massimo di 100.000) lungo la catena lineare del corpus, e ciascuno di essi - vuoi come 'predecessore' o come 'successore' – è trattato come una unità di analisi appartenente ad un cluster tematico (o come non classificato).

Tutti gli output forniti permettono all'utente di esplorare le relazioni sequenziali tra 'temi', sia in modo 'statico' che 'dinamico'. In particolare, tramite alcuni grafici animati che consentono di apprezzare la dinamica temporale delle sequenze, l'utente può verificare quando le persone sono impegnate su temi specifici (vedi, ad es., i punti sulla diagonale delle matrici nelle immagini seguenti) e quando passano da un tema dominante a un altro.

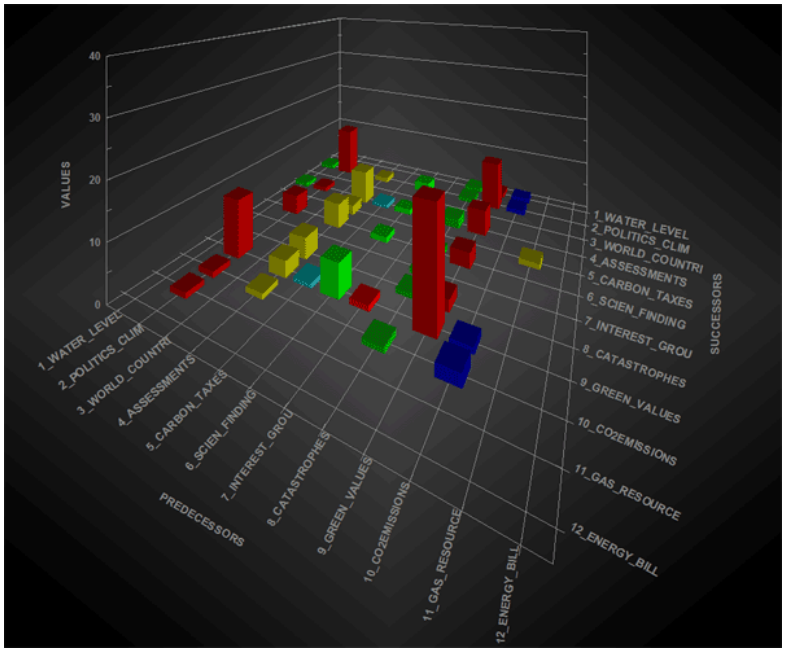
Passo dopo passo, di seguito viene fornita una breve descrizione delle varie opzioni disponibili.

(N.B.: Tutti gli output dell'esempio sono stati ottenuti tramite un'analisi tematica del libro 'The Politics of Climate Change " di Antony Giddens).

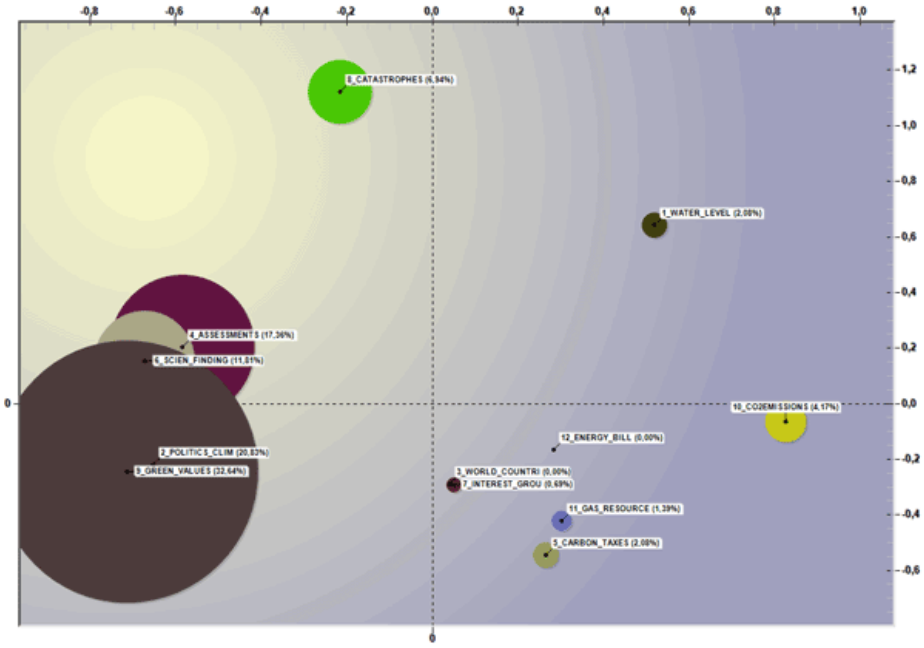
Quando è abilitato il pulsante 'Sequenze di Temi', cliccando su di esso diventa visibile ed attivo il seguente 'player'.

L'opzione '1' (vedi sopra) si riferisce al tipo di grafico scelto per la visualizzazione delle sequenze, sia all'interno dell'intero corpus che all'interno una parte di esso (vedi sopra opzione '2').

L'opzione 'matrice' rende disponibile un grafico 3D che riassume le relazioni tra predecessori e successori tramite barre colorate posizionate ai rispettivi incroci. In questo caso, quando sono visualizzati grafici 3D animati, l'incremento in altezza delle varie barre indica l'aumento delle occorrenze delle rispettive sequenze (vedi relazioni binarie tra 'predecessori' e 'successori' nel grafico seguente).



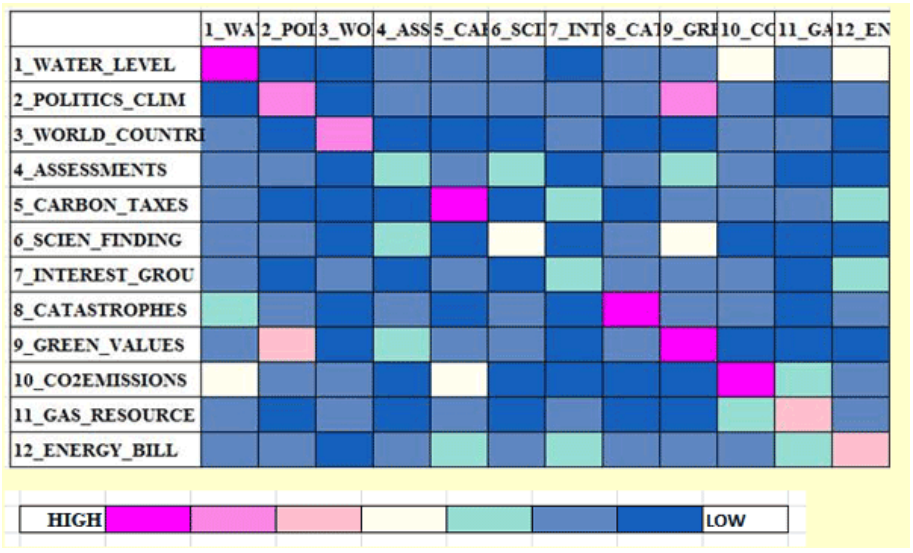
L'opzione 'spazio' rende disponibile un grafico 2d in cui le dimensioni (cioè percentuali) e le relazioni tra gruppi tematici sono rappresentate su un piano organizzato da due assi fattoriali selezionati dall'utilizzatore. In questo caso, quando sono visualizzati grafici animati, le dimensioni delle 'bolle' - che vengono continuamente riadattate a un totale pari al 100 % - indicano come la percentuali degli elementi appartenenti a ogni cluster tematico variano nel tempo e, contemporaneamente, il movimento delle frecce indica la direzione in cui i temi si susseguono.



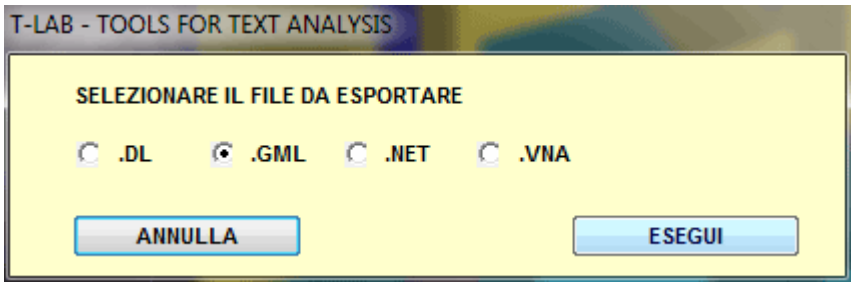
In entrambi i casi appena descritti, dopo l'arresto del video (vedi pulsante 'pausa'), è possibile visualizzare due ulteriori output:

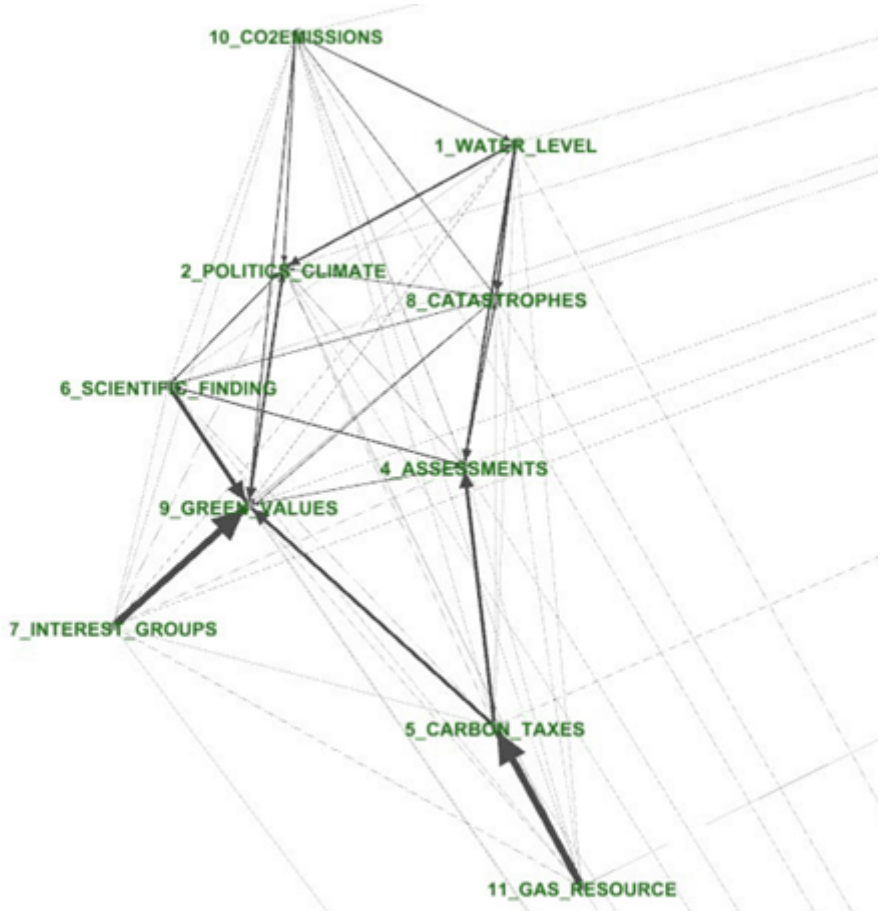
A - tabelle html che riassumono i rapporti tra predecessori e successori (vedi sotto);

	1_WA	2_POI	3_WO	4_ASS	5_CAI	6_SCI	7_INT	8_CA	9_GRI	10_CC	11_GA	12_EN	TOT
1_WATER_LEVEL	41	4	4	8	5	6	3	9	6	15	8	18	127
2_POLITICS_CLIM	4	24	4	9	5	8	5	6	26	5	1	5	102
3_WORLD_COUNTR	5	3	24	2	3	2	6	2	1	6	6	4	64
4_ASSESSMENTS	7	8	3	12	5	13	3	9	10	5	3	3	81
5_CARBON_TAXES	9	3	4	4	31	1	11	3	9	8	8	11	102
6_SCIEN_FINDING	5	9	2	11	1	17	1	9	16	2	0	2	75
7_INTEREST_GROU	8	2	6	1	6	0	10	5	6	5	3	10	62
8_CATASTROPHES	12	9	4	5	3	7	4	30	5	8	2	5	94
9_GREEN_VALUES	6	22	2	12	8	9	3	8	41	3	4	3	121
10_CO2EMISSIONS	18	7	6	2	15	1	2	3	3	48	13	9	127
11_GAS_RESOURCE	7	4	9	4	9	2	5	2	2	12	22	5	83
12_ENERGY_BILL	8	6	2	6	10	6	10	5	5	8	10	21	97



B - file grafici che possono essere importati da software per l'analisi di rete.



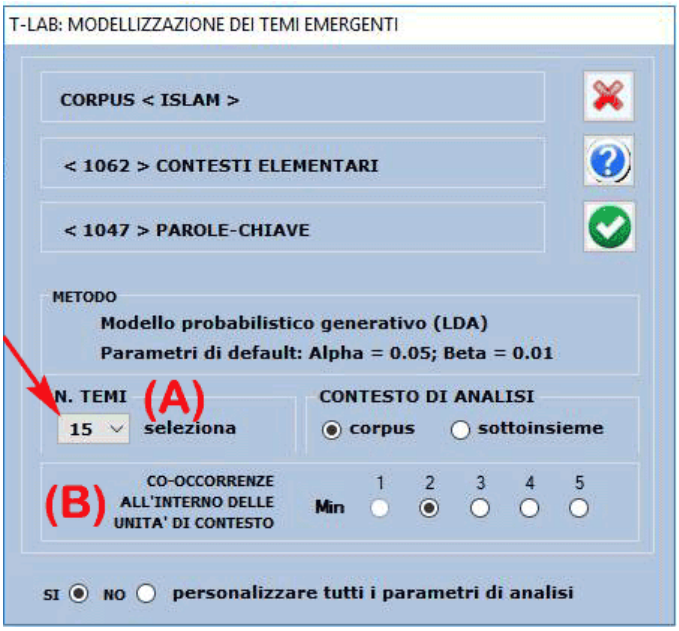


N.B.: Il grafico precedente, che si riferisce al terzo capitolo del libro di Giddens, è stato creato per mezzo del software Gephi (vedi https://gephi.org/).

Modellizzazione dei Temi Emergenti

Questo strumento **T-LAB** consente di **individuare, esaminare e modellare i principali temi che emergono dai testi** per poi – eventualmente - utilizzarli in ulteriori analisi, sia esse di tipo qualitativo (ad. es. per costruire griglie per l'analisi di contenuto) o di tipo quantitativo.

I temi emergenti, che sono descritti tramite il loro vocabolario caratteristico, cioè tramite insiemi di parole chiave (lemmi o categorie) co-occorrenti all'interno delle unità di contesto esaminate, possono essere infatti utilizzati per **classificare** quest'ultime (sia esse documenti o contesti elementari) e **ottenere nuove varia**bili da utilizzare in ulteriori analisi **T-LAB**.



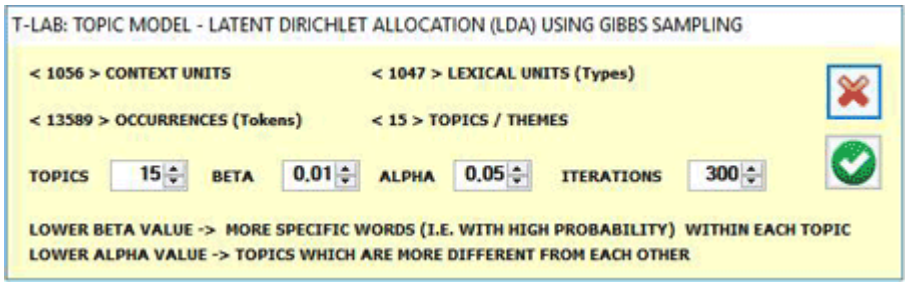
Una finestra di dialogo T-LAB (vedi sopra) consente di impostare due parametri di analisi.

In particolare:

- il parametro (A) consente di definire il numero di temi da ottenere. (Si noti che quanto maggiore è questo numero tanto più consistenti saranno le relazioni di co-occorrenza all'interno di ciascun tema; inoltre, se necessario, alcuni temi, ad esempio quelli ridondanti o difficili da interpretare, potranno essere eliminati successivamente);

- il parametro (B) consente di escludere dall'analisi qualsiasi unità di contesto che non contenga un numero minimo di parole chiave incluse nella lista utilizzata.

Solo quando si sceglie di personalizzare tutti i parametri di analisi (vedi sopra l'opzione 'Sì'), verrà visualizzata la finestra seguente e saranno disponibili ulteriori opzioni. (Si noti che nell'immagine seguente il numero di unità di contesto è determinato dal parametro "B" menzionato in precedenza).



La **procedura automatica di analisi** effettua i seguenti passi:

a – costruzione di una matrice documenti per parole, dove i documenti sono sempre contesti elementari corrispondenti alle unità di contesto (cioè frammenti, frasi, paragrafi) in cui il corpus è stato suddiviso;

b – analisi dei dati tramite un modello probabilistico che usa la Latent Dirichlet Allocation e il Gibbs Sampling (per ulteriori informazioni si vedano le corrispondenti voci di Wikipedia: http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation; http://en.wikipedia.org/wiki/Gibbs_sampling);

c – descrizione di ogni tema mediante i valori di probabilità associati alle sue parole caratteristiche, sia esse "specifiche" o "condivise" da due o più temi.

Al termine del processo di analisi, l'utilizzatore può agevolmente effettuare le seguenti operazioni:

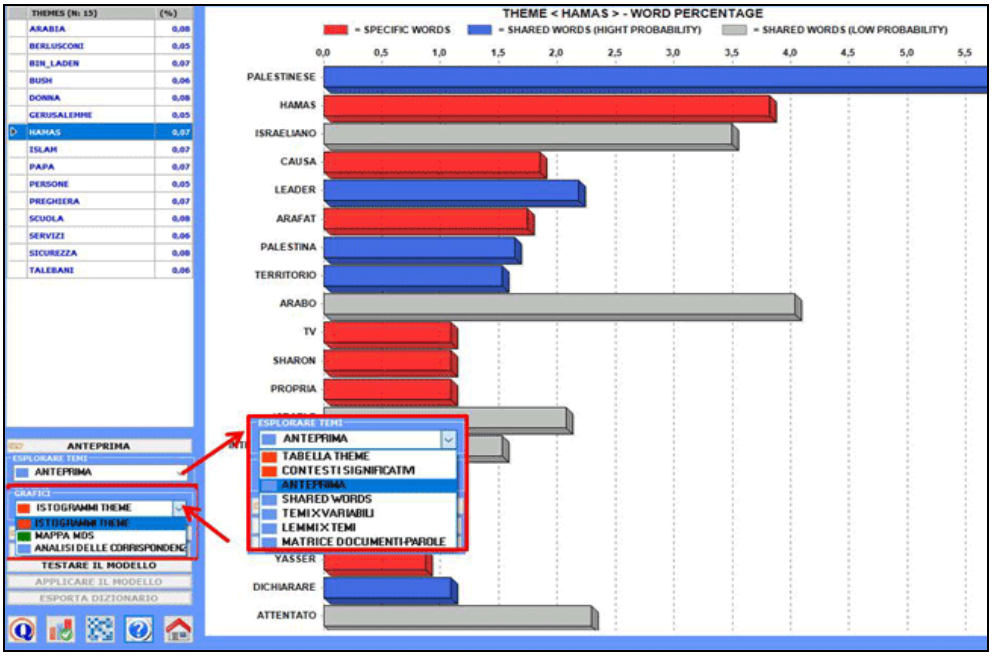
1 – esplorare le caratteristiche di ogni singolo tema;
2 – esplorare le relazioni tra i vari temi;
3 – rinominare o eliminare specifici temi;
4 – verificare la coerenza semantica dei vari temi;
5 – testare il modello ed assegnare i temi alle unità di contesto, sia esse documenti e/o contesti elementari;
6 – applicare il modello e creare una nuova variabile tematica da utilizzare con altri strumenti **T-LAB**;
7 – esportare un dizionario delle categorie che potrà essere utilizzato in ulteriori analisi.

Nel dettaglio:

1 – Esplorare le caratteristiche di ogni singolo tema

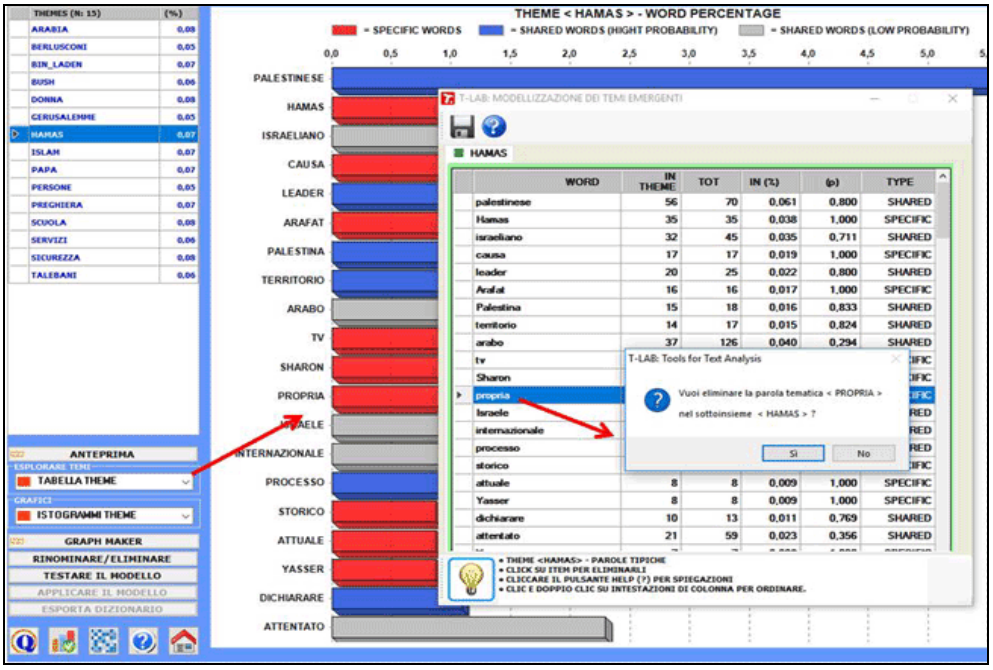
Il primo output che può essere consultato e salvato è costituito da una tabella con una sintesi di tutti i temi. E, quando lo si desideri, la stessa tabella può essere visualizzata usando il pulsante 'Anteprima' (vedi sotto).

Altri tipi di output sono accessibili selezionando una delle opzioni evidenziate nell'immagine seguente.



N.B.: In questo tipo di grafico (vedi sopra) “hight probability” indica una probabilità >=.75.

Quando viene selezionato un tema, facendo clic sull'opzione "Tabella Theme", è possibile verificare le sue caratteristiche; inoltre - facendo clic su qualsiasi parola nella tabella mostrata - diventa disponibile l’opzione per "eliminare" specifiche parole dal tema (vedi immagine seguente).



Le chiavi di lettura di questo tipo di tabella sono le seguenti:

IN THEME = occorrenze (tokens) di ogni parola all’interno del tema selezionato;

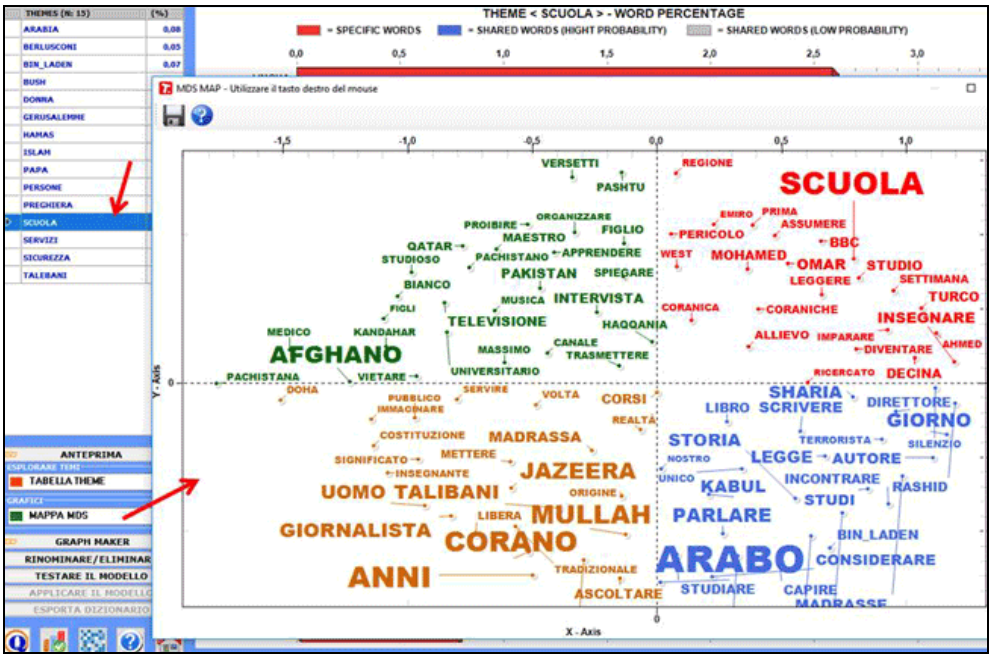
TOT = occorrenze (tokens) di ogni parola all’interno del corpus o del sottoinsieme analizzato;

IN (%) = peso percentuale di ogni parola all’interno del tema selezionato;

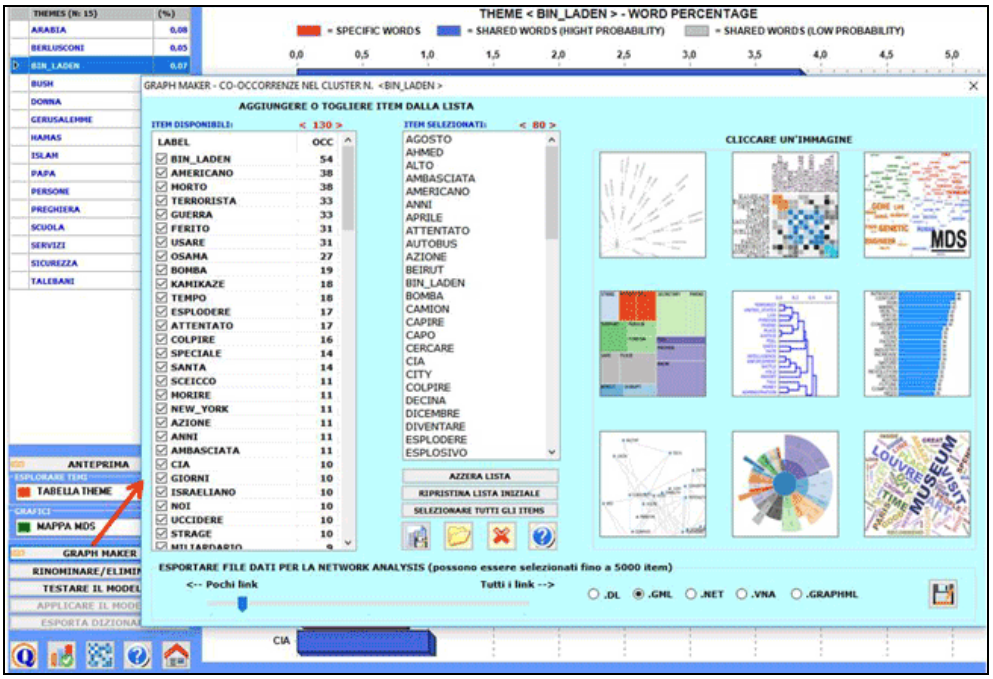
(p) = valore di probabilità associato a ogni relazione parola x tema;

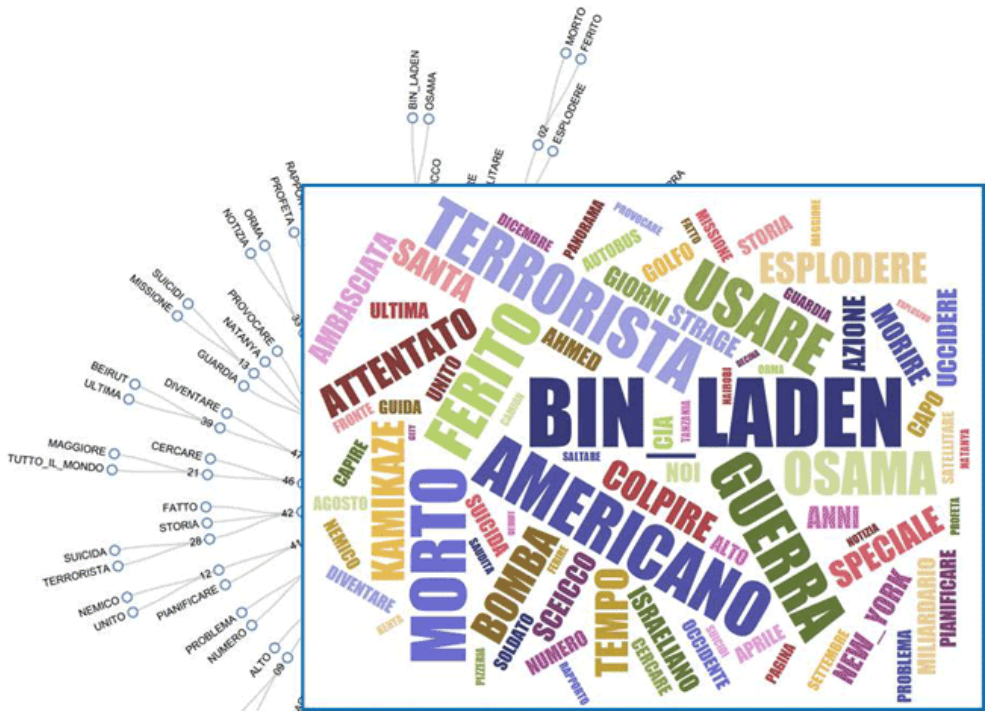
TYPE = contrassegnato con “specific” quando la parola (con p = 1) appartiene solo al tema selezionato, e come “shared” negli altri casi (cioè quando la parola, in diverso modo, è presente in più di un tema).

Quando viene selezionato un tema, facendo clic sull'opzione "Mappa MDS" si possono facilmente esplorare le relazioni semantiche tra le parole che risultano più caratteristiche (vedere l'immagine seguente).

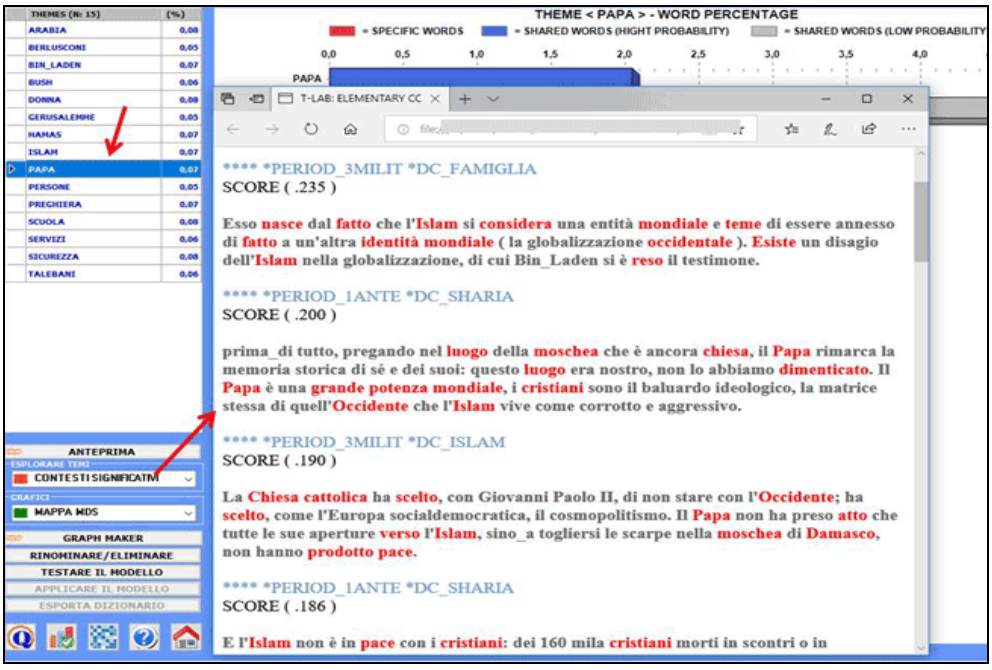


Inoltre, utilizzando lo strumento 'Graph Maker', diventano disponibili ulteriori opzioni grafiche (vedi le immagini seguenti).





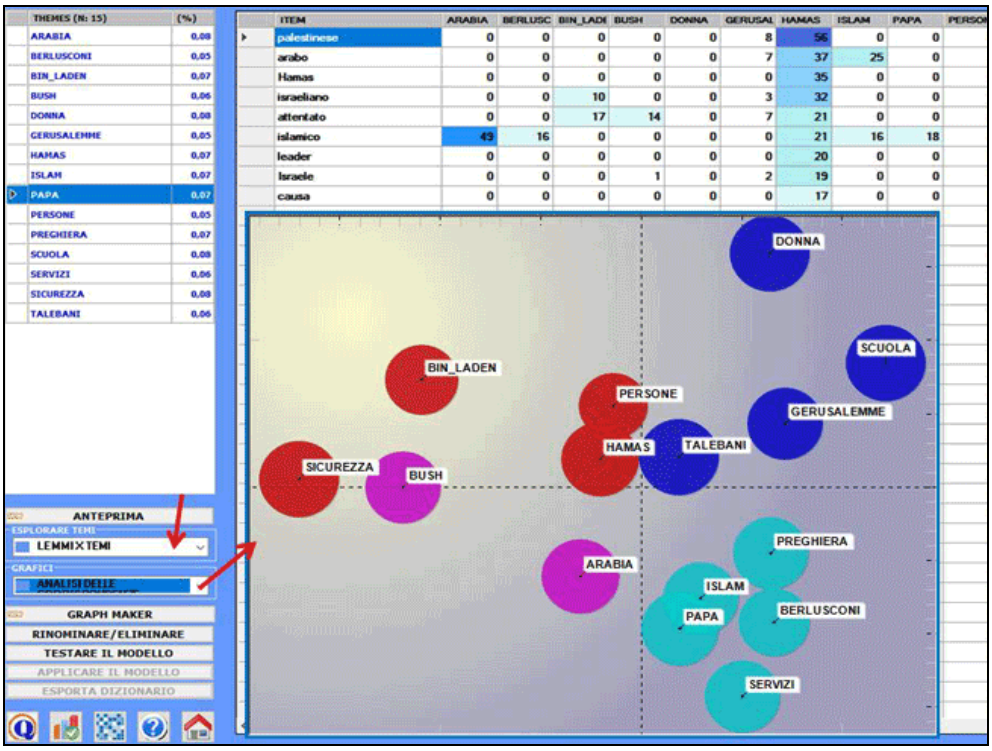
Quando viene selezionato un tema, facendo clic sull'opzione 'contesti significativi', viene creato un file HTML in cui vengono visualizzati i primi 20 segmenti di testo, che corrispondono maggiormente alle caratteristiche del tema in questione (vedere l'immagine seguente).



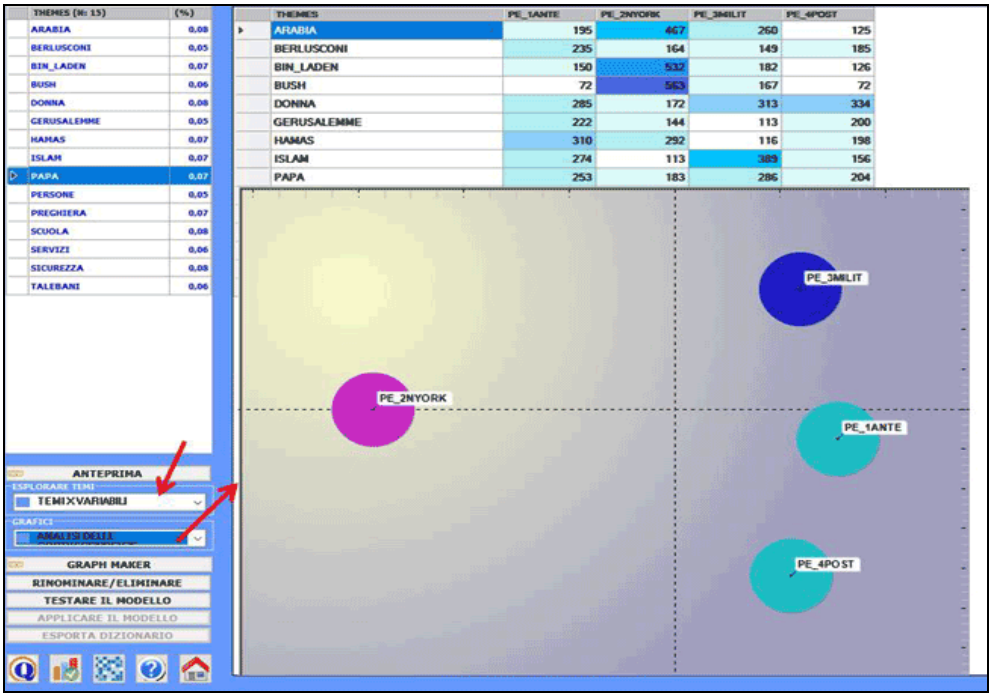
2 – Esplorare le relazioni tra i vari temi

Tramite lo strumento Analisi delle Corrispondenze, è possibile creare ed esplorare due tipi di tabelle di contingenza:

2.1) una tabella parole per temi (vedi sotto)

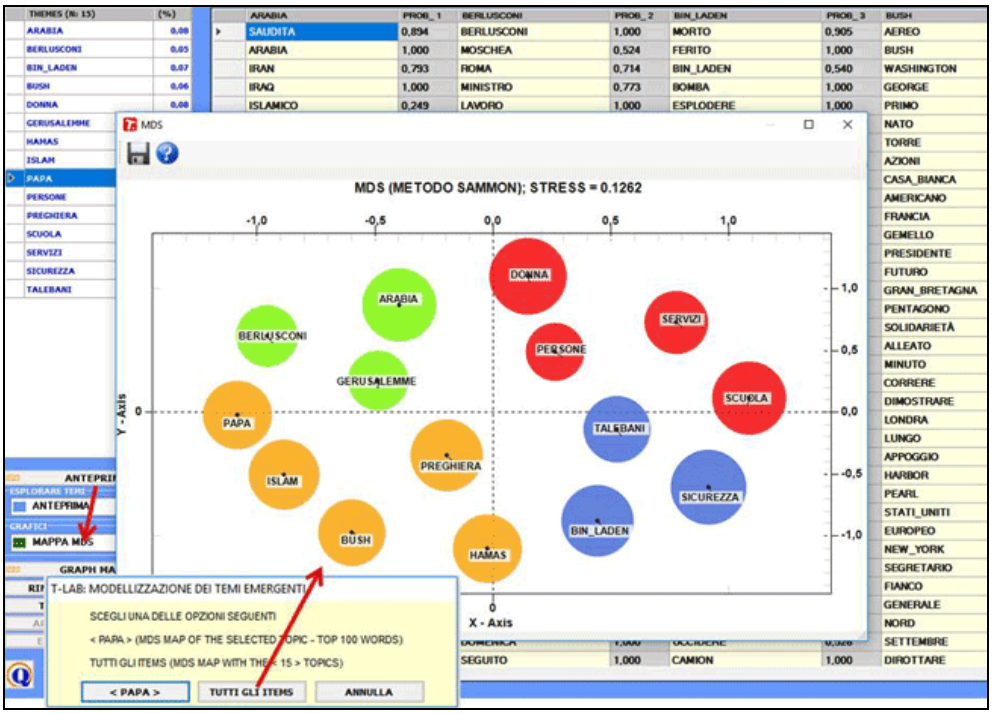


2.2) una tabella che incrocia i temi con le modalità variabile selezionata

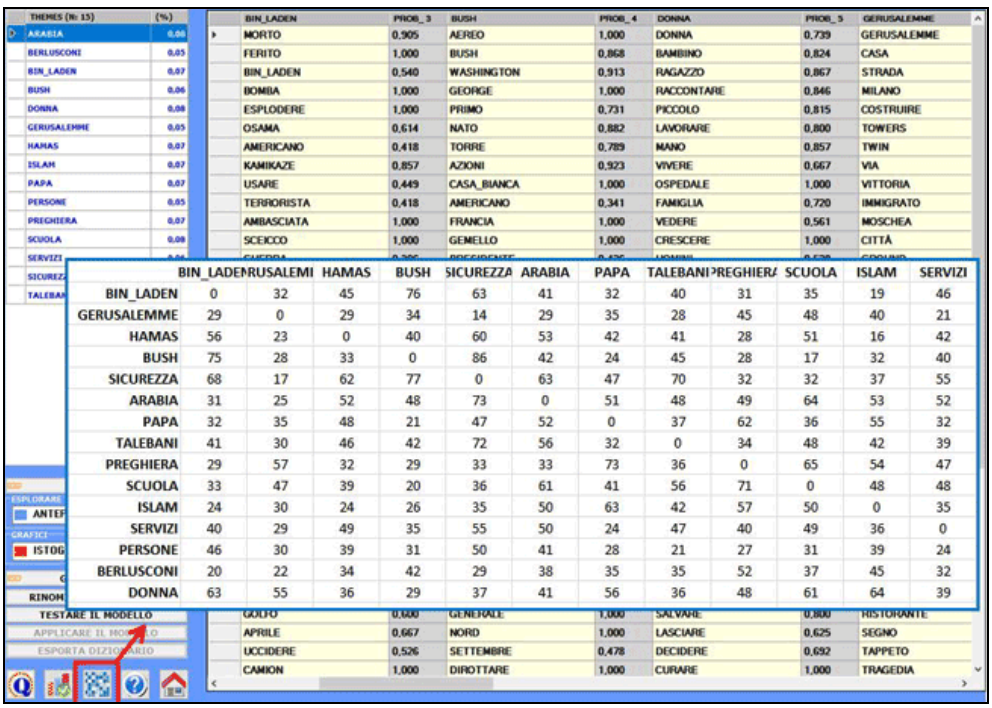


Sono anche disponibili altre due opzioni grafiche che consentono di mappare le relazioni tra i vari temi / topic.

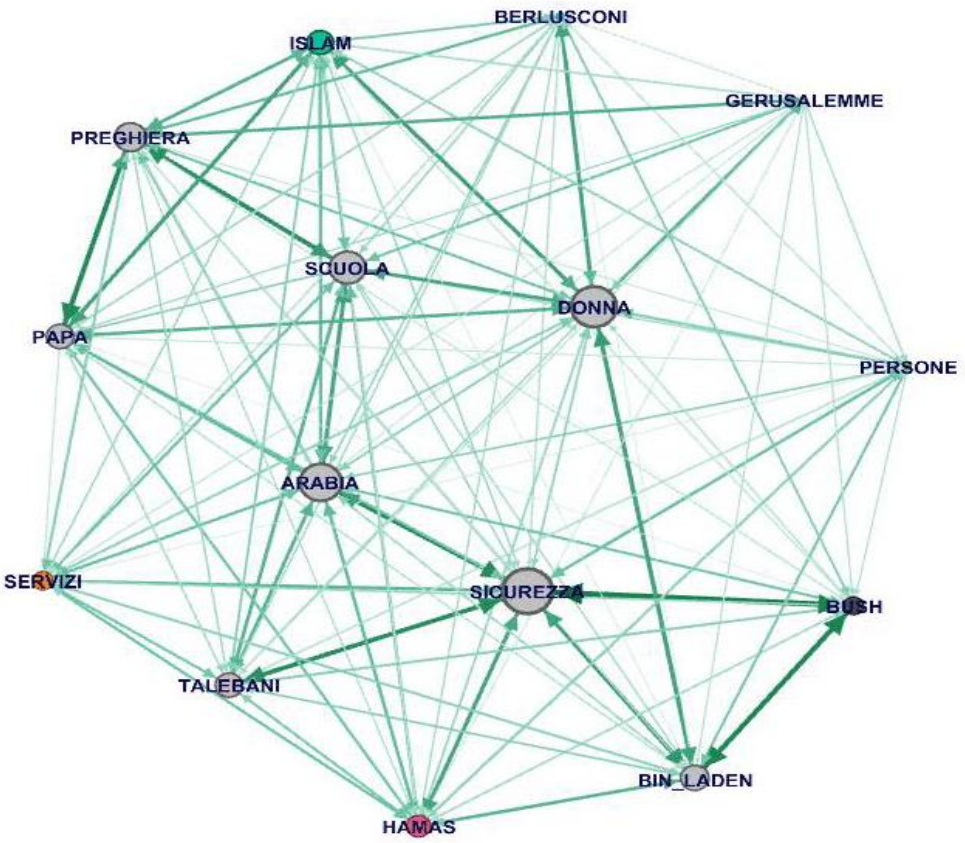
2.3) una Mappa MDS



2.4) grafici di rete ottenuti esportando / importando la tabella di adiacenza creata da **T-LAB** (vedi sotto)



	BIN_LADEN	GERUSALEMME	HAMAS	BUSH	SICUREZZA	ARABIA	PAPA	TALEBANI	PREGHIERA	SCUOLA	ISLAM	SERVIZI
BIN_LADEN	0	32	45	76	63	41	32	40	31	35	19	46
GERUSALEMME	29	0	29	34	14	29	35	28	45	48	40	21
HAMAS	56	23	0	40	60	53	42	41	28	51	16	42
BUSH	75	28	33	0	86	42	24	45	28	17	32	40
SICUREZZA	68	17	62	77	0	63	47	70	32	32	37	55
ARABIA	31	25	52	48	73	0	51	48	49	64	53	52
PAPA	32	35	48	21	47	52	0	37	62	36	55	32
TALEBANI	41	30	46	42	72	56	32	0	34	48	42	39
PREGHIERA	29	57	32	29	33	33	73	36	0	65	54	47
SCUOLA	33	47	39	20	36	61	41	56	71	0	48	48
ISLAM	24	30	24	26	35	50	63	42	57	50	0	35
SERVIZI	40	29	49	35	55	50	24	47	40	49	36	0
PERSONE	46	30	39	31	50	41	28	21	27	31	39	24
BERLUSCONI	20	22	34	42	29	38	35	35	52	37	45	32
DONNA	63	55	36	29	37	41	56	36	48	61	64	39

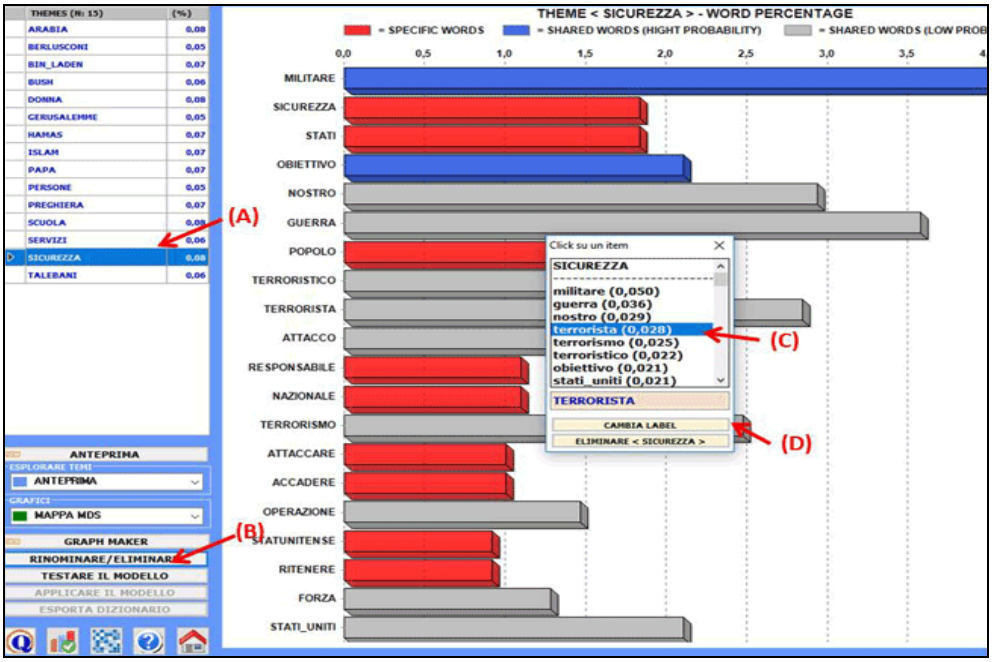


N.B.: Questo grafico è stato creato utilizzando il software open-source Gephi (https://gephi.org/) per importare una tabella esportata tramite **T-LAB**.

3 – **Rinominare o eliminare specifici temi**

Per rinominare o eliminare specifici temi è sufficiente selezionare gli item corrispondenti (vedi sotto "A") e cliccare sul pulsante "rinominare/eliminare" (vedi sotto "B").

Quando compare il box con le varie opzioni (vedi sotto), a seconda dei propri obiettivi, l'utilizzatore può cambiare la label del tema (sia scegliendo tra le parole disponibili che digitandone una nuova; vedi sotto "C") oppure eliminare il tema selezionato con un click sull'apposito pulsante (vedi sotto "D").



4 – Verificare la coerenza semantica dei vari temi

Quando viene cliccato il pulsante 'Indici di Qualità', **T-LAB** calcola le similarità tra le prime dieci (top 10) parole caratteristiche di ogni tema.

Più specificatamente:
- le prime 10 parole sono quelle con il più alto valore di probabilità;
- le misure di similarità sono calcolate usando il coefficiente del coseno;
- come nel caso dello strumento **Associazioni di Parole,** il coefficiente del coseno è calcolato verificando le co-occorrenze di ogni coppia di parole all'interno dei segmenti di testo definiti come contesti elementari.

Come risultato, **T-LAB** crea un file HTML in cui i 'k' temi solo elencati con il rispettivo indice di 'coerenza semantica'.

N.B.: Poiché le misure di similarità variano con il variare delle parole selezionate, si raccomanda di ripetere la procedura ogni volta che qualcuna delle prime dieci parole di un qualche tema venga eliminata dall'utilizzatore.

5 – **Testare il modello**

Al termine dell'analisi dei dati (vedi sopra i punti "a" e "b" relativi alla procedura di analisi) ogni unità di contesto (es. un documento o un contesto elementare) risulta costituito da un una

"mistura" di temi (o topics). Diversamente, il processo di classificazione utilizzato in questa fase consente di associare ogni unità di contesto al tema che più lo caratterizza. Ne risulta che, a questo punto, ogni tema diventa di fatto un cluster di unità di contesto.

Per questa ragione, quando viene selezionata l'opzione "Testare il modello" **T-LAB** produce due file XLS (vedi sotto) che consentono all'utilizzatore di verificare l'appartenenza di ogni unità di contesto a uno specifico tema.

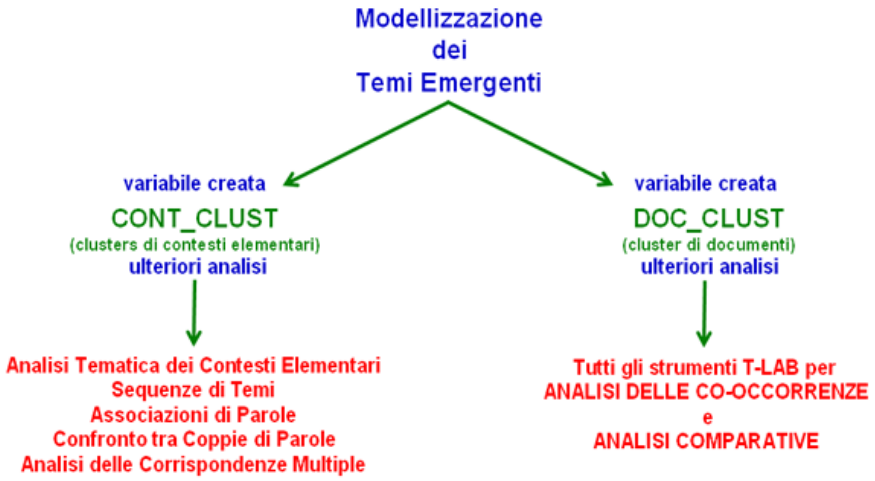
ID_DOC	BEST	ARABIA	RLUSCONI	BIN_LADEN	BUSH	DONNA	RUSALEMME	HAMAS	ISLAM	PAPA	PERSONE	PREGHIERA	SCUOLA	SERVIZI	ICUREZZA	TALEBANI
1	15	0,071	0,061	0,129	0,018	0,010	0,000	0,029	0,057	0,048	0,041	0,056	0,000	0,083	0,079	0,292
2	7	0,321	0,020	0,217	0,000	0,024	0,099	0,937	0,064	0,100	0,429	0,469	0,080	0,215	0,241	0,029
3	12	0,919	0,274	0,546	0,140	0,299	0,249	0,178	0,569	0,275	0,340	0,561	3,101	0,668	0,135	0,977
4	8	0,055	0,206	0,000	0,041	0,017	0,000	0,108	0,359	0,019	0,017	0,261	0,102	0,038	0,054	0,140
5	11	0,270	1,525	0,051	0,081	0,540	1,318	0,056	0,489	0,181	0,192	1,961	0,356	0,351	0,039	0,186
6	5	0,077	0,311	0,000	0,053	0,825	0,186	0,016	0,674	0,235	0,253	0,345	0,093	0,030	0,006	0,000
7	9	0,076	0,216	0,000	0,000	0,023	0,308	0,064	0,368	0,835	0,041	0,107	0,015	0,033	0,013	0,000
8	2	0,000	0,119	0,000	0,000	0,075	0,009	0,000	0,000	0,000	0,011	0,070	0,000	0,052	0,000	0,075
9	7	0,145	0,162	0,028	0,007	0,041	0,049	1,623	0,061	0,024	0,194	0,153	0,046	0,079	0,054	0,132
10	9	0,241	0,026	0,000	0,000	0,094	0,087	0,007	0,352	0,693	0,118	0,314	0,051	0,173	0,065	0,059
11	7	0,186	0,285	0,983	0,000	1,352	0,384	1,956	0,197	0,336	0,324	0,275	0,067	0,375	0,431	0,276
12	8	0,009	0,015	0,012	0,000	0,000	0,000	0,005	0,791	0,386	0,012	0,054	0,038	0,074	0,020	0,002
13	11	0,000	0,280	0,008	0,058	0,018	0,007	0,007	0,058	0,006	0,049	0,521	0,000	0,103	0,032	0,129
14	4	0,798	0,261	0,283	1,050	0,007	0,027	0,202	0,099	0,066	0,119	0,038	0,011	0,609	0,625	0,033
15	3	0,675	0,082	4,001	0,576	0,091	0,037	0,397	0,072	0,081	0,250	0,368	0,140	0,506	0,657	1,158
16	14	0,448	0,074	0,934	1,108	0,240	0,054	0,210	0,102	0,298	0,284	0,079	0,122	0,320	1,171	0,369
17	14	0,409	0,175	0,459	0,789	0,112	0,106	0,460	0,127	0,238	0,824	0,028	0,067	0,244	1,007	0,234
18	14	0,175	0,082	0,602	0,803	0,024	0,125	0,197	0,015	0,036	0,133	0,228	0,000	0,222	1,072	0,208
19	14	0,380	0,269	0,323	0,624	0,007	0,060	0,354	0,031	0,143	0,116	0,000	0,000	0,156	0,955	0,173
20	4	0,461	0,067	0,199	0,525	0,013	0,010	0,016	0,264	0,070	0,003	0,004	0,000	0,000	0,041	0,133
21	4	0,603	0,284	0,857	1,636	0,008	0,010	0,068	0,152	0,098	0,151	0,170	0,146	0,484	1,150	0,444
22	4	0,216	0,031	0,132	0,454	0,031	0,053	0,088	0,084	0,151	0,024	0,199	0,011	0,278	0,357	0,142
23	7	0,414	0,087	0,079	0,166	0,040	0,047	0,760	0,104	0,486	0,315	0,090	0,006	0,148	0,303	0,331

N.B.: Nella tabella sopra riportata, ogni documento ha un valore di probabilità associato ad ogni tema.

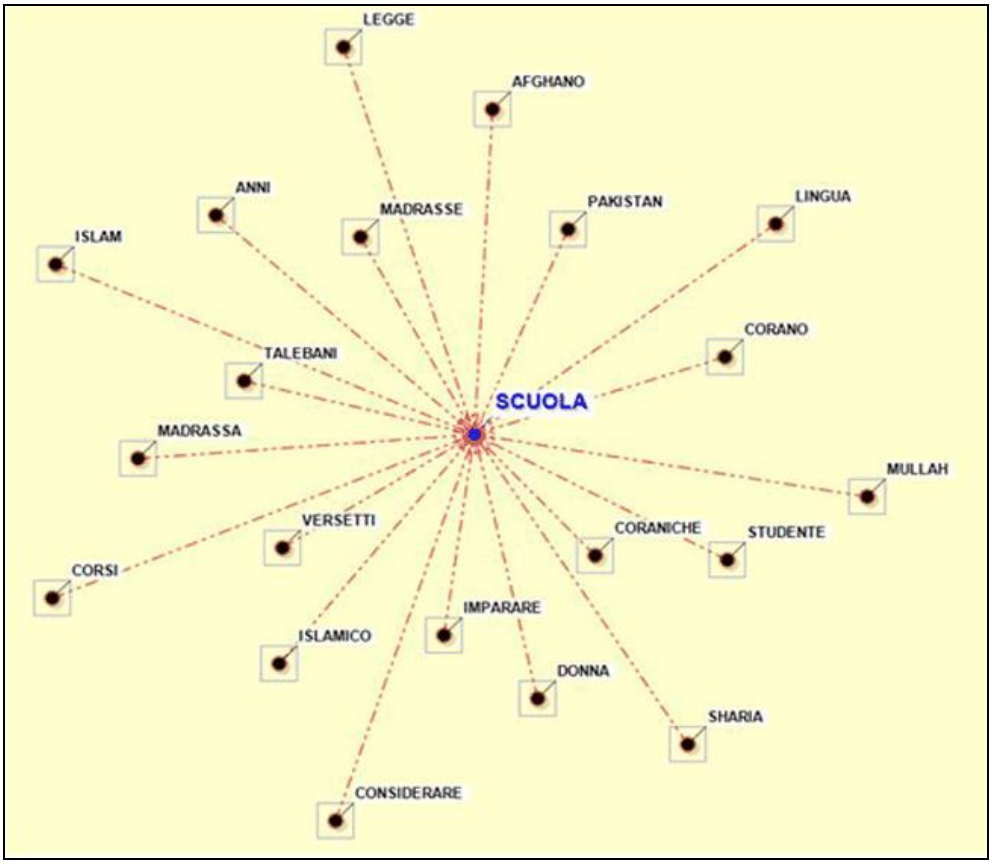
IdDoc	IdSeg	Topic	Score	Segm
15	334	BIN_LADEN	0,363	7 agosto 1998 a_distanza_di soli dieci minuti due esplosioni sventrano le ambasciate americane di Nairobi , in Kenya e Dar-es-Salaam , capi
9	186	HAMAS	0,326	E se Ahmad Tibi , un arabo israeliano eletto deputato alla Knesset e insieme consigliere di Arafat , diceva alla tv israeliana che gli arabi dev
45	1006	BIN_LADEN	0,311	il 9 agosto 2001 un terrorista suicida si fa saltare per_aria nella pizzeria Sbarro , 15 morti , donne , bambini , una famiglia intera di cinque per
25	572	ARABIA	0,290	500 DOLLARI ARABIA SAUDITA Re Fahd É uno dei paesi più caldi , per la forte presenza di organizzazioni fondamentaliste , inferocite per la
15	318	BIN_LADEN	0,279	TUTTI GLI ASSALTI CONTRO GLI USA Cronologia di 25 anni di attentati , con migliaia di morti e feriti Negli ultimi 25 anni gli stati_uniti sono_
26	585	BIN_LADEN	0,258	Mustafà Mahamud , saudita , secco come un ' aringa : ' ' Bin_Laden ? Che mi frega ? Terrorismo è solo americano , russo o israeliano '
39	874	ARABIA	0,256	IL BARATTO TRA FEDE E PETROLIO Perché a Roma sorge una moschea mentre in Arabia Saudita i cristiani sono perseguitati ? Leggo che durar
17	386	BIN_LADEN	0,252	Bin_Laden ha già fatto colpire il World_Trade_Center nel 1993 . Ha fatto saltare le ambasciate americane in Kenya e in Tanzania . Ha silurat
47	1060	BERLUSCONI	0,251	Ma se l ' attività di un governo è una corsa a tappe , non c ' è dubbio che nell ' ultima settimana un paio di mosse siano andate molto_ben
25	563	SICUREZZA	0,250	Si sente nel mirino del Pentagono e degli americani ? Se gli stati_uniti vogliono lottare contro il terrorismo , io non rischio di essere un obi
19	432	HAMAS	0,248	L ' organizzazione di Bin_Laden attaccava bersagli statunitensi anche quando il processo di pace di Oslo stava dando buoni frutti e i leader p
15	331	BIN_LADEN	0,246	13 novembre 1995 Un ' autobomba distrugge l ' edificio della Guardia nazionale saudita di Riyadh dove lavorano consiglieri Usa : sette mor
20	437	BUSH	0,236	L ' alleato più fedele degli stati_uniti è in questo momento la Russia . L ' Europa si chiama fuori nei_fatti e offre una solidarietà verbale . I
3	38	SCUOLA	0,235	Il pashtu , tra l ' altro , si scrive in caratteri persiani e non arabi come il Corano . Comunque sia , gli allievi delle madrassa , salvo rare ecce
38	871	PAPA	0,235	Esso nasce dal fatto che l ' Islam si considera una entità mondiale e teme di essere annesso di fatto a un ' altra identità mondiale (la globa
32	717	SCUOLA	0,232	Sarebbe come se scoprissimo che i musulmani ci considerano tutti inglesi . L ' ARABO SI IMPARA COSÍ I corsi consentono di apprendere la li
17	373	BUSH	0,228	Bush , due aerei di linea sequestrati dai kamikaze islamici a Boston si sono scagliati contro le torri gemelle del World_Trade_Center di New
38	865	PREGHIERA	0,228	IL VERO , GRANDE PROBLEMA DELL ' ISLAM SPECIALE A un difficile bivio i seguaci di Allah Un islamico non può condannare un altro islamicc
27	615	DONNA	0,227	E in questa seconda battaglia di Algeri le donne sono_state sempre in prima fila . Donne letterate , studentesse , ma anche povere massai
44	989	DONNA	0,223	E così una divorziata perde i suoi figli anche_se a lasciarla è il marito . Una donna tradita deve sopportare di vivere con la sua rivale , e forse
15	332	BIN_LADEN	0,221	16 giugno 1996 A Dahran , in Arabia Saudita , un ' altra autobomba esplode tra i militari della base aeronautica Usa : 19 morti e 446 feriti (:
44	974	PREGHIERA	0,220	Ci sono musulmani equilibrati e seri , ma dovremo convivere anche con chi si porta dietro subculture e fantasmi di religioni che di divino n
45	997	RUSALEMME	0,219	Storie dra se questa è la città d ' oro culla del monoteismo , la magnifica capitale di re David , lo stupefacente tempio di re Salomone
15	323	BIN_LADEN	0,218	27 dicembre 1985 Nell ' aeroporto romano di Fiumicino un commando di terroristi islamici assalta con bombe a_mano e mitragliatrici i banc
3	28	PREGHIERA	0,216	la guerra santa contro gli infedeli o gli stessi musulmani che si sono distaccati dall ' insegnamento del Profeta . Ovviamente non è così in tu
25	554	SICUREZZA	0,216	L ' esercito israeliano uccide ogni giorno i civili palestinesi con i missili , gli F-16 , gli elicotteri americani Apache . Lo squilibrio delle forze
25	555	SICUREZZA	0,214	Nella pizzeria di Gerusalemme , il 9 agosto , e prima nella discoteca di Tel Aviv sono morti decine di ragazzi , passanti , bambini . Anche lo
43	949	HAMAS	0,213	Molto interessante , tuttavia , mi è parso il confronto tra Sharon e Hamas a cui Silvia Antonucci dedica la sua lettera . Fra il primo ministro i
27	612	DONNA	0,210	Masa , la piccola sorella di Marian , bambina di 16 anni morta ' ' suicida ' ' nel deserto di Baghdad racconta e piange . Marian aveva un p
37	832	SICUREZZA	0,210	Gli stati_uniti sono nemici di coloro che aiutano i terroristi e dei criminali barbari che profanano una grande religione commettendo crimini
12	256	ISLAM	0,209	E ancora : il musulmano sta innanzi al Dio coranico come sottomesso . L ' interiorizzazione islamica è l ' interiorizzazione della sottomissic
21	466	ARABIA	0,209	I russi , antitalebani , sono alla ricerca di un aiuto nella guerra contro la Cecenia che per loro è un conflitto contro il fondamentalismo islam
15	291	BIN_LADEN	0,208	Osama Bin_Laden , alias Osama Muhammad al-Wahad , alias Abu Abdallah , alias Al-Qaqa , 44 anni , 17° dei 52 figli di Muhammad Bin Awc

6 – **Applicare il modello**

Dopo aver applicato e salvato il modello, poiché i temi sono archiviati da **T-LAB** come modalità di due nuove variabili che si riferiscono a cluster di contesti elementari (**CONT_CLUST**) e/o a cluster di documenti (**DOC_CLUST**), le relazioni tra gli stessi temi e/o tra le loro caratteristiche possono essere ulteriormente esplorati con diversi strumenti di analisi (vedi sotto).



Ad esempio, utilizzando lo strumento **Associazioni di Parole** e selezionando il sottoinsieme (cioè il tema) "Religione" è possibile creare il grafico seguente.



7 – Esportare un dizionario

Quando viene selezionata questa opzione, **T-LAB** crea un file dizionario con estensione .dictio pronto per essere importato tramite uno degli strumenti per l'analisi tematica. In tale dizionario ciascun categoria è descritta tramite le sue parole caratteristiche.

Classificazione Tematica di Documenti

Questa funzione è abilitata solo quando il corpus in analisi comprende da un minimo di 20 a un massimo 99.999 documenti primari.

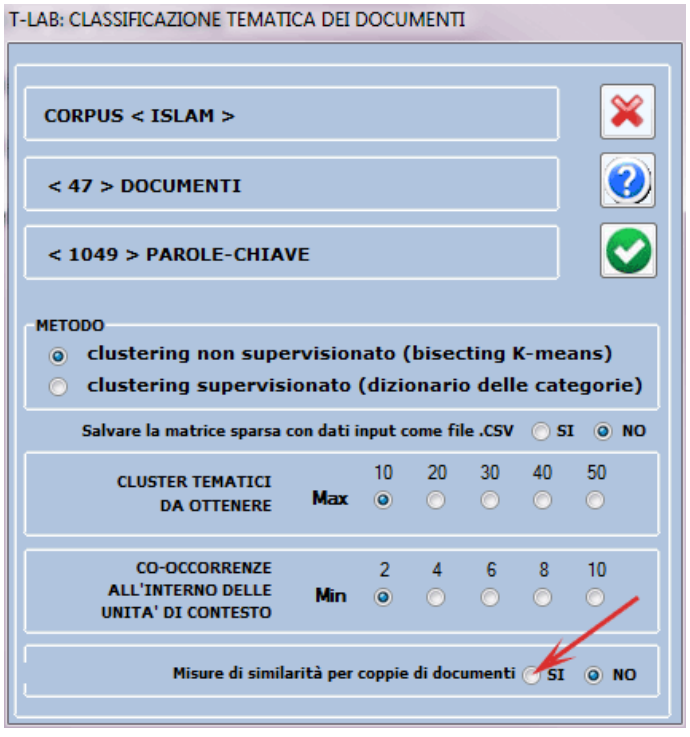
Il processo di analisi può essere effettuato tramite un metodo di clustering 'non supervisionato' (nel caso specifico, un algoritmo di bisecting K-Means) o tramite una classificazione supervisionata (vale a dire approccio top-down). Quando si sceglie il secondo (cioè classificazione supervisionata), viene richiesto di importare un dizionario delle categorie, sia esso creato tramite una precedente analisi **T-LAB** che costruito dall'utilizzatore.

Il suo uso consente di costruire cluster di documenti e di esplorare le loro caratteristiche attraverso operazioni/opzioni simili a quelle descritte nella sezione dedicata all'**Analisi Tematica dei Contesti Elementari.**

La sua specificità consiste nel fatto che la tabella analizzata è costituita da tante righe quanti sono i documenti del corpus, ciascuno dei quali è rappresentato come un vettore con valori che indicano le occorrenze delle parole in esso presenti.

Inoltre, quando i documenti analizzati non superano i 3000, è possibile ottenere misure di similarità (indice del coseno) tra ciascuno di essi e tutti gli altri (vedi sotto).

N.B.: In questo caso la soglia minima dell'indice di similarità è fissata a 0.05.



Gli output che differenziano questa funzione sono quindi i seguenti:

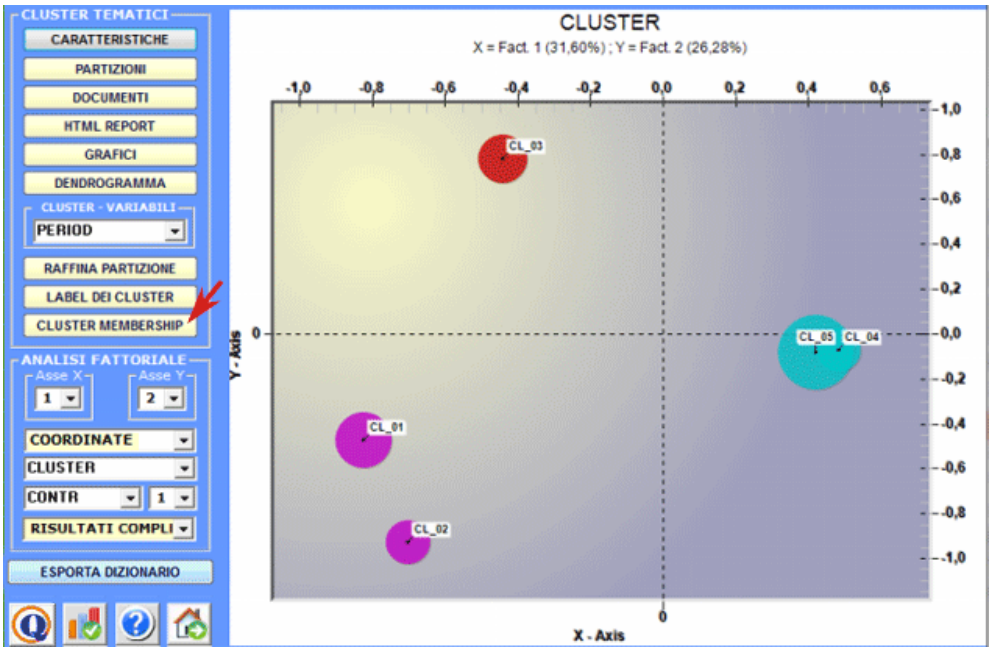
I documenti appartenenti ad ogni cluster sono ordinati secondo il valore decrescente del loro score (vedi sopra) e possono essere esplorati nel formato HTML.

In questo caso il valore di rilevanza (score) assegnato ad ogni documento (i) del cluster (k) è ottenuto applicando la seguente formula:

$$score_{i,k} = \cos(d_i, c_k)$$

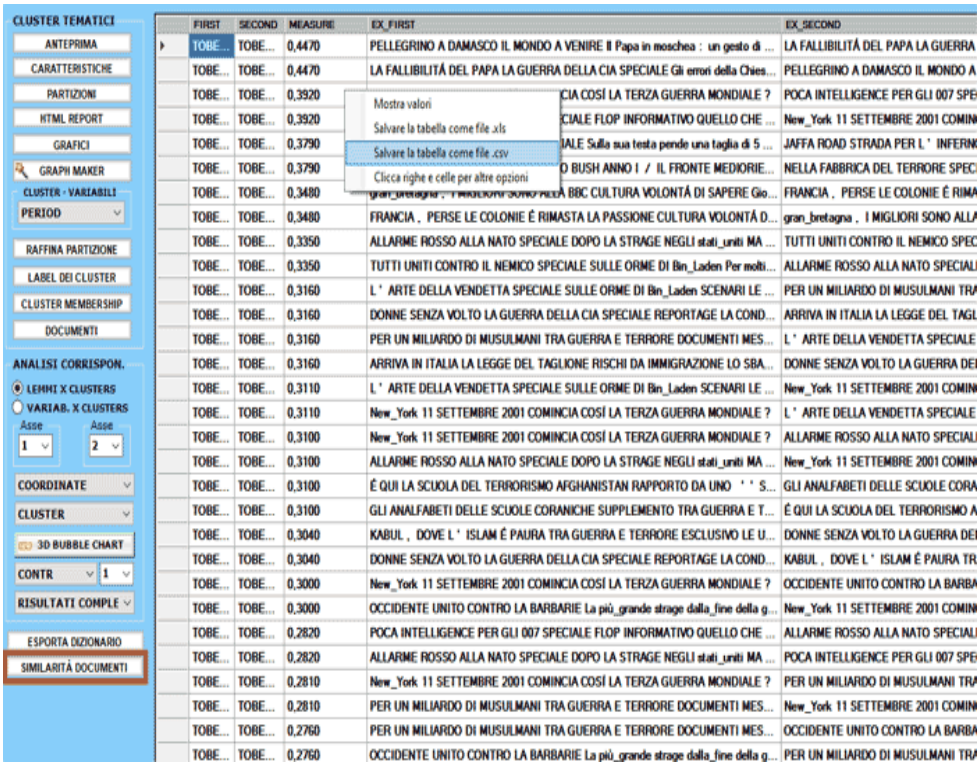
Dove:
i - si riferisce al documento ***i***;
k – si riferisce cluster ***k***;
cos - è il simbolo di coseno;
$\mathbf{d}_i$ - è il vettore normalizzato del $TF_{j,i}IDF_j$, dove **j** si riferisce a una parola del documento **i**
$\mathbf{c}_k$ - è il vettore normalizzato del $TF_{j,k}IDF_j$, dove **j** si riferisce una parola del cluster **k**

Usando gli score ottenuti dalla suddetta formula, T-LAB rende disponibile il file "Document_Membership_Degree.xls " (vedi sotto) che contiene i cluster a cui sono stati assegnati i vari documenti, sia mediante il metodo bisecting K-Means (appartenenza esclusiva di ogni documento a un cluster) che mediante il valore del TF-IDF (appartenenza "mista" – in formato percentuale - di ogni documento ai vari cluster).

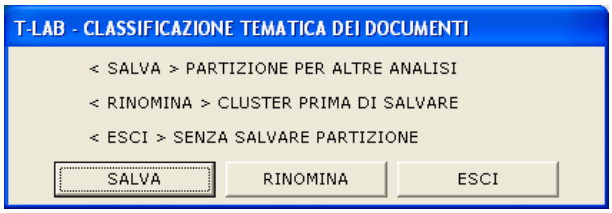


DOC_ID	VAR_01	CLUST_K-	BEST_TF-	MATCHIN(	CLUST-1	CLUST-2	CLUST-3	CLUST-4	CLUST-5
1	PE_1ANTE	1	1	1	0,465	0,023	0,154	0,167	0,19
2	PE_1ANTE	5	5	1	0,163	0,093	0,171	0,249	0,323
3	PE_1ANTE	5	5	1	0,115	0,128	0,219	0,115	0,423
4	PE_1ANTE	1	1	1	0,451	0,185	0,135	0,088	0,141
5	PE_1ANTE	1	1	1	0,539	0,11	0,186	0,071	0,095
6	PE_1ANTE	3	3	1	0,179	0,118	0,494	0,082	0,127
7	PE_1ANTE	3	3	1	0,151	0,066	0,651	0,044	0,086
8	PE_1ANTE	1	1	1	0,439	0,048	0,282	0,085	0,145
9	PE_1ANTE	5	5	1	0,179	0,096	0,144	0,179	0,402
10	PE_1ANTE	5	5	1	0,143	0,142	0,224	0,121	0,371
11	PE_1ANTE	5	5	1	0,132	0,048	0,171	0,172	0,477
12	PE_1ANTE	5	3	0	0,178	0,142	0,255	0,172	0,252
13	PE_1ANTE	1	1	1	0,67	0,054	0,123	0,073	0,08
14	PE_2NYORK	4	4	1	0,156	0,077	0,119	0,512	0,136
15	PE_2NYORK	5	5	1	0,093	0,033	0,125	0,226	0,523
16	PE_2NYORK	4	4	1	0,106	0,078	0,138	0,503	0,175
17	PE_2NYORK	4	4	1	0,101	0,044	0,19	0,45	0,215
18	PE_2NYORK	4	4	1	0,079	0,046	0,106	0,599	0,169
19	PE_2NYORK	4	4	1	0,137	0,053	0,107	0,523	0,18
20	PE_2NYORK	4	4	1	0,137	0,069	0,211	0,419	0,163
21	PE_2NYORK	4	4	1	0,108	0,055	0,129	0,542	0,166
22	PE_2NYORK	4	4	1	0,137	0,065	0,167	0,466	0,165
23	PE_2NYORK	5	5	1	0,148	0,069	0,196	0,273	0,315
24	PE_2NYORK	3	3	1	0,162	0,036	0,437	0,178	0,189
25	PE_2NYORK	5	5	1	0,126	0,06	0,136	0,269	0,409
26	PE_2NYORK	3	3	1	0,181	0,087	0,301	0,224	0,207
27	PE_3MILIT	3	3	1	0,153	0,077	0,473	0,104	0,192
28	PE_3MILIT	1	1	1	0,428	0,329	0,108	0,073	0,061
29	PE_3MILIT	2	2	1	0,066	0,772	0,043	0,06	0,059
30	PE_3MILIT	3	3	1	0,126	0,068	0,564	0,138	0,103
31	PE_3MILIT	5	5	1	0,123	0,104	0,149	0,168	0,456
32	PE_3MILIT	2	2	1	0,149	0,582	0,134	0,048	0,087
33	PE_3MILIT	3	3	1	0,198	0,064	0,508	0,141	0,089
34	PE_3MILIT	2	2	1	0,107	0,563	0,094	0,137	0,099
35	PE_3MILIT	3	3	1	0,105	0,112	0,535	0,082	0,166
36	PE_3MILIT	5	5	1	0,129	0,105	0,254	0,163	0,349

Quando il pulsante 'Documento Similarity' è abilitato, cliccando su di esso è possibile verificare in che misura ogni documento è simile a ciascuno degli altri. In questo caso la misura di similarità è il coefficiente del coseno e il suo valore varia in funzione di quante parole sono state utilizzate per la classificazione tematica. L'immagine seguente descrive le opzioni disponibili per questo tipo di verifica.



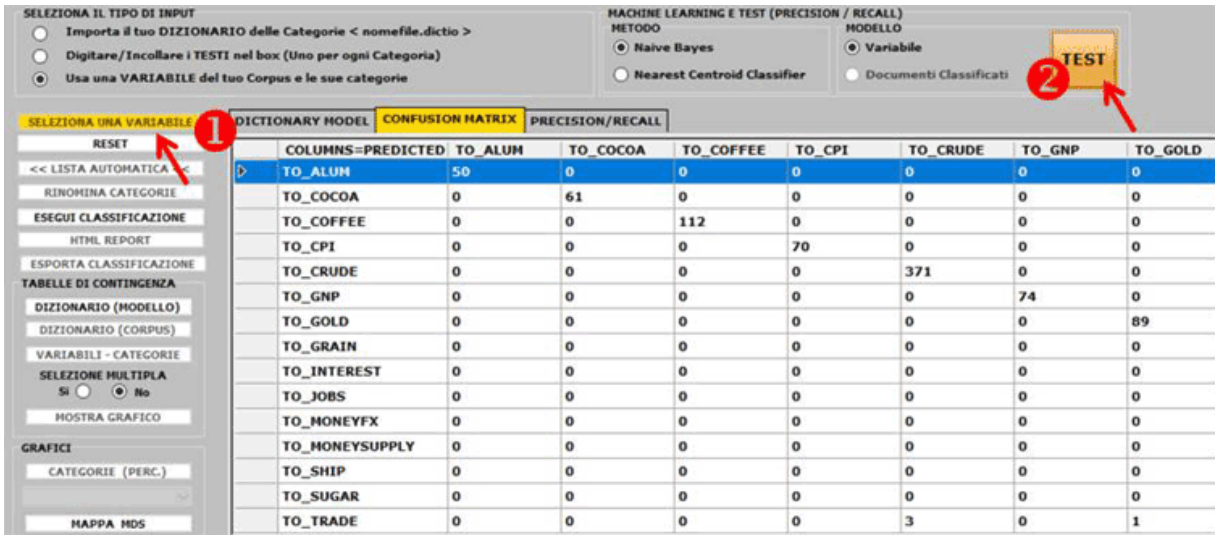
All'uscita di questa funzione, alcuni messaggi ricordano che è possibile esplorare i cluster ottenuti con altri strumenti **T-LAB**.



Scegliendo l'opzione "**SALVA**", la variabile < **DOC_CLUST**> (cluster di documenti) resta disponibile in tutte le successive analisi dello stesso corpus realizzate con altri strumenti **T-LAB.**

Classificazione basata su Dizionari

N.B.: Le immagini di questa sezione fanno riferimento a una versione precedente di T-LAB. In **T-LAB 10** l'aspetto è leggermente diverso. In particolare, a partire dalla versione 2021, una nuova funzionalità consente di testare facilmente qualsiasi modello su dati etichettati (es. dati che includono temi ottenuti da una precedente analisi qualitativa) e ottenere output come matrici di confusione e metriche di precision / recall (vedi immagine seguente).



COLUMNS=PREDICTED	TO_ALUM	TO_COCOA	TO_COFFEE	TO_CPI	TO_CRUDE	TO_GNP	TO_GOLD
TO_ALUM	50	0	0	0	0	0	0
TO_COCOA	0	61	0	0	0	0	0
TO_COFFEE	0	0	112	0	0	0	0
TO_CPI	0	0	0	70	0	0	0
TO_CRUDE	0	0	0	0	371	0	0
TO_GNP	0	0	0	0	0	74	0
TO_GOLD	0	0	0	0	0	0	89
TO_GRAIN	0	0	0	0	0	0	0
TO_INTEREST	0	0	0	0	0	0	0
TO_JOBS	0	0	0	0	0	0	0
TO_MONEYFX	0	0	0	0	0	0	0
TO_MONEYSUPPLY	0	0	0	0	0	0	0
TO_SHIP	0	0	0	0	0	0	0
TO_SUGAR	0	0	0	0	0	0	0
TO_TRADE	0	0	0	0	3	0	1

Questo strumento **T-LAB** permette di eseguire una **classificazione automatica** delle **unità lessicali** (cioè parole e lemmi, incluse multiwords) o delle **unità di contesto** (cioè frasi, paragrafi o documenti brevi) presenti in un corpus applicando un insieme di categorie pre-definite o scelte dall'utilizzatore.

A seconda del tipo di categorie usate, le quali possono essere contenute in un dizionario opportunamente importato o generate da **T-LAB**, tale classificazione può essere considerata un tipo di **analisi del contenuto** o un tipo di **sentiment analysis**.

Poiché il processo di analisi consente di creare nuove variabili e altri dizionari che possono essere esportati e importati in ulteriori progetti di analisi, tale strumento può essere usato anche per esplorare lo stesso corpus da prospettive diverse così come per analizzare due o più insiemi di testi applicando gli stessi modelli.

Tra i **possibili usi** di questo strumento, segnaliamo i seguenti:

- Codifica automatica di risposte a domande aperte;
- Analisi top-down dei discorsi politici;
- Sentiment Analysis di commenti concernenti specifici prodotti;

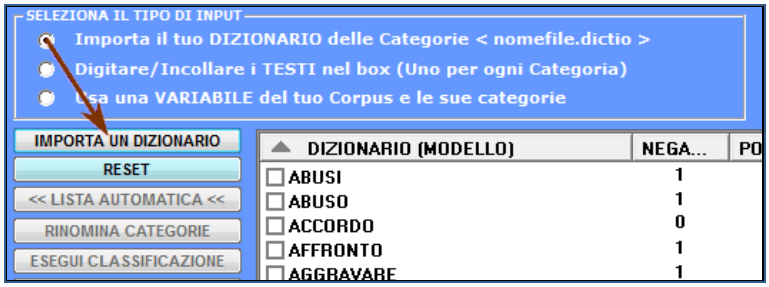
- Verifica del processo psicoterapeutico;
- Validazione di metodi per l'analisi qualitativa.

Di seguito viene fornita una breve descrizione delle quattro fasi principali del processo di analisi, che - tuttavia - sono da considerarsi indipendenti l'una dall'altra. Infatti, il ricercatore può utilizzare questo strumento anche solo per personalizzare i suoi dizionari o per esplorare il suo set di dati.

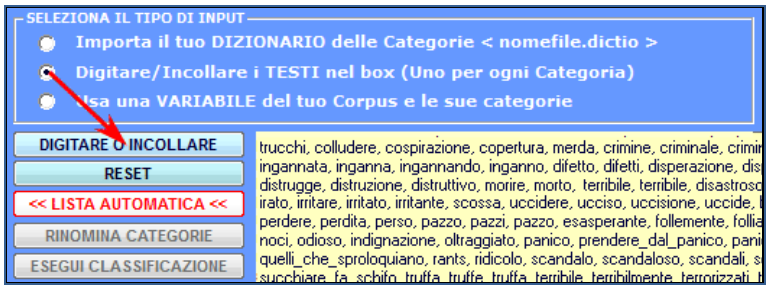
A) - FASE DI PRE-PROCESSING

I punti di partenza e i corrispondenti **tipi di input** della fase di pre-processing possono essere tre:

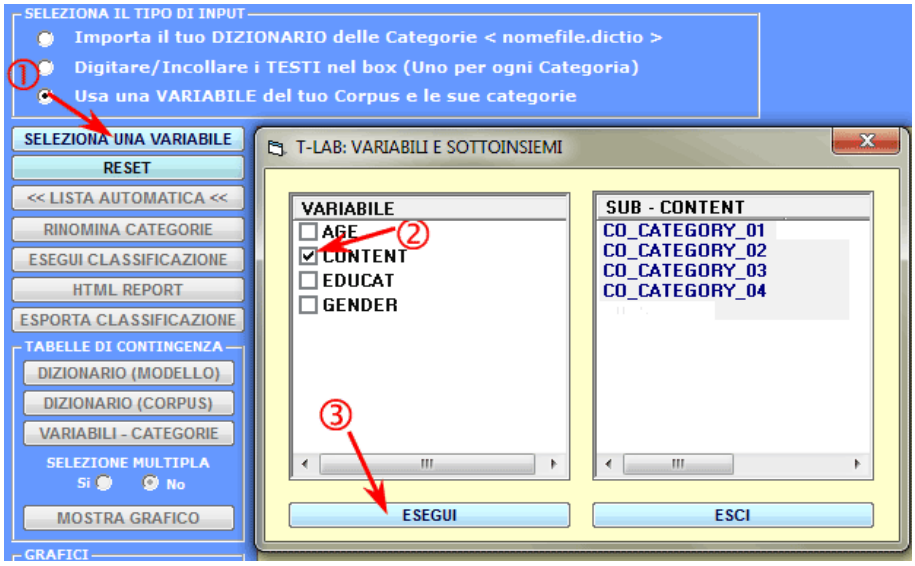
1 - un **dizionario** delle categorie nel formato appropriato è già disponibile (vedere le relative informazioni nella sezione 'E' di questo documento). In questo caso basta cliccare l'opzione '**Importa un Dizionario**' (vedi sotto);



2 - un dizionario delle categorie deve essere ricavato da **esempi** di testo o da **liste di parole** fornite dall'utilizzatore. In questo caso, è sufficiente digitare o copiare / incollare i testi nella casella appropriata (un esempio per ogni categoria, uno dopo l'altro, max 100.000 caratteri ciascuno);

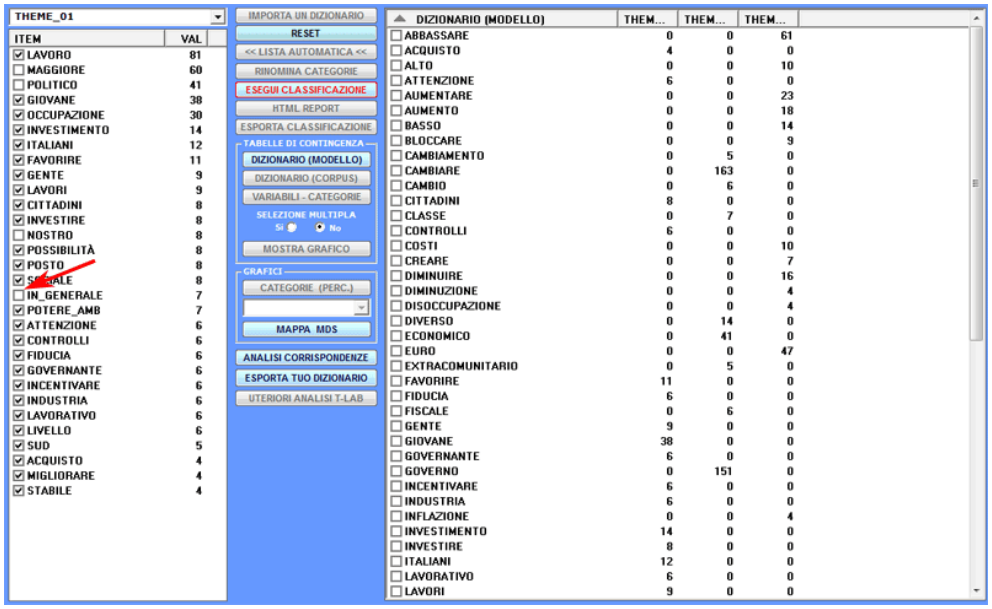


3 - un dizionario delle categorie deve essere ricavato da una **variabile** derivante da una precedente analisi di contenuto. In questo caso, basta cliccare l'opzione '**Seleziona una Variabile**' ed effettuare le scelte appropriate (vedi sotto).

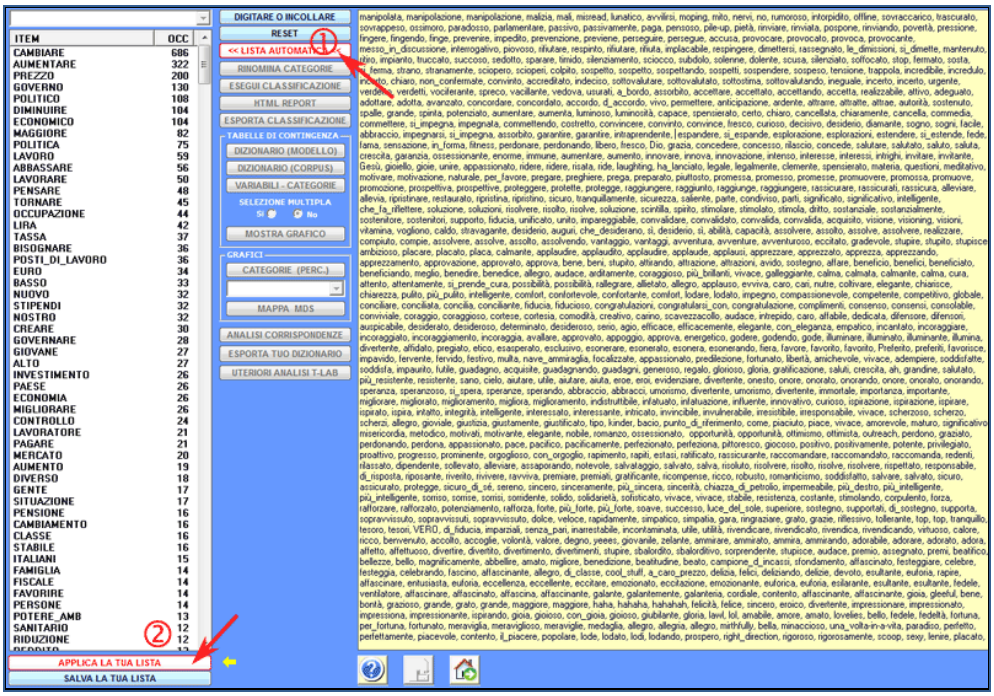


A seconda dei tre casi sopra elencati, prima di abilitare l'opzione 'Esegui Classificazione', **T-LAB** funziona nel modo seguente:

1 - il dizionario importato viene trasformato in una tabella di contingenza che l'utilizzatore può esplorare in vari modi (vedere la sezione 'C' del presente documento); inoltre, selezionando ogni categoria, uno o più degli elementi corrispondenti possono essere eliminati (vedi immagine seguente).

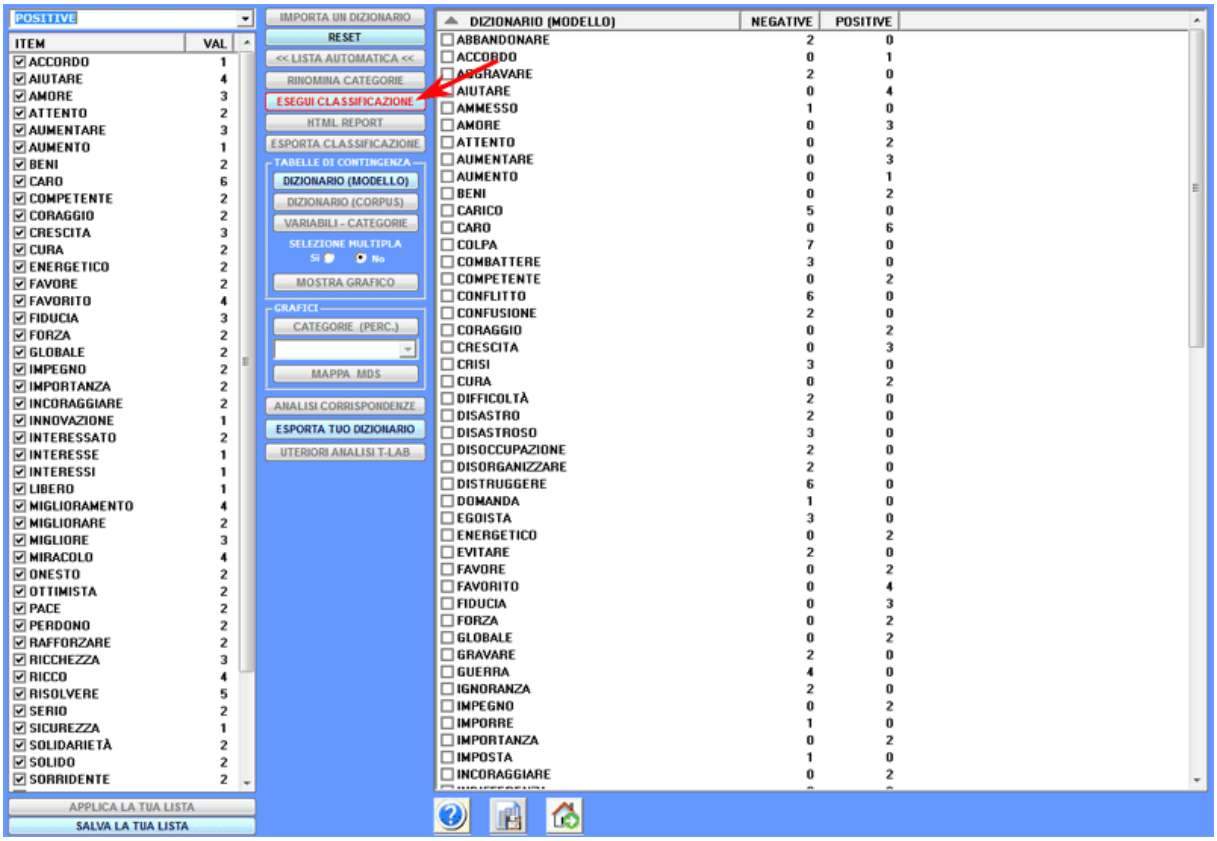


2 - quando i testi di esempio sono inseriti nella casella corrispondente, dopo aver cliccato il pulsante '**Lista Automatica**' (vedi sotto), **T-LAB** esegue uno specifico tipo di lemmatizzazione che utilizza solo il vocabolario del corpus selezionato (vedi la lista di parole nella zona sinistra dell'immagine seguente), quindi trasforma ogni testo in un elenco i cui elementi possono essere selezionati e deselezionati. Successivamente, per convalidare ogni lista di parole (cioè ogni categoria del dizionario), bisogna cliccare l'opzione '**Applica la tua lista**' (vedi sotto). Tutte le suddette operazioni devono essere ripetute per ogni categoria del dizionario, dopodiché l'utilizzatore viene abilitato ad eseguire le operazioni descritte nella sezione 'C' di questo documento.



3 - quando viene selezionata una variabile risultante da una precedente analisi del contenuto, **T-LAB** visualizza la relativa tabella di contingenza parole per categorie e l'utilizzatore può eseguire tutte le operazioni di esplorazione dei dati (vedere la sezione 'C' del presente documento).

B) - PROCESSO DI CLASSIFICAZIONE



Dopo aver cliccato l'opzione '**Esegui classificazione**' (vedi sopra), a seconda del tipo di corpus in analisi, l'utilizzatore può effettuare le scelte seguenti:

A questo punto, se l'utente decide di **classificare le 'parole'**, non sono disponibili altre scelte; infatti, in tal caso, le occorrenze di ogni parola (cioè i word tokens) sono semplicemente conteggiati come occorrenze della categoria corrispondente. Per esempio, se una categoria del nostro dizionario è 'religione' e questa include parole come 'fede' e 'preghiera', quando si analizza un documento che contiene le due parole in questione, **T-LAB** si limita a raggruppare le loro occorrenze. Ad esempio, 2 occorrenze di 'fede' e 3 occorrenze di 'preghiera' diventano 5 occorrenze di 'religione'.

Diversamente, se l'utente decide di **classificare le unità di contesto** (e cioè 'contesti elementari' come frasi e paragrafi o 'documenti'), **T-LAB** considera sia le categorie dizionario che le unità di contesto da classificare come profili di co-occorrenze (cioè term vectors) e calcola le loro misure di similarità. A questo scopo, i profili di co-occorrenze posssono essere filtrati tramite una 'Lista T-LAB' (cioè da una lista che include tutte parole-chiave con valori di occorrenza maggiori o uguali alla soglia minima di 4) o tramite una lista personalizzata (cioè da una lista che include tutte parole-chiave derivanti da scelte dell'utilizzatore) , le quali liste - tuttavia -possono a volte risultare uguali. Inoltre in questi casi **T-LAB** consente di escludere dall'analisi unità di contesto che non contengano un numero minimo di parole chiave al loro interno (vedi sopra il parametro 'co-occorrenze all'interno delle unità di contesto').

Quando, come nel caso appena descritto, gli 'oggetti' da classificare sono le unità di contesto, **T-LAB** procede nel modo seguente:

a) normalizza i vettori corrispondenti alle 'k' categorie del dizionario utilizzato, cioè i relativi profili colonna;
b) normalizza i vettori corrispondenti alle unità di contesto da analizzare;
c) calcola misure di similarità (coseno) e differenza (distanza euclidea) tra ogni 'i' vettore corrispondente a una unità di contesto e ogni 'k' vettore corrispondente a una categoria del dizionario utilizzato;
d) assegna ogni unità di contesto ('i') alla classe o categoria ('k') con la quale ha la relazione di somiglianza più elevata. (NB: In tutti i casi, per ogni coppia 'unità di contesto' / 'categoria' deve esserci una corrispondenza tra il massimo valore del coseno e il minimo valore della distanza euclidea, altrimenti T-LAB considera la 'i' unità di contesto come 'non classificata').

In In altre parole, nel caso appena descritto **T-LAB** utilizza una sorta di metodo K-means in cui i 'k' centroidi sono definiti priori ed essi non vengono aggiornati durante il processo di analisi.

Poiché in questo caso la classificazione è di tipo top-down, la qualità dei risultati ottenuti dipende essenzialmente da due fattori:

1 - la 'pertinenza' del dizionario utilizzato (vedi relazione tra lessico del corpus e dizionario delle categorie);
2 - la capacita 'discriminante' di ciascuna delle categorie (vedi relazione tra le varie categorie del dizionario).
Infatti, quando tali due fattori sono ottimali, entrambi i parametri di 'precision' e 'recall' (vedi http://en.wikipedia.org/wiki/Precision_and_recall) hanno valori compresi tra 80% e 95%.

Si ricordi che, al momento, **T-LAB** non tiene conto delle formule di negazione; di conseguenza, effettuando una sentiment analysis, una frase come 'Non odiare il tuo nemico' può risultare classificata come a tonalità 'negativa'. Gli utilizzatori esperti possono gestire questo problema durante l'importazione corpus (vedi l'uso di liste per stop-words e multi-words). Ad esempio, l'espressione 'non odiare' può essere trasformata in 'non_odiare' e, se lo si ritiene opportuno, può essere inclusa nella categoria 'positivo'.

C) - ESPLORAZIONE DEI DATI

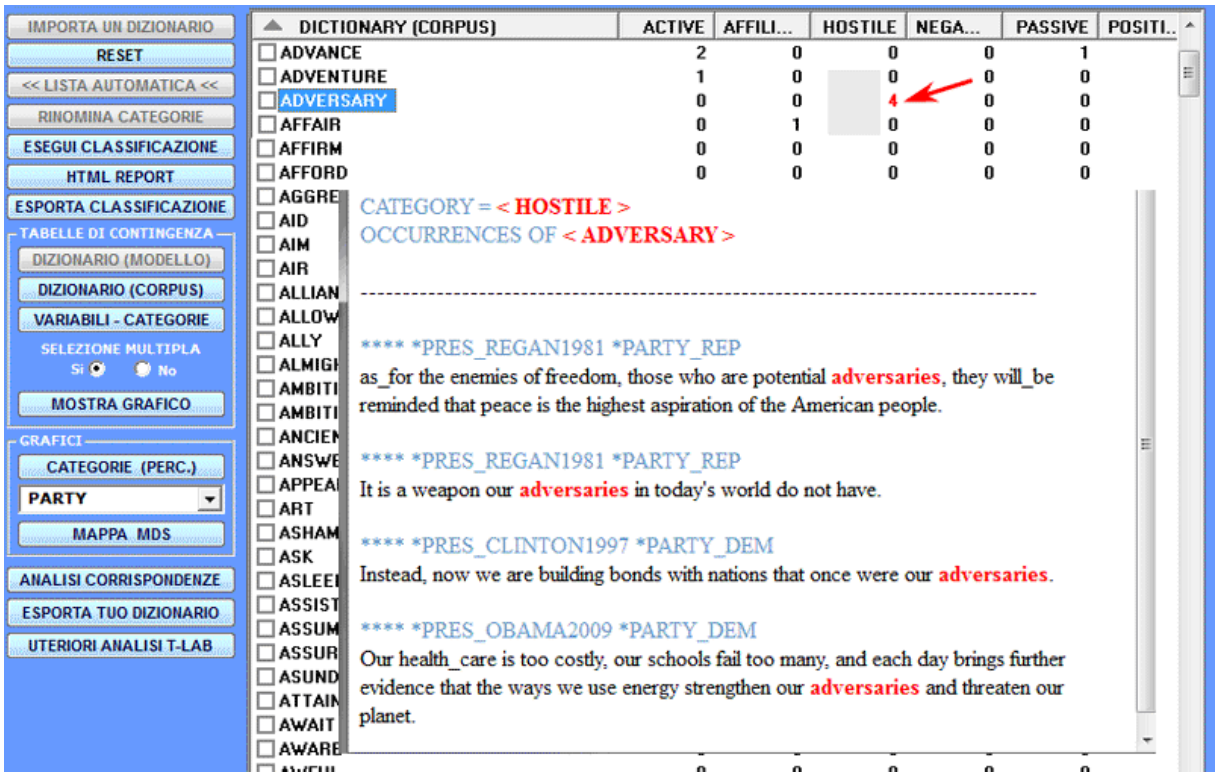
Nell'uso di questo strumento qualsiasi attività di esplorazione fa riferimento a **tabelle di contingenza** in cui, a seconda dei casi, possono essere rappresentati sia i dati in input (ad esempio un dizionario di categorie) che i dati in output (ad esempio i risultati del processo di classificazione).

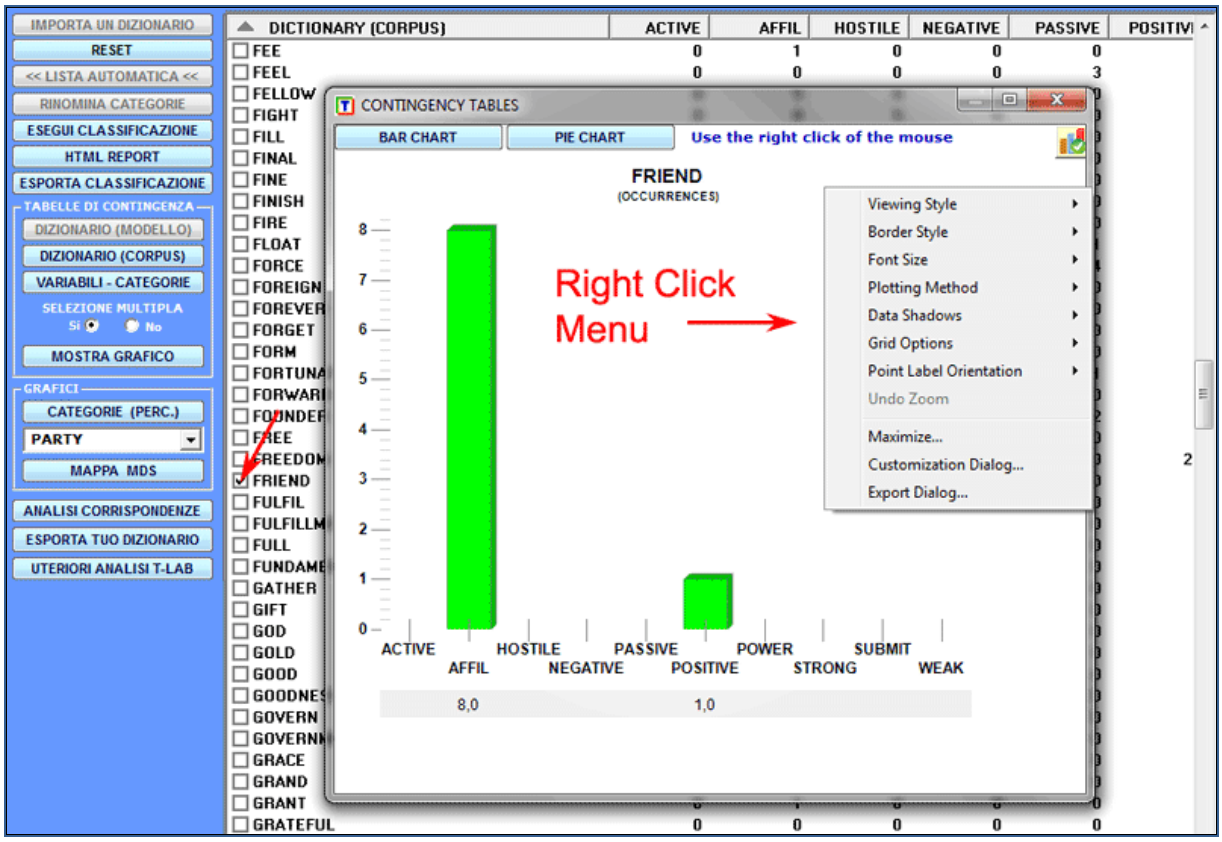
In particolare, per quanto riguarda i risultati dell'analisi, a seconda delle unità testuali classificate - rispettivamente (a) 'parole', (b) 'contesti elementari' o (c) 'documenti' - le celle delle tabelle visualizzate contengono i seguenti valori:

a) totale delle occorrenze di ogni parola che, all'interno del corpus analizzato o di un suo sottoinsieme, è stata classificata come appartenente ad una categoria predefinita (ovvero alla 'j' colonna della rispettiva tabella di contingenza). Si noti che in questo tipo di classificazione le parole appartenenti contemporaneamente a due o più categorie hanno gli stessi valori ripetuti nelle colonne corrispondenti;
b) totale dei contesti elementari assegnati ad una determinata categoria (vale a dire la 'j' colonna) in cui è presente la parola nella riga ('i') corrispondente;
c) totale delle occorrenze di ogni parola (vedi righe della relativa tabella di contingenza) all'interno dei documenti assegnati a ciascuna categoria (vedi colonne della tabella di contingenza).

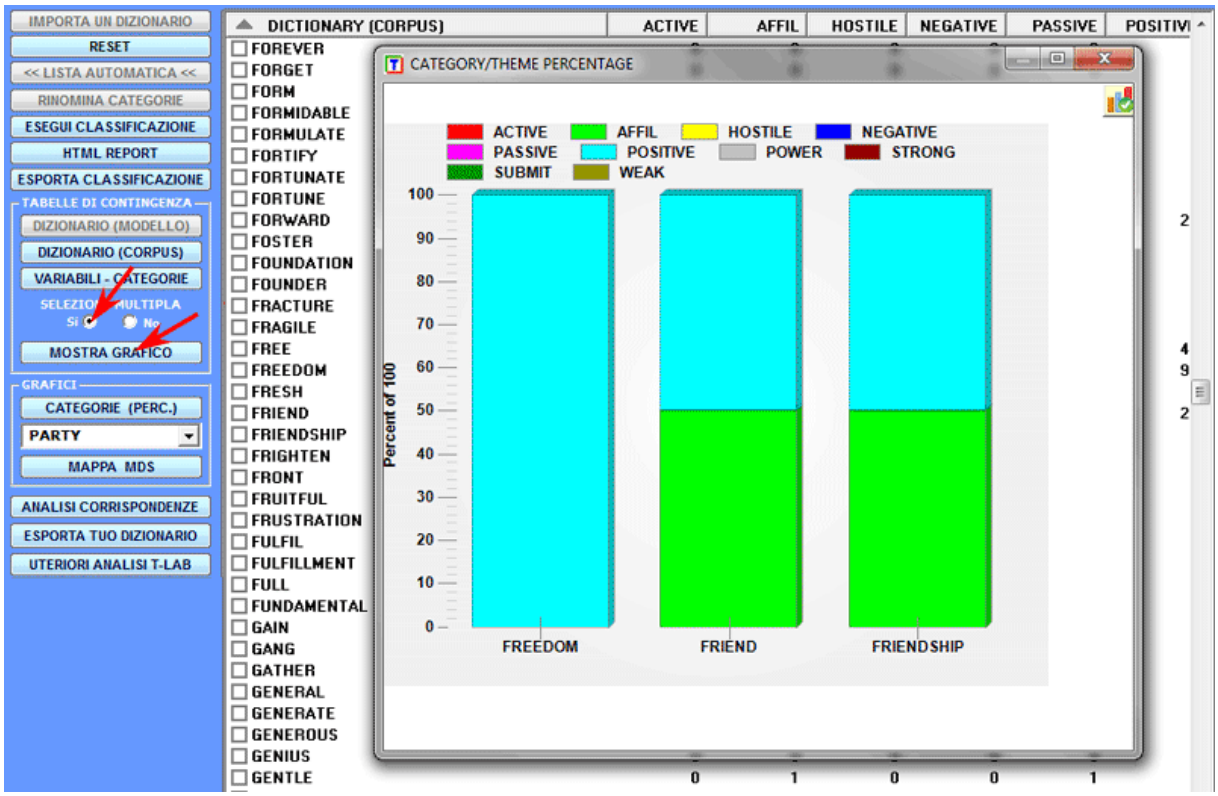
Cliccando i check-box corrispondenti ai vari item in riga è possibile ottenere grafici che possono essere personalizzati in vari modi; inoltre, ma solo nel caso della classificazione di tipo 'b' (vedi sopra), cliccando i valori contenuti nelle celle è possibile visualizzare i contesti di occorrenza di ogni parola.

Di seguito vengono riportati alcuni output risultanti da un processo di analisi in cui alcune categorie di un 'classico' dizionario per l'analisi di contenuto (Harvard IV-4) sono state applicate ai discorsi inaugurali dei presidenti degli Stati Uniti.

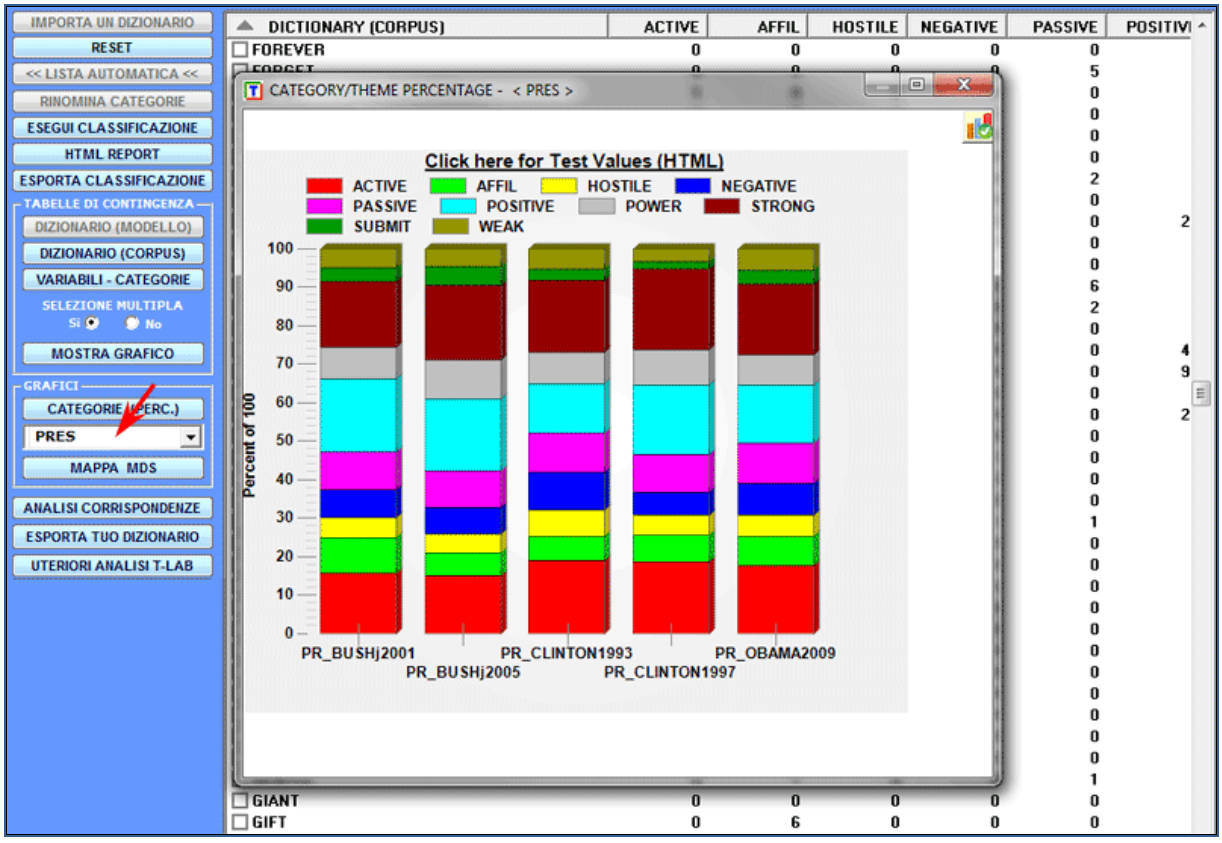




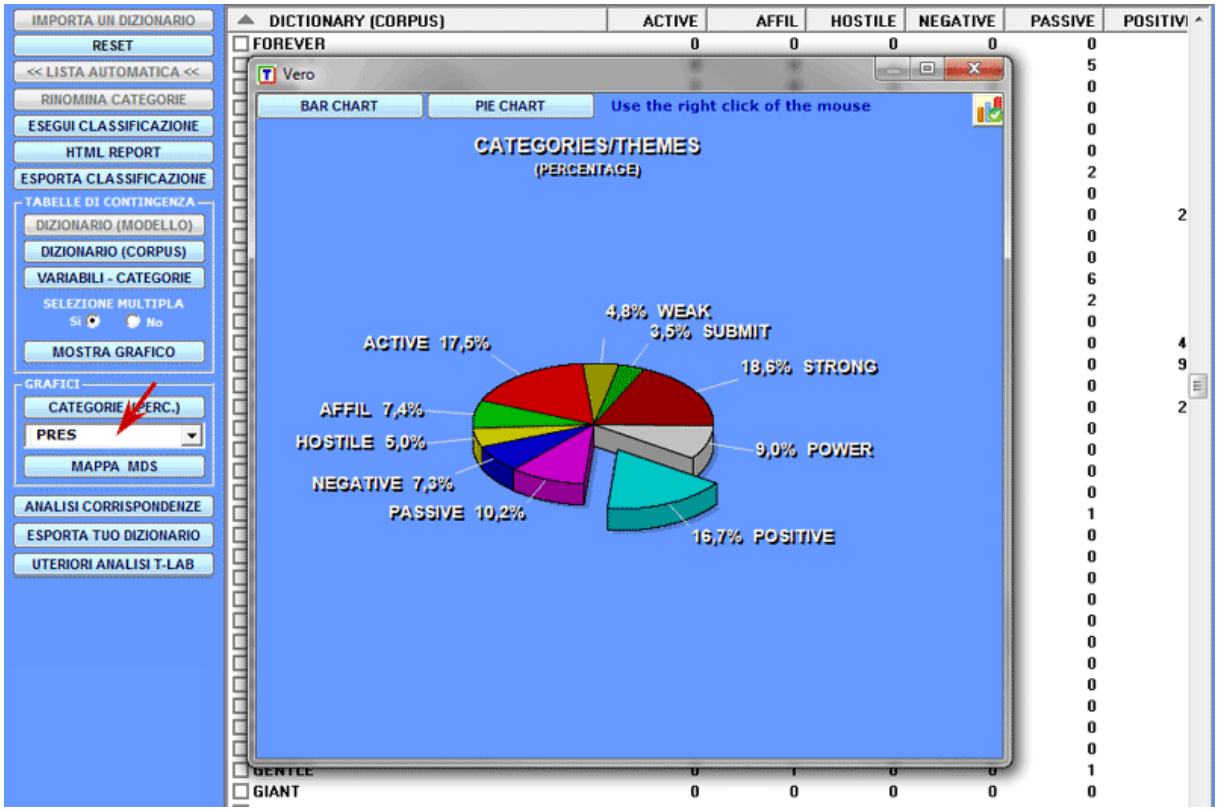
Per realizzare grafici con più serie di dati corrispondenti a più righe delle tabelle di contingenza, basta scegliere '**Selezione multipla**' (opzione 'Si'), selezionare fino a 20 elementi e cliccare il pulsante '**Mostra Grafico**' (vedi sotto).



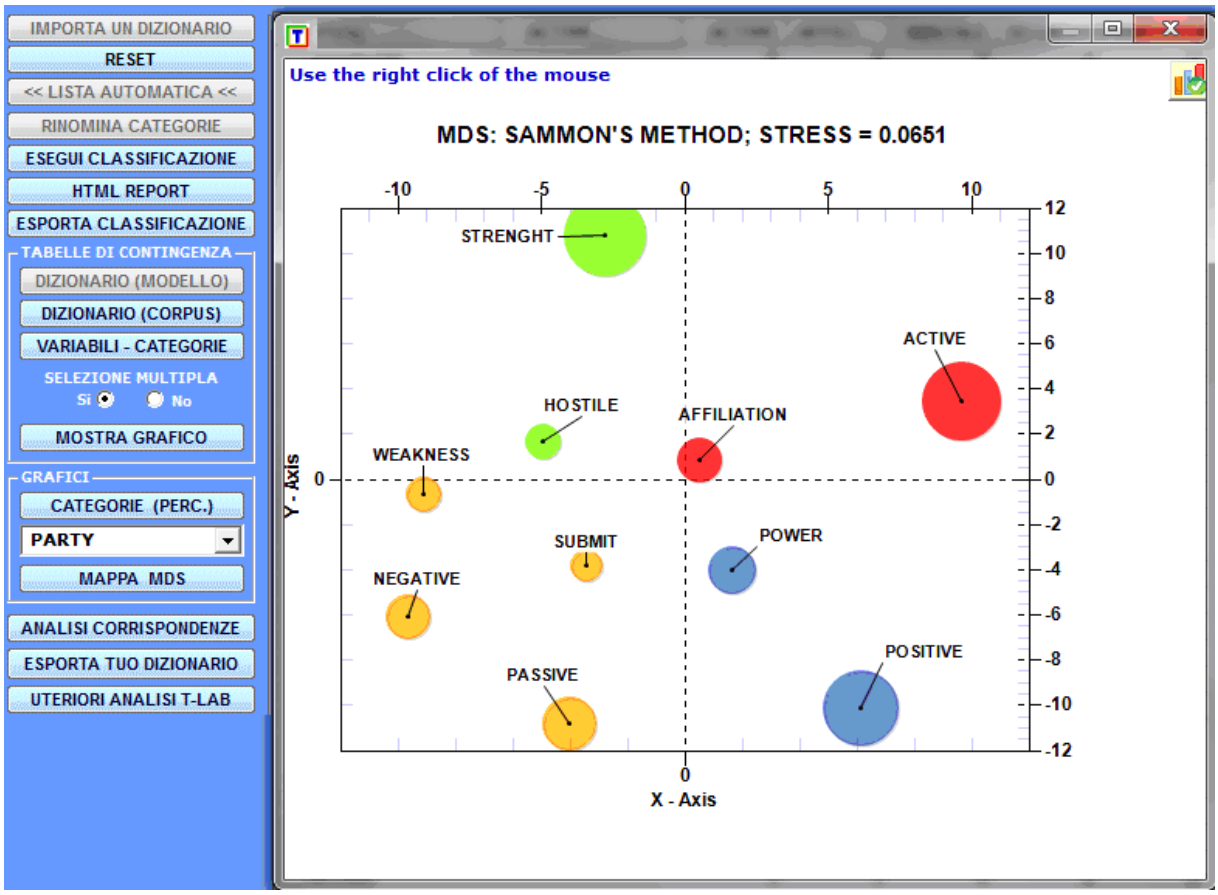
Le due opzioni di cui sopra sono anche disponibili per le tabelle con i valori delle variabili.

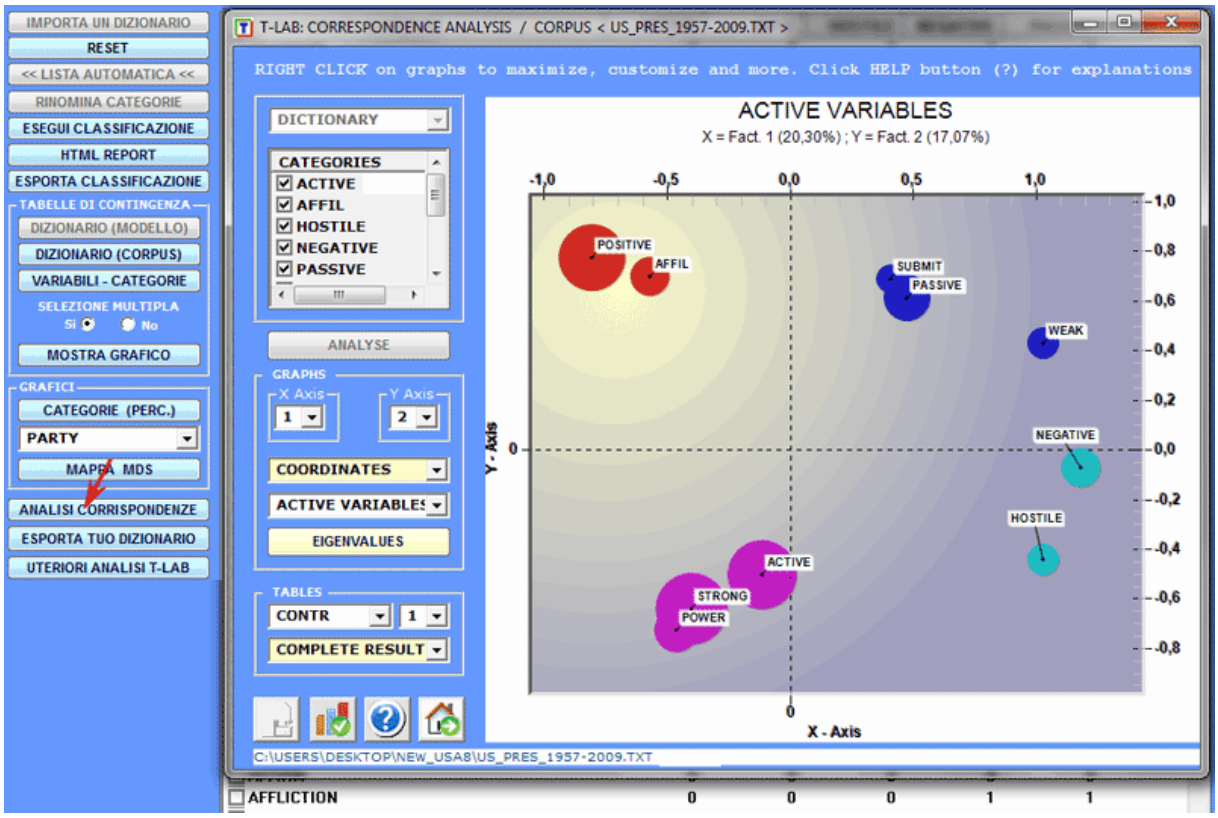


Le percentuali delle categorie possono essere verificate in vari modi (vedi sotto)



Per esplorare la struttura complessiva dei dati riportati nelle tabelle di contingenza è possibile utilizzare sia l'opzione '**MDS**' che l'opzione '**Analisi delle Corrispondenze**' (vedi sotto).





Solo nel caso in cui siano state classificate unità di contesto è possibile visualizzare ed esportare ulteriori output con i dati corrispondenti; inoltre, in tal caso, è anche possibile salvare i risultati dell'analisi in una nuova variabile e proseguire l'esplorazione con altri strumenti del menu **T-LAB**.

In dettaglio, cliccando sul pulsante '**HTML Report**' è possibile visualizzare alcuni risultati del processo di classificazione in cui un punteggio di somiglianza (Coseno) è assegnato a tutti i 'contesti elementari' o 'documenti' appartenenti alle varie categorie (N.B.: le immagini che seguono sono relative un corpus di documenti contenenti brevi descrizioni di aziende).

THEME < MEDICAL >

SCORE (.143)

Cytokinetics, Incorporated (Cytokinetics) is a biopharmaceutical company focused on developing small molecule therapeutics for the treatment of cardiovascular diseases and cancer. The Company's development efforts are directed to advancing multiple drug candidates through clinical trials to demonstrate proof-of-concept in humans in two markets: heart failure and cancer.

SCORE (.119)

Pharmacopeia, Inc. is a clinical development stage biopharmaceutical company dedicated to discovering and developing small molecule therapeutics to address medical needs. It has a portfolio of clinical and preclinical candidates under development internally or by partners, including eight clinical compounds in Phase II or Phase I development addressing multiple indications.

SCORE (.115)

Dyax Corp. (Dyax) is a clinical stage biotechnology company focused on the discovery, development and commercialization of biotherapeutics for unmet medical needs, with an emphasis on oncology and inflammatory indications. Dyax uses the drug discovery technology, known as phage display, to identify antibody, small protein and peptide compounds for clinical development.

SCORE (.111)

Rigel Pharmaceuticals, Inc. (Rigel) is a clinical-stage drug development company that discovers and develops small molecule drugs for the treatment of inflammatory/autoimmune diseases, cancer and viral diseases. The Company's research focuses on intracellular signalling pathways and related targets that are critical to disease mechanisms.

SCORE (.111)

It is also awaiting a decision from the United States Food and Drug Administration (FDA) regarding its application to market VELCADE for patients with diagnosed multiple myeloma. Millennium Pharmaceuticals, Inc. has a development pipeline of clinical and preclinical product candidates in its therapeutic focus areas of cancer and inflammatory diseases.

DOCUMENT	THEME	SCORE	BEGINNING
00001	SEMICONDUCTOR	0,051	2Wire , or not 2Wire , that is the question ...
00002	SEMICONDUCTOR	0,125	3Com Corporation (3Com) provides secure ...
00003	SEMICONDUCTOR	0,059	3D Systems Corporation is a holding company ...
00004	CHEMICAL	0,065	3M Company (3M) is a diversified technology ...
00005	SEMICONDUCTOR	0,095	What We Build 3PAR® (NYSE Arca : PAR ...
00006	MEDICAL	0,102	Abbott Laboratories is engaged in the discovery ...
00007	MEDICAL	0,071	ABIOMED , Inc . (ABIOMED) , provides ...
00008	CHEMICAL	0,046	Manufactures turbines & turbine generator ...
00009	CHEMICAL	0,085	ACCO Brands Corporation is a supplier of ...
00010	MEDICAL	0,013	focused on the casino industry . Developing ...
00011	CHEMICAL	0,078	Slides rule at Accuride International . ...
00012	MEDICAL	0,102	Established : Acorn Cardiovascular™ is ...
00013	SEMICONDUCTOR	0,094	Actel Corporation is a supplier of low-power ...
00014	MEDICAL	0,120	ActivBiotics , Inc . (ActivBiotics) ...
00015	SEMICONDUCTOR	0,129	ActivIdentity Corp . is a provider of digital ...
00016	CHEMICAL	0,126	Actuant Corporation (Actuant) is a manufacturer ...
00017	CHEMICAL	0,094	Acuity Brands , Inc . (Acuity Brands ...
00018	CHEMICAL	0,041	The Adams Manufacturing Company cares for ...
00019	SEMICONDUCTOR	0,145	Adaptec , Inc (Adaptec) , designs ...
00020	SEMICONDUCTOR	0,183	ADC Telecommunications , Inc . (ADC ...
00021	SEMICONDUCTOR	0,118	Adobe Systems Incorporated is a diversified ...
00022	MEDICAL	0,089	Adolor Corporation is a development-stage ...
00023	SEMICONDUCTOR	0,159	ADTRAN , Inc . (ADTRAN) designs , ...
00024	SEMICONDUCTOR	0,124	Advanced Analogic Technologies Incorporated ...
00025	MEDICAL	0,033	Advanced Ceramic Research was founded in ...

Dati analoghi possono essere esportati in file XLS (vedi sotto) che contengono tutte le informazioni riguardanti i contesti elementari ('Context_Classification.xls') o i documenti ('Document_Classification.xls') correttamente classificati;

(1) - Context_Classification.xls

'IDNUMBER	THEME	SCORE	CONTEXT
'0000100001	SEMICONDUCTOR	0,017	2Wire , or not 2Wire , that_is the question : Whether ' tis nobler in networks to suffer the slings a
'0000100002	SEMICONDUCTOR	0,044	2Wire ' s HomePortal and OfficePortal networking devices combine router and firewall functions , ar
'0000100003	SEMICONDUCTOR	0,01	2Wire also makes DSL filters and adapters . Alcatel-Lucent owns one-quarter of 2Wire . For in bro
'0000200001	SEMICONDUCTOR	0,065	3Com Corporation (3Com) provides secure , converged networking solutions on a global scale to
'0000200002	SEMICONDUCTOR	0,081	3Com ' s long-term , technology-based strategy centers on enterprises and public_sector organizat
'0000300001	CHEMICAL	0,033	3D Systems Corporation is a holding company that operates through subsidiaries in the United_Sta
'0000300002	SEMICONDUCTOR	0,043	The Company ' s systems are used by its customers to produce physical objects from digital data u
'0000400001	CHEMICAL	0,035	3M Company (3M) is a diversified technology company with a presence in various businesses , i
'0000400002	CHEMICAL	0,024	3M manages its operations in six business segments : Industrial and Transportation ; health_care
'0000400003	CHEMICAL	0,032	The Company ' s products are sold through numerous distribution channels , including directly to u
'0000500001	SEMICONDUCTOR	0,018	What We Build 3PAR® (NYSE Arca : PAR) is the leading global provider of utility storage , a c
'0000500002	SEMICONDUCTOR	0,008	Next-generation storage is a category of arrays developed to address the limitations of traditional st
'0000500003	SEMICONDUCTOR	0,03	The Problem We Solve 3PAR Utility Storage is designed to address the problem of costly , comple
'0000500004	SEMICONDUCTOR	0,066	Our Customers 3PAR customers are organizations for whom delivering IT as a service is mission-cr
'0000500005	SEMICONDUCTOR	0,038	The Value We Bring 3PAR Utility Storage enables customers to cut Total Cost of Data by up_to 50
'0000600001	MEDICAL	0,033	Abbott Laboratories is engaged in the discovery , development , manufacture and sale of a diversifi
'0000600002	MEDICAL	0,042	The Diagnostic Products segment ' s products include diagnostic systems and tests for blood bank
'0000600003	MEDICAL	0,034	The Vascular Products segment ' s products include a line of coronary , endovascular and vessel c
'0000700001	MEDICAL	0,022	ABIOMED , Inc . (ABIOMED) , provides medical products and services in the area of circulator
'0000700002	MEDICAL	0,044	The Company ' s products can be used in a range of clinical settings , including by heart surgeons
'0000700004	MEDICAL	0,008	intra-aortic balloons (IABs) , and ventricular assist devices (VADs) .
'0000800001	CHEMICAL	0,046	Manufactures turbines & turbine generator sets & parts ; manufactures motor vehicle parts & acce
'0000900001	CHEMICAL	0,052	ACCO Brands Corporation is a supplier of select categories of branded office products (excluding f
'0000900002	CHEMICAL	0,03	personal computer accessory products , paper-based time management products , presentation a
'0000900003	CHEMICAL	0,013	During the year ended December 31 , 2007 , these markets represented 61% , 28% and 8% of its
'0001000001	MEDICAL	0,013	focused on the casino industry . Developing innovative new games , dazzling visual environments ,
'0001100001	CHEMICAL	0,017	Slides rule at Accuride International . Accuride International designs and makes ball bearing slides
'0001100002	CHEMICAL	0,072	The company ' s slides are also found in automotive accessories , including storage units and arm
'0001200001	MEDICAL	0,009	Establishe Acorn Cardiovascular™ is a privately held medical device company that was incorporate
'0001200002	MEDICAL	0,047	Mission : Acorn Cardiovascular develops innovative solutions to successfully treat patients with he
'0001200003	MEDICAL	0,031	Backgroun Heart failure (HF) is a condition that_is caused by damage to the heart muscle , whic
'0001200004	MEDICAL	0,027	An estimated 550 , 000 new HF cases are diagnosed each year in the United States alone . Heart
'0001200006	MEDICAL	0,033	It is intended to prevent and reverse the progression of heart failure by improving the heart ' s structu
'0001300001	SEMICONDUCTOR	0,065	Actel Corporation is a supplier of low-power field-programmable gate arrays (FPGAs) and prograr
'0001300002	SEMICONDUCTOR	0,039	programming hardware and starter kits ; and a variety of design services . Its Flash-based solutions
'0001400001	MEDICAL	0,063	ActivBiotics , Inc . (ActivBiotics) is a biopharmaceutical company focused on the discovery , d

(2) - Document_Classification.xls

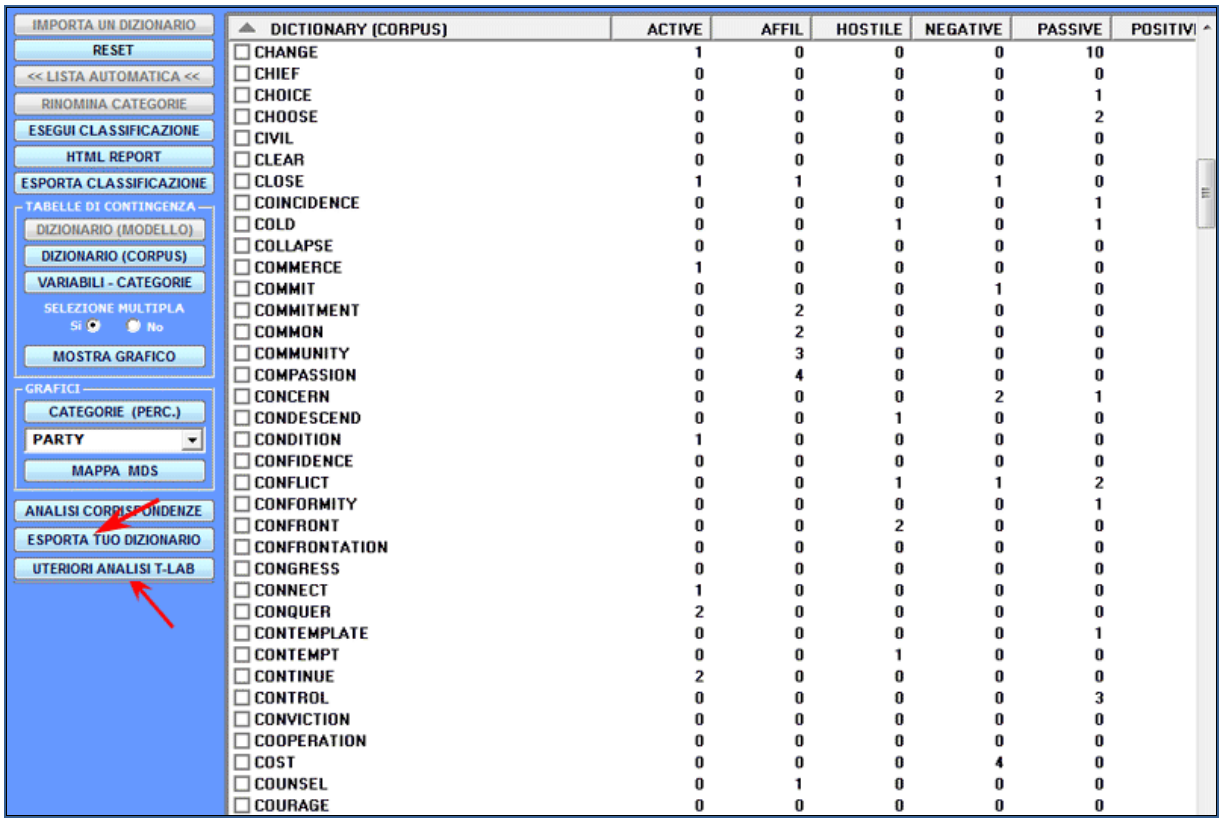
	'IDNUMBER	THEME	SCORE
2	'00001	SEMICONDUCTOR	0,051
3	'00002	SEMICONDUCTOR	0,125
4	'00003	SEMICONDUCTOR	0,059
5	'00004	CHEMICAL	0,065
6	'00005	SEMICONDUCTOR	0,095
7	'00006	MEDICAL	0,102
8	'00007	MEDICAL	0,071
9	'00008	CHEMICAL	0,046
10	'00009	CHEMICAL	0,085
11	'00010	MEDICAL	0,013
12	'00011	CHEMICAL	0,078
13	'00012	MEDICAL	0,102
14	'00013	SEMICONDUCTOR	0,094
15	'00014	MEDICAL	0,12
16	'00015	SEMICONDUCTOR	0,129
17	'00016	CHEMICAL	0,126
18	'00017	CHEMICAL	0,094
19	'00018	CHEMICAL	0,041
20	'00019	SEMICONDUCTOR	0,145
21	'00020	SEMICONDUCTOR	0,183
22	'00021	SEMICONDUCTOR	0,118
23	'00022	MEDICAL	0,089
24	'00023	SEMICONDUCTOR	0,159
25	'00024	SEMICONDUCTOR	0,124
26	'00025	MEDICAL	0,033
27	'00026	SEMICONDUCTOR	0,045
28	'00027	SEMICONDUCTOR	0,046
29	'00028	CHEMICAL	0,057
30	'00029	MEDICAL	0,082
31	'00030	SEMICONDUCTOR	0,058
32	'00031	CHEMICAL	0,051
33	'00033	MEDICAL	0,138
34	'00034	CHEMICAL	0,129
35	'00035	CHEMICAL	0,035
36	'00036	SEMICONDUCTOR	0,064

D) - ULTERIORI FASI DEL PROCESSO DI ANALISI

Quando il processo di classificazione ha prodotto i suoi output, sono disponibili due ulteriori opzioni:

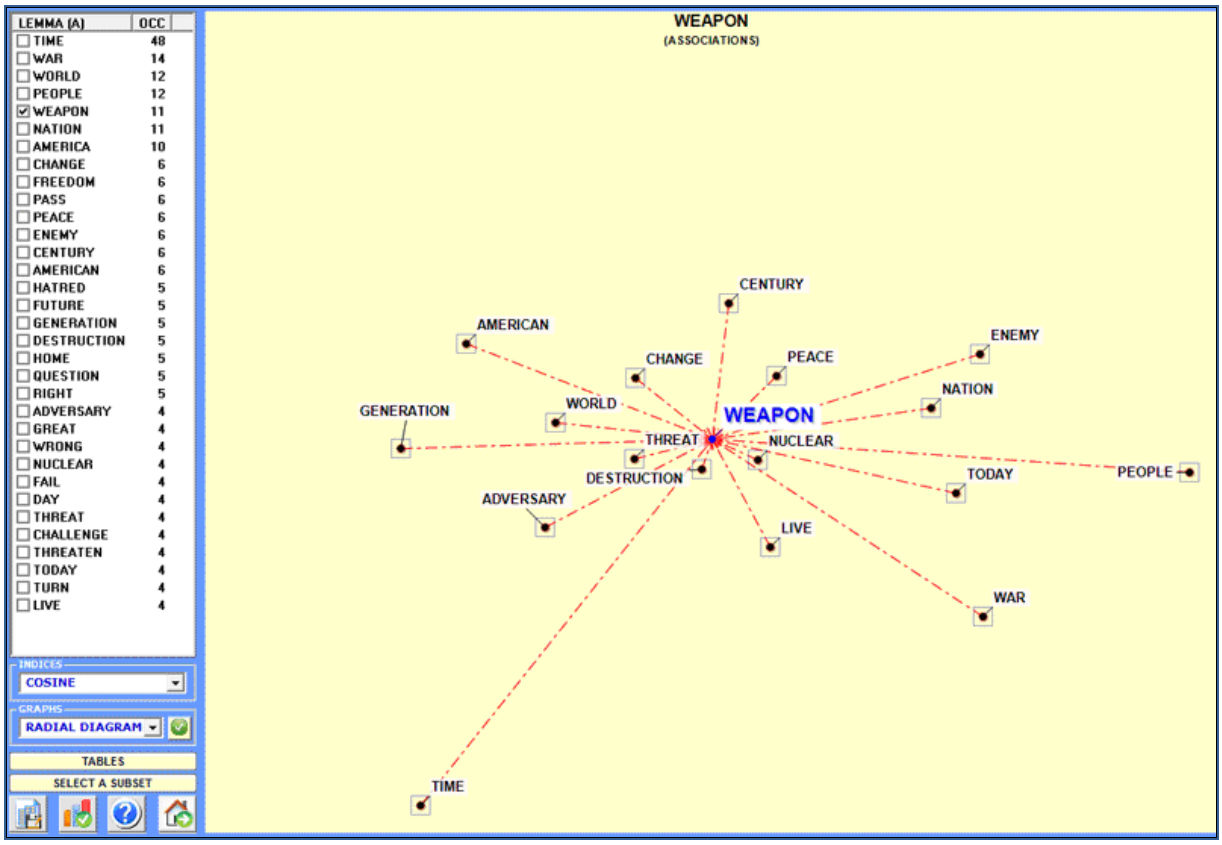
- '**Esporta il tuo Dizionario**', che crea un dizionario pronto per essere importato e utilizzato con altri strumenti **T-LAB** per le analisi tematiche;

- '**Ulteriori analisi T-LAB** ', che, a seconda della struttura del corpus analizzato, del tipo di classificazione eseguita e del numero di categorie applicate, produce una nuova variabile che può essere utilizzata da altri strumenti T-LAB (vedi sotto).



DICTIONARY (CORPUS)	ACTIVE	AFFIL	HOSTILE	NEGATIVE	PASSIVE	POSITIV
CHANGE	1	0	0	0	10	
CHIEF	0	0	0	0	0	
CHOICE	0	0	0	0	1	
CHOOSE	0	0	0	0	2	
CIVIL	0	0	0	0	0	
CLEAR	0	0	0	0	0	
CLOSE	1	1	0	1	0	
COINCIDENCE	0	0	0	0	1	
COLD	0	0	1	0	1	
COLLAPSE	0	0	0	0	0	
COMMERCE	1	0	0	0	0	
COMMIT	0	0	0	1	0	
COMMITMENT	0	2	0	0	0	
COMMON	0	2	0	0	0	
COMMUNITY	0	3	0	0	0	
COMPASSION	0	4	0	0	0	
CONCERN	0	0	0	2	1	
CONDESCEND	0	0	1	0	0	
CONDITION	1	0	0	0	0	
CONFIDENCE	0	0	0	0	0	
CONFLICT	0	0	1	1	2	
CONFORMITY	0	0	0	0	1	
CONFRONT	0	0	2	0	0	
CONFRONTATION	0	0	0	0	0	
CONGRESS	0	0	0	0	0	
CONNECT	1	0	0	0	0	
CONQUER	2	0	0	0	0	
CONTEMPLATE	0	0	0	0	1	
CONTEMPT	0	0	1	0	0	
CONTINUE	2	0	0	0	0	
CONTROL	0	0	0	0	3	
CONVICTION	0	0	0	0	0	
COOPERATION	0	0	0	0	0	
COST	0	0	0	4	0	
COUNSEL	0	1	0	0	0	
COURAGE	0	0	0	0	0	

Di seguito è riportato un esempio ottenuto analizzando un 'sottoinsieme' dei contesti classificati mediante lo strumento **Associazioni di Parole** (vedi il menu principale **T-LAB**).



E) - FORMATO INPUT/OUTPUT DEI DIZIONARI T-LAB

Di seguito vengono riportate tutte le informazioni sul formato dei dizionari che possono essere importati da questo strumento **T-LAB**.

- tutti i dizionari devono essere file testo (ASCII/ANSI) con estensione 'dictio.' (e.s.: Mycategories.dictio);
- tutti i dizionari creati da strumenti **T-LAB** per le analisi tematiche, inclusi quelli creati dallo strumento 'Classificazione Basata su Dizionari', sono pronti per essere importati senza ulteriori interventi da parte dell'utilizzatore;
- altri dizionari, sia essi 'standard' che personalizzati devono essere prodotti seguendo le indicazioni riportate di seguito:

1 - ciascun dizionario è costituito da 'n' righe e non può superare il limite di 100.000 record;
2 - ogni riga del dizionario include due o tre 'stringhe' separate dal segno di punto e virgola (ad es.: economico; credito);
3 - per ogni linea, la prima stringa deve essere una 'categoria', la seconda una 'parola' (o lemma), la terza - se presente - deve essere un numero reale positivo (cioè un numero intero) da '1' a '999' che rappresenta il 'peso' di ogni parola all'interno della categoria corrispondente;
4 - la lunghezza massima di una stringa (parola, lemma o categoria) è di 50 caratteri e non deve contenere né gli spazi vuoti né apostrofi;
5 - quando il dizionario include multi-words (es. Governo Federale), gli spazi vuoti devono essere sostituiti con il carattere '_' (es. Governo_Federale);
6 - in ogni dizionario, il numero delle categorie utilizzate possono variare da un minimo di 2 a un massimo di 50. Quando il numero di categorie è superiore a 50 si consiglia di utilizzare un dizionario di formato diverso e di importarlo tramite lo strumento **Personalizzazione del**

Dizionario (vedi 'Strumenti Lessico' nel menu **T-LAB**). In tal caso si ricorda che ogni parola deve essere in corrispondenza univoca con una (sola) categoria.

Di seguito sono riportati due estratti di file .dictio, rispettivamente con due e tre stringhe per riga.

a) caso con due stringhe (vale a dire 'coppie' di categorie e parole)

```
...
negativo;catastrofico
negativo;nocivo
...
positivo;fantastico
positivo;soddisfatto
...
```

b) caso con tre stringhe (cioè categorie, parole e numeri)

```
...
negativo;catastrofico;10
negativo;nocivo;8
...
positivo;fantastico;9
positivo;soddisfatto;7
```

Testi e Discorsi come Sistemi Dinamici

N.B .: Questa sezione dell'help è disponibile solo in inglese.

This **T-LAB** tool provides several **integrated analysis options** (see picture below) which can be used in various combinations for obtaining measures and graphical representations concerning **texts treated as dynamic systems**.

In particular this tool allows us to verify how texts are organized in time, how the **recurring themes** and the **sequential order** of utterances relate to each other and how **similarities** and **differences** between them evolve in time. For these reasons this tool – more than other **T-LAB** tools - challenges the divide between qualitative and quantitative approaches in text analysis.

In principle the objects of this type of integrated analysis should be texts in which – like discourses and conversations – the **sequence** and the temporal flow of utterances is important (i.e. transcripts of focus group sessions, interviews, speeches, debates, doctor/patient iterations, novels etc.).

However, as this tool provides us with **similarity measures** concerning all pairs of text segments (both within the whole corpus and within its subsets), it may be also useful in other cases. Just remember that - when text segments are not in sequential order – the use of RQA Analysis and/or Sequence Analysis options does not produce proper results.

To begin with, two things must be taken into consideration:

-as the granularity is important, the key-word list chosen before using this tools should contain as many items as possible;
-at the moment, this tool allows us to analyse a corpus which includes up to 30,000 text segments (i.e. about 5,000 pages), which can even be organized in two or more sub-sections (i.e. corpus subsets). However, due to some limitations concerning the visualization of recurrence plots, both the RQA Analysis and the Similarities Measures are available only for corpuses consisting of up to 3,000 text segments (i.e. about 500 pages, and a bit more when the corpus has been segmented into paragraphs).

The **analysis procedure** consists of the several steps, some of which are automatic and others which – when desired - can be manually performed by the user.

The **initial steps** performed automatically by **T-LAB** are the following:

a - construction of a **document-term matrix**, where documents are always text segments (i.e. text fragments, sentences, paragraphs) into which the corpus has been subdivided (see the **T-LAB** initial settings options);
b - **topic analysis** based on a probabilistic model which uses the Latent Dirichlet Allocation and the Gibbs Sampling (see the related information on Wikipedia);

c – use of a **Naïve Bayes classifier** for estimating the probability values of each topic within each text segment, and for assigning each text segment to the topic (or theme **) it most closely resembles.

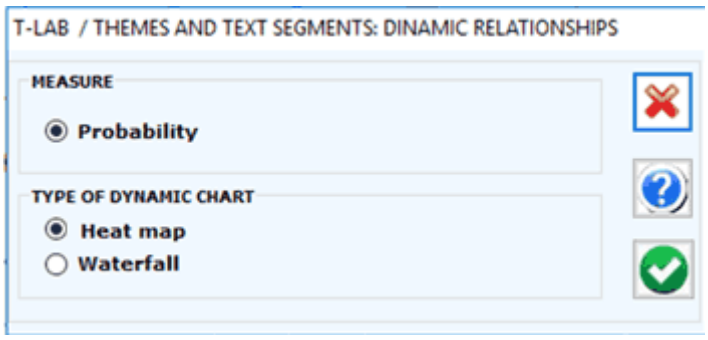
(**) 'Topic' and 'Theme' will be hereafter treated as synonymous terms.

Please note that the main goal of the above automatic steps is to extract 'k' latent dimensions (where 'k' varies from 20 to 30) which determine the content structure of the analysed text and which – like a mixture model - can be used for exploring both text dynamics and similarities between text segments. For this reason the segments used for building the model are only those in which at least two key-terms included in the user list are present. Differently, after building the model, every text segment – even by maintaining the mixed nature of its content - is assigned to the topic to which it most closely resembles.

At the end of automatic steps, **five options** are made available, two of which correspond to two analysis tools already present in the **T-LAB** menu – namely the Topic Analysis (i.e. Modelling of Emerging Themes) and the Sequence Analysis of themes – and which, for this very reason, do not need further explanations. Just consult the parts of this help/manual where the main options depicted in the below section 'F' are commented.

Regarding the **new tools**, here is – for each of them - the required information.

A) Real Time Charts



When plotting real time charts, which allow us to **dynamically visualize** the time sequence of the text segments from the beginning to the end, the measures used are always the probability values that the Bayes classifier has assigned – for each of the 'k' topics - to each text segment.

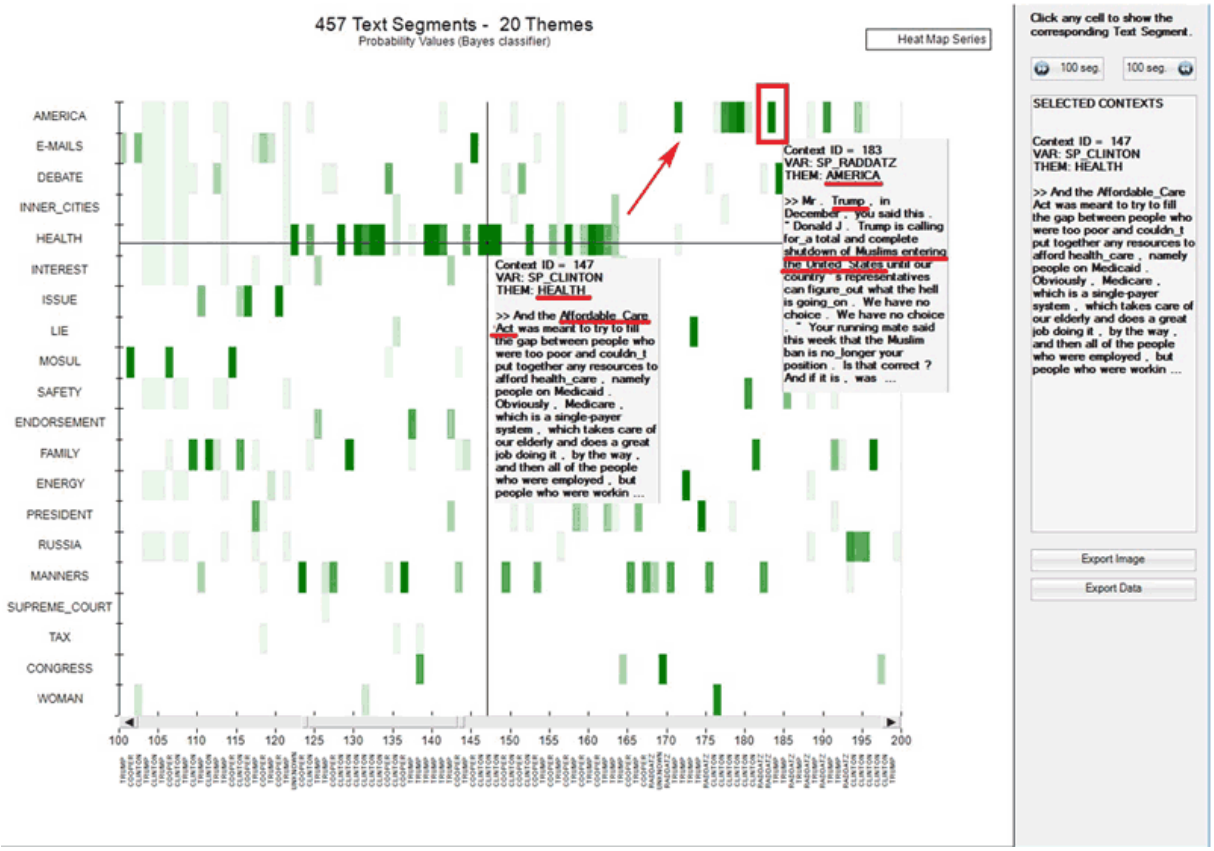
Two complementary charts allows us to easily appreciate various types of events, including the **strong recurrences** of some themes or the **shifts** from a theme to another (see the below pictures, obtained by analysing a presidential debate between Hillary Clinton and Donald Trump which took place on October 2016. N.B.: In this case the corpus was automatically segmented into paragraphs and a multi-word list was applied).

From a semiotic point of view, we may argue that both these types of charts deal with the relationships between **paradigm** and **syntagm** or – in other words – between the synchronic and diachronic axes, where paradigm/synchronic refers to the various themes and syntagm/diachronic refers to the temporal sequence of the 'N' text segments.

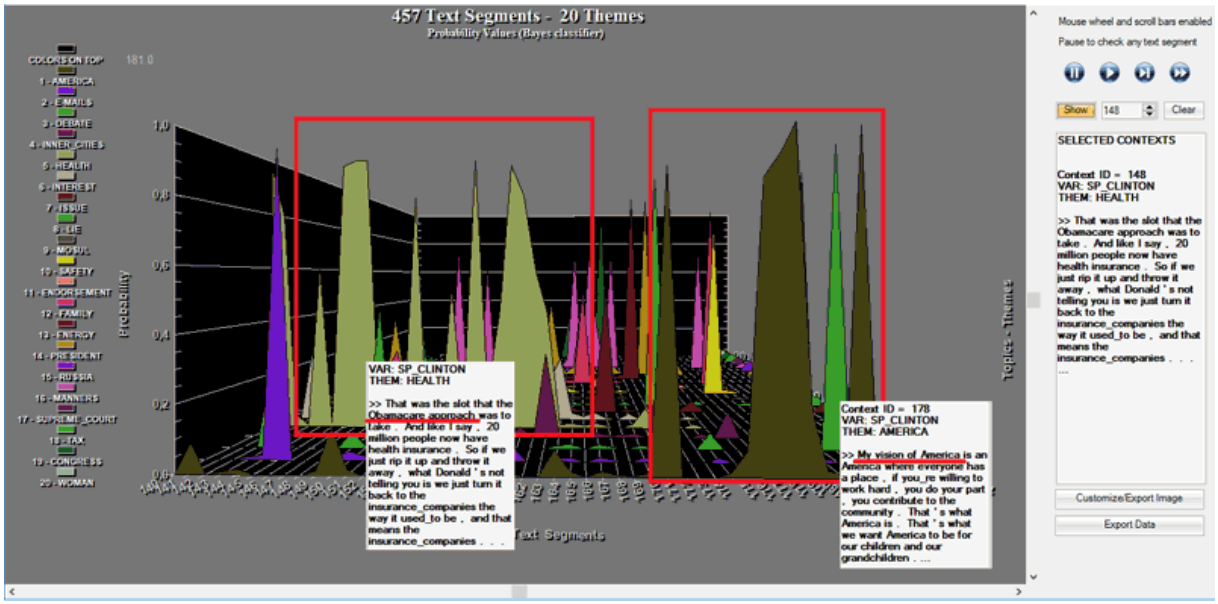
As the information summarized by these types of charts mainly refers to formal aspects of text contents, the same charts may be regarded as some sort of musical scores where the sequence of themes and their 'intensity' (i.e. probability) vary in time.

Anytime, in order to check 'who' is speaking and about 'what', just click the corresponding point.

A.1 - Heat map



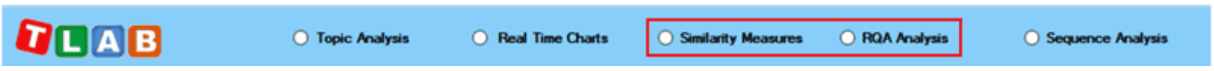
A.2 - Waterfall



Please note that in the real time charts all text segments are present, and each of them is represented as a mixture of probability values associated with the various topics which the model consists of. In fact, when clicking the 'Export Data' option, all this information is made available in a data table in CSV format like the following.

SPEAKER	THEME	ID_Segm	Selected	AMERICA	E-MAILS	DEBATE	INNER_CITIES	HEALTH	INTEREST	ISSUE	LIE	
SP_RADDATZ	MANNERS	1	16	0.0159	0.0003	0.0003	0.0027	0.0029	0.0006	0.0003	0.0003	...
SP_COOPER	MANNERS	2	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_UNKNOWN	DEBATE	3	3	0.0062	0.0000	0.9929	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_CLINTON	AMERICA	4	1	0.5593	0.1448	0.0002	0.0002	0.0006	0.0055	0.0148	0.0002	...
SP_CLINTON	AMERICA	5	1	0.9999	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	...
SP_CLINTON	AMERICA	6	1	0.9997	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_CLINTON	E-MAILS	7	2	0.1328	0.4183	0.3872	0.0130	0.0005	0.0003	0.0005	0.0001	...
SP_CLINTON	AMERICA	8	1	0.9969	0.0000	0.0000	0.0026	0.0000	0.0000	0.0000	0.0001	...
SP_COOPER	MANNERS	9	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_TRUMP	FAMILY	10	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_TRUMP	LIE	11	8	0.0000	0.0000	0.0000	0.0001	0.0244	0.0000	0.0000	0.9740	...
SP_TRUMP	LIE	12	8	0.0000	0.0000	0.0000	0.2745	0.0000	0.0001	0.0000	0.7248	...
SP_TRUMP	FAMILY	13	12	0.0003	0.0000	0.0252	0.0000	0.0000	0.0000	0.0000	0.0028	...
SP_TRUMP	INNER_CITIES	14	4	0.0016	0.0001	0.0001	0.7819	0.0002	0.0001	0.0007	0.1364	...
SP_COOPER	ISSUE	15	7	0.0000	0.0000	0.0071	0.0000	0.0000	0.0000	0.8903	0.0000	...
SP_TRUMP	E-MAILS	16	2	0.0002	0.7197	0.0000	0.0038	0.0000	0.0028	0.0000	0.0000	...
SP_TRUMP	FAMILY	17	12	0.0000	0.0000	0.0003	0.0046	0.0014	0.0769	0.0003	0.0001	...
SP_TRUMP	INNER_CITIES	18	4	0.0319	0.0004	0.0001	0.7348	0.0015	0.0152	0.0001	0.0835	...
SP_TRUMP	ENDORSEMENT	19	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_COOPER	MANNERS	20	16	0.0143	0.0139	0.0139	0.0161	0.0161	0.0245	0.0113	0.0117	...
SP_TRUMP	SUPREME_COURT	21	17	0.0230	0.0062	0.0017	0.0019	0.0154	0.0017	0.0014	0.0014	...
SP_COOPER	WOMAN	22	20	0.0004	0.0003	0.0003	0.0030	0.0027	0.0043	0.0003	0.0003	...
SP_TRUMP	WOMAN	23	20	0.0087	0.0011	0.0011	0.0013	0.0013	0.0011	0.0009	0.0352	...
SP_COOPER	ENDORSEMENT	24	11	0.0410	0.0398	0.0398	0.0460	0.0460	0.0398	0.0323	0.0336	...
SP_TRUMP	WOMAN	25	20	0.0002	0.0000	0.0000	0.0004	0.0000	0.0002	0.0000	0.0000	...
...	...	...	...	...	...	...	...	...	...	...	...	...

B) Preliminary information about the Recurrence plots



Both the 'Recurrence Quantification Analysis (RQA)' and the 'Similarity Measures' tools use the **recurrence plot** technique. That is to say they build a N × N matrix, the rows and columns of which – in our case - are text segments ordered according to their temporal sequence. However in the two cases the recorded information is different. In fact, in the first case (i.e. RQA) any **recurrence** – marked with an unshaded dot - refers to the presence (absence in the case of white spaces) of the same theme in the 'i' and 'j' items (i.e. where the 'X' and 'Y' values are the same) and uses a categorical time series as input; differently, in the second case (i.e. Similarity Measures) any recurrence – marked with a shaded dot - refers to the similarity (i.e. Cosine) concerning the 'i' and 'j' items, the values of which are continuous (i.e. they vary from 0 to 1).

N.B.: In the case of recurrence plots with similarity measures the cut-off limit used by **T-LAB** is 0.0001 (Cosine measure). This because many scholars tend to count all nonzero entries of the similarity matrix.

Though the two types of recurrence plots may highlight similar patterns (see the below Fig. 1 and Fig. 2, which have been obtained by analysing a legislative text), by default **T-LAB** uses the first (i.e. Fig. 1) for computing the RQA measures and it uses the second (i.e. Fig. 2) for exploring similarities and differences concerning text segments.

However, by clicking the appropriate button, the user is also allowed to obtain the RQA measures for the recurrence plots with the similarity measures. Just remember that, as in this case the percentage of recurrent points is higher, all RQA measures are somehow inflated.
The fact remains that, like the 2D barcodes used for marketing purposes, both the below recurrence plots can be seen as unique fingerprints of the analysed text.

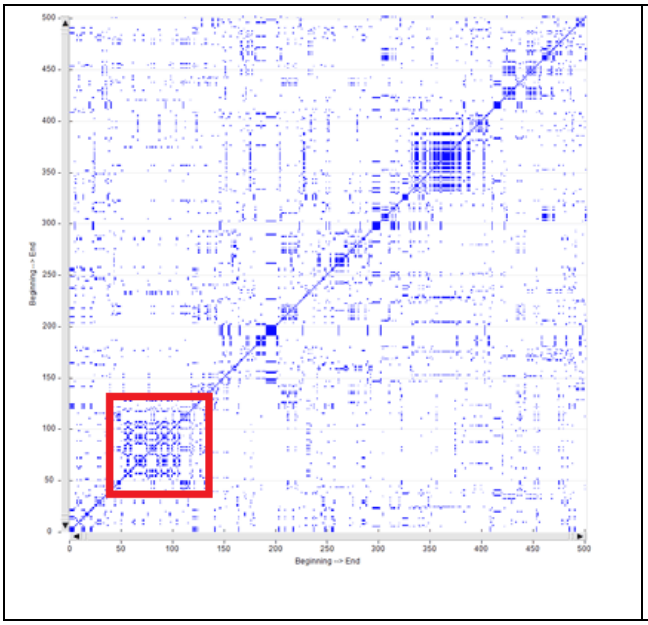


Fig. 1 - Time series

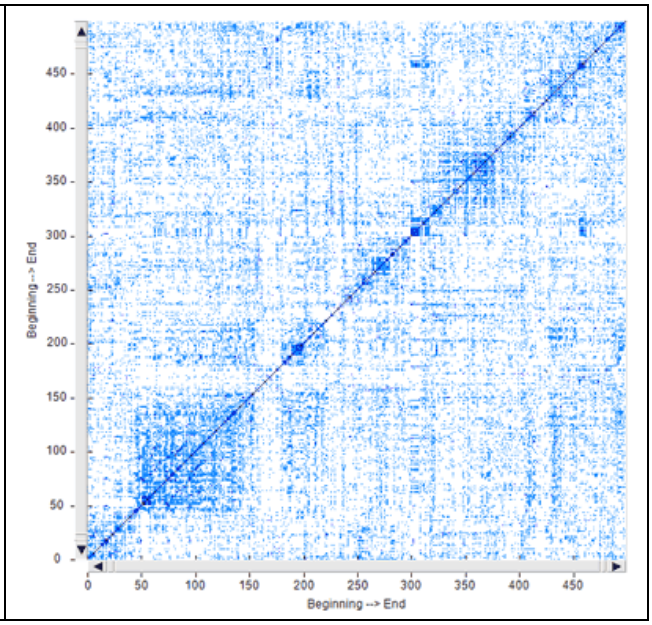
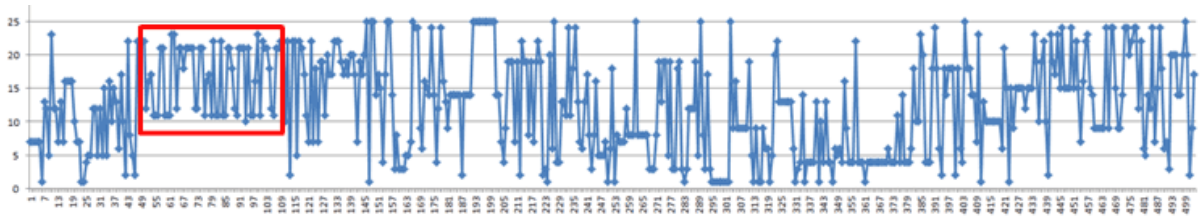
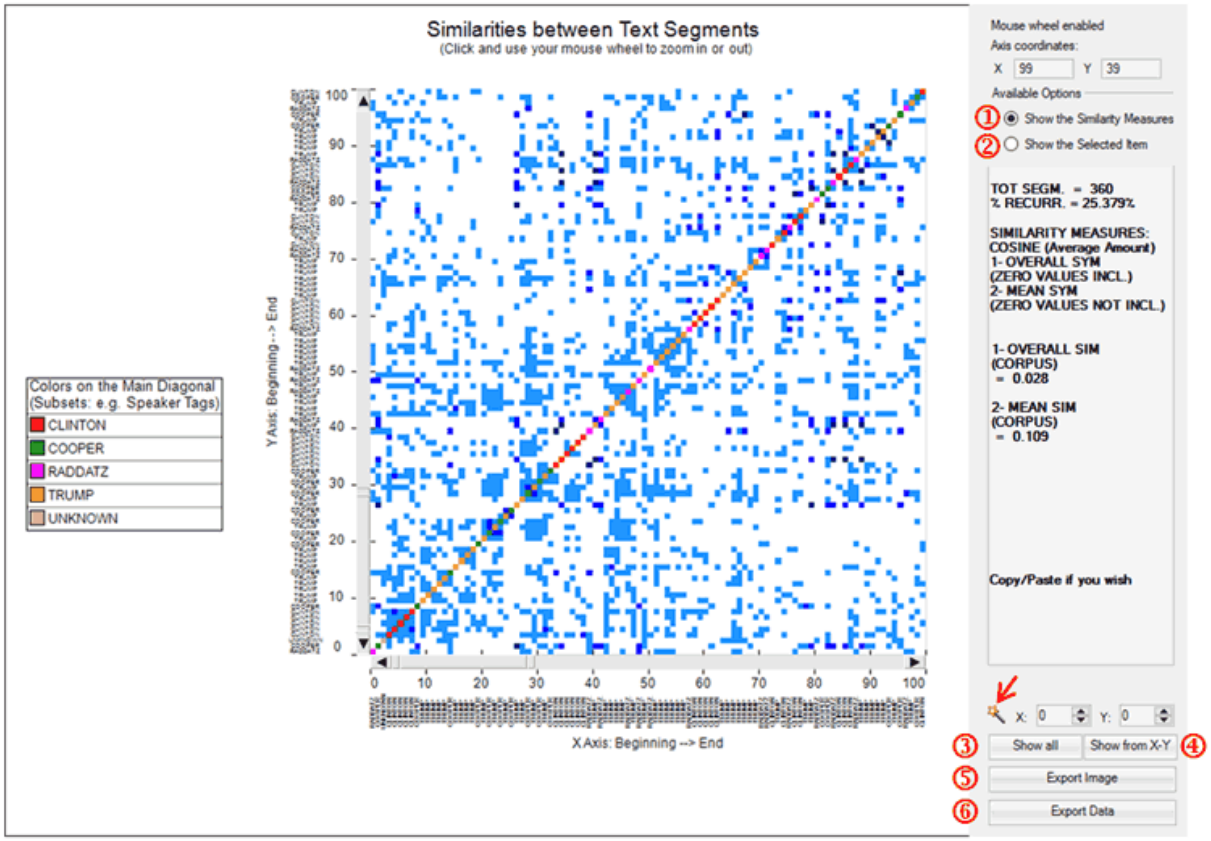


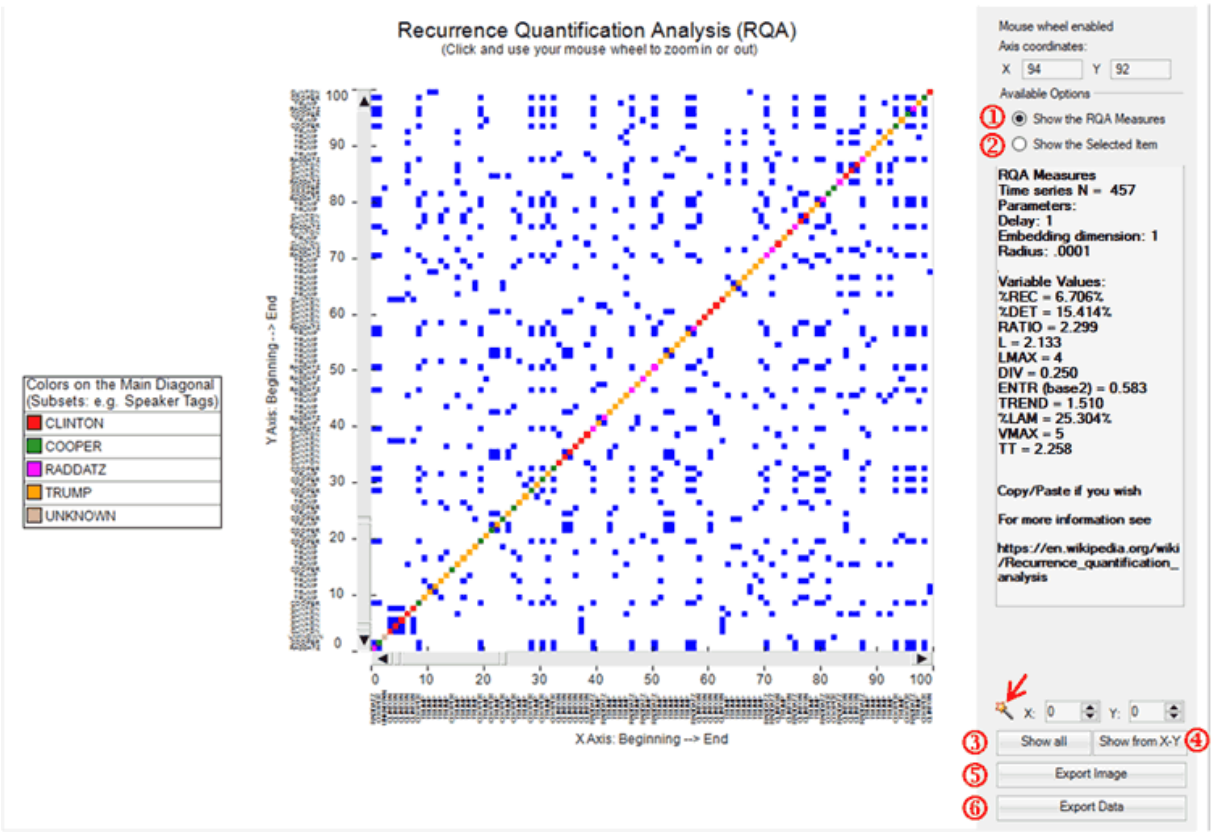
Fig2 - Similarities

N.B. The time series used for the recurrence plot in Fig. 1 is the following:



Both when clicking 'Similarity Measures' and 'Recurrence Quantification Analysis (RQA)' the default **T-LAB** chart shows a 100x100 recurrence plot which however **can be zoomed in and out** by using the mouse wheel. Moreover in both cases **six different options** allow us to perform different operations (see pictures below).



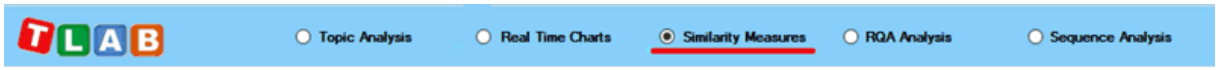


In particular:
-options '1' and '2' allow us to visualize the general measures ('1') or the transcript of the selected segment ('2');
-options '3' and '4' allow us to visualize the complete recurrence plot ('3') or a subsection of it ('4');
-options '5' and '6' allow us to export the image in different formats ('5') or to export a data table with all the analysed values ('6').

Please note:
-in the RQA case the magic wand button (🪄) allows us to check some characteristics which will be explained in the below section 'D'. Differently, in the case of similarities, the same button may be used for obtaining the RQA measures for the shown recurrence plot;
-when exporting the similarity data, all measures concerning 'Self-Similarity' and 'Other-Similarity' are included (see table below).

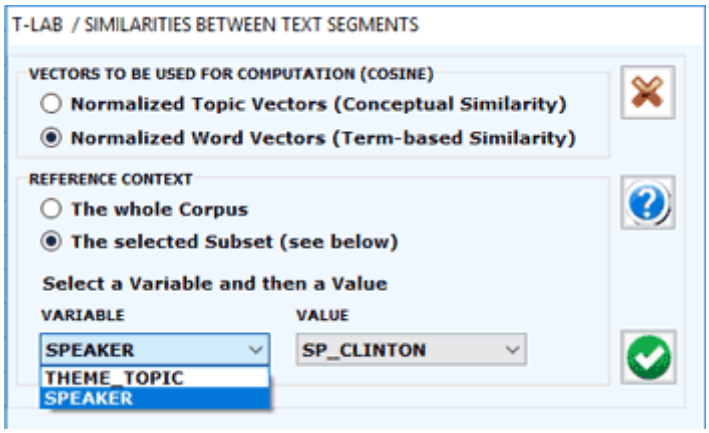
FIRST	SECOND	Cosine
SP_CLINTON	SP_CLINTON	0.0961
SP_CLINTON	SP_COOPER	0.1099
SP_CLINTON	SP_RADDATZ	0.1025
SP_CLINTON	SP_TRUMP	0.0847
SP_CLINTON	SP_UNKNOWN	0.1087
SP_COOPER	SP_CLINTON	0.1099
SP_COOPER	SP_COOPER	0.3106
SP_COOPER	SP_RADDATZ	0.2359
SP_COOPER	SP_TRUMP	0.1432
SP_COOPER	SP_UNKNOWN	0.1446
SP_RADDATZ	SP_CLINTON	0.1025
SP_RADDATZ	SP_COOPER	0.2359
SP_RADDATZ	SP_RADDATZ	0.2121
SP_RADDATZ	SP_TRUMP	0.1103
SP_RADDATZ	SP_UNKNOWN	0.1161
SP_TRUMP	SP_CLINTON	0.0847
SP_TRUMP	SP_COOPER	0.1432
SP_TRUMP	SP_RADDATZ	0.1103
SP_TRUMP	SP_TRUMP	0.1003
SP_TRUMP	SP_UNKNOWN	0.0958
...	...	...

C) Similarity Measures

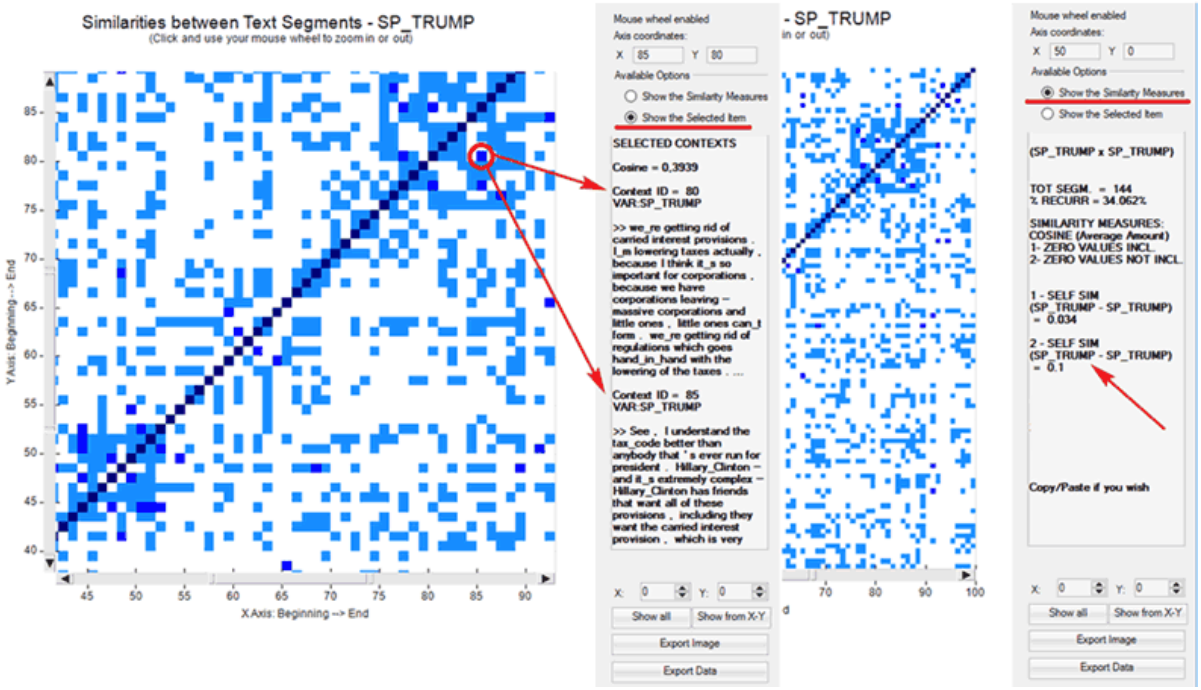


When choosing 'Similarity Measures', several options are made available (see picture below) which allow the user to select both the vectors to be used for the similarity computation and the reference context to be analysed (i.e. either the entire corpus or a subset of it).

N.B.: The difference between 'conceptual' (1) and 'term-based'(2) similarities is that in the first case (1) each text segment is represented by a feature vector concerning topics, whereas in the second case (2) each text segment is represented by a feature vector concerning words. In both cases the similarity measure used is the Cosine coefficient.

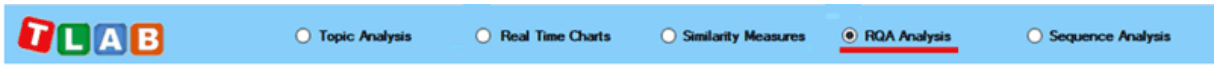


According to the design of the user interface, in this case - like in the RQA analysis (see section 'D' below) - the user can choose between visualizing the global measures or the transcripts of recurrent segments (see picture below). Moreover, when a corpus subset is selected, two further measures are provided concerning the 'self-similarity' (i.e. averaged cosine similarity) between all pairs of text segments within the chosen corpus subset, one (1) with and the other (2) without zero values included. Other measures concerning similarities between all pairs of corpus subsets can be exported by clicking the 'Export Data' button.



Please remember that, unlike the RQA, the 'Similarity Measures' option considers only those text segments in which at least two key-terms included in the user list are present. This is in order to reduce biases in the Cosine computation.

D) Recurrence Quantification Analysis (RQA)



RQA is a method of nonlinear data analysis for the investigation of dynamical systems which quantifies the information contained in a recurrence plot and detects the transitions in the systems by analysing time series (see https://en.wikipedia.org/wiki/Recurrence_quantification_analysis).

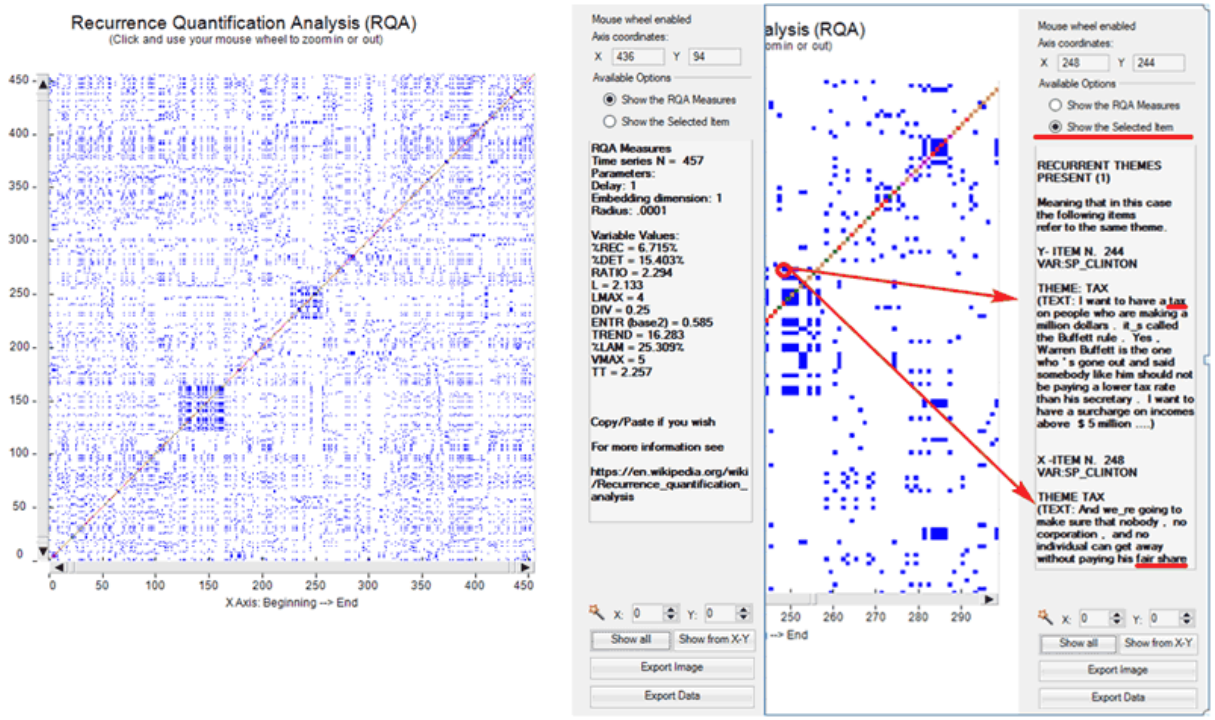
In this **T-LAB** tool, both in the case of the RQA Analysis and in the case of the Sequence Analysis (i.e. Markovian Analysis), a time series is represented by a categorical vector where each element is an integer which corresponds to the topic assigned to the 'i' text segment. However only in the case of the RQA a square matrix is built where the time series is both in rows and in columns.

When using the RQA tool, two main options are made always available (see pictures below):

1-Show the RQA Measures;
2-Show the Selected Item.

In the first case, the **standard measures** of RQA are provided (e.g. %REC, %DET, ENTR etc.**). In the second case the excerpts of recurring text segments are displayed. In both cases, the mouse wheel allows zooming in and out. Moreover two buttons allow the user to export both the picture and the analysed data.

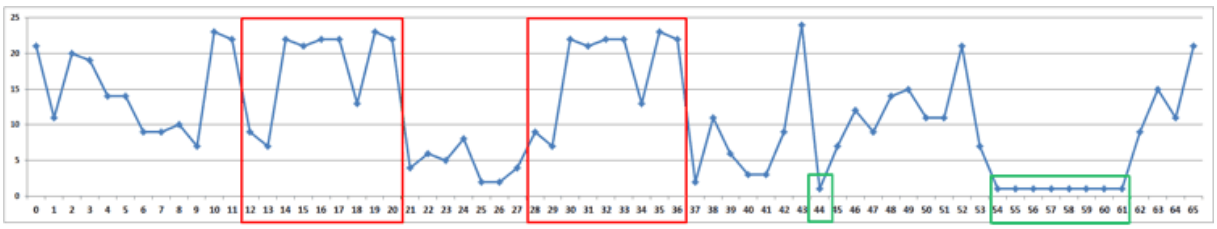
(**) For more information about the RQA measures see section 'E' below.



Please note that in the recurrence plot analysed with RQA the representation is symmetric across the main diagonal and two types of lines are particularly important: the **diagonals** parallel to the main diagonal and the **vertical lines** (**). In fact these lines mark the **transitions** present in the system and they are the base for obtaining the various RQA measures.

(**) In any recurrence plot vertical lines and horizontal lines mirror each other. In fact vertical lines in the upper part of the plot correspond to horizontal lines in the lower part, and vice versa.

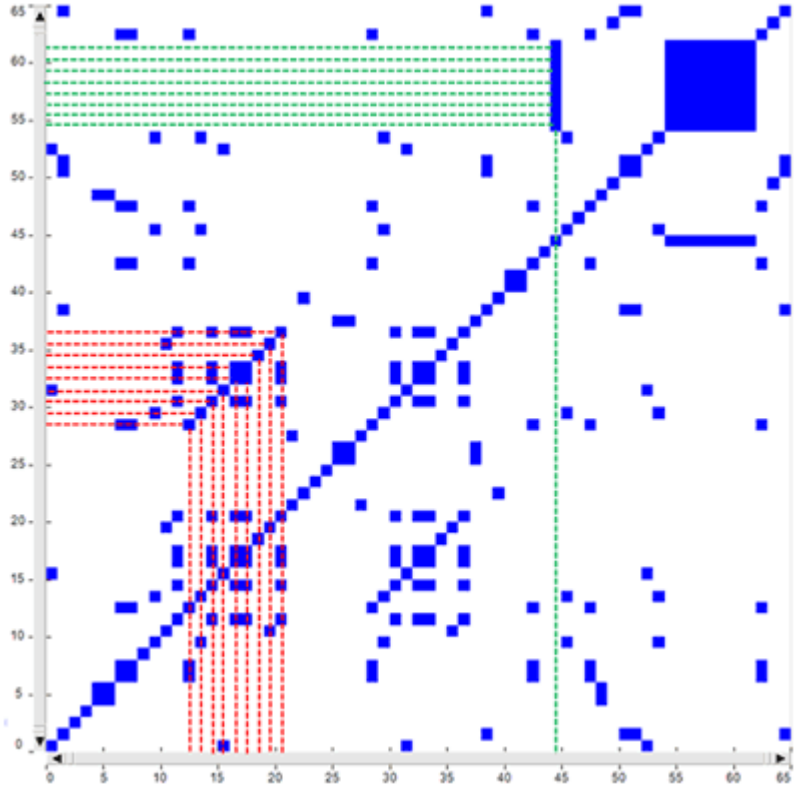
In particular, the distribution of diagonal lines allows for the investigation of **determinism** (i.e. the predictability of the system) and the distribution of vertical lines allows for the investigation of **intermittency** (i.e. the sequences which are interspersed by erratic breaks).



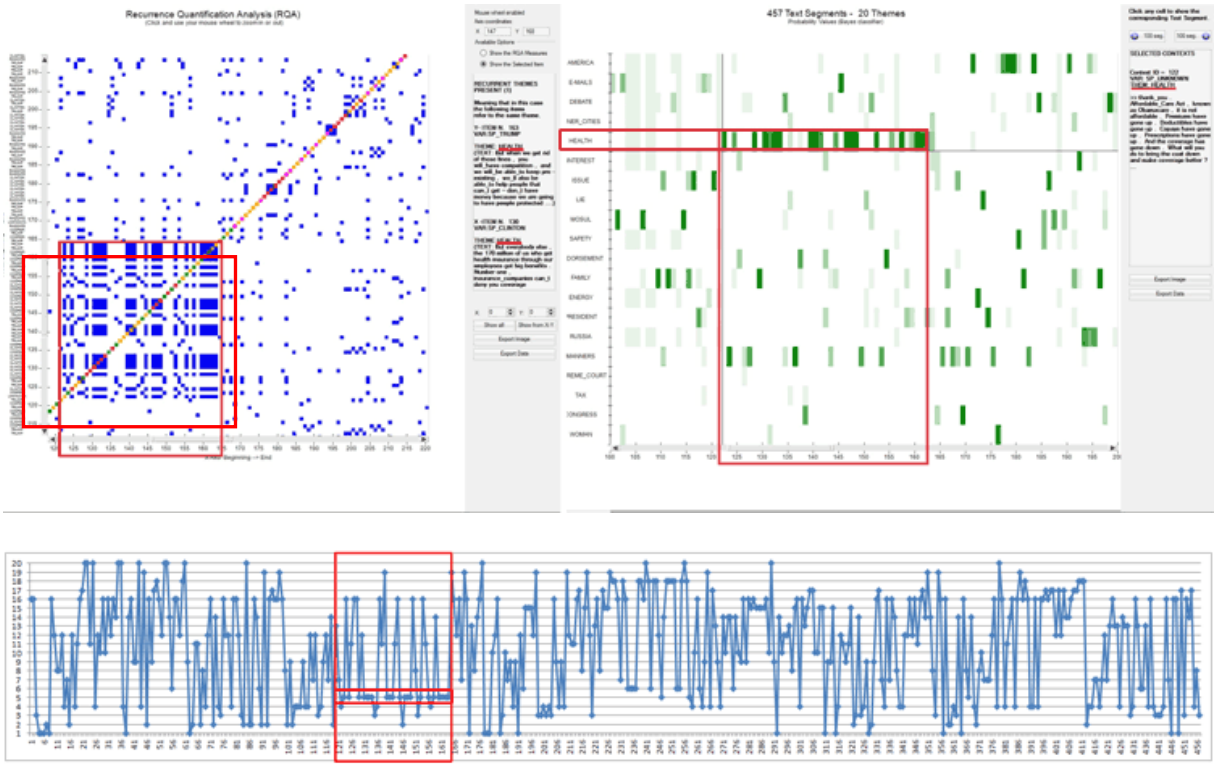
As an example, just consider the above fictitious time series. In it the same sequence of nine points/themes is repeated two times in different time spans (see the above red rectangles), respectively from t-12 to t-20 and from t-28 to t-36, where each 't' stands for a different text segment. In the same series there is also a sequence – from t-54 to t-61 - in which the same theme which appears at t-44 is repeated eight times (see the above green rectangle).

The corresponding recurrence plot (RP) - which has the same time series on the 'X' and the 'Y' axes - is that depicted in the image below.
Please note that in the case of diagonal line each point on the 'X' axis (i.e. from t-12 to t-20) recurs with the corresponding point on the 'Y' axis (i.e. from t-28 to t-36); differently the eight points which form the vertical line recur with just one point (i.e. t-44). Accordingly, in musical terms we may say that diagonal lines refer to a restatement of a motif (i.e. a pattern is repeated), whereas vertical lines refer to a repetition of a single note which somehow breaks the thematic variation.

Please note that when a monothematic sequence like that form t-54 to t-61 is repeated two or more times, usually in the recurrence plot it is represented by a square or by a rectangle.

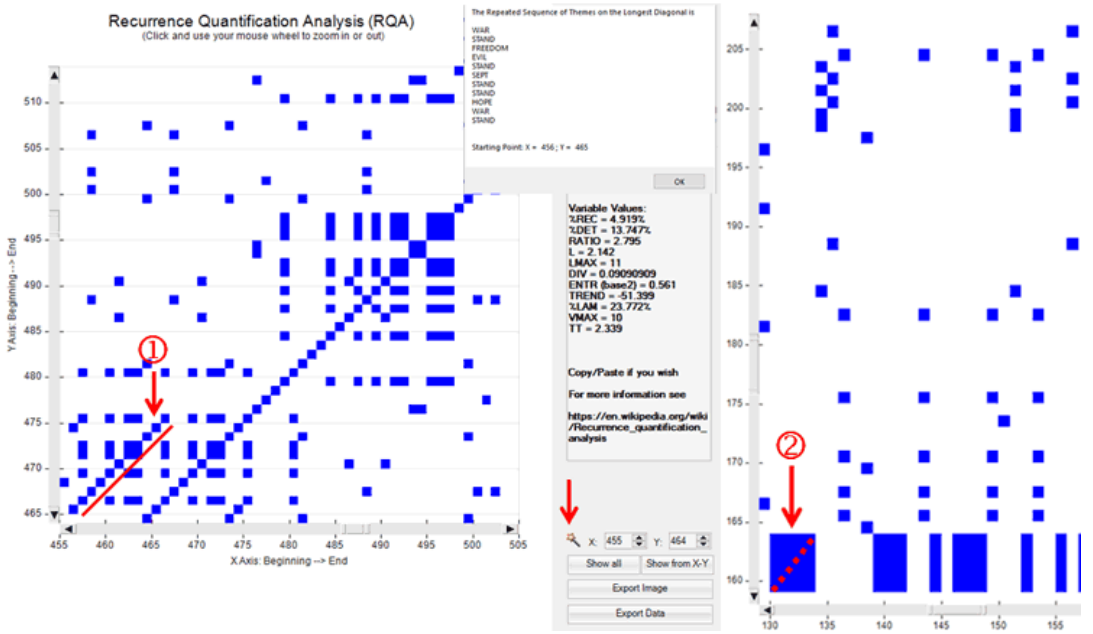


Regarding the **rectangular block** structures – which actually include both vertical and diagonal lines - they can be seen as referring to recurrences of the same topics in sub sections of the time series, i.e. to groups of overall similar feature vectors. In fact each dot in the graph represents a revisit of the same state and there is a correspondence between the rectangular blocks of the recurrence plot, the rectangles highlighted in the real time heat map and the chart of the time series (see pictures below). In other words we may say that in this cases speakers are repeatedly engaged on the same topic/theme, which appears to be 'hot'.

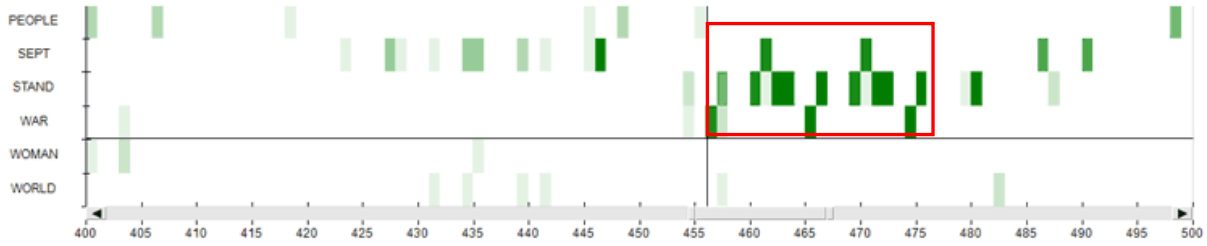


As stated above, in the RQA outputs the **longest diagonals** parallel to the main diagonal allow us to detect interesting repetitions of the same thematic sequence. However their shapes are not so evident as the rectangular block structures, also because sometimes they can be hidden inside one of them (see the below case marked with '2'). For this reason T-LAB includes a specific option (see the magic wand below) which automatically detects the longest diagonal, informs the user about the sequence of repeated themes included in it and automatically positions the cursor in the corresponding X-Y coordinates.

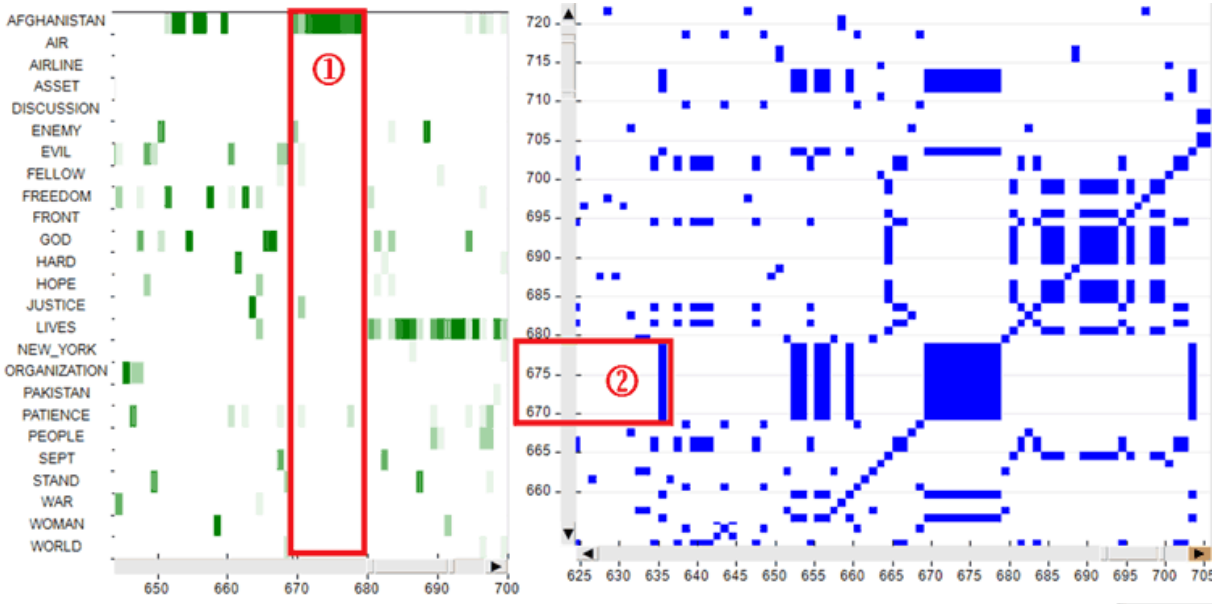
N.B.: Soon after the longest diagonal is detected **T-LAB** allows the user to export a file with the most frequent **repeated sequences**, each one of them including at least three concatenated themes. Such a file can be considered a sort of summary of the main themes - and of the corresponding variations - present in the corpus.



N.B.: In the case of the above diagonal '1', one of the corresponding patterns on the heat map is the following.



Regarding the vertical/horizontal lines they can be easily checked by exploring the heat map first (see case '1' in the image below) and then the recurrence plot (see case '2' in the image below).



E) Some notes about the RQA measures

When talking about the RQA measures, we have to make a clear distinction between their technical definitions (1) and their relevance in a thematic text analysis (2).
In fact the technical definitions correspond to formulas and are the same in all sciences using RQA for the study of dynamic systems and their time series (e.g. physics, physiology, meteorology, finance, etc.). Differently, the relevance – and also the meaning – of the RQA measures in text analysis is a matter of debate.

Starting with the technical definitions (1), here is a table which summarizes the relevant information for the most used RQA measures.

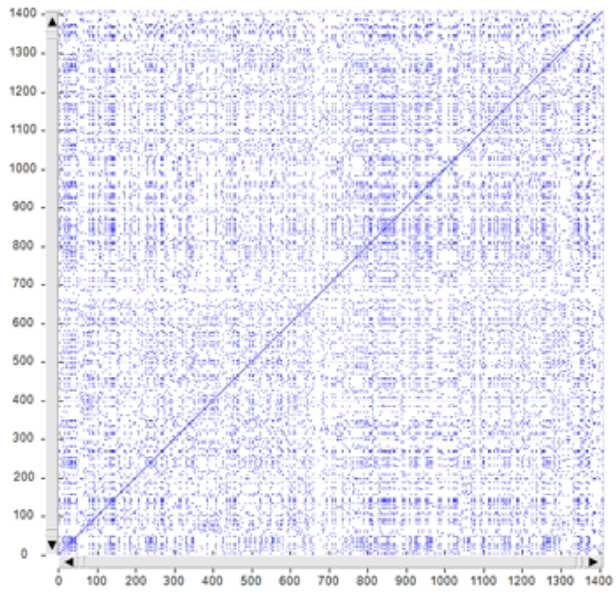
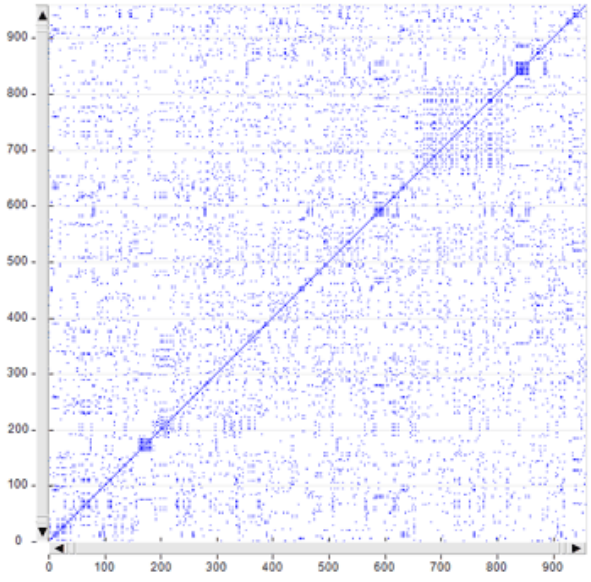
Measure	**Definition**
%REC - Recurrence Rate	The percentage of recurrence points in a Recurrence Plot which fall within a specified radius.
%DET - Determinism	The percentage of recurrence points which form diagonal line structures, main diagonal not included (N.B.: In RQA the main diagonal is also called LOI, i.e. Line of Identity, because in it each point recurs with itself).
RATIO	The ratio between %DET and %REC.
L	The average length of the diagonal lines.
LMAX	The length of the longest diagonal line.
DIV - Divergence	The inverse of LMAX.
ENTR - Entropy	The Shannon entropy of all diagonal line lengths distributed over integer bins in a histogram (Webber, C. L., & Zbilut, J. P. , 2005, p. 48). Accordingly, if there are lots of diagonal lines with varying lengths, the entropy will be high. Please note that, as in the RQA case entropy reflects the complexity of the RP in respect of the diagonal lines, here the definition of entropy does not correspond to the entropy of physical systems, where the higher the entropy the greater the disorder.
TREND	The degree of system stationarity . Accordingly, when recurrent points are homogeneously distributed across the recurrence plot, TREND value will be close to zero. Differently, when points 'fade away' from the central diagonal, the trend will have a negative value.
%LAM - Laminarity	The percentage of recurrence points which form vertical lines.
VMAX	The length of the longest vertical line.
TT – Trapping time	The average length of the vertical lines.

Regarding the relevance of RQA measures in text analysis (2) both **%DET** and **TREND** deserve special attention. In fact higher determinism (%DET) values indicates that the same thematic patterns are repeated more often and that – accordingly – the dynamic of analysed system is somehow more predictable. On the other hand TREND can be interpreted as a measure referring to how quick the transitions are from some themes to others, where lower TREND values indicate quicker transitions.

For example, when comparing RQA measures obtained by analysing a scientific essay ('a') and a novel ('b'), we can find out that in the first case ('a') the %DET value is higher than 'b' and that in the second case ('b') the TREND value is very low (often below zero).
Below is a comparison of the RQA measures obtained by analysing the essay 'On the Origin of Species' (C. Darwin) and the novel 'The adventures of Pinocchio' (C. Collodi).

On the Origin of Species (a)	The adventures of Pinocchio (b)
%REC = 8.201%	**%REC = 3.525%**
%DET = 16.474%	**%DET = 9.676%**
RATIO = 2.009	RATIO = 2.745
L = 2.093	L = 2.089
LMAX = 6	LMAX = 5
DIV = 0.167	DIV = 0.2
ENTR (base2) = 0.460	ENTR (base2) = 0.435
TREND = 4.705	**TREND = -5.599**
%LAM = 30.717%	%LAM = 23.194%
VMAX = 7	VMAX = 6
TT = 2.263	TT = 2.267

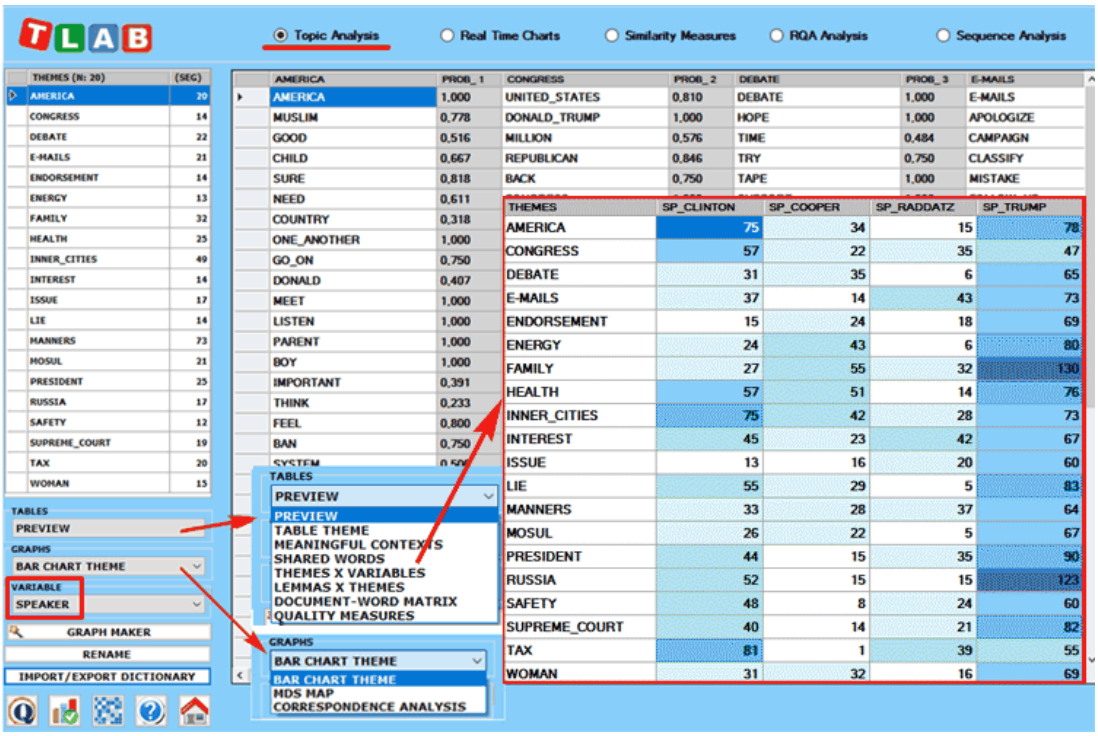
Here are the two corresponding recurrence plots.

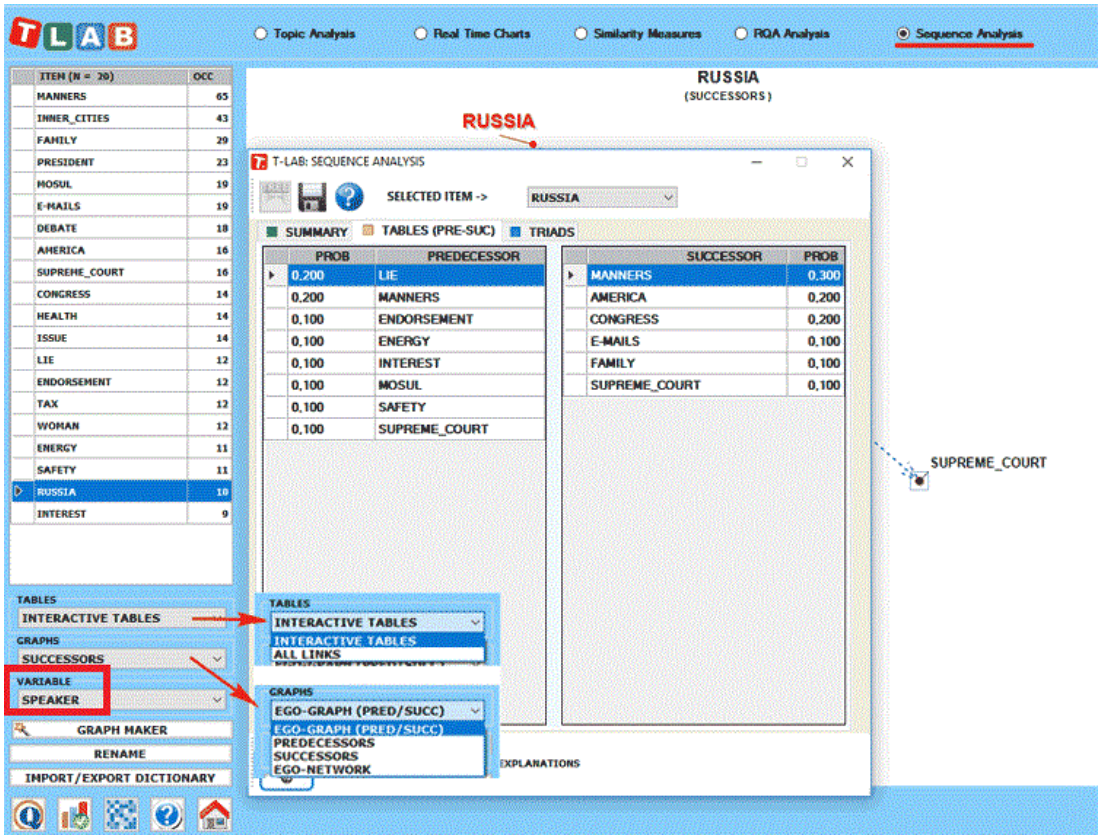
On the Origin of Species (a)	The adventures of Pinocchio (b)
	

N.B.: A table which summarizes the meanings of typical patterns in recurrence plots can be found at page 251 of the following article:
N. Marwan, M. Romano, M. Thiel and J. Kurths, "Recurrence Plots for the Analysis of Complex Systems", Phys. Rep. 438, 240-329 (2007).

F) Topic Analysis and Sequence Analysis

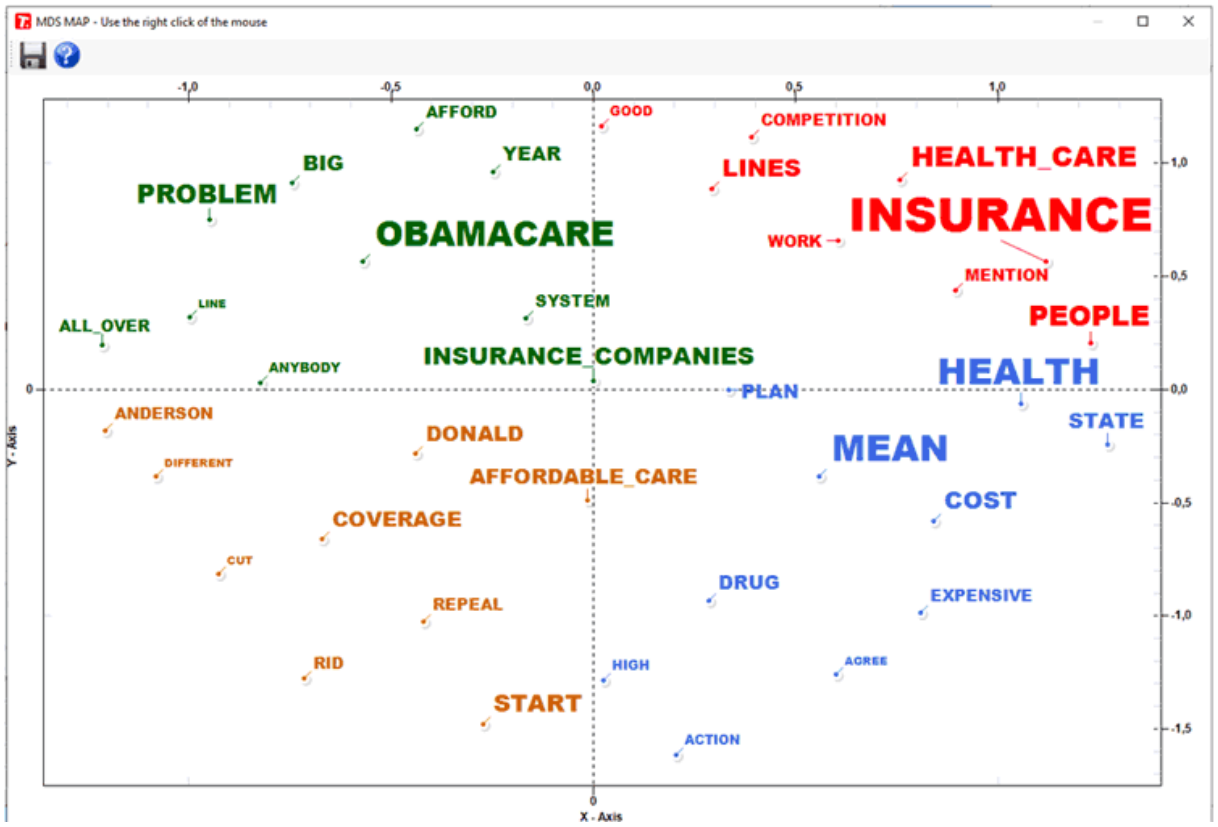
The below pictures summarize the main options of two tools already present in the **T-LAB** menu, which are integrated with the new ones and which are explained in the corresponding sections of this manual/help, i.e. 'Modeling of Emerging Themes' and 'Sequence and Network Analysis'.





N.B.:
-Any variable selected in the above forms (see the label highlighted by a red rectangle) will be used in the outputs provided by the various tools (Please note that only categorical variables with up to 20 values are made available) ;
-The 'Export/Import Dictionary' option, which is no longer available after performing a Sequence Analysis, is intended to allow the user to save time when repeating the same analysis by using topic labels manually assigned previously. In other words: just export the topic dictionary after completing - if desired - all renaming operations and import the same dictionary when repeating the same analysis with the same corpus, the same key-word list and the same parameters;
-While the Correspondence Analysis option allows us to explore the relationships between the various topics and the various speakers, the 'Graph Maker' tool allows us to explore the relationships between key-terms within each selected topic (see pictures below).

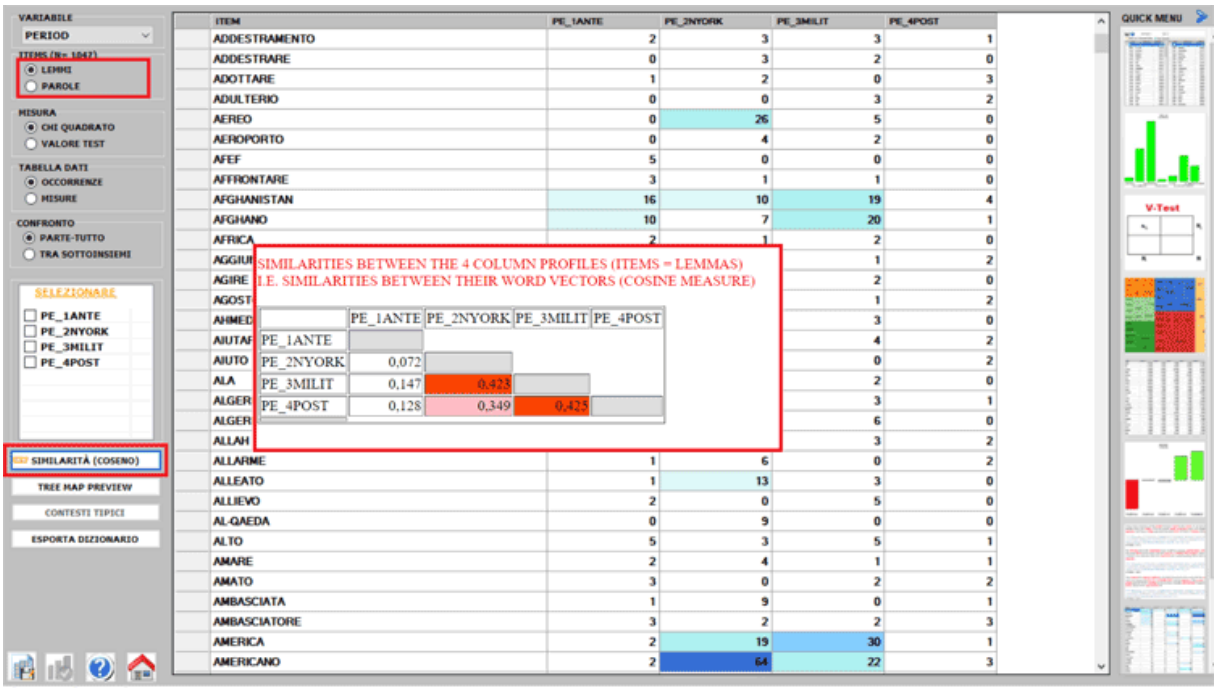

Topic Analysis
Real Time Charts
Similarity Measures
RQA Analysis
Sequence Analysis
GRAPH MAKER - CO-OCCURRENCES WITHIN THE CLUSTER N. <HEALTH>
ADD/REMOVE ITEMS TO BE USED
CLICK A PICTURE TO DISPLAY THE GRAPH
EXPORT DATA FILES FOR NETWORK ANALYSIS (up to 5.000 items can be selected at a time)
MDS

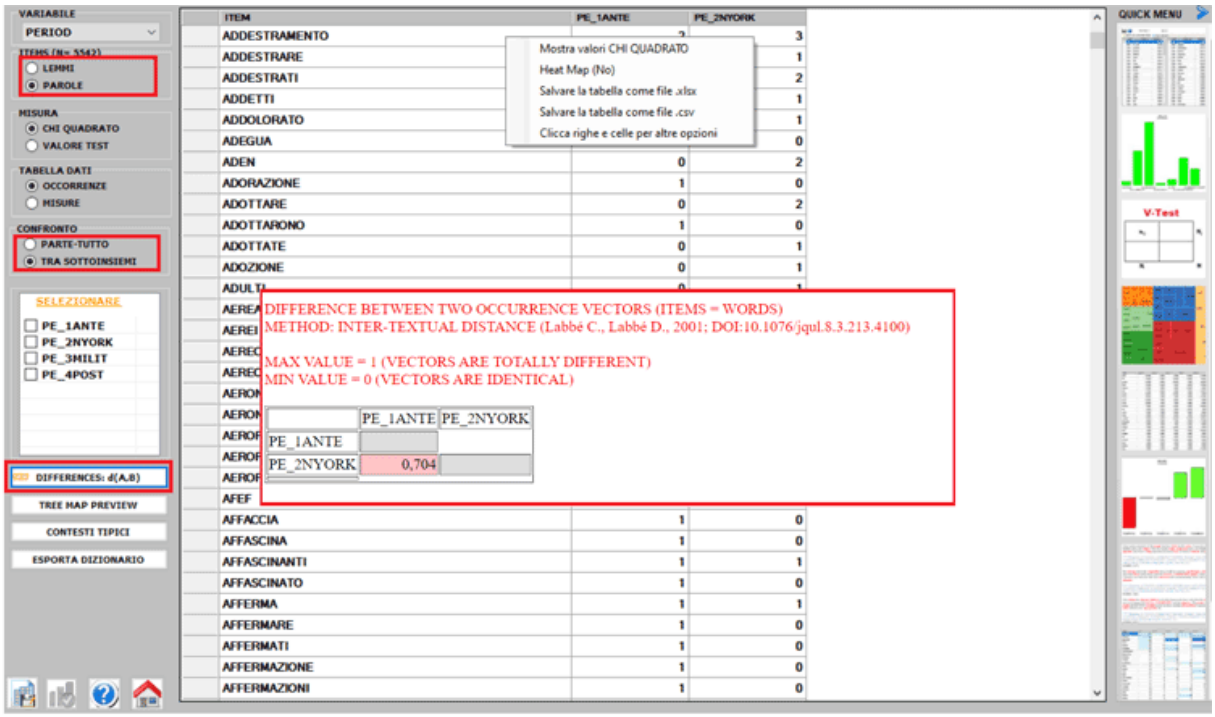

MDS MAP - Use the right click of the mouse
AFFORD
GOOD
COMPETITION
BIG
YEAR
LINES
HEALTH_CARE
PROBLEM
INSURANCE
OBAMACARE
WORK
MENTION
LINE
SYSTEM
PEOPLE
ALL_OVER
INSURANCE_COMPANIES
ANYBODY
HEALTH
PLAN
ANDERSON
STATE
DONALD
MEAN
DIFFERENT
AFFORDABLE_CARE
COST
COVERAGE
CUT
DRUG
EXPENSIVE
REPEAL
RID
HIGH
AGREE
START
ACTION
X - Axis
Y - Axis

ANALISI COMPARATIVE

Specificità

N.B.: Le immagini di questa sezione fanno riferimento a una versione precedente di T-LAB. In **T-LAB 10** l'aspetto è leggermente diverso. In particolare, a partire dalla versione 2021, una galleria di immagini ad accesso rapido funziona come un menu aggiuntivo che permette di passare da un output all'altro con un solo clic. Inoltre l'utente può ora valutare facilmente somiglianze (es.Coseno) e differenze (es.Distanza intertestuale) tra sottoinsiemi di corpus (da 2 a 150), e quindi anche per rilevare la presenza di documenti (vedi immagini seguenti).





Questo strumento **T-LAB** permette di verificare quali unità lessicali (cioè parole, lemmi o categorie) sono **tipici** o **esclusivi** in un testo o in un **sottoinsieme** del corpus definito da una variabile categoriale; inoltre esso consente di individuare le **unità di contesto caratteristiche** dei vari sottoinsiemi in esame (ad esempio le frasi 'tipiche' che meglio differenziano i discorsi dei vari leader politici).

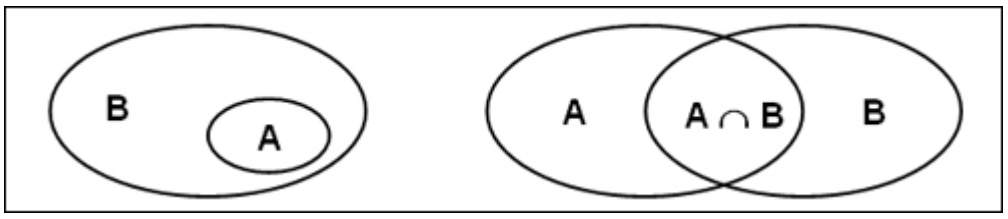
Le **tipiche unità lessicali**, definite dalla proporzione delle rispettive occorrenze (vale a dire dal loro sovra/sotto utilizzo), sono individuate tramite il calcolo **chi-quadrato** o del **valore test**.

Le **unità di contesto caratteristiche** vengono individuate calcolando e sommando i valori **TF-IDF normalizzati** assegnati alle parole di cui ogni frase o paragrafo è costituito.

Per mezzo di questo strumento, analizzando i profili delle occorrenze corrispondenti a righe e colonne delle tabelle di contingenza, è possibile effettuare due tipi di confronti concernenti i profili delle occorrenze:

1- tra una **parte** (es. il sottoinsieme "A") e il **tutto** (es. l'intero corpus in analisi, "B");

2- tra coppie di **sottoinsiemi** del corpus ("A" e "B").



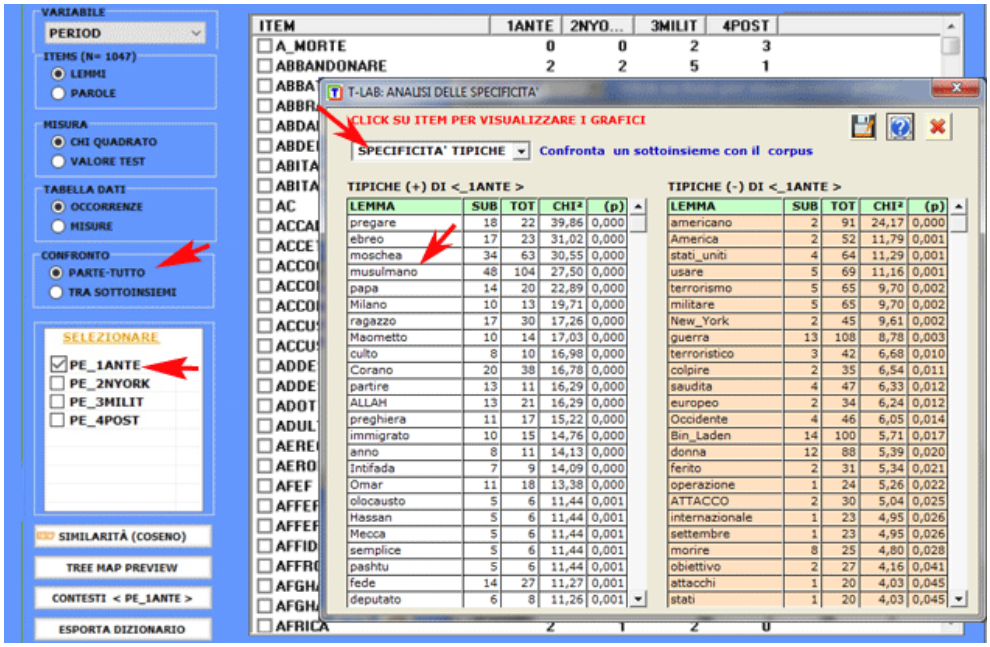
In entrambi i casi possono essere analizzate sia le Specificità relative alle **intersezioni** (parole "tipiche") sia quelle relative alle **differenze** (parole "esclusive").

Le modalità del calcolo sono illustrate nella corrispondente voce del **glossario**.

Le unità lessicali considerate possono essere tutte (impostazioni automatiche) o solo quelle selezionate dall'utilizzatore (impostazioni personalizzate).

Di seguito vengono illustrati i quattro tipi di confronti possibili.

1.1 - **parte/tutto**: unità lessicali "tipiche"
1.2



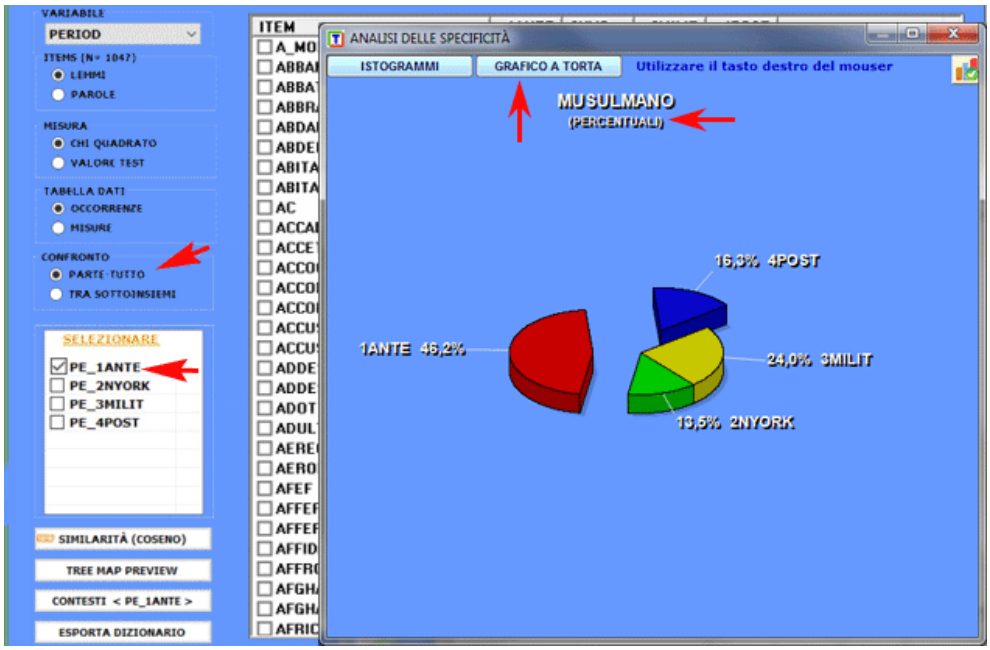
LEMMA	SUB	TOT	CHI²	(p)
pregare	18	22	39,86	0,000
ebreo	17	23	31,02	0,000
moschea	34	63	30,55	0,000
musulmano	48	104	27,50	0,000
papa	14	20	22,89	0,000
Milano	10	13	19,71	0,000
ragazzo	17	30	17,26	0,000
Maometto	10	14	17,03	0,000
culto	8	10	16,98	0,000
Corano	20	38	16,78	0,000
partire	13	11	16,29	0,000
ALLAH	13	21	16,29	0,000
preghiera	11	17	15,22	0,000
immigrato	10	15	14,76	0,000
anno	8	11	14,13	0,000
Intifada	7	9	14,09	0,000
Omar	11	18	13,38	0,000
olocausto	5	6	11,44	0,001
Hassan	5	6	11,44	0,001
Mecca	5	6	11,44	0,001
semplice	5	6	11,44	0,001
pashtu	5	6	11,44	0,001
fede	14	27	11,27	0,001
deputato	6	8	11,26	0,001

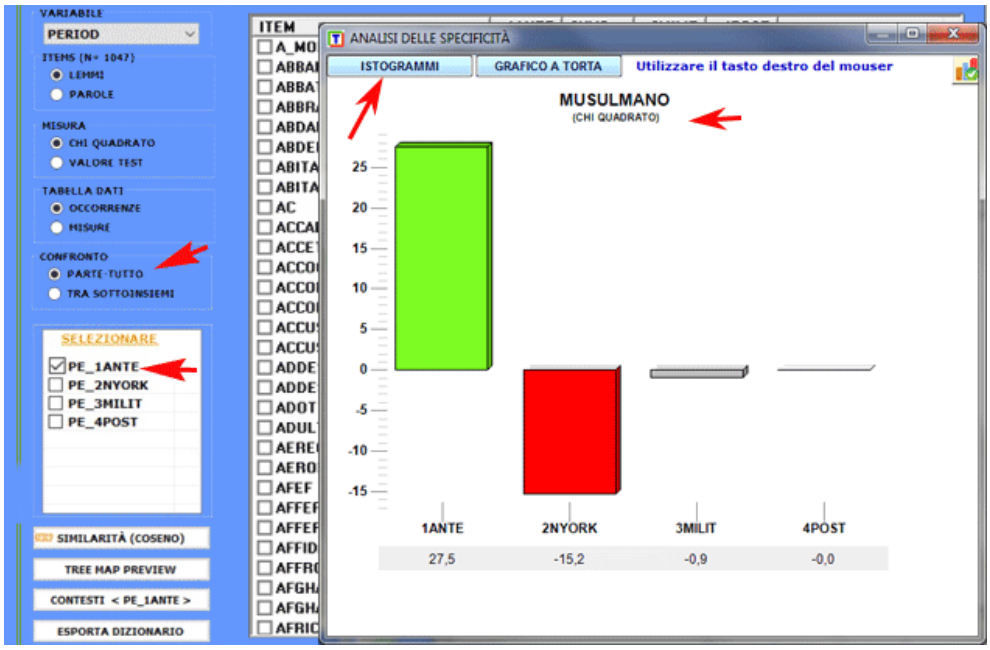
LEMMA	SUB	TOT	CHI²	(p)
americano	2	91	24,17	0,000
America	2	52	11,79	0,001
stati_uniti	4	64	11,29	0,001
usare	5	69	11,16	0,001
terrorismo	5	65	9,70	0,002
militare	5	65	9,70	0,002
New_York	2	45	9,61	0,002
guerra	13	108	8,78	0,003
terroristico	3	42	6,68	0,010
colpire	2	35	6,54	0,011
saudita	4	47	6,33	0,012
europeo	2	34	6,24	0,012
Occidente	4	46	6,05	0,014
Bin_Laden	14	100	5,71	0,017
donna	12	88	5,39	0,020
ferito	2	31	5,34	0,021
operazione	1	24	5,26	0,022
ATTACCO	2	30	5,04	0,025
internazionale	1	23	4,95	0,026
settembre	1	23	4,95	0,026
morire	8	25	4,80	0,028
obiettivo	2	27	4,16	0,041
attacchi	1	20	4,03	0,045
stati	1	20	4,03	0,045

Le chiavi di lettura della tabella sono le seguenti:

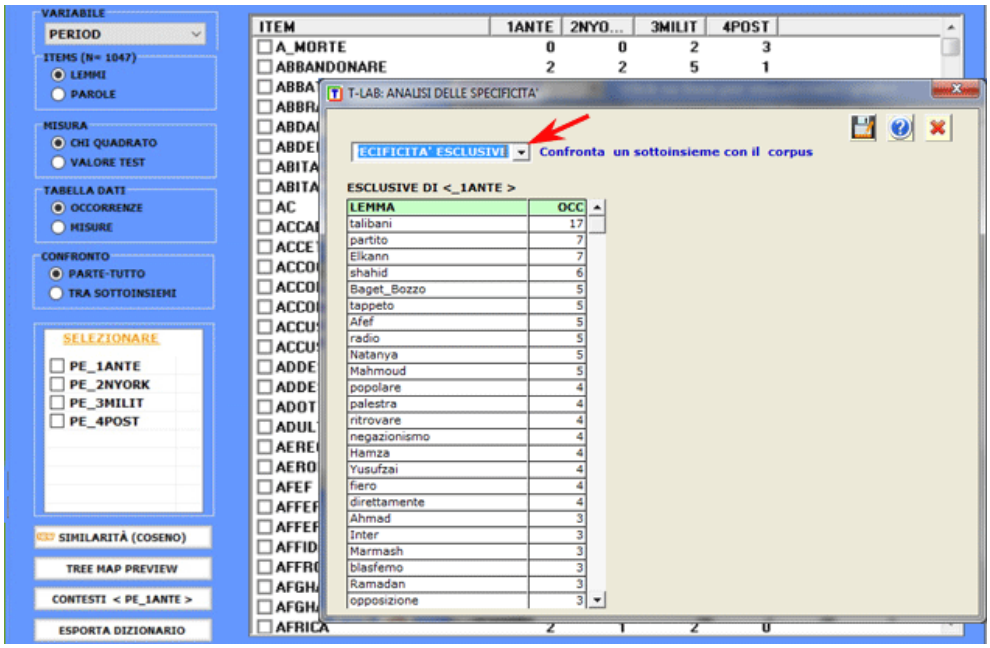
- LEMMA = unità lessicali "tipiche" (per eccesso o per difetto);
- SUB = occorrenze di ogni LEMMA nel sottoinsieme considerato;
- TOT = occorrenze di ogni LEMMA nel corpus o nei due sottoinsiemi confrontati (vedi punto 2.1);
- CHI2 = valori del CHI quadro (o VTEST = Valore Test);
- (p) = probabilità associata al valore del CHI quadro.

Un click sugli item delle tabelle consente di visualizzare vari tipi di grafici (vedi sotto).

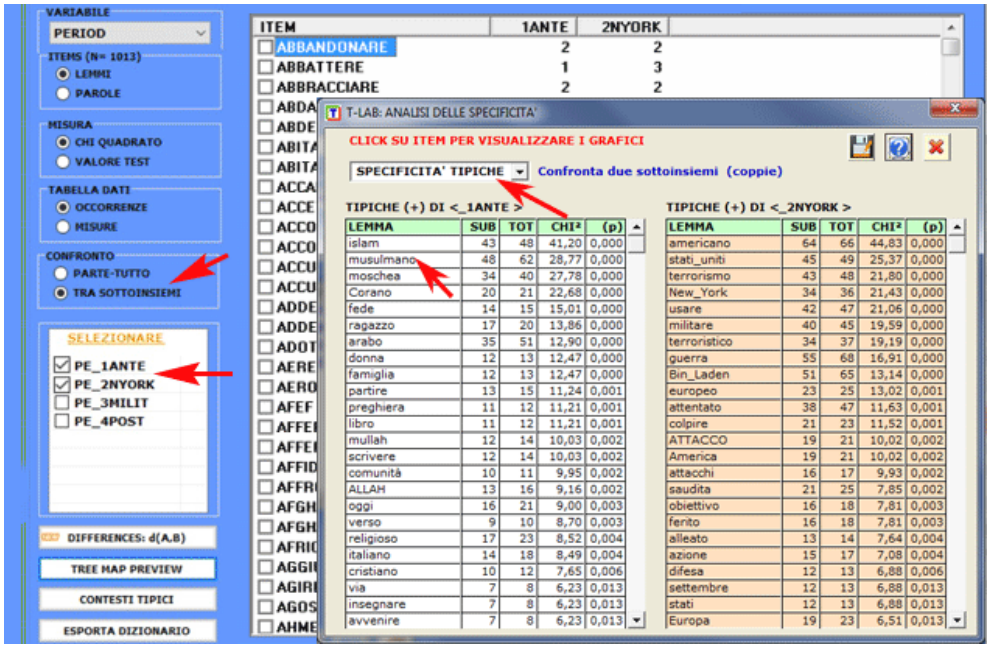

VARIABILE
PERIOD
ITEMS (N= 1047)
LEMMI
PAROLE
MISURA
CHI QUADRATO
VALORE TEST
TABELLA DATI
OCCORRENZE
MISURE
CONFRONTO
PARTE-TUTTO
TRA SOTTOINSIEMI
SELEZIONARE
PE_1ANTE
PE_2NYORK
PE_3MILIT
PE_4POST
SIMILARITÀ (COSENO)
TREE MAP PREVIEW
CONTESTI < PE_1ANTE >
ESPORTA DIZIONARIO
ITEM
ANALISI DELLE SPECIFICITÀ
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouser
MUSULMANO
(PERCENTUALI)
16,3% 4POST
1ANTE 46,2%
24,0% 3MILIT
13,5% 2NYORK

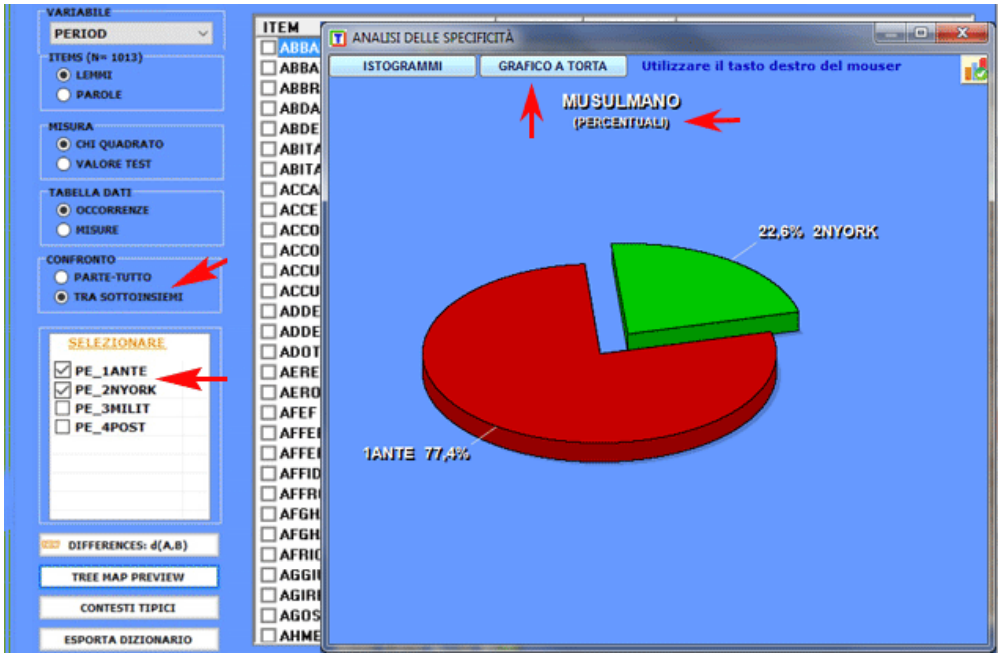

VARIABILE
PERIOD
ITEMS (N= 1047)
LEMMI
PAROLE
MISURA
CHI QUADRATO
VALORE TEST
TABELLA DATI
OCCORRENZE
MISURE
CONFRONTO
PARTE-TUTTO
TRA SOTTOINSIEMI
SELEZIONARE
PE_1ANTE
PE_2NYORK
PE_3MILIT
PE_4POST
SIMILARITÀ (COSENO)
TREE MAP PREVIEW
CONTESTI < PE_1ANTE >
ESPORTA DIZIONARIO
ITEM
ANALISI DELLE SPECIFICITÀ
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouser
MUSULMANO
(CHI QUADRATO)
25
20
15
10
5
0
-5
-10
-15
1ANTE
2NYORK
3MILIT
4POST
27,5
-15,2
-0,9
-0,0

1.2 - **parte/tutto**: unità lessicali "esclusive"

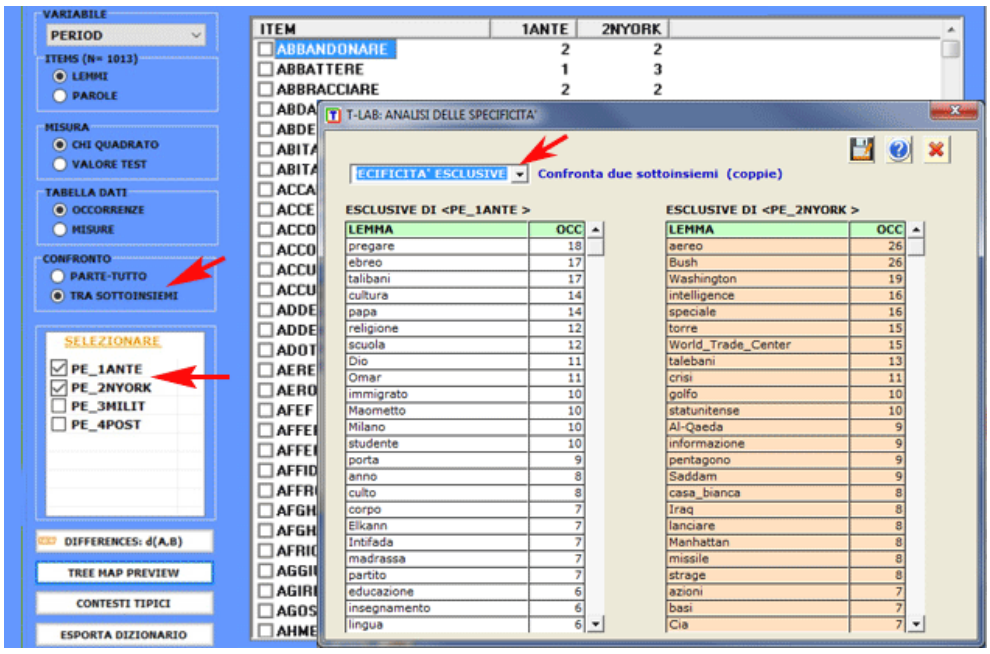


2.1- **sottoinsieme/sottoinsieme**: unità lessicali "tipiche"

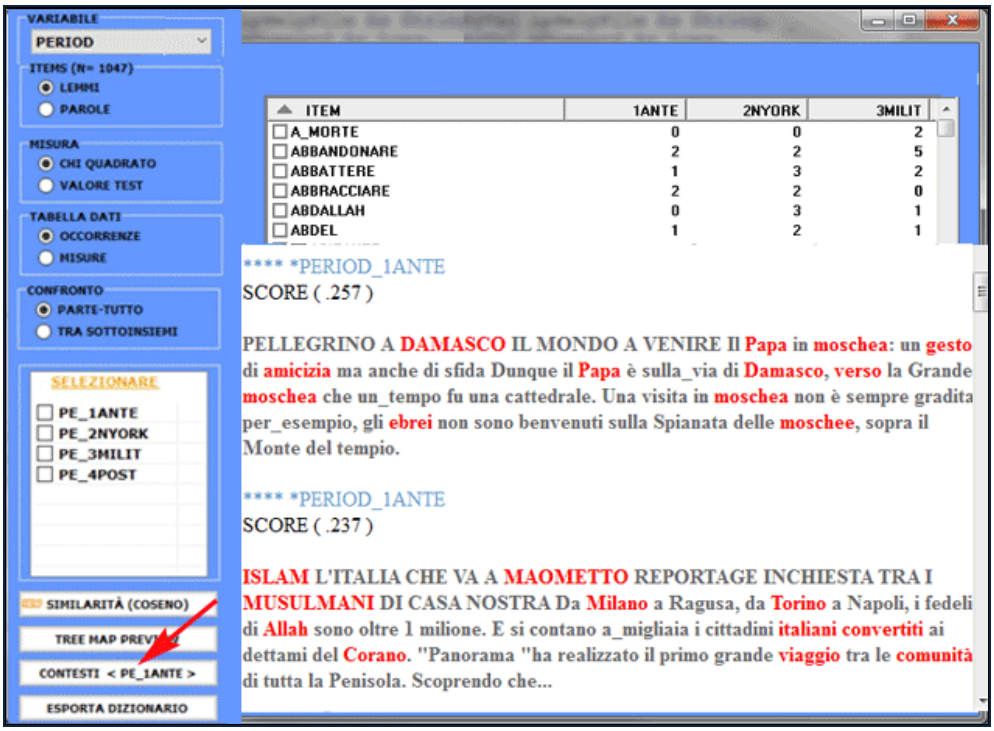




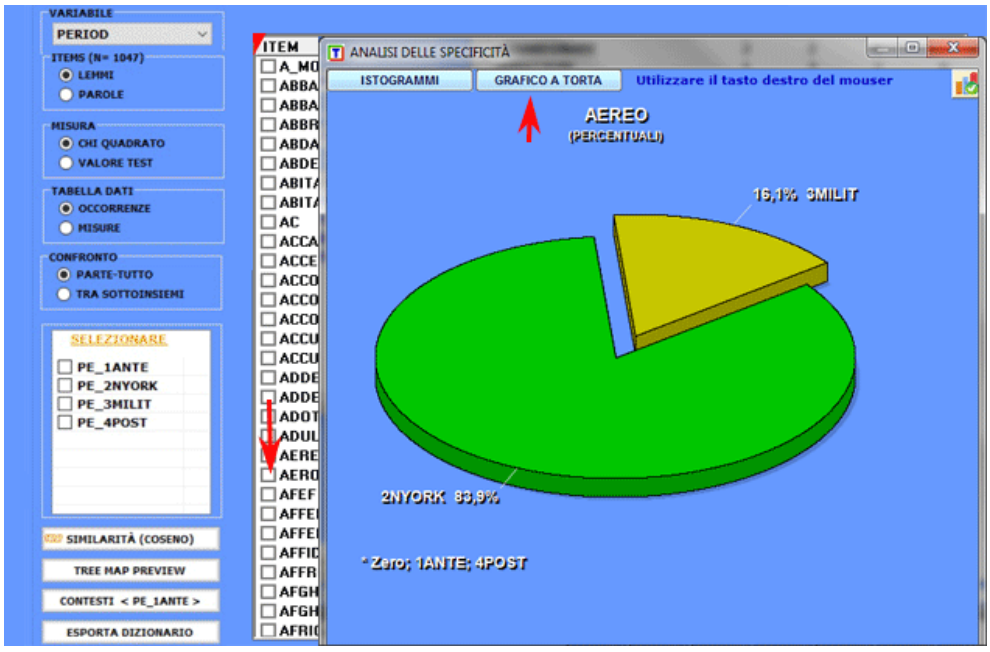
2.2 - **sottoinsieme/sottoinsieme**: unità lessicali "esclusive"

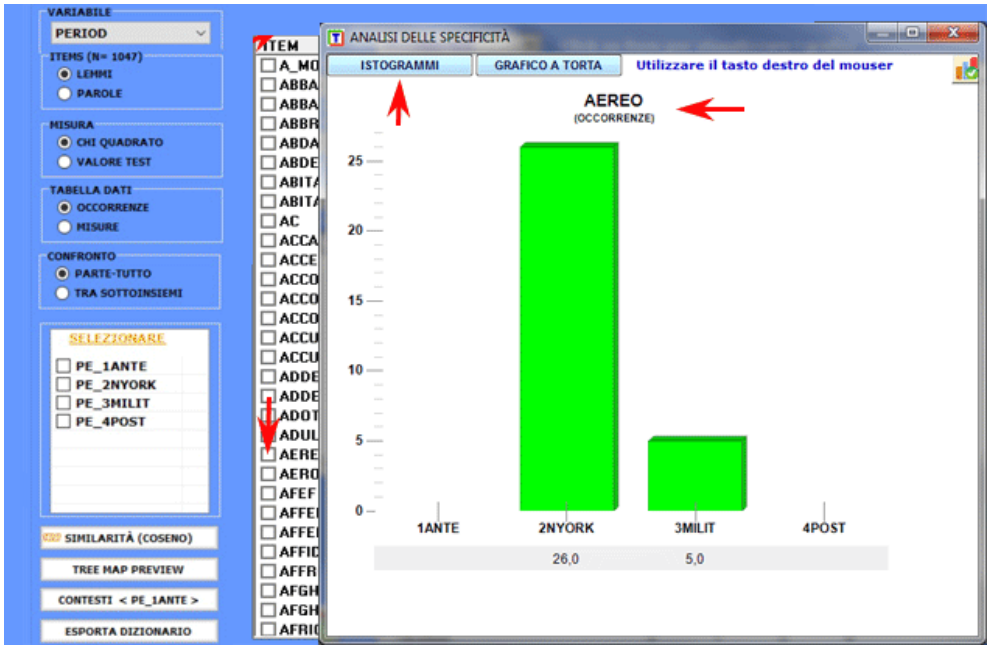


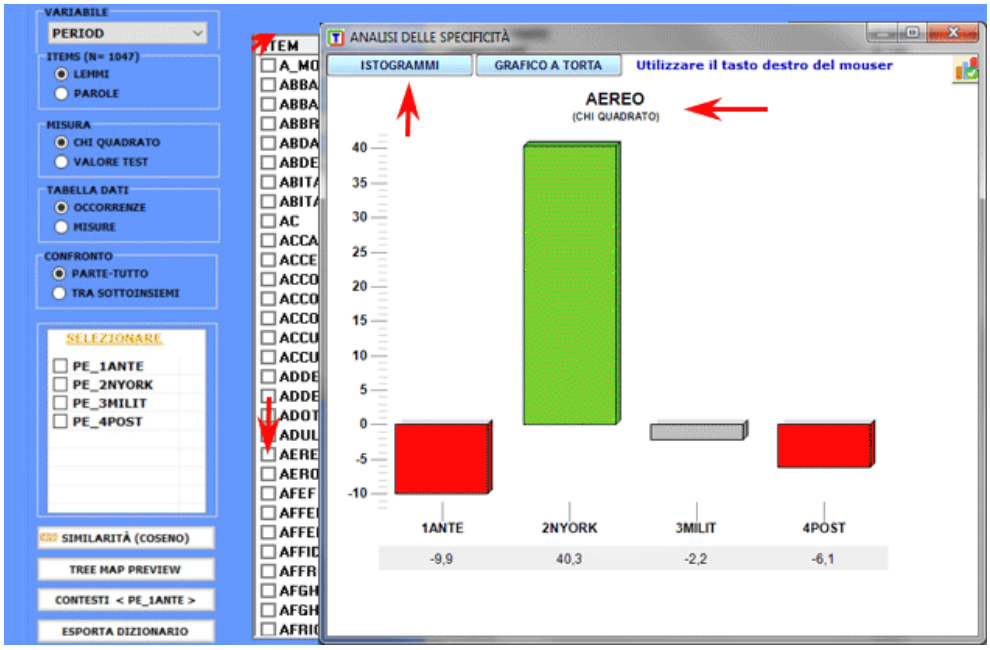
Per ogni sottoinsieme analizzato è anche possibile verificare i contesti elementari (cioè frasi o paragrafi) che meglio lo distinguono dagli altri. In questo caso, la 'specificità' risulta dal calcolo di **valori TF-IDF normalizzati**; più in particolare, lo 'score' assegnato a ciascun contesto elementare (vedi immagine seguente) risulta dalla somma dei valori TF-IDF assegnati alle parole che lo compongono.

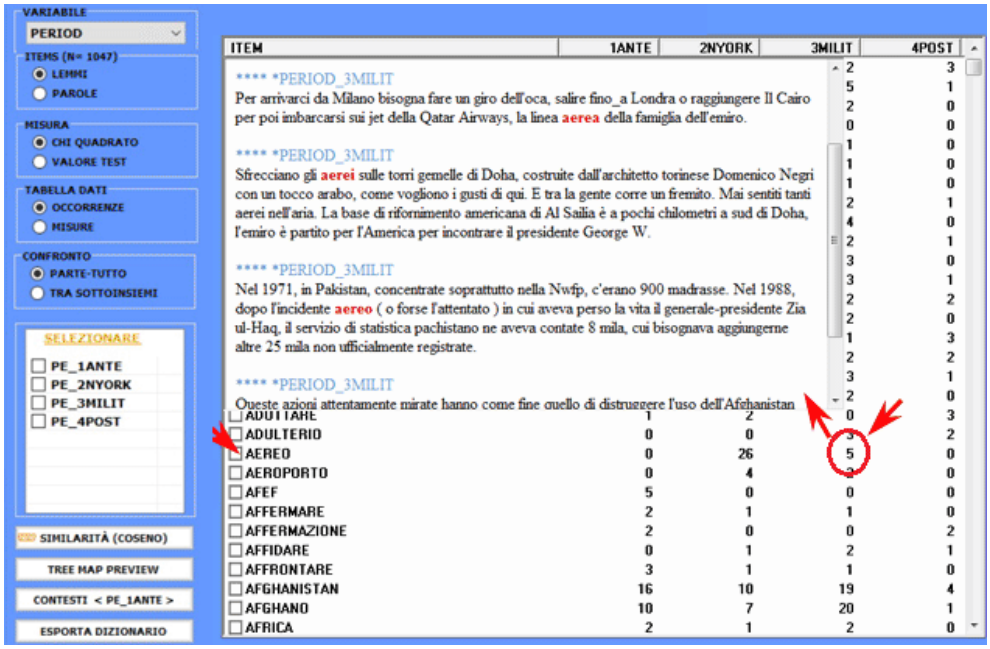


Le tabelle di contingenza utilizzate per i vari confronti possono essere facilmente esportate ed utilizzate per realizzare vari tipi grafici. Inoltre, cliccando su specifiche celle (vedi sotto), è possibile creare file HTML con i tutti i contesti elementari in cui la parola in riga è presente nel sottoinsieme in colonna.

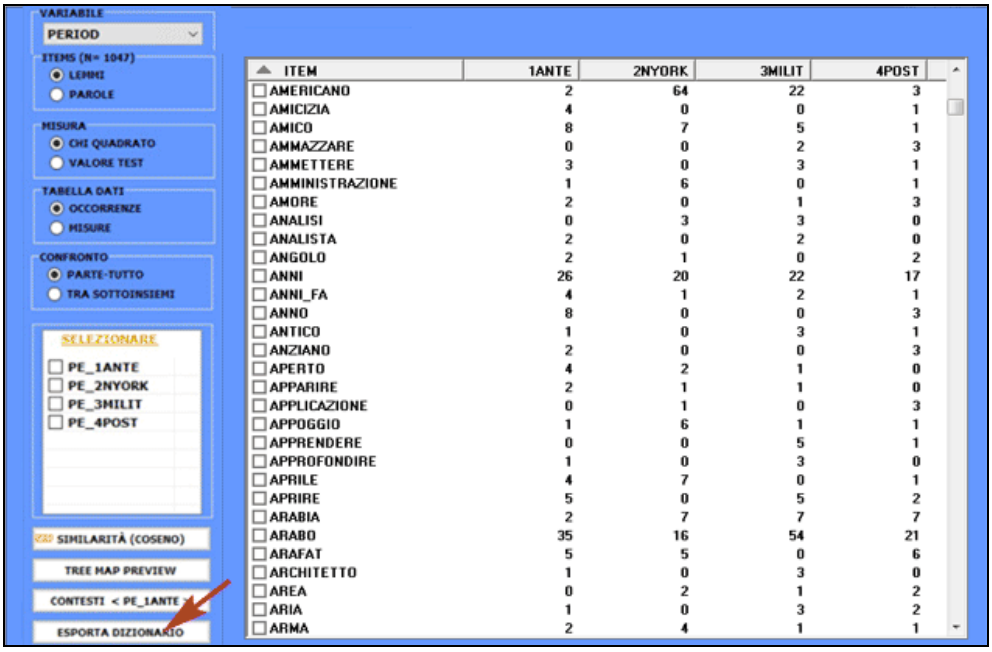



VARIABILE
PERIOD
ITEMS (N= 1047)
LEMMI
PAROLE
MISURA
CHI QUADRATO
VALORE TEST
TABELLA DATI
OCCORRENZE
MISURE
CONFRONTO
PARTE-TUTTO
TRA SOTTOINSIEMI
SELEZIONARE
PE_1ANTE
PE_2NYORK
PE_3MILIT
PE_4POST
SIMILARITÀ (COSENO)
TREE MAP PREVIEW
CONTESTI < PE_1ANTE >
ESPORTA DIZIONARIO
ITEM
ANALISI DELLE SPECIFICITÀ
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouser
AEREO
(OCCORRENZE)
25
20
15
10
5
0
1ANTE
2NYORK
3MILIT
4POST
26,0
5,0


VARIABILE
PERIOD
ITEMS (N= 1047)
LEMMI
PAROLE
MISURA
CHI QUADRATO
VALORE TEST
TABELLA DATI
OCCORRENZE
MISURE
CONFRONTO
PARTE-TUTTO
TRA SOTTOINSIEMI
SELEZIONARE
PE_1ANTE
PE_2NYORK
PE_3MILIT
PE_4POST
SIMILARITÀ (COSENO)
TREE MAP PREVIEW
CONTESTI < PE_1ANTE >
ESPORTA DIZIONARIO
ITEM
ANALISI DELLE SPECIFICITÀ
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouser
AEREO
(CHI QUADRATO)
40
35
30
25
20
15
10
5
0
-5
-10
1ANTE
2NYORK
3MILIT
4POST
-9,9
40,3
-2,2
-6,1

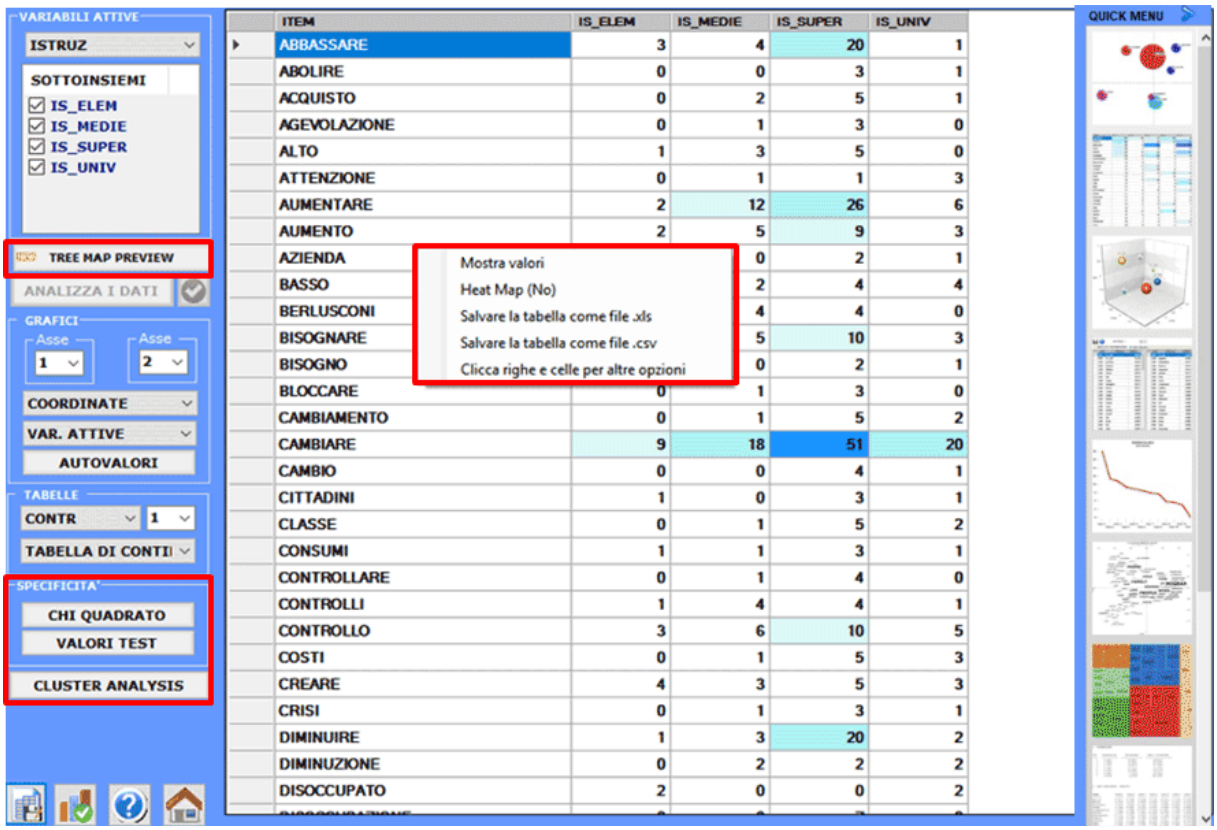


Infine, cliccando l'apposito pulsante (vedi sotto), viene creato un file **dizionario** con l'estensione .dictio che è pronto per essere importato da qualsiasi strumento **T-LAB** per l'**analisi tematica**. Tale dizionario include tutte le parole tipiche della variabile categoriale selezionata.



Analisi delle Corrispondenze

N.B.: Le immagini di questa sezione fanno riferimento a una versione precedente di T-LAB. In **T-LAB 10** l'aspetto è leggermente diverso. Inoltre: a) il **tasto destro** sulle tabelle con le parole chiave rende disponibili opzioni supplementari; b) un nuovo pulsante (TREE MAP PREVIEW) consente di creare grafici dinamici in formato HTML; c) due nuovi pulsanti permettono di verificare le specificità di ciascun valore della variabile sia usando il test chi-quadrato o il valore test; d) è presente un pulsante che consente di effettuare una **cluster analisi** che utilizza le coordinate degli oggetti (a seconda dei casi, unità lessicali o unità di contesto) sui primi assi fattoriali (fino a un massimo di 10); e) una galleria di immagini funziona come un menu aggiuntivo e consente di passare da un output all'altro con un solo clic.
Alcune di queste nuove funzionalità sono evidenziate nell'immagine seguente.



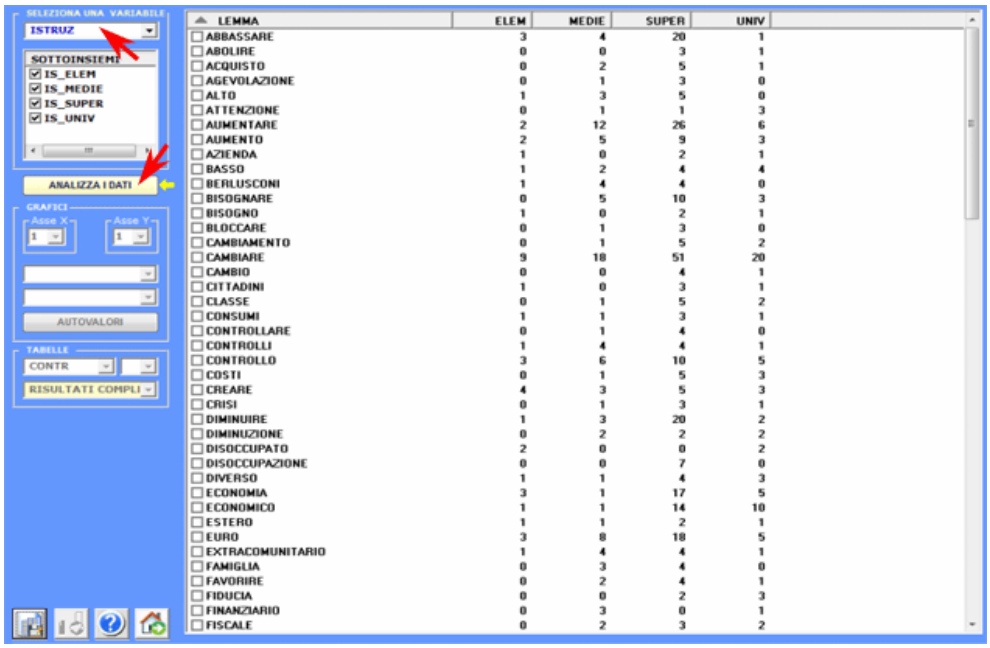
Questo strumento **T-LAB** ha l'obiettivo di evidenziare **somiglianze** e **differenze** tra unità di contesto.

Più precisamente, in **T-LAB**, l'analisi delle corrispondenze consente di analizzare tre tipi di

tabelle:

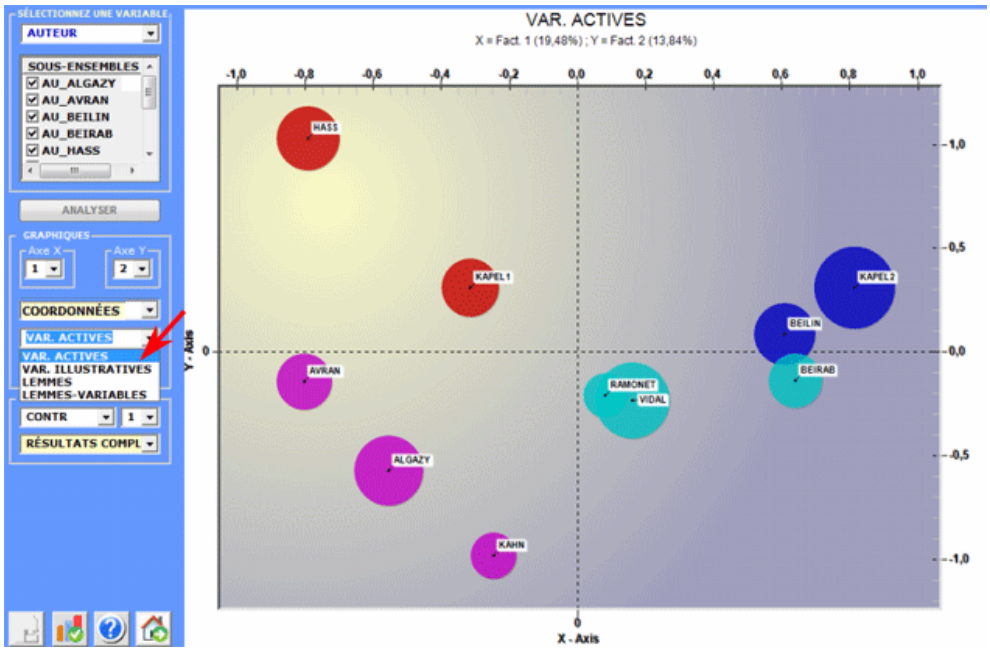
(**A**) quelle parole per modalità di una variabile, con i valori di **occorrenza**;
(**B**) quelle contesti elementari per parole, con i valori di **co-occorrenza**;
(**C**) quelle documenti per parole, con i valori di **occorrenza**.

Per analizzare le tabelle (**A**) si richiede che il corpus sia costituito da almeno tre testi oppure che sia codificato con qualche **variabile** (non meno di tre modalità).
Le variabili sono elencate in un apposito box e possono essere utilizzate una alla volta.
Dopo ogni selezione - in sequenza – viene mostrata la tabella di contingenza e ci viene chiesto di cliccare sul pulsante **analizza i dati** (vedi sotto).

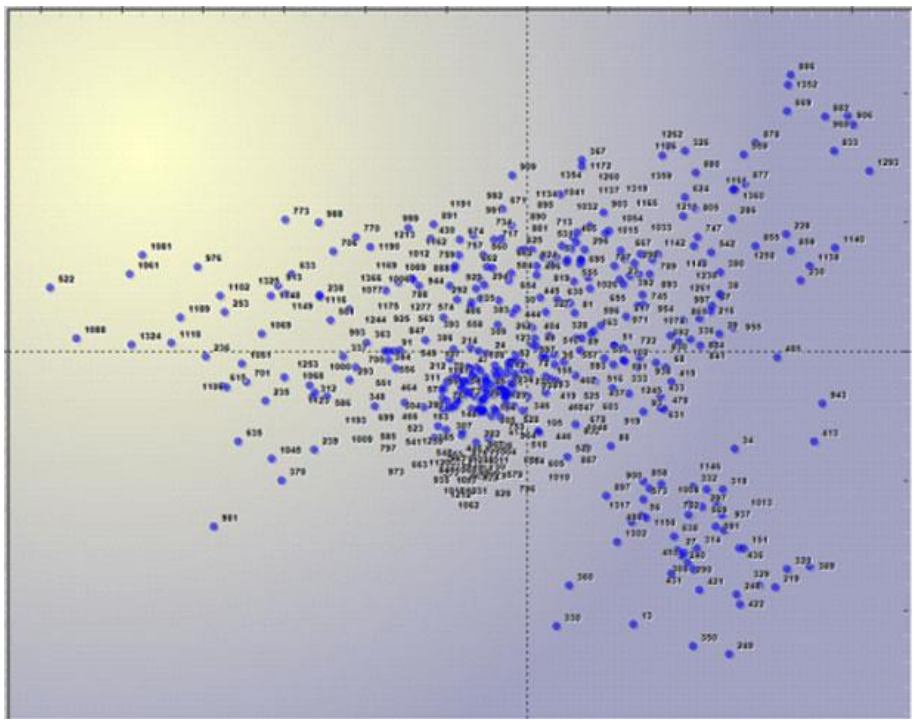


Il risultato dell'analisi è costituito da tabelle che consentono la realizzazione di grafici in cui - su piani cartesiani - sono rappresentati sia le relazioni tra sottoinsiemi del corpus sia quelle tra

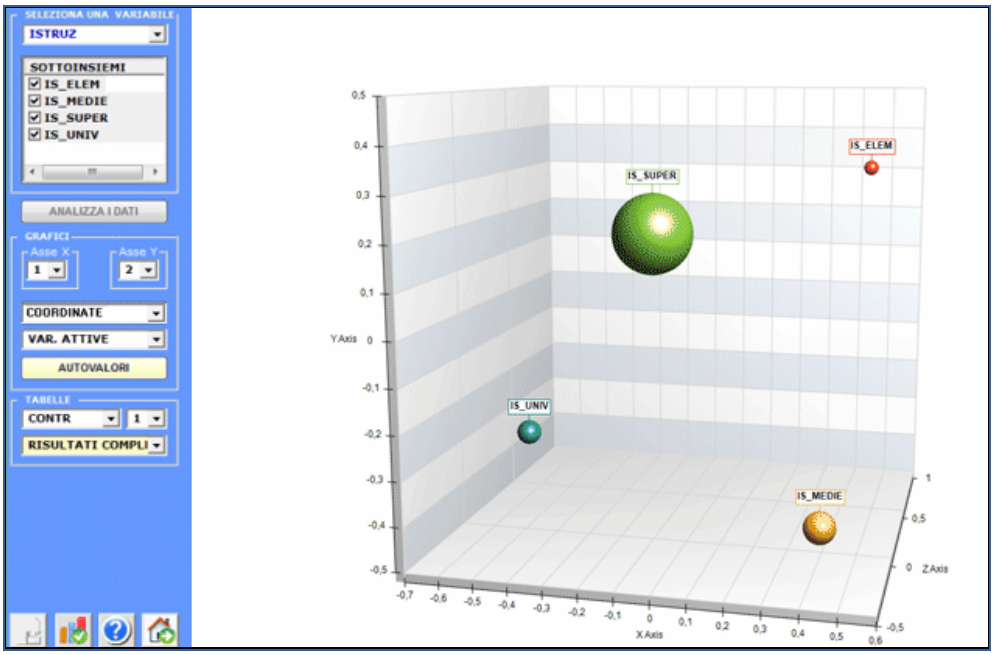
le unità lessicali che li costituiscono. Più precisamente, a seconda dei casi, i tipi di grafici disponibili mostrano le relazioni tra **variabili attive**, tra **variabili illustrative**, tra lemmi, tra lemmi e variabili.

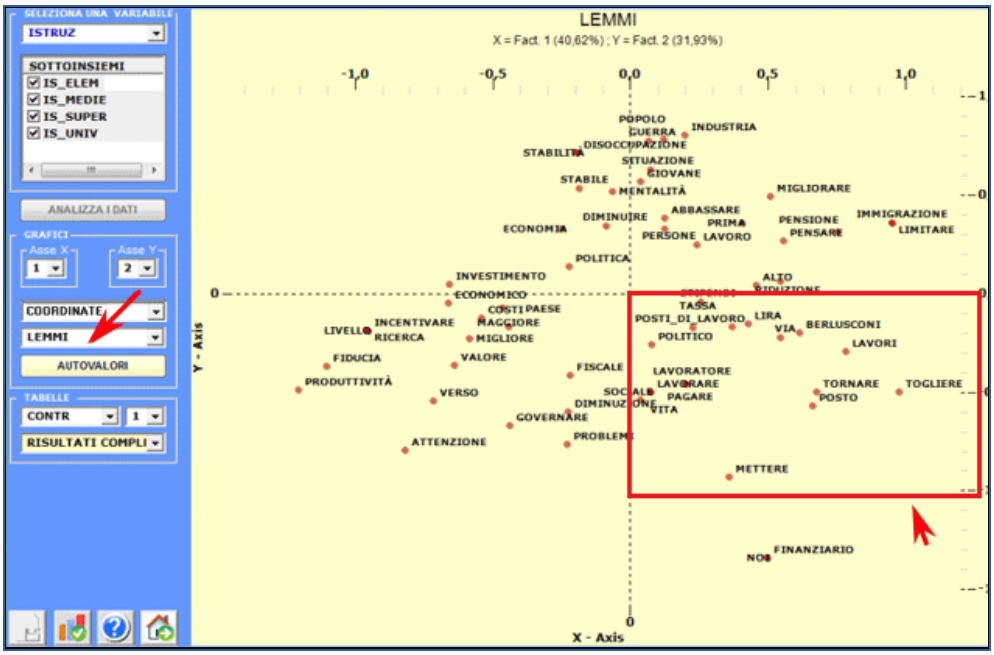


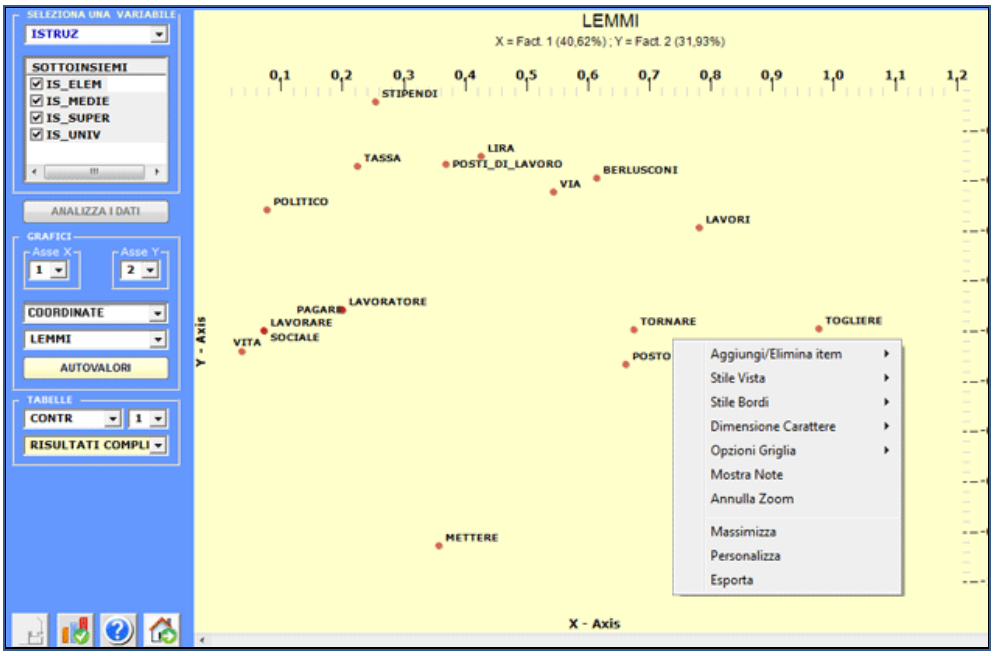
Inoltre, quando la tabella analizzata è del tipo documenti per parole, è possibile visualizzare i punti (Max 3000) corrispondenti a ciascun documento.



Tutti i grafici possono essere massimizzati e personalizzati usando l'apposita finestra di dialogo (apertura tramite il tasto destro del mouse). Inoltre, quando le modalità della variabile in analisi sono più di tre, le loro relazioni possono essere esplorate attraverso grafici in **3D**.

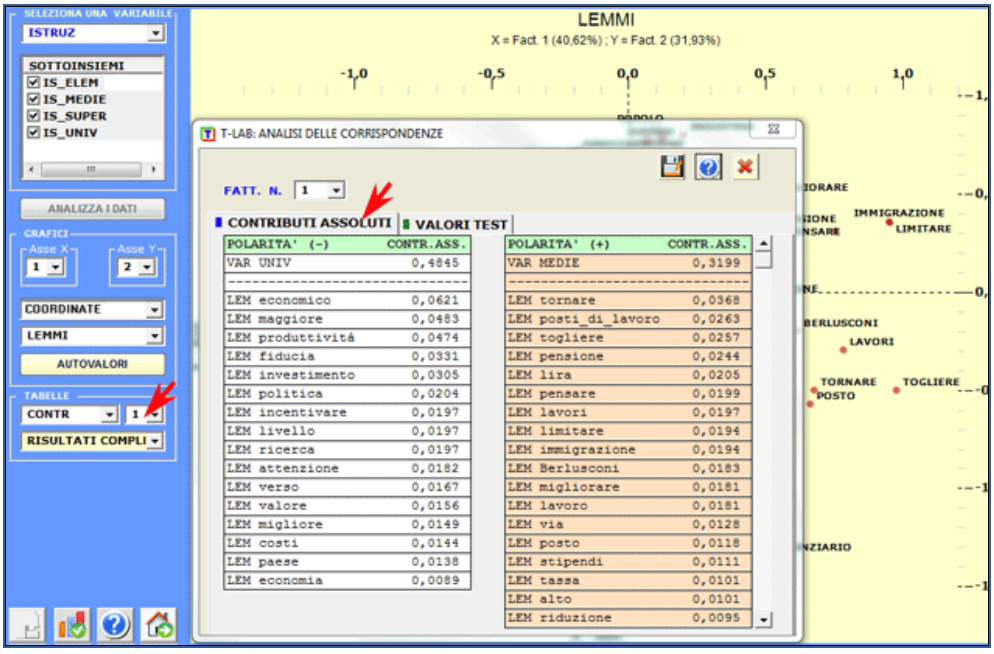




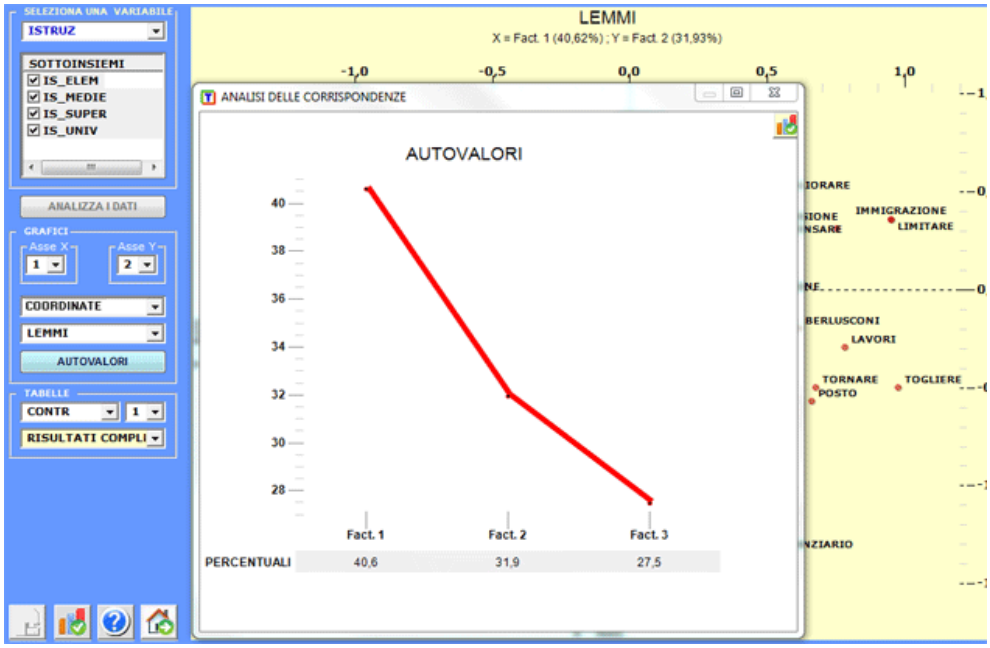


Per esplorare le varie combinazioni degli assi fattoriali è sufficiente selezionarli negli appositi box ("Asse X", "Asse Y").

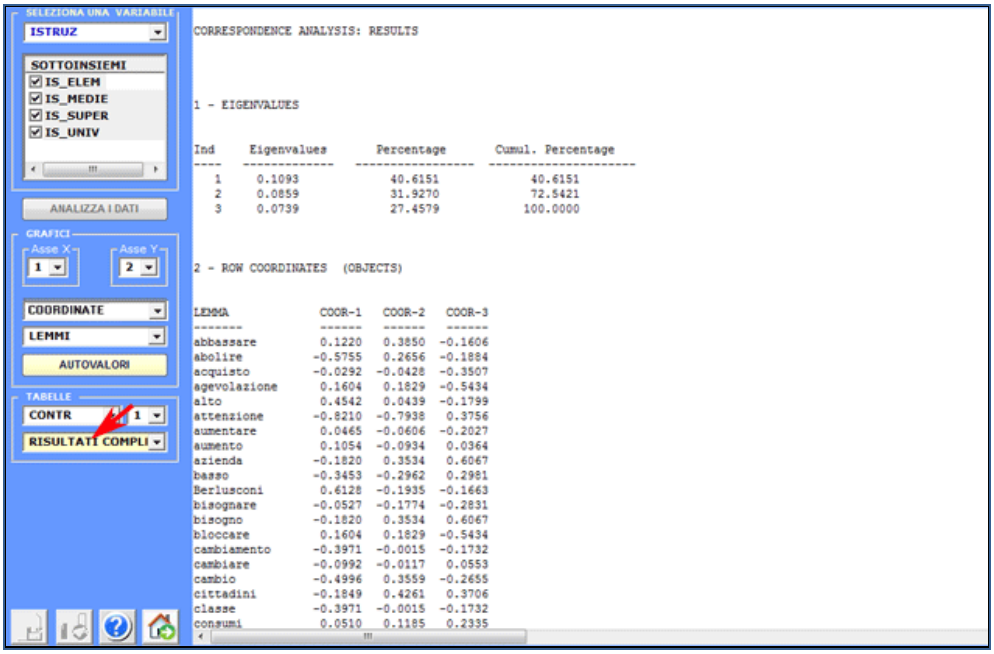
In **T-LAB** le caratteristiche di ogni **polarità fattoriale** (cioè le opposizioni sugli assi orizzontale e verticale dei grafici) possono essere verificate tramite alcune tabelle interattive che mostrano i **Contributi Assoluti** con valore di soglia maggiore o uguale a 1/N (N = righe della tabella analizzata) o i **Valori Test** (in questo caso la soglia di significatività è +/- 1.96).



Il grafico degli **autovalori** (scree plot) consente di apprezzare il peso relativo di ogni fattore, ovvero la percentuale di varianza spiegata da ciascuno di essi.



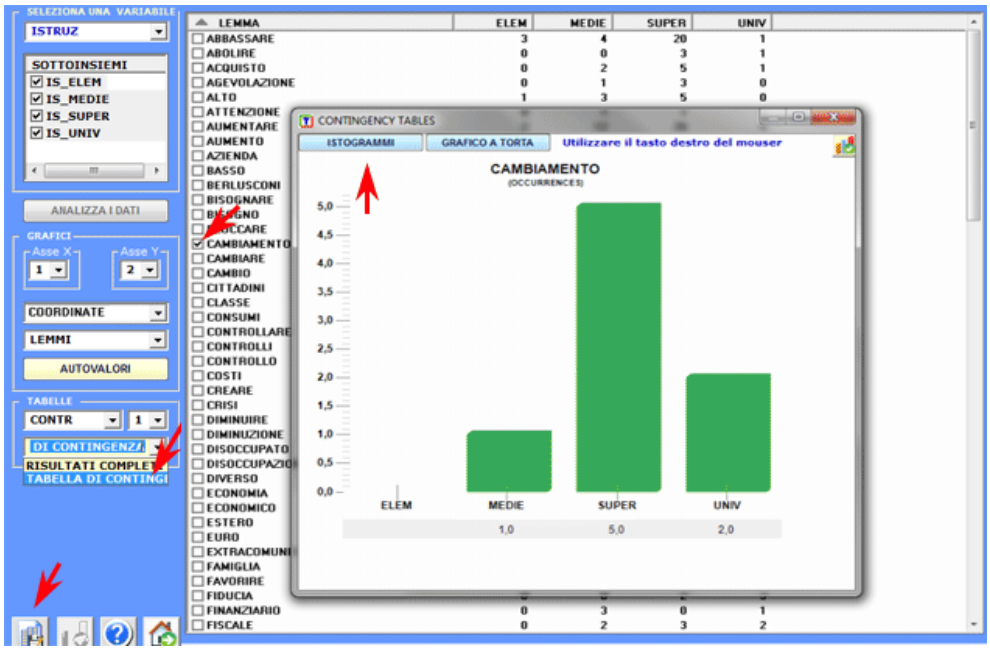
Un click sul pulsante **Risultati Completi** consente di visualizzare e di esportare il file che riporta tutti i risultati dell'analisi: autovalori, coordinate, contributi assoluti e relativi, valori test.

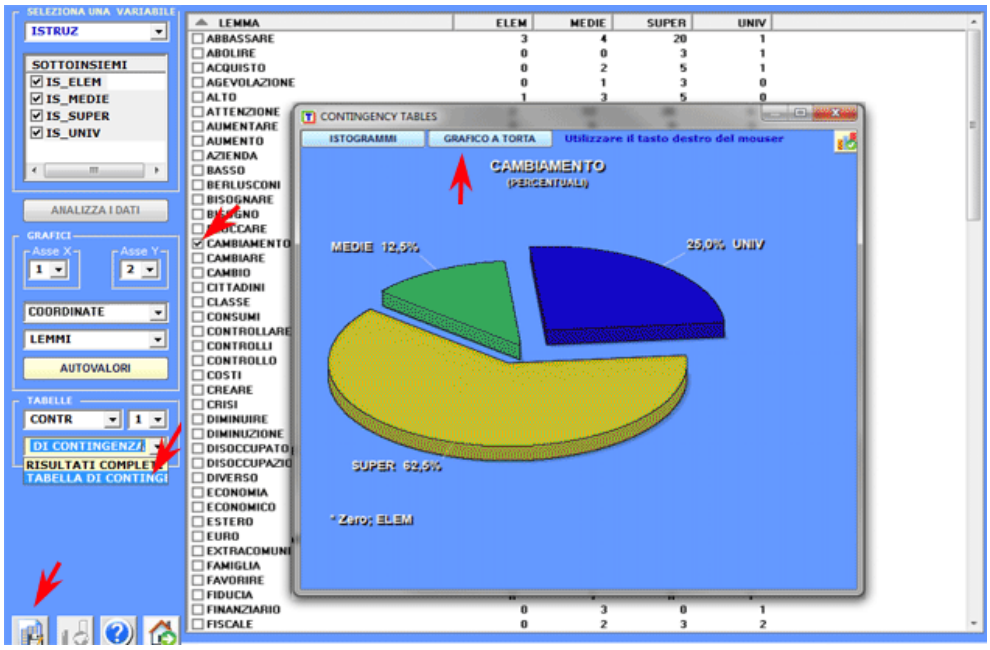


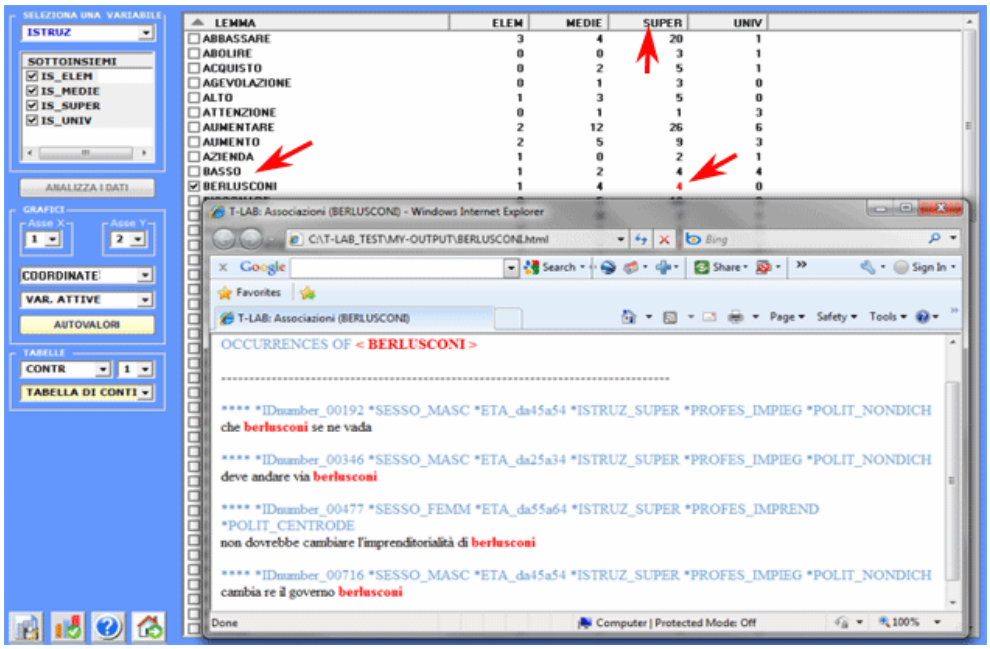
Tutte tabelle di contingenza analizzate possono essere facilmente visualizzate ed esportate.

Le occorrenze delle parole in esse presenti possono essere esplorate con vari tipi di grafici.

Inoltre, cliccando su specifiche celle, è possibile creare file HTML con i tutti i contesti elementari in cui la parola in riga è presente nel sottoinsieme in colonna. (vedi sotto).







Nel caso delle tabelle (**B**) e (**C**), esse sono costituite da tante righe quanti sono le unità di contesto (max 10.000) e tante colonne quante sono le parole chiave selezionate (max 1.500).

L'algoritmo di calcolo e gli output sono analoghi a quelli dell'analisi di tabelle (**A**) unità lessicali per variabili, solo che - in questo caso - per contenere i tempi di elaborazione, **T-LAB** si limita ad estrarre i primi 10 fattori: un numero più che sufficiente per riassumere la variabilità dei dati.

Inoltre, in successione è possibile effettuare due tipi di **Cluster Analysis** i cui "oggetti" sono rispettivamente costituiti dalle parole chiave o dai contesti elementari.

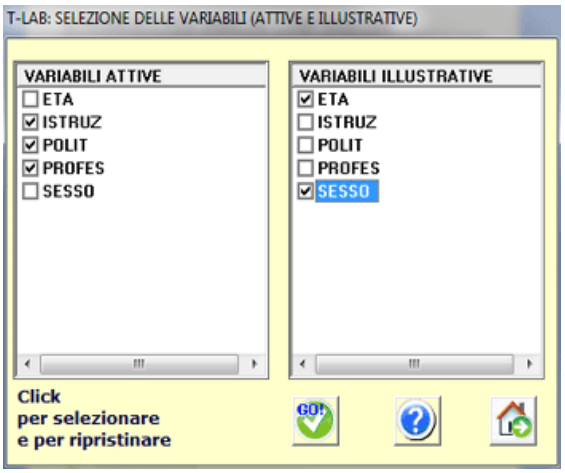
Analisi delle Corrispondenze Multiple

N.B.: Le immagini di questa sezione fanno riferimento a una precedente versione di T-LAB. In **T-LAB 10** l'aspetto è leggermente diverso. Inoltre è presente un pulsante che consente di effettuare una **cluster analisi** che utilizza le coordinate degli oggetti sui primi assi fattoriali (fino a un massimo di 10).

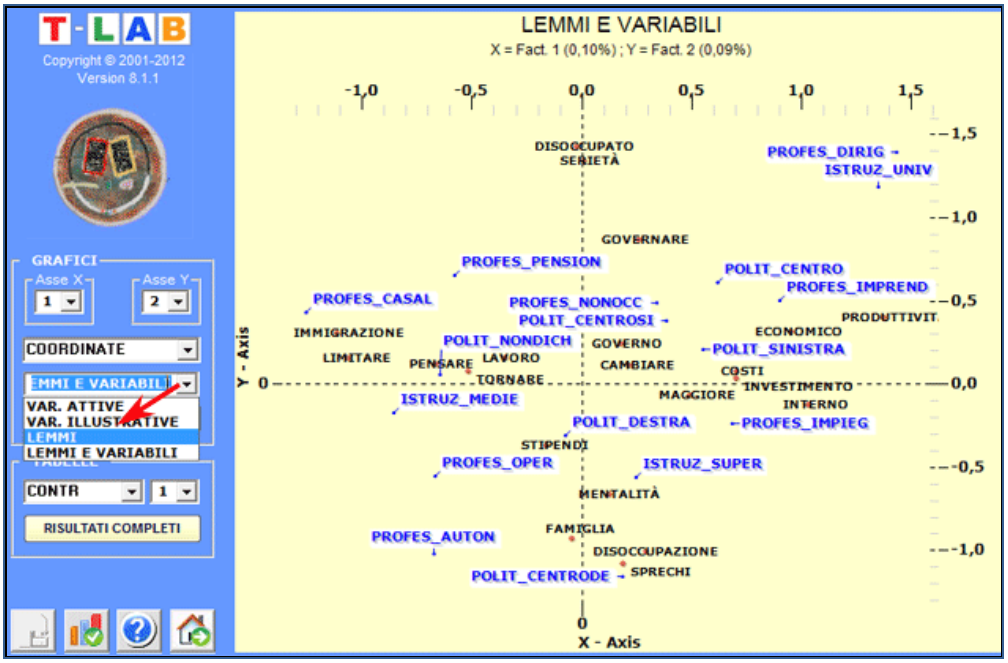
L'Analisi delle Corrispondenze Multiple, che può essere considerata un'estensione dell'Analisi delle Corrispondenze semplice (vedi sopra), consente di analizzare le relazioni tra le modalità di due o più variabili categoriali.

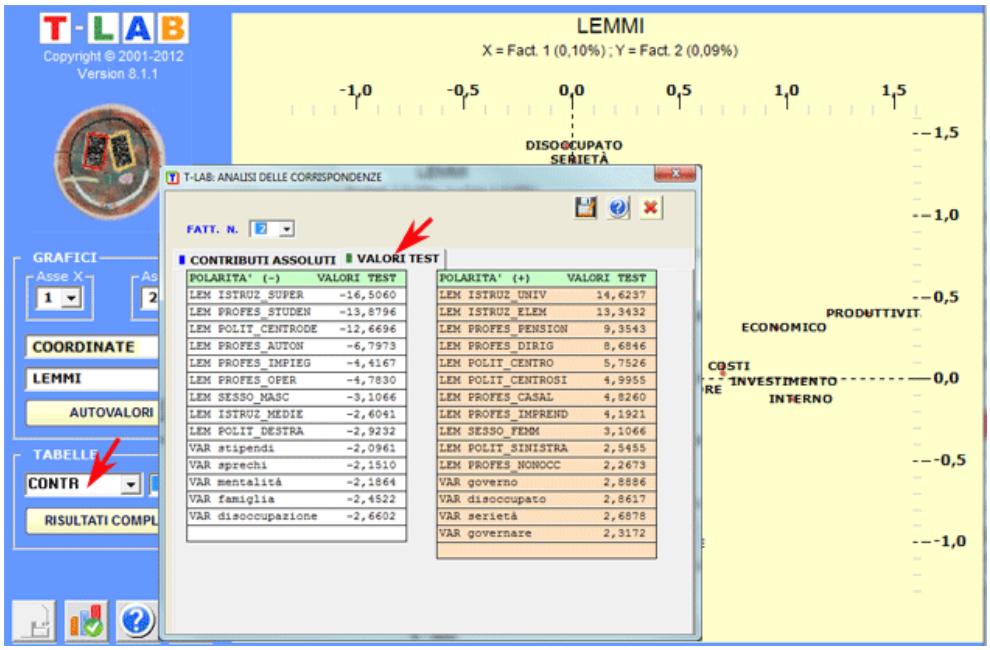
Nel caso di **T-LAB**, le limitazioni di questo tipo di analisi sono le seguenti:
- 150.000 contesti elementari (le righe della tabella analizzata);
- 250 modalità delle variabili (le colonne della tabella analizzata);
- 3.000 parole-chiave, corrispondenti ad altrettante colonne illustrative (vedi Lebart L., Salem A., 1994).

Per effettuare questo tipo di analisi, disponibile solo se il corpus include almeno due variabili e non supera le limitazioni previste, **T-LAB** richiede che l'utilizzatore definisca le sue scelte all'interno della finestra seguente:



Al termine dell'analisi, gli output **T-LAB** sono analoghi a quelli dell'analisi delle corrispondenze (vedi sotto), con la differenza che viene prodotta anche una tabella (Burt_Table.xls) che include tutti gli incroci delle categorie analizzate:

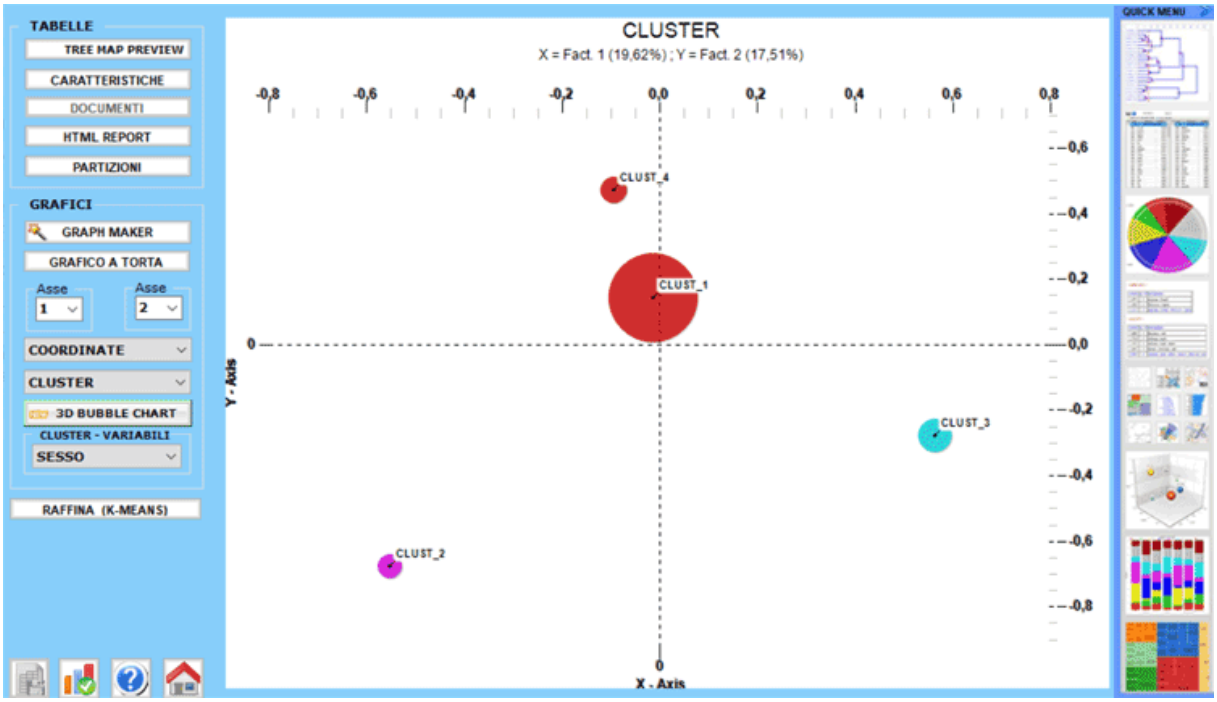




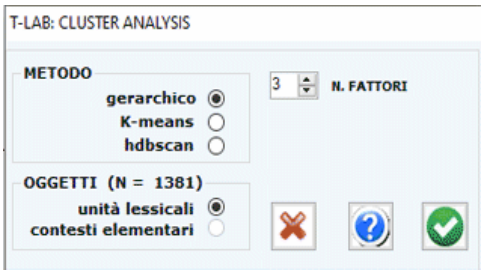
Dopo l'analisi delle corrispondenze multiple, solo se i contesti elementari corrispondono ai contesti iniziali (es. risposte a domande aperte), è possibile effettuare una **cluster analysis**.

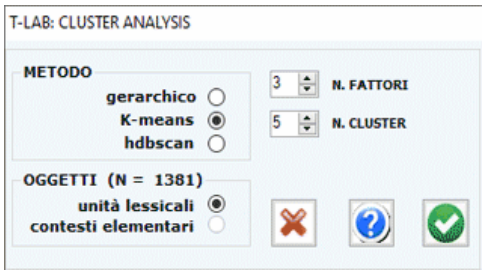
Cluster Analysis

N.B.: Le immagini di questa sezione fanno riferimento a una versione precedente di **T-LAB**. In **T-LAB 10** l'aspetto è leggermente diverso. Inoltre: a) un nuovo pulsante (**TREE MAP PREVIEW**) consente di creare grafici dinamici in formato HTML; b) il pulsante DENDROGRAMMA è stato sostituito con lo strumento **GRAPH MAKER**; c) una galleria di immagini funziona come un menu aggiuntivo e consente di passare da un output all'altro con un solo clic (vedi immagine seguente).



L'opzione **Cluster Analysis** attiva una procedura di calcolo che utilizza i risultati di una precedente **Analisi delle Corrispondenze**; in particolare utilizza le coordinate degli oggetti (unità lessicali o unità di contesto) sui primi assi fattoriali (fino a un massimo di 10).





A seconda dei casi, l'utilizzatore può scegliere tra tre tecniche di clusterizzazione:

a) **gerarchica** (metodo Ward);
b) **K-means** (metodo MacQueen);
c) **hdbscan** (hierarchical DBSCAN).

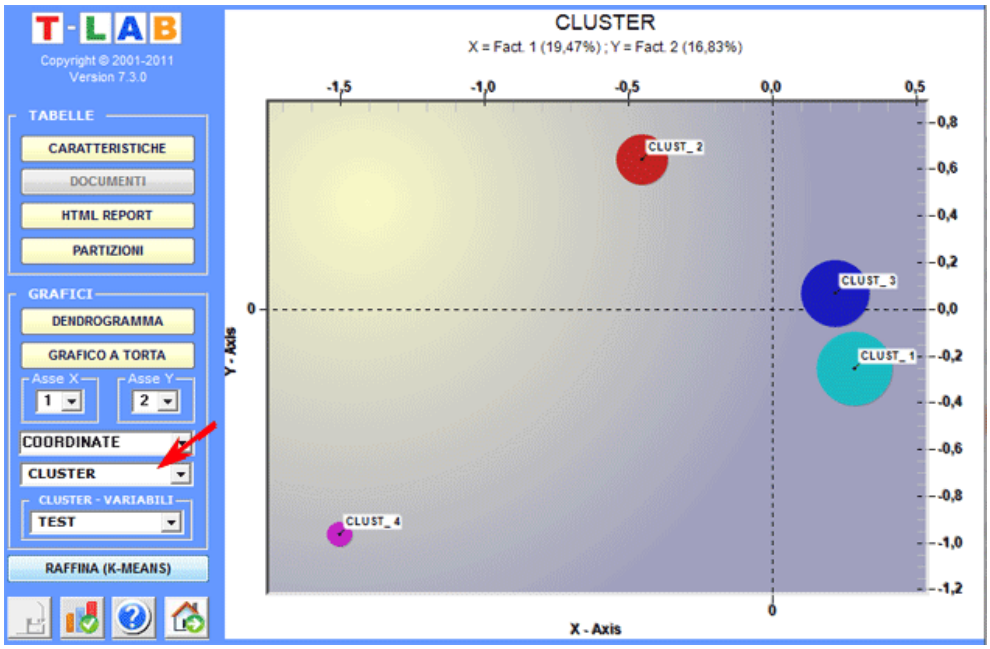
Le prime due (a, b) consentono di esplorare (tabelle e grafici) soluzioni da 3 a 20 cluster; mentre la terza (c), che richiede un parametro aggiuntivo (ovvero il numero minimo di parole all'interno di un cluster), consente all'utente di esplorare solo una soluzione.

N.B.: Quando viene selezionato il metodo gerarchico **T-LAB** abilita un'opzione (vedi pulsante 'RAFFINA') che consente di combinare i metodi Ward e K-Means.

Una breve descrizione delle tre tecniche è contenuta nel **glossario** di questo manuale.

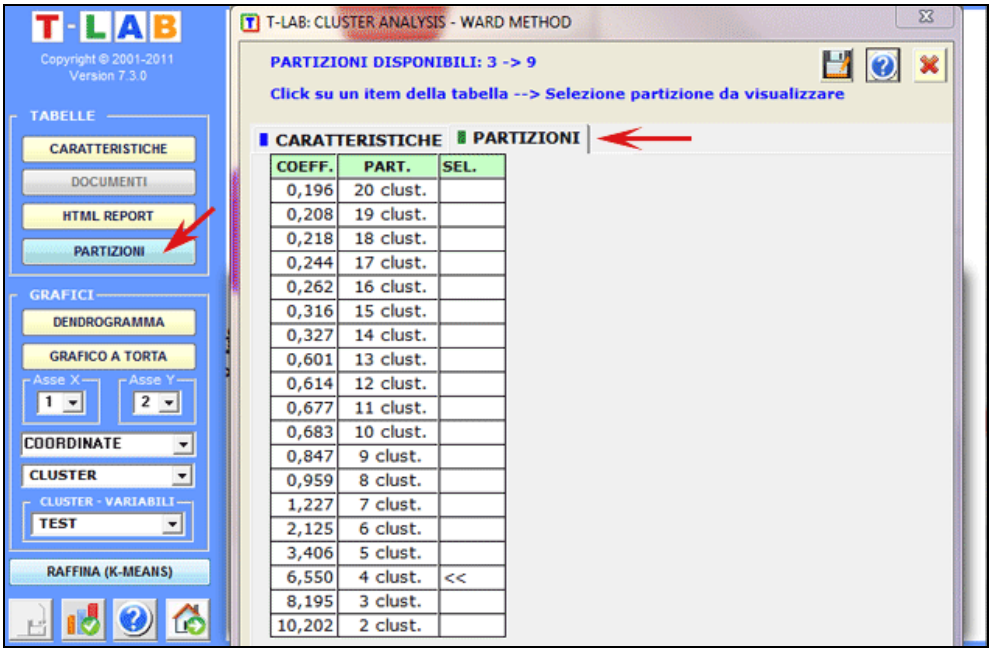
Al termine dell'elaborazione **T-LAB** rende disponibili grafici e tabelle.

Alcuni grafici rappresentano i cluster nello stesso spazio individuato tramite l'analisi delle corrispondenze (vedi sotto).



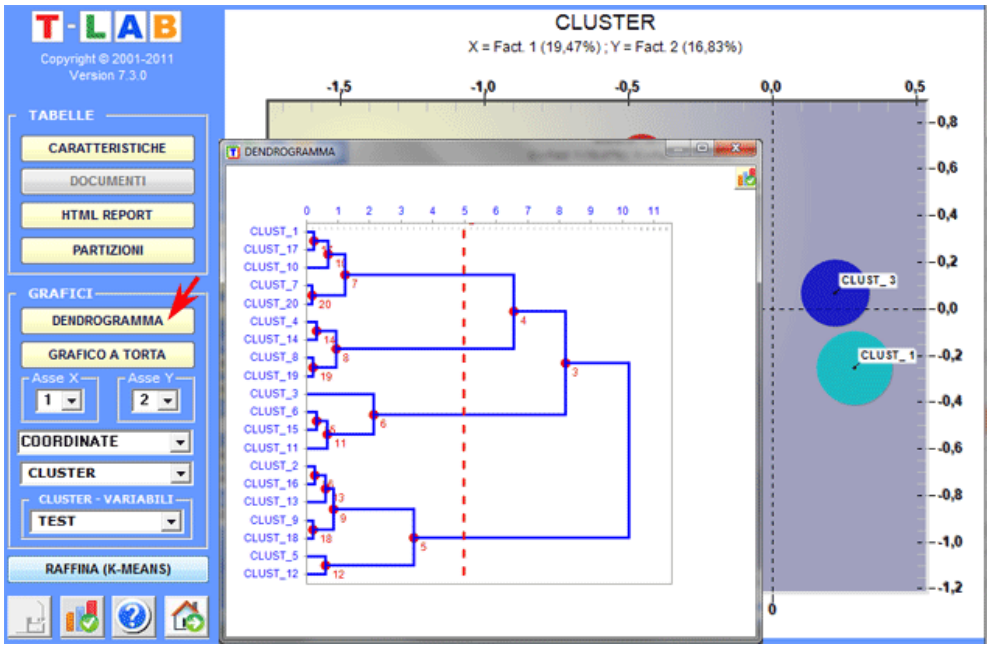
N.B.: Per esplorare le varie combinazioni degli assi fattoriali è sufficiente selezionarli negli appositi box ("Asse X", "Asse Y").

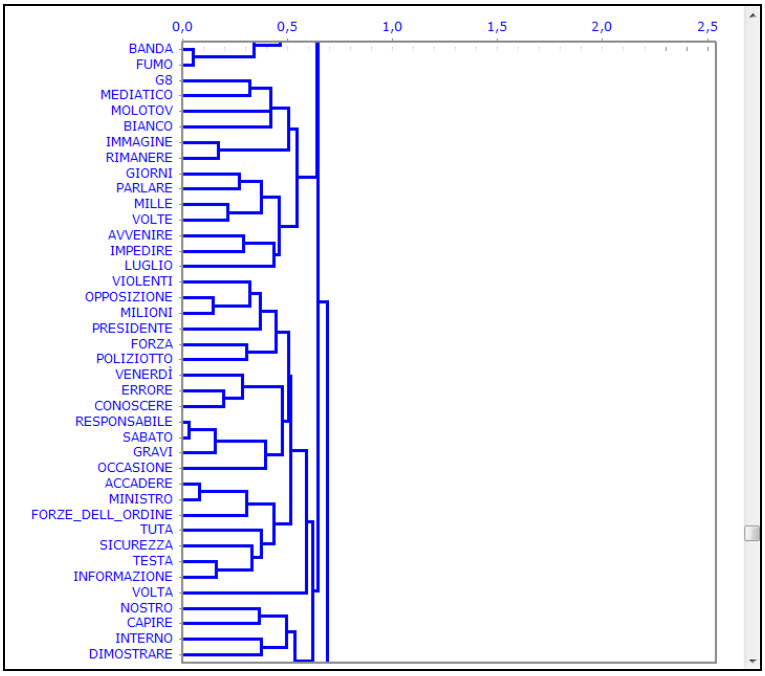
Nel caso di clusterizzazione gerarchica, l'utilizzatore può agevolmente esplorare (grafici e tabelle) le diverse partizioni.

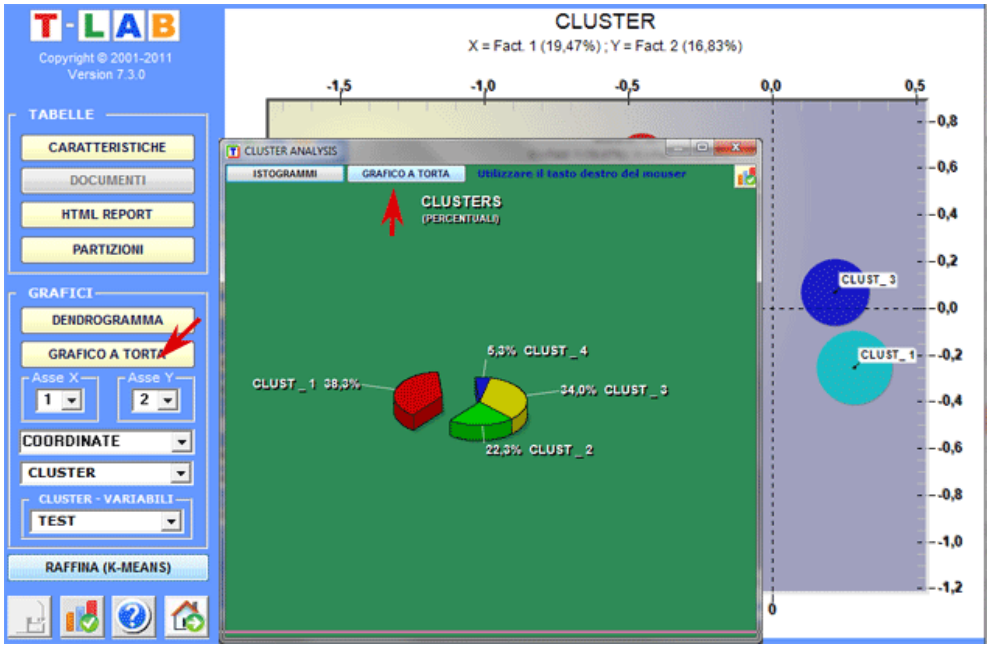


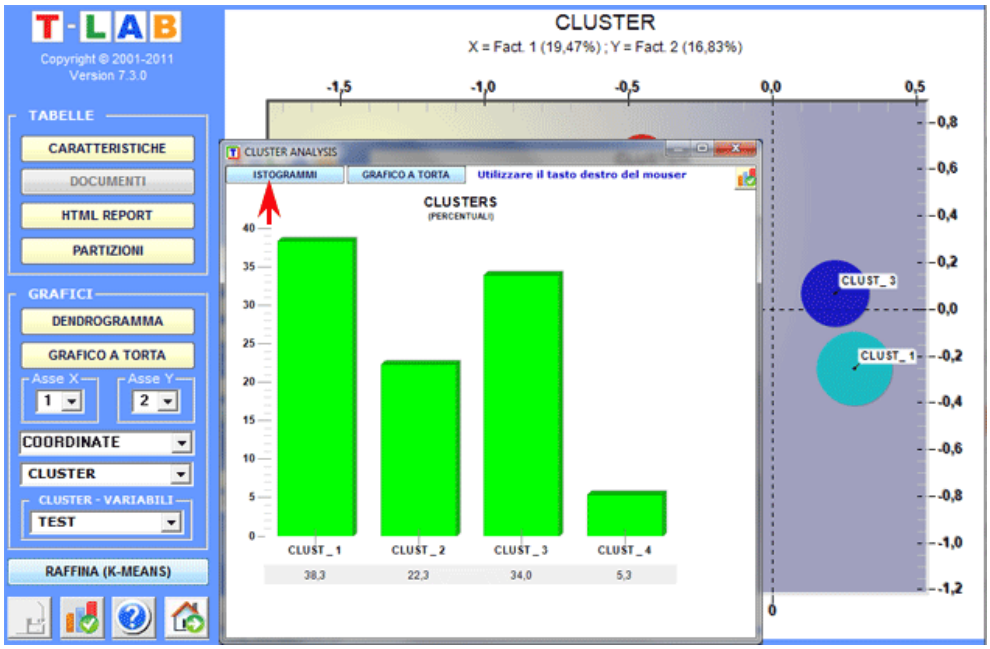
COEFF.	PART.	SEL.
0,196	20 clust.	
0,208	19 clust.	
0,218	18 clust.	
0,244	17 clust.	
0,262	16 clust.	
0,316	15 clust.	
0,327	14 clust.	
0,601	13 clust.	
0,614	12 clust.	
0,677	11 clust.	
0,683	10 clust.	
0,847	9 clust.	
0,959	8 clust.	
1,227	7 clust.	
2,125	6 clust.	
3,406	5 clust.	
6,550	4 clust.	<<
8,195	3 clust.	
10,202	2 clust.	

Dendrogrammi, grafici a torta e istogrammi consentono di verificare le caratteristiche di ogni partizione.

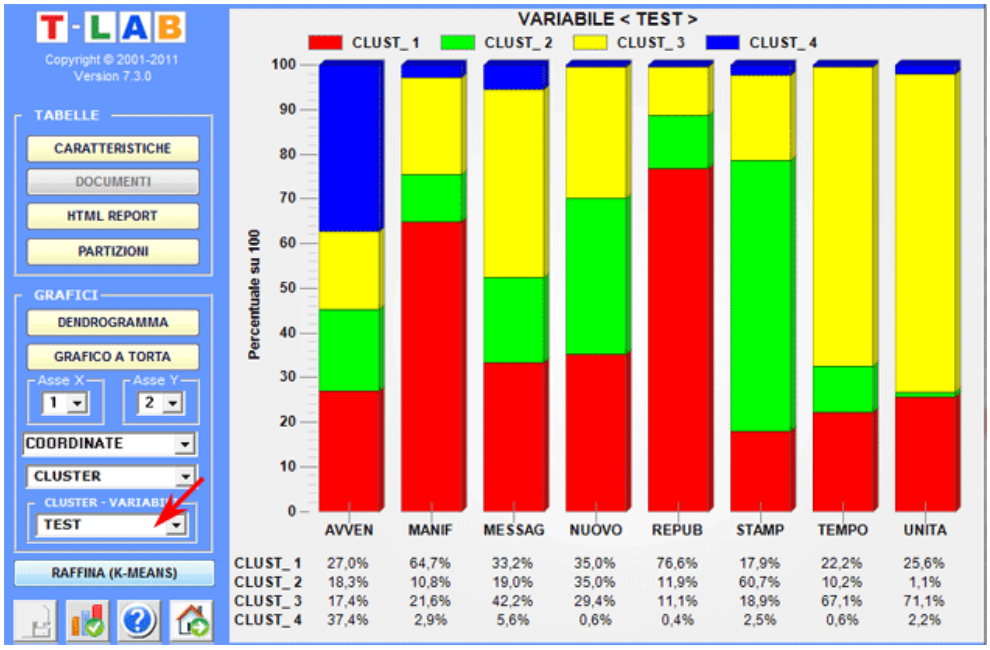



0,0
0,5
1,0
1,5
2,0
2,5
BANDA
FUMO
G8
MEDIATICO
MOLOTOV
BIANCO
IMMAGINE
RIMANERE
GIORNI
PARLARE
MILLE
VOLTE
AVVENIRE
IMPEDIRE
LUGLIO
VIOLENTI
OPPOSIZIONE
MILIONI
PRESIDENTE
FORZA
POLIZIOTTO
VENERDÌ
ERRORE
CONOSCERE
RESPONSABILE
SABATO
GRAVI
OCCASIONE
ACCADERE
MINISTRO
FORZE_DELL_ORDINE
TUTA
SICUREZZA
TESTA
INFORMAZIONE
VOLTA
NOSTRO
CAPIRE
INTERNO
DIMOSTRARE


T-LAB
Copyright © 2001-2011
Version 7.3.0
TABELLE
CARATTERISTICHE
DOCUMENTI
HTML REPORT
PARTIZIONI
GRAFICI
DENDROGRAMMA
GRAFICO A TORTA
Asse X
1
Asse Y
2
COORDINATE
CLUSTER
CLUSTER - VARIABILI
TEST
RAFFINA (K-MEANS)
CLUSTER
X = Fact. 1 (19,47%) ; Y = Fact. 2 (16,83%)
-1,5
-1,0
-0,5
0,0
0,5
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
-1,0
-1,2
CLUSTER ANALYSIS
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouse
CLUSTERS
(PERCENTUALI)
5,3% CLUST _ 4
CLUST _ 1 38,3%
34,0% CLUST _ 3
22,3% CLUST _ 2
CLUST _ 3
CLUST _ 1

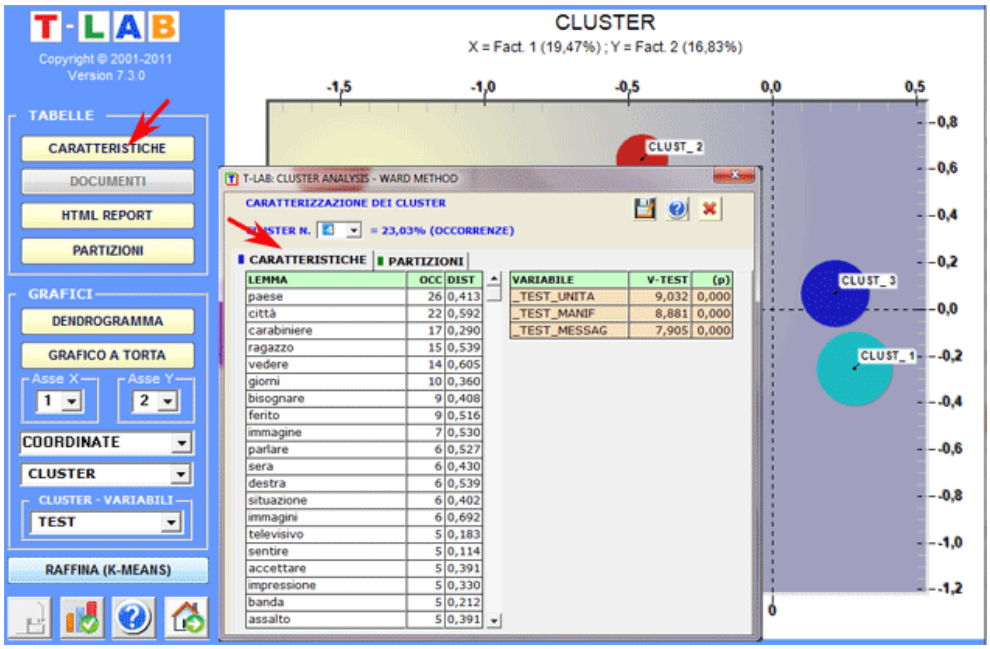


Alcuni istogrammi consentono di verificare le relazioni tra cluster e modalità delle variabili.

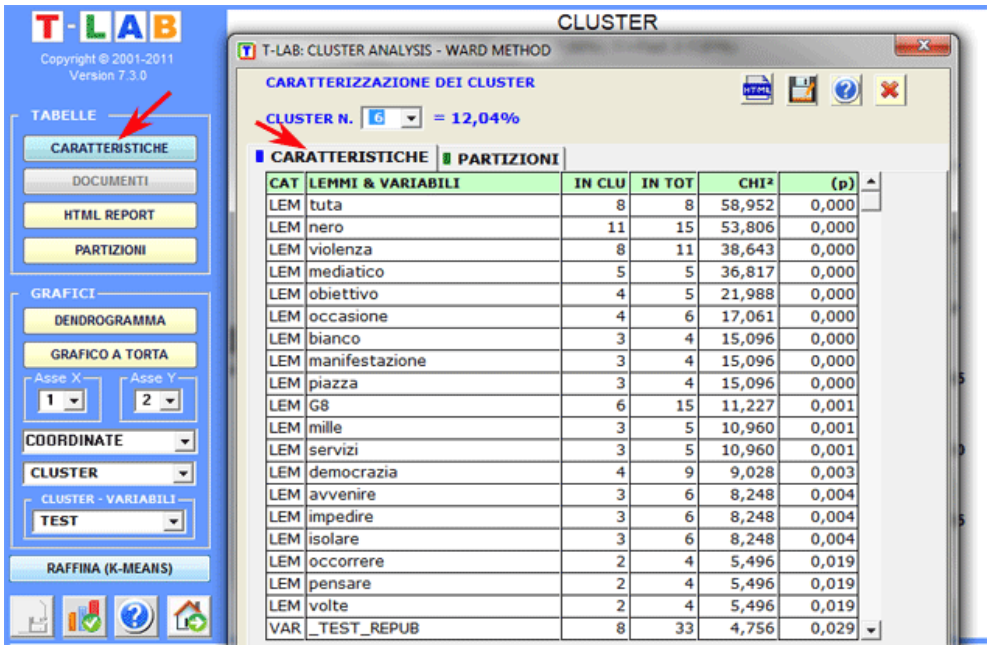


Le tabelle sono di due tipi:

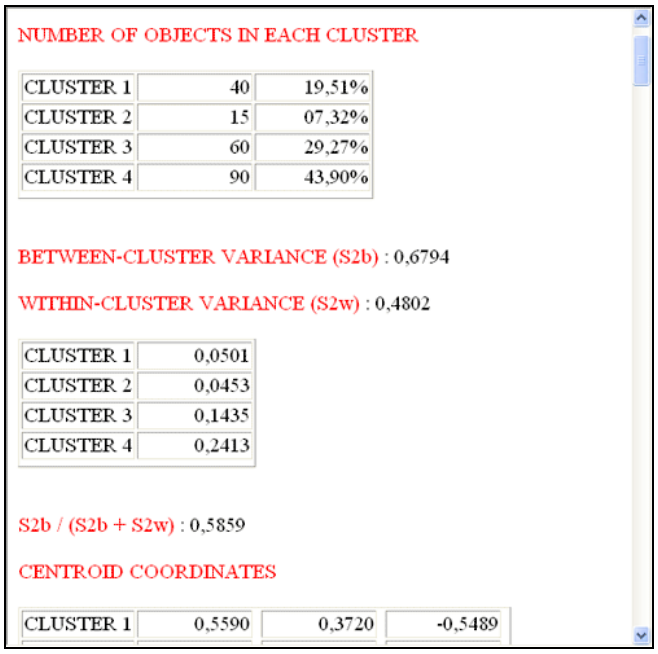
(**A**) se gli oggetti clusterizzati sono le unità lessicali, per ciascuna di esse (e per ogni cluster) vengono riportate le relative occorrenze ('OCC') e le loro distanze ('DIST') dal centroide; inoltre per le variabili che in modo significativo risultano associate al cluster in esame vengono mostrati i relativi Valori Test.



(**B**) se gli oggetti clusterizzati sono i contesti elementari, le caratteristiche di ogni cluster (unità lessicali e variabili) sono descritte con lo stesso metodo usato nella funzione Analisi Tematica dei Contesti Elementari.



Nel caso di analisi realizzate con metodi gerarchici o K-means **T-LAB** consente di visualizzare ed esportare un file (vedi pulsante "Output HTML") in cui sono riportate le caratteristiche dei cluster e alcune misure concernenti la qualità della partizione in esame.



NUMBER OF OBJECTS IN EACH CLUSTER

CLUSTER 1	40	19,51%
CLUSTER 2	15	07,32%
CLUSTER 3	60	29,27%
CLUSTER 4	90	43,90%

BETWEEN-CLUSTER VARIANCE (S2b) : 0,6794

WITHIN-CLUSTER VARIANCE (S2w) : 0,4802

CLUSTER 1	0,0501
CLUSTER 2	0,0453
CLUSTER 3	0,1435
CLUSTER 4	0,2413

S2b / (S2b + S2w) : 0,5859

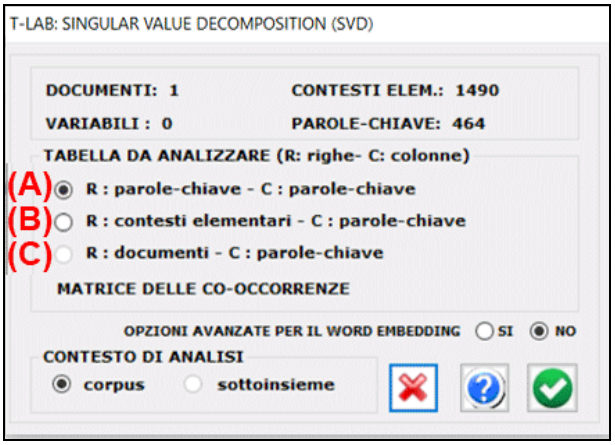
CENTROID COORDINATES

CLUSTER 1	0,5590	0,3720	-0,5489

Singular Value Decomposition (SVD)

La **Singular Value Decomposition** (SVD) è una tecnica per la riduzione delle dimensioni, che - in Text Mining - può essere utilizzata per verificare le **dimensioni latenti** (o componenti) che determinano le **somiglianze semantiche** tra parole (cioè unità lessicali) o tra documenti (cioè unità di contesto).

T-LAB ci consente di eseguire una Singular Value Decomposition di **tre tipi di tabelle dati**. Nel primo caso (vedi 'A' sotto), la tabella dati è una matrice delle co-occorrenze con - in riga e in colonna - le parole chiave selezionate. Nel secondo caso (vedi 'B' sotto), la tabella dati contesti elementari X parole chiave conterrà valori di presenza / assenza (cioè '1' e '0'). Nel terzo caso (vedi "C" sotto), la tabella dati documenti X parole chiave conterrà valori di occorrenza.
N.B.: Si noti che, quando vengono analizzate matrici di co-occorrenze, le cui righe e colonne sono termini chiave (vedere 'A' di seguito), **T-LAB** fornisce vettori densi di alta qualità (cioè word embeddings).



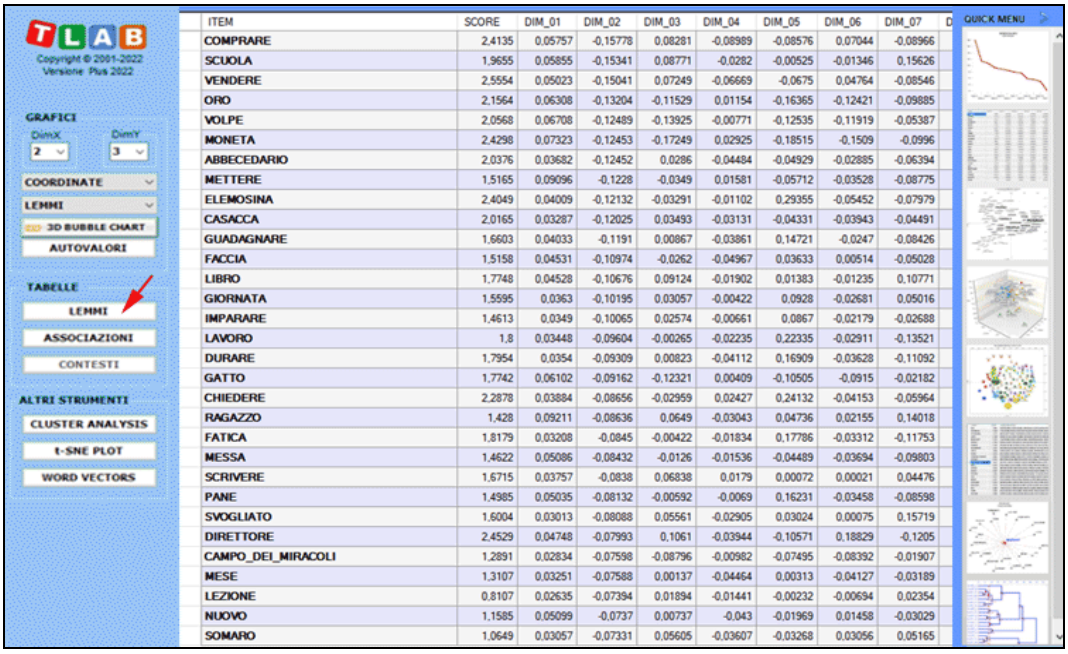
La procedura di analisi consiste dei seguenti passaggi:
1 - costruzione della tabella dati da analizzare (fino a 300.000 righe x 5.000 colonne);
2 - normalizzazione TF-IDF e applicazione della norma euclidea (i.e. trasformazione di tutti i vettori a lunghezza '1');
3 - estrazione delle prime 20 'dimensioni latenti' attraverso l'algoritmo di Lanczos.
N.B.:
- Nel caso della matrici di co-occorrenze (vedi 'A' sopra), la normalizzazione dei dati è ottenuta mediante la misura del coseno;
- Quando sono selezionate le opzioni avanzate per il word embedding, T-LAB calcola i valori PPMI (Positive Pointwise Mutual Information) e rende possibile l'utilizzo delle prime 50 dimensioni della SVD.

I risultati dell'analisi sono sintetizzati in **tabelle** e **grafici**.

Nel dettaglio:

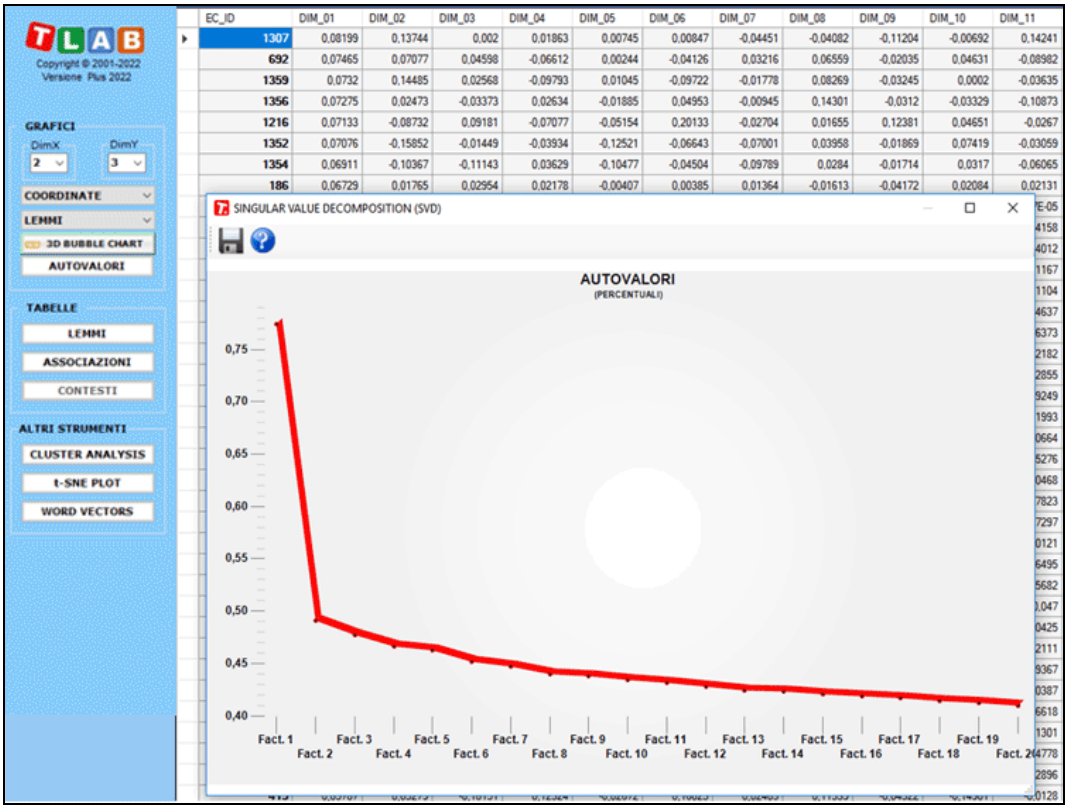
Due tabelle - le cui righe possono essere unità lessicali o unità di contesto - hanno tante colonne quante sono le dimensioni estratte.

Nel caso della tabella LEMMI (cioè unità lessicali), viene visualizzata un'ulteriore colonna in cui vengono riportati punteggi di importanza (vedi 'score' nella tabella seguente).



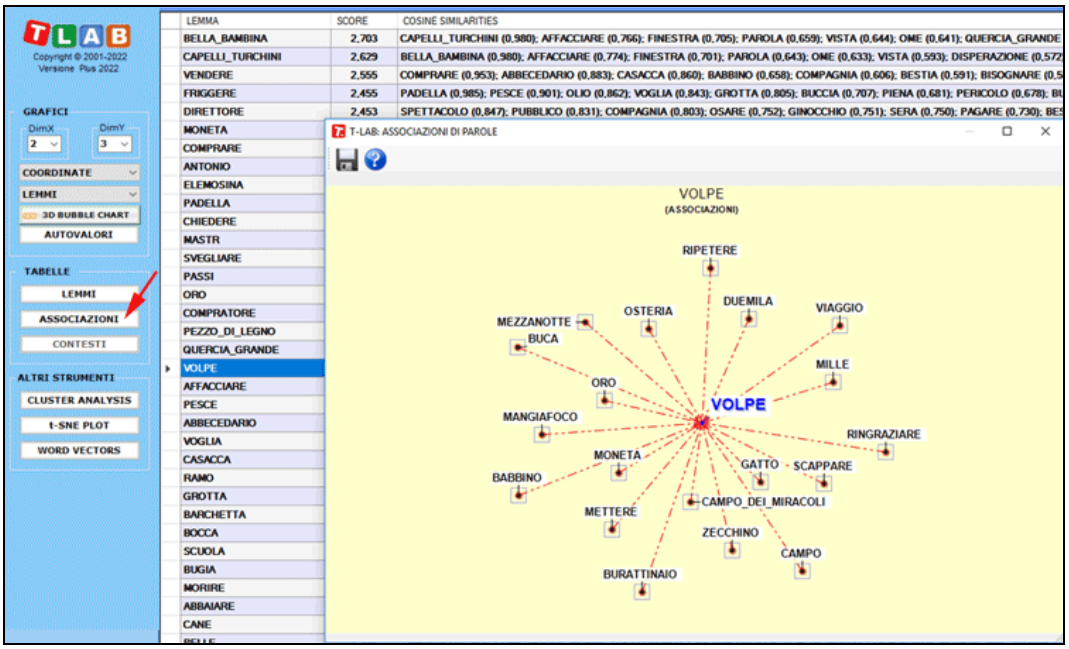
ITEM	SCORE	DIM_01	DIM_02	DIM_03	DIM_04	DIM_05	DIM_06	DIM_07
COMPRARE	2,4135	0,05757	-0,15778	0,08281	-0,08989	-0,08576	0,07044	-0,08966
SCUOLA	1,9655	0,05855	-0,15341	0,08771	-0,0282	-0,00525	-0,01346	0,15626
VENDERE	2,5554	0,05023	-0,15041	0,07249	-0,06669	-0,0675	0,04764	-0,08546
ORO	2,1564	0,06308	-0,13204	-0,11529	0,01154	-0,16365	-0,12421	-0,09885
VOLPE	2,0568	0,06708	-0,12489	-0,13925	-0,00771	-0,12535	-0,11919	-0,05387
MONETA	2,4298	0,07323	-0,12453	-0,17249	0,02925	-0,18515	-0,1509	-0,0996
ABBECEDARIO	2,0376	0,03682	-0,12452	0,0286	-0,04484	-0,04929	-0,02885	-0,06394
METTERE	1,5165	0,09096	-0,1228	-0,0349	0,01581	-0,05712	-0,03528	-0,08775
ELEMOSINA	2,4049	0,04009	-0,12132	-0,03291	-0,01102	0,29355	-0,05452	-0,07979
CASACCA	2,0165	0,03287	-0,12025	0,03493	-0,03131	-0,04331	-0,03943	-0,04491
GUADAGNARE	1,6603	0,04033	-0,1191	0,00867	-0,03861	0,14721	-0,0247	-0,08426
FACCIA	1,5158	0,04531	-0,10974	-0,0262	-0,04967	0,03633	0,00514	-0,05028
LIBRO	1,7748	0,04528	-0,10676	0,09124	-0,01902	0,01383	-0,01235	0,10771
GIORNATA	1,5595	0,0363	-0,10195	0,03057	-0,00422	0,0928	-0,02681	0,05016
IMPARARE	1,4613	0,0349	-0,10065	0,02574	-0,00661	0,0867	-0,02179	-0,02688
LAVORO	1,8	0,03448	-0,09604	-0,00265	-0,02235	0,22335	-0,02911	-0,13521
DURARE	1,7954	0,0354	-0,09309	0,00823	-0,04112	0,16909	-0,03628	-0,11092
GATTO	1,7742	0,06102	-0,09162	-0,12321	0,00409	-0,10505	-0,0915	-0,02182
CHIEDERE	2,2878	0,03884	-0,08656	-0,02959	0,02427	0,24132	-0,04153	-0,05964
RAGAZZO	1,428	0,09211	-0,08636	0,0649	-0,03043	0,04736	0,02155	0,14018
FATICA	1,8179	0,03208	-0,0845	-0,00422	-0,01834	0,17786	-0,03312	-0,11753
MESSA	1,4622	0,05086	-0,08432	-0,0126	-0,01536	-0,04489	-0,03694	-0,09803
SCRIVERE	1,6715	0,03757	-0,0838	0,06838	0,0179	0,00072	0,00021	0,04476
PANE	1,4985	0,05035	-0,08132	-0,00592	-0,0069	0,16231	-0,03458	-0,08598
SVOGLIATO	1,6004	0,03013	-0,08088	0,05561	-0,02905	0,03024	0,00075	0,15719
DIRETTORE	2,4529	0,04748	-0,07993	0,1061	-0,03944	-0,10571	0,18829	-0,1205
CAMPO_DEI_MIRACOLI	1,2891	0,02834	-0,07598	-0,08796	-0,00982	-0,07495	-0,08392	-0,01907
MESE	1,3107	0,03251	-0,07588	0,00137	-0,04464	0,00313	-0,04127	-0,03189
LEZIONE	0,8107	0,02635	-0,07394	0,01894	-0,01441	-0,00232	-0,00694	0,02354
NUOVO	1,1585	0,05099	-0,0737	0,00737	-0,043	-0,01969	0,01458	-0,03029
SOMARO	1,0649	0,03057	-0,07331	0,05605	-0,03607	-0,03268	0,03056	0,05165

N.B.: Il punteggio di **importanza** di ciascun lemma è calcolato sommando i valori assoluti delle sue prime 20 coordinate (cioè gli autovettori), ciascuno moltiplicato per l' autovalore corrispondente.

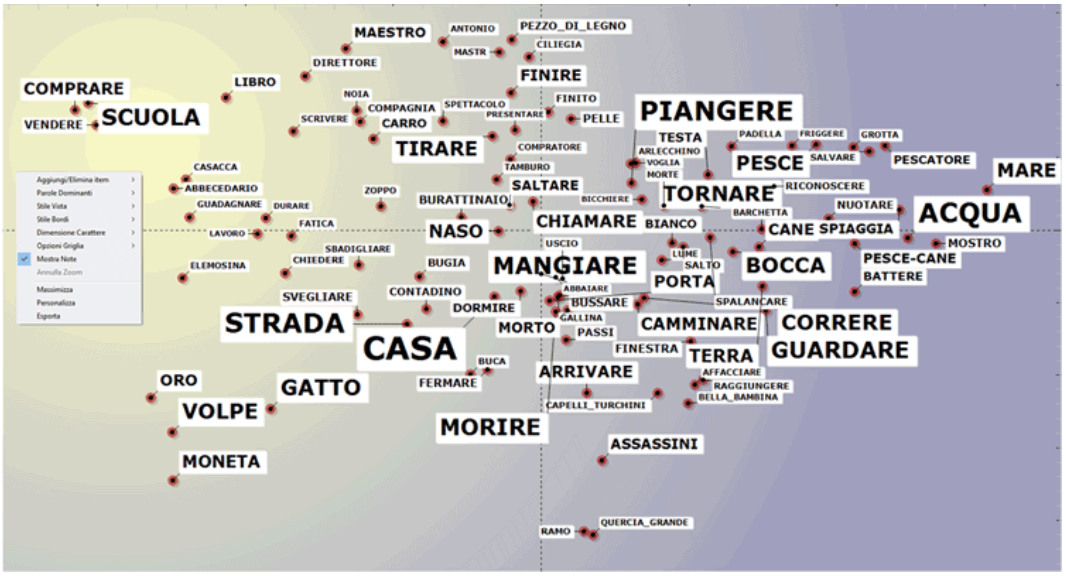


Qualsiasi tabella può essere **ordinata** in ordine crescente o decrescente facendo clic su qualsiasi intestazione di colonna.
Per **esportare** qualsiasi tabella, basta usare il tasto destro del mouse quando vengono visualizzati i relativi dati. Si noti che, la prima volta che viene esportata una tabella di questo tipo, vengono esportati anche gli autovalori. In questo modo l'utente può agevolmente valutare il peso relativo di ogni dimensione, cioè la percentuale di varianza spiegata da ciascuna delle 20 dimensioni.

Facendo clic sul pulsante **Associazioni**, viene visualizzata un'ulteriore tabella con le misure di somiglianza (cioè i coefficienti del coseno) relative ad ogni parola. Inoltre, quando si fa clic su una qualsiasi riga di tale tabella, viene visualizzato un grafico con i dati corrispondenti.



I **grafici** principali mostrano le relazioni tra i termini chiave (cioè i lemmi) sulle dimensioni selezionate (vedi sotto).



Per impostazione predefinita, il grafico di cui sopra include i 100 lemmi più importanti. Tuttavia, l'utilizzatore può personalizzare sia il numero di lemmi sia le caratteristiche del grafico.

OPERAZIONI PRELIMINARI

Preparazione del Corpus

Nel caso di un unico documento (o di un corpus trattato come unico testo) **T-LAB** non richiede ulteriori accorgimenti: basta selezionare l'opzione 'Importare un singolo file…' e procedere (vedi la sezione corrispondente di questo manuale).

Quando, invece, il corpus è costituito da più testi e vengono utilizzate **codifiche** che rinviano all'uso di qualche **variabile**, nella fase di preparazione bisogna utilizzare il modulo **Corpus Builder** che – in maniera automatica – procede alla trasformazione di vari materiali testuali in un file corpus pronto per essere importato da **T-LAB**.

N.B.:

- E' consigliabile una revisione ortografica del materiale da analizzare. Inoltre, se alcune sigle rilevanti sono intervallate da punteggiatura (ad es. "O.N.U." o "M.P.I.") se ne raccomanda la trasformazione in stringhe unitarie (as es. "ONU" o "O_N_U", "MPI" o "M_P_I"); ciò in quanto, nella fase di **normalizzazione**, **T-LAB** interpreta i segni di punteggiatura come separatori;
- Al termine della fase di preparazione si raccomanda di creare una nuova cartella di lavoro con al suo interno il solo file corpus da importare.

Criteri Strutturali

I **criteri strutturali** da rispettare riguardano le **dimensioni** del **corpus** e la sua suddivisone in parti.

Quanto alle dimensioni, tutti gli strumenti **T-LAB** sono stati testati con un corpus di **90** Megabytes, pari a circa 55.000 pagine in formato solo testo.

I limiti per la **grandezza minima** richiedono criteri di valutazione diversi; questo perché - sotto una certa soglia - le dimensioni del corpus possono compromettere l'attendibilità di molte analisi statistiche. A questo proposito, basta attenersi alle seguenti indicazioni: un minimo di 5.000 occorrenze (circa 30 K); oppure, nel caso di risposte a domande aperte, un minimo di 50 risposte. In quest'ultimo caso, infatti, ogni risposta costituisce una diversa unità di contesto.

Ai fini del trattamento, il corpus può essere costituito da un unico testo senza ulteriori partizioni, da un unico testo ripartito secondo criteri stabiliti dall'utilizzatore (ad es. un libro suddiviso in capitoli), da più testi (ad es. diverse interviste o risposte a domande aperte) classificati attraverso l'uso di etichette che rinviano ad altrettante **variabili** o **IDnumber**. In tutti questi casi, il corpus è suddiviso in parti che devono essere individuate con precisi **criteri formali**.

Criteri Formali

Nel caso di un **corpus** costituito da **unico testo**, e comunque quando l'utilizzatore non fa ricorso all'uso di variabili, **non sono richiesti altri tipi di interventi** e si può passare direttamente alla fase di **importazione**.

Quando invece il corpus è costituito da più testi e/o si fa uso di variabili, la preparazione del corpus va essere realizzata tramite il modulo **Corpus Builder** che, in modo automatico, rispetta i seguenti criteri:

Ogni testo o sottoinsieme di esso (le "parti" individuate da variabili e/o Idnumber) è preceduto da **una riga di codifica**.

Ogni riga di codifica ha il seguente formato:

- **Inizia** con una stringa di **quattro asterischi** (****) seguita da uno spazio (blank). Da **T-LAB** questa stringa viene interpretata nel modo seguente: "qui inizia un testo o una unità di contesto definita dall'utilizzatore";
- **Continua**, con l'aggiunta di stringhe costituite da **singoli asterischi** ed etichette che individuano **casi** (**IDnumber**), **variabili** e rispettive **modalità**.
- **Termina** con un ritorno di carrello ("a capo").

Ecco qualche esempio.

La riga seguente introduce un testo (o parte del corpus) codificato con tre variabili - ETA (età), SES (sesso) e PROF (professione) - e relative modalità (ADUL, FEM, OPER)

**** *ETA_ADUL *SES_FEM *PROF_OPER

La riga seguente introduce un testo (o parte del corpus) codificato con le stesse variabili e con l'etichetta **IDnumber**

**** *IDnumber_0001 *ETA_ADUL *SES_FEM *PROF_OPER

La riga seguente introduce un testo (o parte del corpus) codificato con due variabili: ANNO, TEST (testata giornalistica)

**** *ANNO_98 *TEST_REPUB

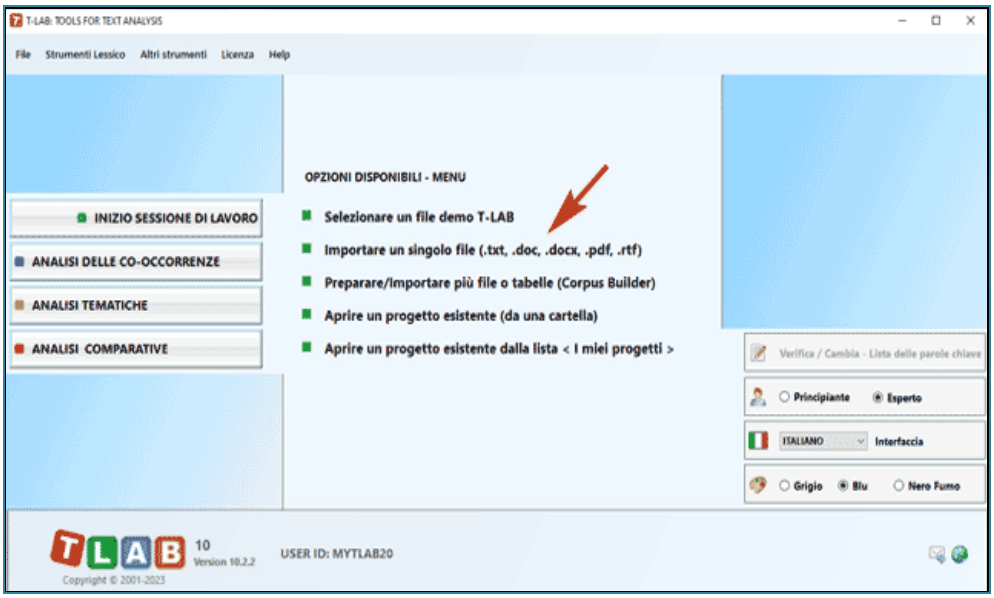
In ogni riga di codifica, le norme T-LAB da rispettare sono le seguenti**:**

- ogni etichetta (IDnumber, variabile o modalità) non deve essere intervallata da spazi vuoti;
- ogni etichetta, sia nel caso delle variabili che delle modalità, non deve superare la lunghezza di **25** caratteri (min. 2);
- ogni etichetta delle variabili va congiunta alla rispettiva modalità attraverso l'uso del trattino basso "_" (underscore);
- tra una variabile e l'altra, cioè prima del successivo asterisco, va inserito uno spazio vuoto (blank);
- per ogni parte del corpus, la riga di codifica deve includere tutte le variabili usate;
- il numero massimo di variabili utilizzabili è 50, quello delle modalità (per ogni variabile) è di 150;
- il numero massimo di IDnumber è fissato a 99.999 per i testi brevi (Max. 2.000 caratteri ciascuno, es. risposte a domande aperte, messaggi twitter etc.) e a 30.000 per gli altri casi.

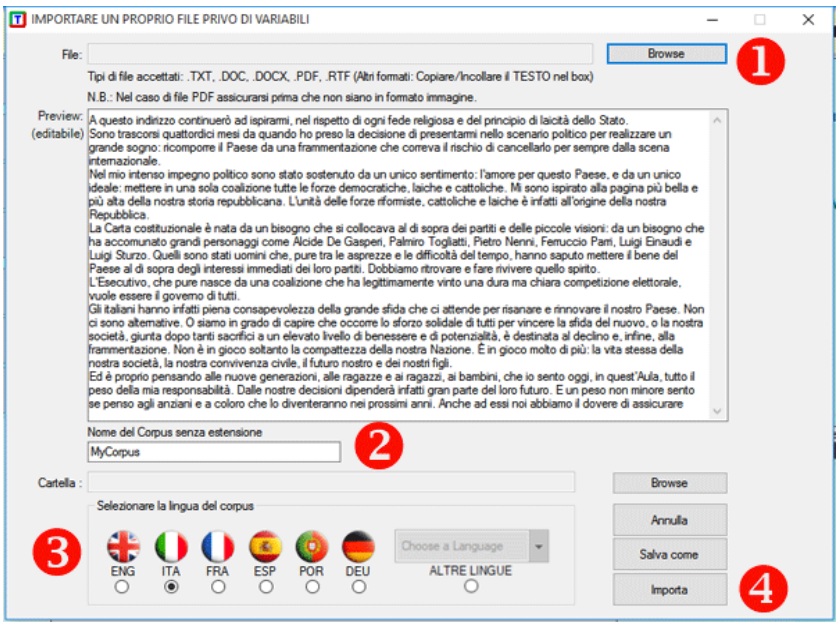
OPERAZIONI SUI FILE

Importare un singolo file..

Nel caso di un unico testo (o di un corpus trattato come unico testo) **T-LAB** non richiede ulteriori accorgimenti: basta selezionare l'opzione 'Importare un singolo file…' (vedi sotto).

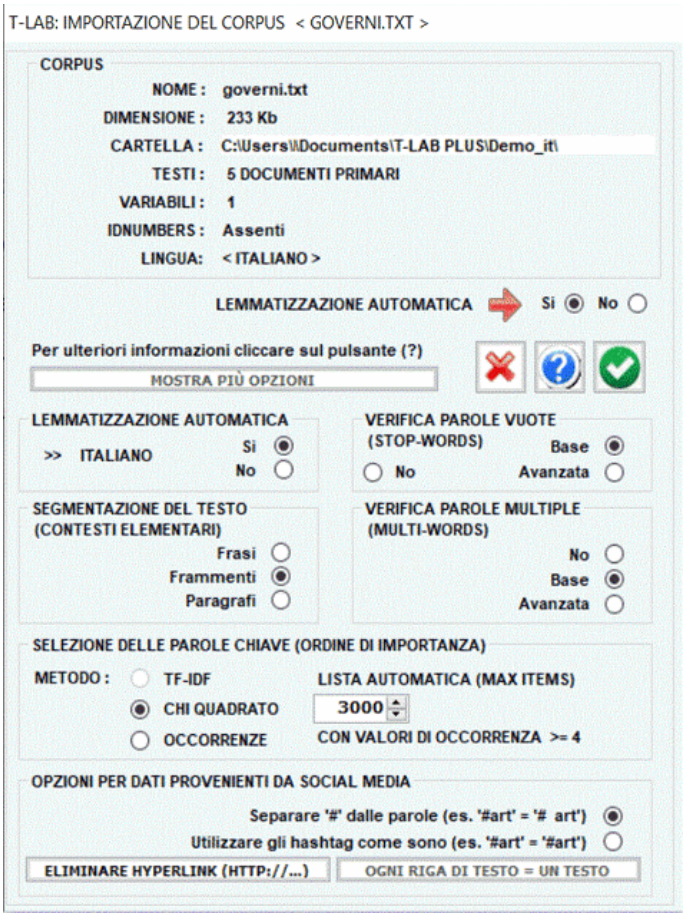


Quindi si richiedono quattro passaggi (vedi immagine seguente) : (1) selezionare un qualsiasi file; (2) scegliere il nome del progetto; (3) selezionare la lingua del testo; (4) cliccare su 'Importa' .



Successivamente compare una finestra di riepilogo (vedi immagine seguente) in cui possono essere effettuate alcune scelte.

N.B.:
- Poiché i trattamenti preliminari determinano il tipo e la quantità delle unità di analisi (cioè quali e quante unità di contesto e quali e quante unità lessicali), scelte diverse in questa fase comportano risultati diversi delle successive analisi (vedi sotto opzioni avanzate). Per questa ragione, tutti gli output **T-LAB** mostrati nel manuale e nell'help hanno solo valore indicativo;
- Tutte le fasi di pre-processing vengono eseguite durante l'importazione di qualsiasi tipo di corpus.



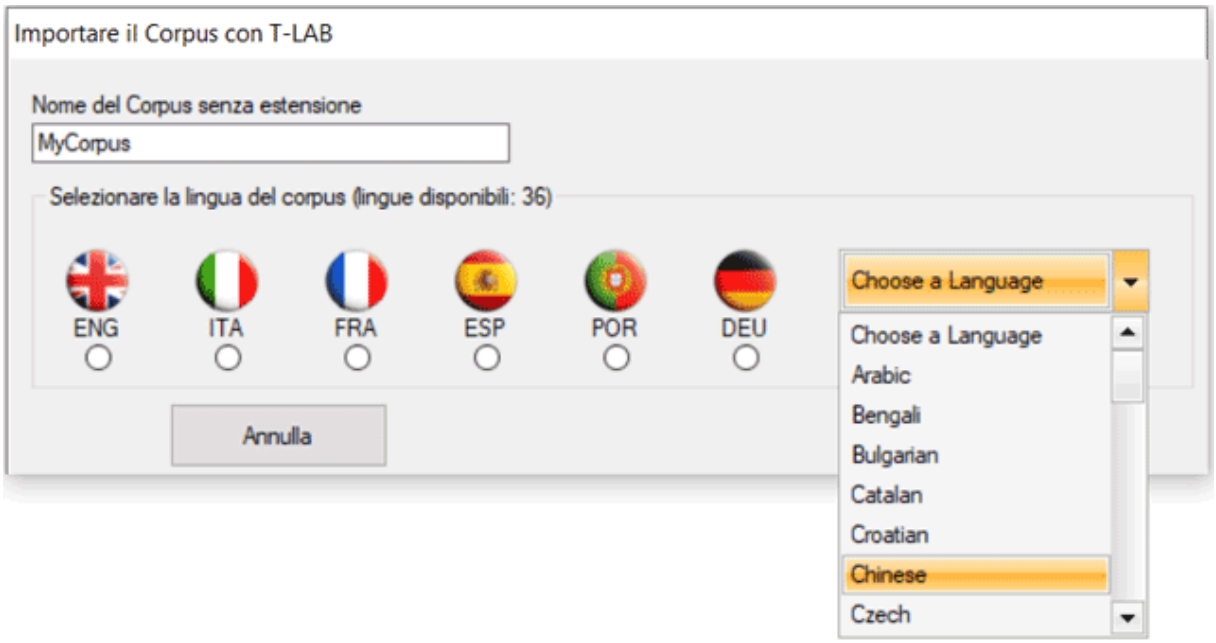
1 - LEMMATIZZAZIONE AUTOMATICA O STEMMING

Di seguito la lista complete delle trenta (30) lingue per le quali **T-LAB** supporta la lemmatizzazione automatica o lo stemming.

LEMMATIZZAZIONE: catalano, croato, francese, inglese, italiano, latino, polacco, portoghese, rumeno, russo, serbo, slovacco, spagnolo, svedese, tedesco, ucraino.

STEMMING: arabo, bengali, bulgaro, ceco, danese, finlandese, greco, hindi, indonesiano, marathi, norvegese, olandese, persiano, turco, ungherese.

In ogni caso, senza lemmatizzazione automatica e/o usando dizionari personalizzati, possono essere analizzati testi in tutte le lingue le cui parole siano separate da spazi e/o da punteggiatura.



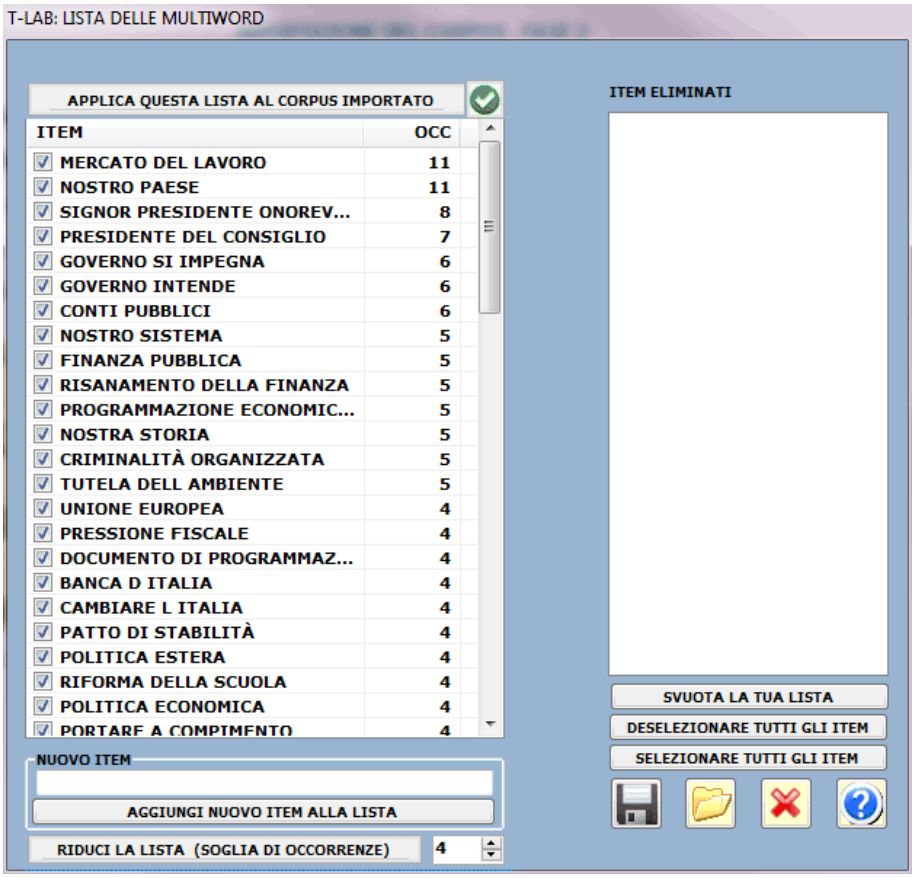
Il risultato del processo di lemmatizzazione può essere verificato tramite la funzione **Vocabolario** e può essere modificato tramite la funzione **Personalizzazione del Dizionario**.

2 - SEGMENTAZIONE DEI TESTI IN CONTESTI ELEMENTARI

A seconda della scelta dell'utilizzatore, i **contesti elementari** per il calcolo delle **co-occorrenze** possono essere di quattro tipi: frasi, frammenti di lunghezza comparabile, paragrafi e testi brevi (es. risposte a domande aperte).
Il risultato del processo di segmentazione può essere verificato tramite il file corpus_segments.dat.

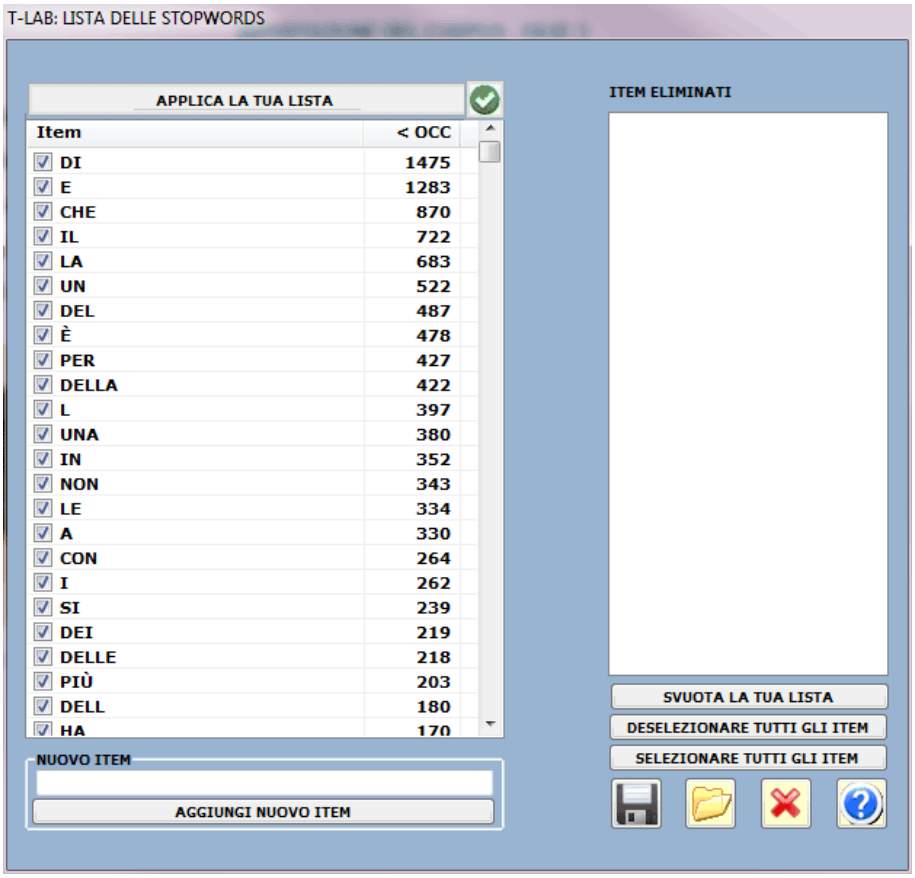
3 - VERIFICA DELLE PAROLE MULTIPLE (MULTI-WORDS)

L'opzione "**Base**" abilita l'uso automatico della lista **multi-words** di T-LAB.
Diversamente, l'opzione "**Avanzata**", abilitata solo in caso di lemmatizzazione automatica, consente di verificare e modificare la lista delle multi-words presenti nel corpus e non incluse nel dizionario **T-LAB** (vedi immagine seguente). Inoltre è possibile importare e usare altre **liste predisposte dall'utilizzatore** (file Multiwords.txt).



4 - VERIFICA DELLE PAROLE VUOTE (STOP-WORDS)

L'opzione "**Base**" abilita l'uso automatico della lista **parole vuote** di T-LAB.
Diversamente, l'opzione "**Avanzata**" consente di verificare e modificare la lista delle **parole vuote** presenti nel corpus.
Inoltre è possibile importare e usare altre **liste predisposte dall'utilizzatore** (file Stopwords.txt).



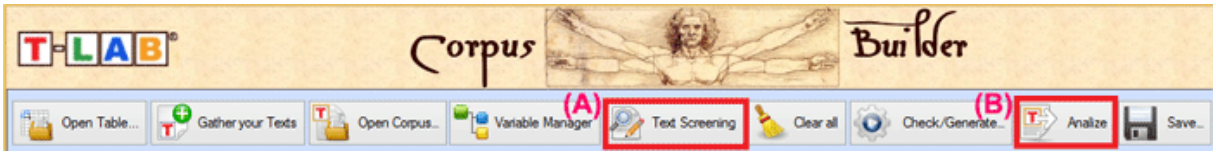
5 - SELEZIONE DELLE PAROLE CHIAVE

Le opzioni disponibili consentono di scegliere il metodo di selezione (**TF-IDF** o **Chi-quadro**) e il numero massimo di **unità lessicali** da includere nella lista usata da **T-LAB** per analizzare i testi con **impostazioni automatiche**.

N.B.: Al termine della fase d'importazione, mediante le **impostazioni personalizzate,** l'utilizzatore può rivedere la selezione delle parole e costruire varie **liste** da applicare.

Preparare un Corpus (Corpus Builder)

N.B.: Le immagini di questa sezione fanno riferimento a una versione precedente di T-LAB. In **T-LAB 10** questo strumento include due pulsanti aggiuntivi: a) uno che, per corpus di dimensioni non superiori a 20 MB, attiva l'opzione **Text Screening**; b) l'altro che consente di procedere immediatamente con l'**importazione** dei materiali testuali selezionati (vedi immagine seguente).



Questo strumento software è stato progettato per facilitare la preparazione e la trasformazione di vari materiali testuali in un file **corpus** pronto per essere importato da **T-LAB**. Più specificamente, tale strumento consente di eseguire rapidamente le seguenti operazioni:

1. **Importare** automaticamente vari tipi di file;
2. **Editare** e modificare i testi dei file importati;
3. Gestire l'uso di **variabili categoriali**;
4. **Salvare** il risultato del lavoro in un file pronto per essere importato da **T-LAB**;
5. **Verificare** e **modificare** qualsiasi file corpus che corrisponda al formato richiesto da **T-LAB**.

Mentre il modo di importare i file (vedi sopra '1') varia in base al loro formato, tutte le altre operazioni seguono la stessa logica.

Di seguito una breve descrizione dei modi per importare i vari tipi di file.

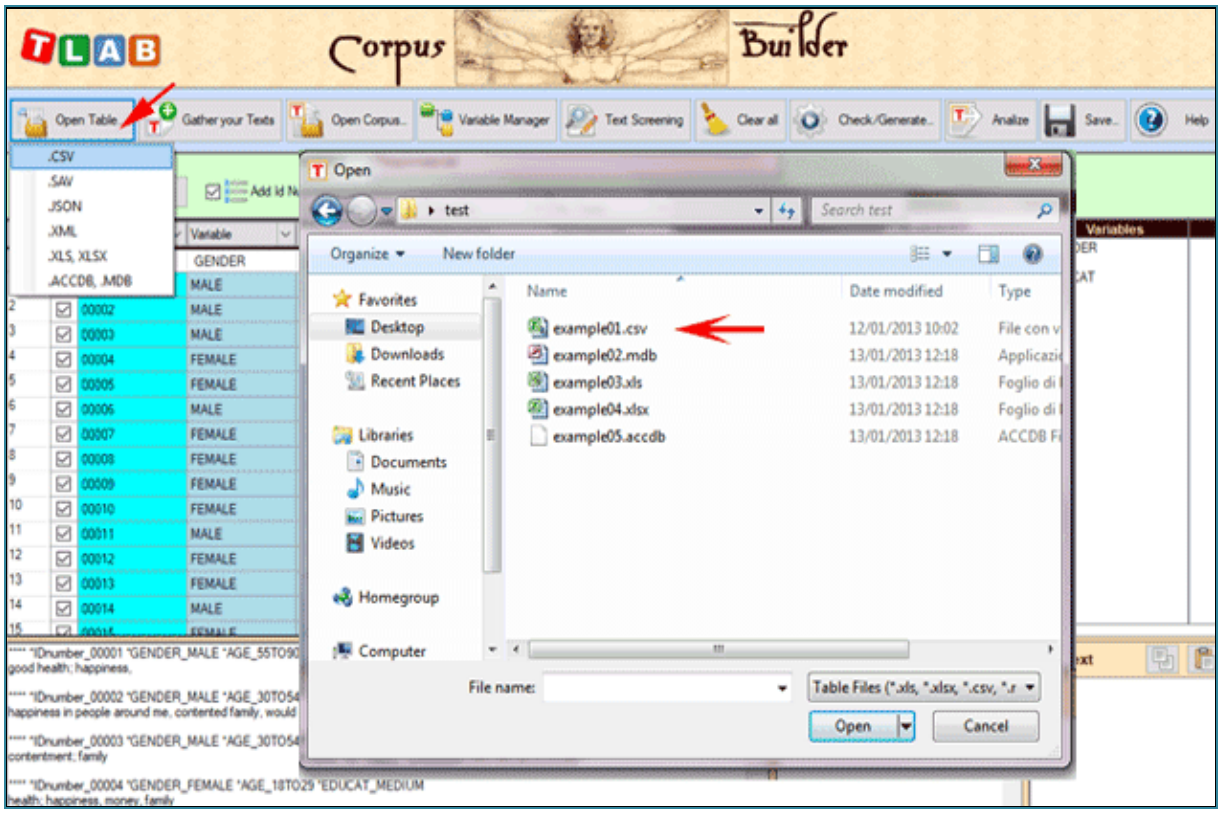
A - **Importazione di file in formato tabellare** (CSV, .SAV, .JSON, .XML, .XLS, XLSX, .MDB, .ACCDB).

Un **singolo file** che includa fino a 30.000 record può essere importato usando l'opzione 'Open Table' o tramite il metodo drag and drop (N.B.: quando nessuno dei testi supera i 2.000 caratteri, il limite dei record da importare è esteso a 99.999).

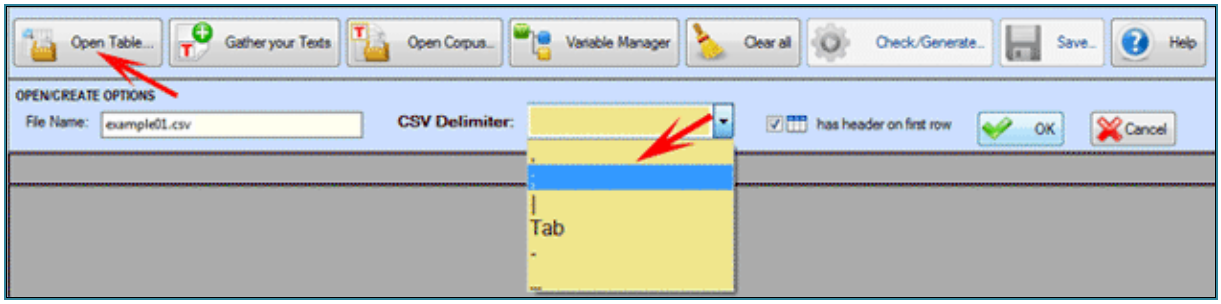
Tale file può essere costituito da varie colonne contenenti i seguenti dati:

- Variabili categoriali (una per ogni colonna, fino a un massimo di 50)
- Testi da analizzare (una sola colonna);
- IDnumbers, cioè identificativi di unità di contesto o di soggetti/casi.

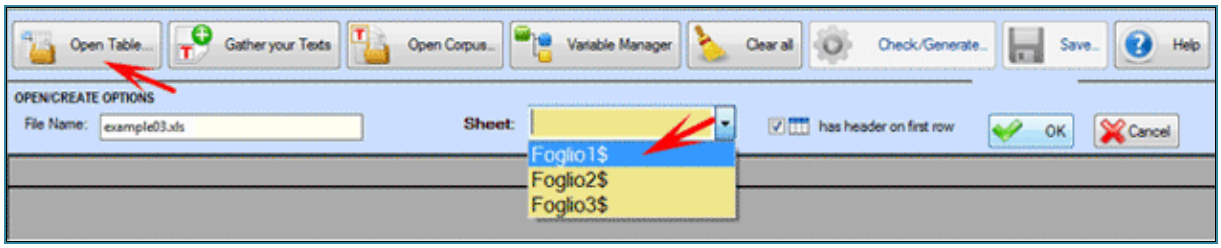
N.B.: Mentre la presenza di variabili categoriali e IDnumbers è opzionale, la presenza di almeno una colonna contenente i testi da analizzare è obbligatoria.



Quando viene importato un file .CSV, deve essere opportunamente selezionato il delimitatore usato (vedi sotto).



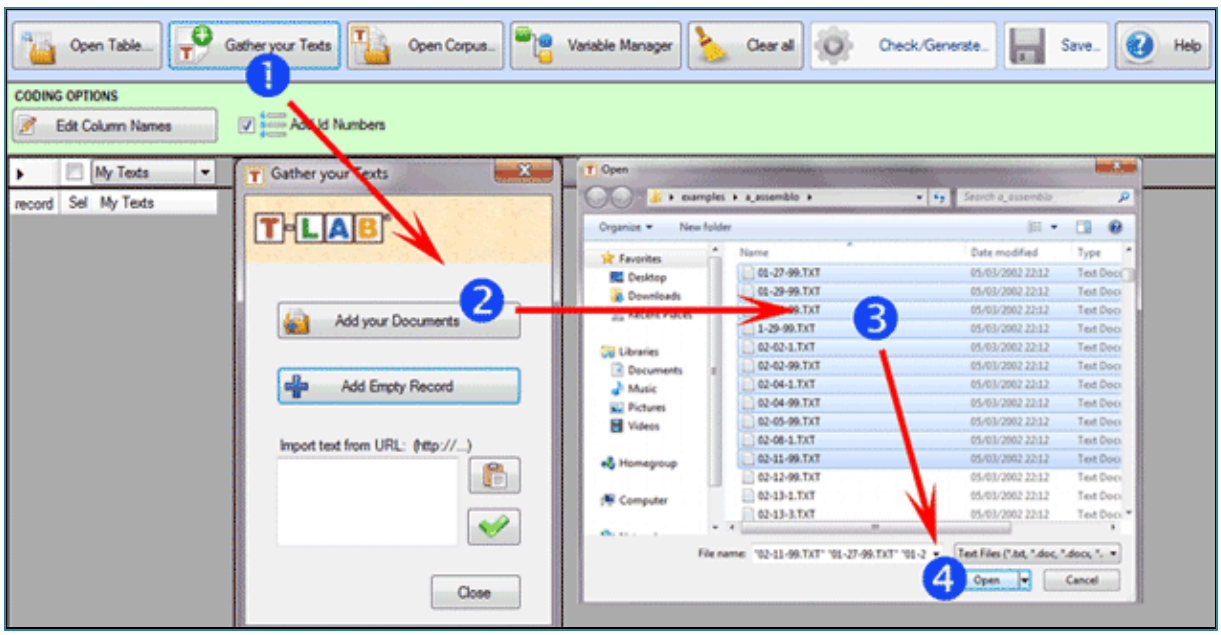
Quando vengono importati file Excel o Access, è possibile selezionare solo una tabella (vedi sotto).



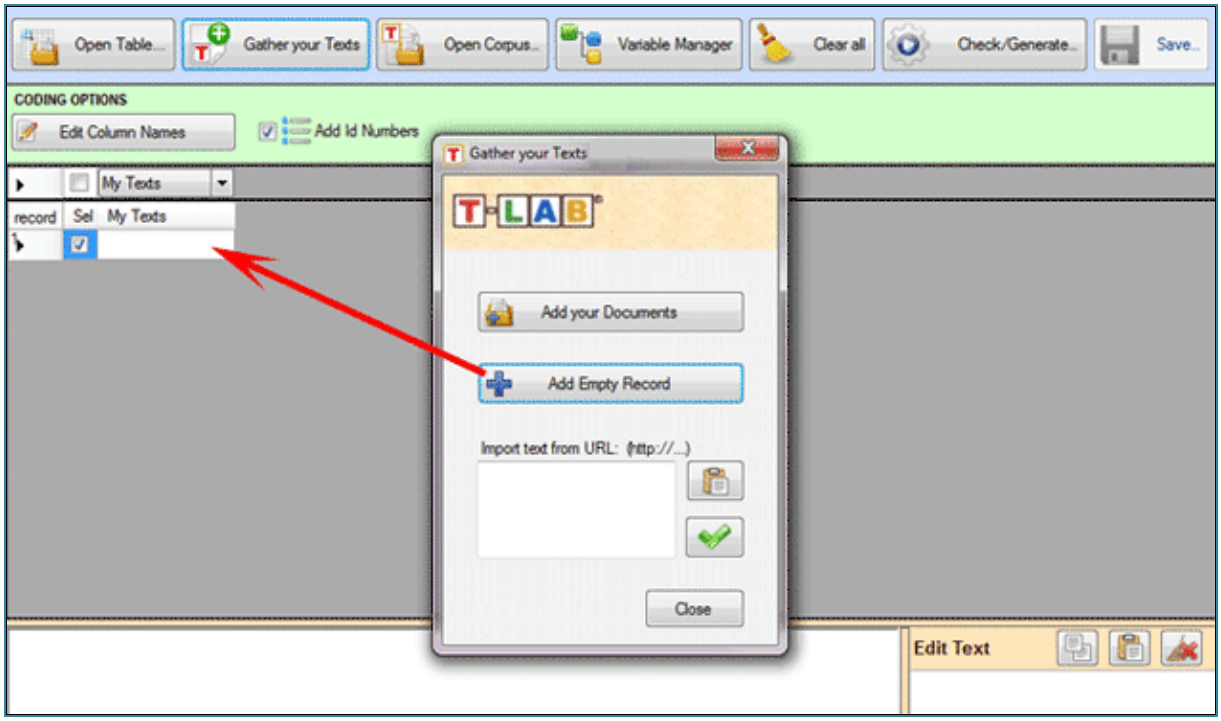
B - **Importazione testi e documenti in vari formati** (.TXT, .DOC, .DOCX, .PDF, .RTF, .HTML).

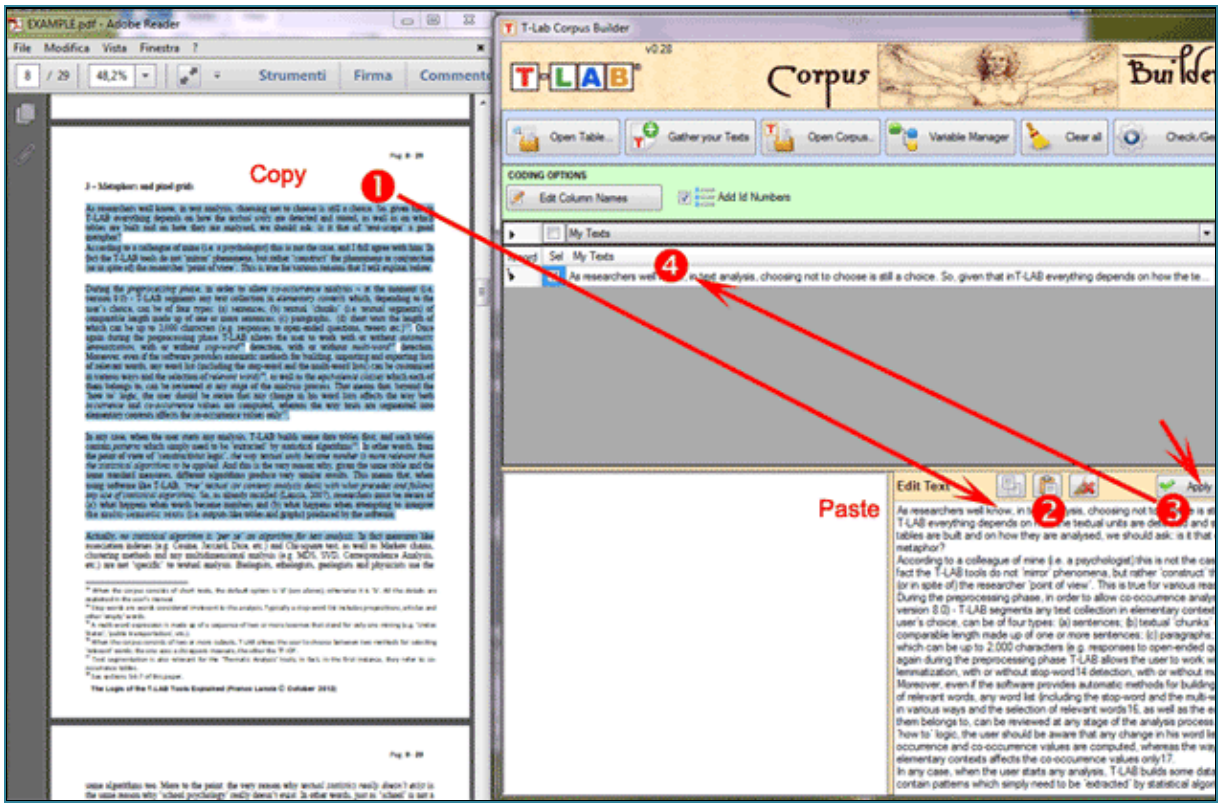
L'opzione 'Gather your Texts' (vedi sotto) consente di importare fino a 30.000 documenti, sia uno per volta che tramite selezione multipla, utilizzando **tre diversi metodi**.

Il **primo metodo** ('Add your Documents') prevede l'importazione automatica di file tipo .TXT, .DOC, .DOCX, .PDF, .RTF.



Il **secondo metodo** ('Add EmptyRecord') consente di aggiungere singoli record in cui è possibile copiare/incollare qualsiasi tipo di testo (vedi sotto).



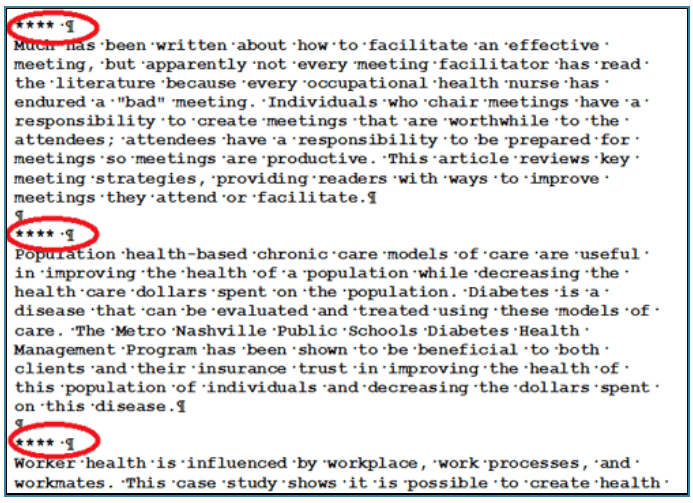


Il **terzo metodo** (‘Import Text from URL’).consente di scaricare direttamente singoli file HTML da internet, di editarne il contenuto per eventuali modifiche e - quindi - di importarli (vedi sotto).

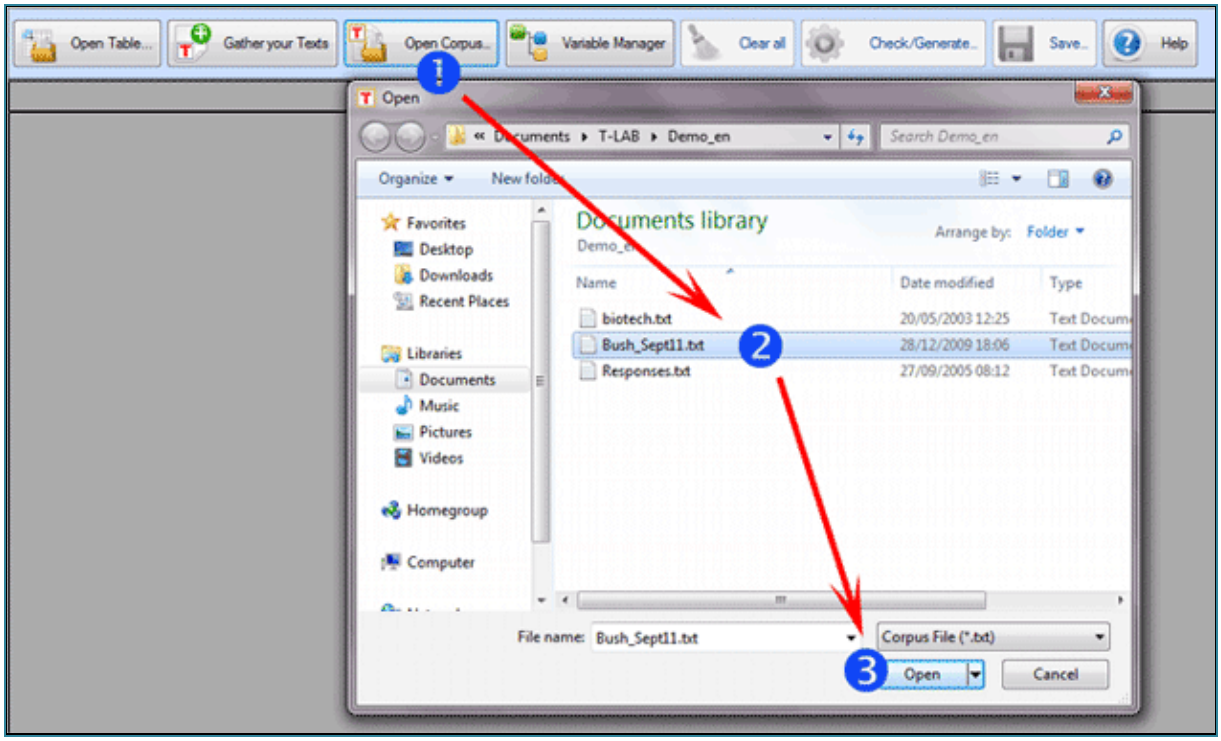
C - **Importazione di un corpus** già codificato secondo le specifiche di **T-LAB.**

Si consiglia l’uso dell’opzione ‘Open Corpus’ in tre tipi di casi:

1 – l’utilizzatore intende modificare la struttura di un file corpus già codificato (es. aggiungere degli altri testi tramite i metodi spiegati nella precedente sezione ‘B’, modificare le denominazioni delle variabili e/o delle modalità, etc.);
2 – l’utilizzatore intende verificare/correggere gli eventuali errori contenuti in una codifica del corpus effettuata manualmente e senza l’ausilio del modulo Corpus Builder;
3 – l’utilizzatore intende importare un file corpus con una codifica ‘grezza’ (vedi immagine seguente), cioè un file corpus le cui parti (documenti o record) siano tutte precedute solo da una riga con quattro asterischi seguiti da uno spazio (‘**** ’).



In tutti i tre casi sopra menzionati (1,2,3) è sufficiente selezionare un singolo file tramite l'opzione 'Open Corpus' o trascinarlo con il metodo drag and drop (vedi sotto).

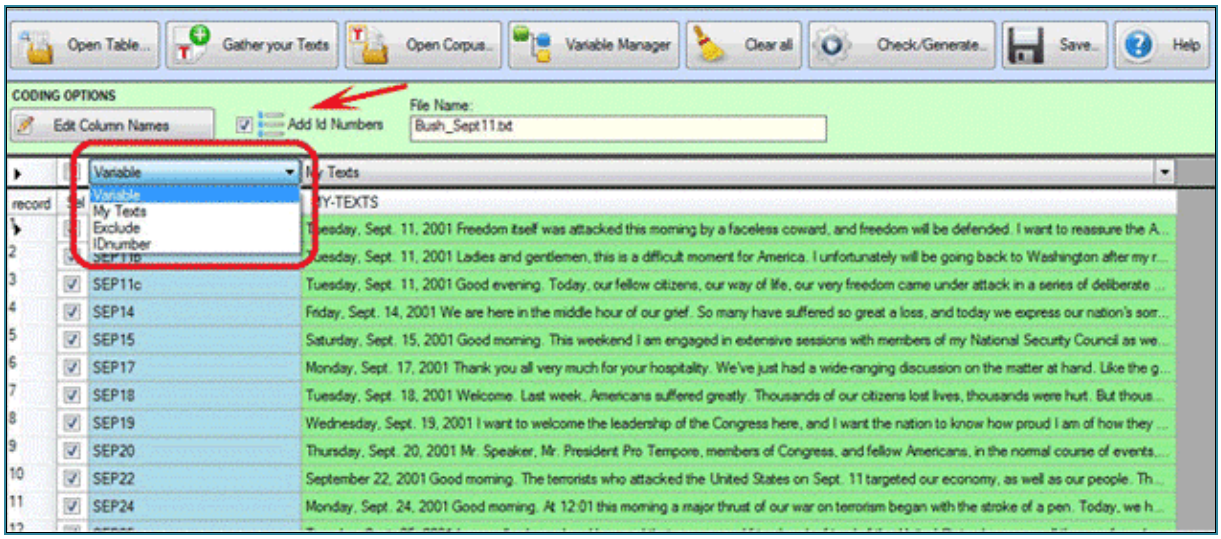


Operazioni successive all'importazione dei file

Al termine della fase attraverso la quale i file sono stati importati in Corpus Builder, sia nel caso in cui 'non' si sia interessati all'uso di variabili, sia nel caso in cui le operazioni di codifica siano state già effettuate, si può procedere con l'opzione 'Check /Generate' e – successivamente – con l'importazione del corpus in **T-LAB**.

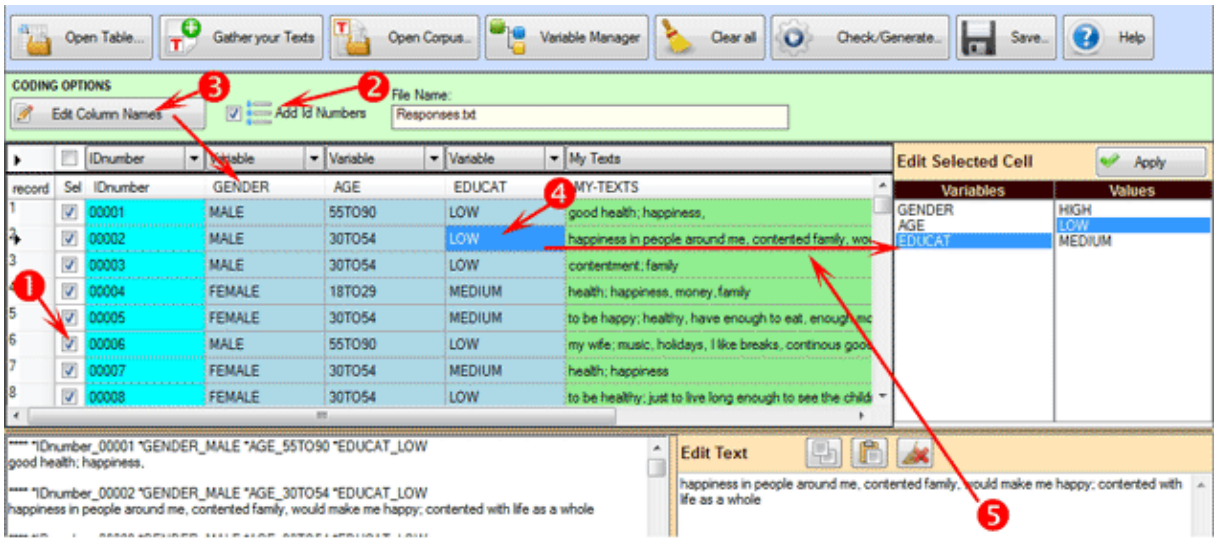
Quando il corpus contenga codifiche, va tenuto presente che in tutti e i tre i tipi di importazione menzionati nelle precedenti sezioni di questo documento ('A', 'B', 'C'), i dati vengono visualizzati in diverse colonne, le cui intestazioni possono essere le seguenti:

- 'Variable', cioè variabili categoriali, il cui uso è necessario quando si intendano analizzare le caratteristiche e le relazioni di distinti sottoinsiemi del corpus;
- 'IDnumber', cioè identificatori di casi / record, il cui uso è opzionale;
- 'My Texts', cioè il testi da analizzare, il cui uso è possibile in una sola colonna ed è obbligatorio;
- 'Exclude', da usarsi per segnalare a Corpus Builder che i dati contenuti nella corrispondente colonna non vanno utilizzati.



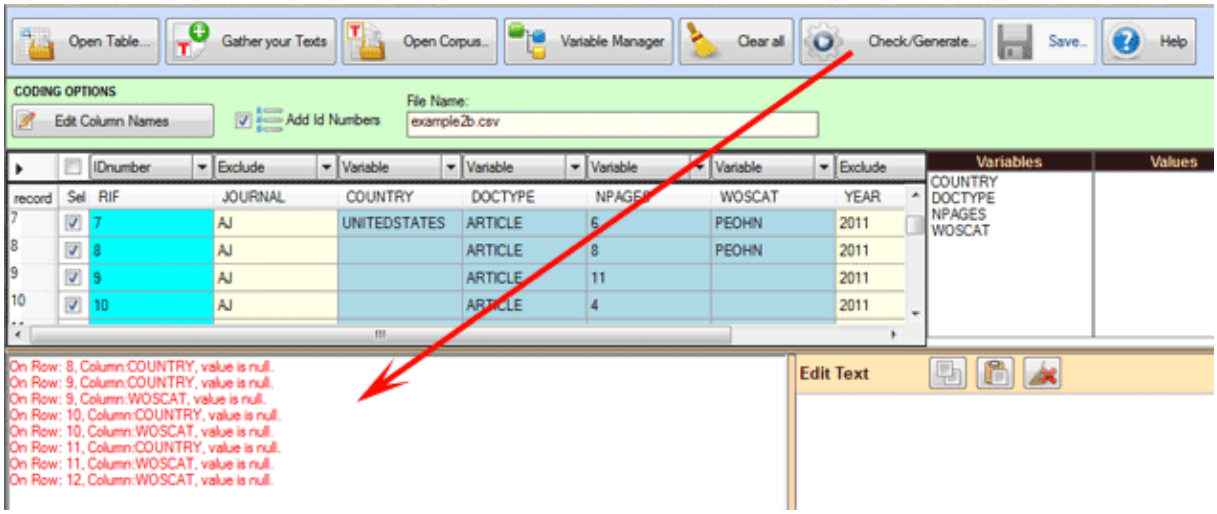
In **tutti i casi**, valgono le seguenti indicazioni:

- ogni record può essere selezionato o deselezionato (vedi sotto '1 ');
- gli IDnumber possono essere aggiunti automaticamente (vedi sotto '2');
- i nomi delle variabili possono essere editati e modificati (vedi sotto '3');
- ogni valore di variabile può essere editato e modificato (vedi sotto '4');
- ogni campo 'My Texts' può essere editato e modificato (vedi sotto '5').



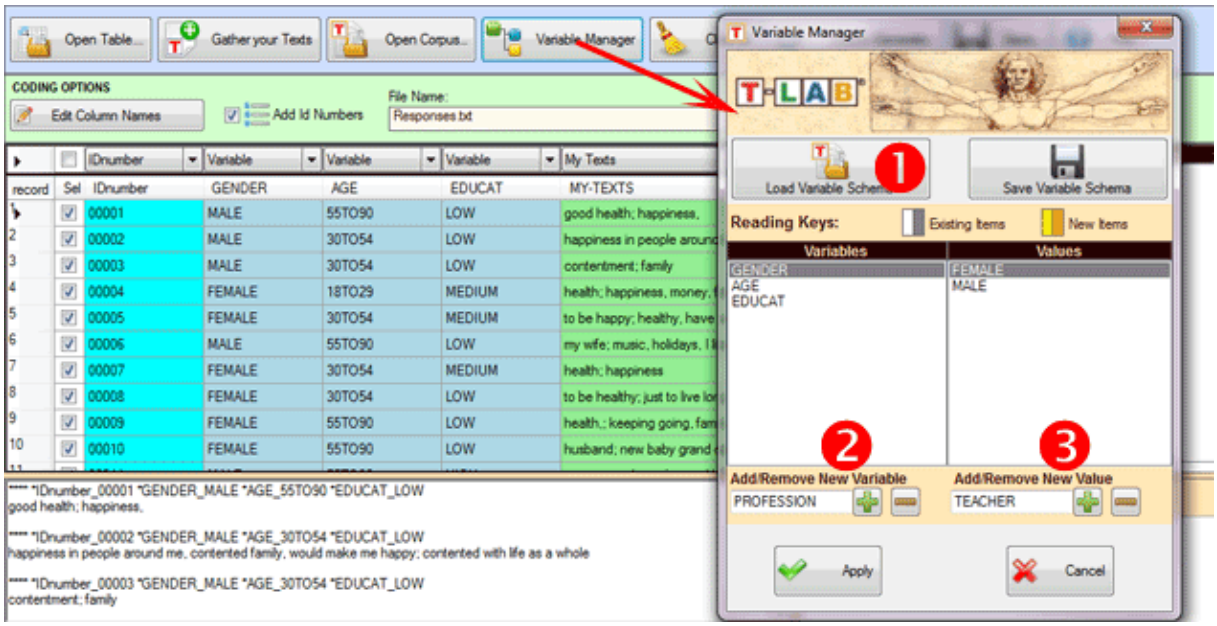
Si ricorda inoltre che:

- Il numero delle colonne con variabili categoriali non deve superare i 50, e ciascuna di esse deve avere minimo 2 massimo 150 valori;
- I valori degli IDnumber, se usati, devono essere progressivi a partire da 1 (es., 1, 2, 3, etc.);
- Ogni etichetta, sia nel caso delle variabili che delle modalità, non deve superare la lunghezza di 25 caratteri alfanumerici (min. 2) e non deve essere intervallata da spazi vuoti;
- Nel modulo Corpus Builder tutti gli errori rilevati vengono visualizzati nel box in basso a sinistra (vedi sotto).



Uso dello strumento Variable Manager

Lo strumento 'Variable Manager' consente di costruire, editare, modificare e salvare qualsiasi schema di codifica, anche proveniente da un corpus diverso.
Ogni schema include l'elenco delle variabili e quello dei rispettivi valori (vedi sotto).

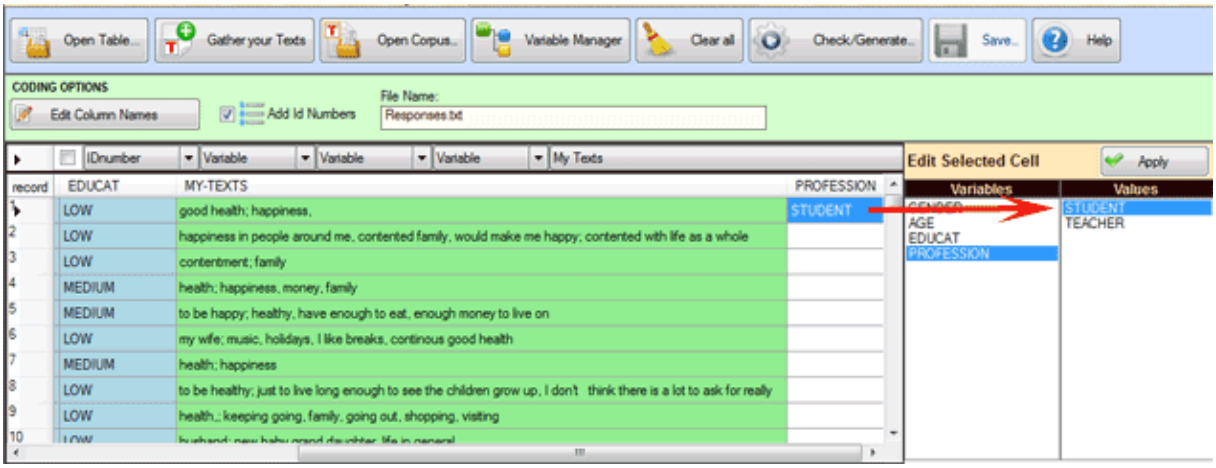


Per aggiungere variabili provenienti da un altro corpus o da uno schema precedentemente salvato, bisogna selezionare l'opzione '1' (vedi sopra). Diversamente, per aggiungere manualmente variabili e relativi valori, bisogna usare in sequenza l'opzione '2' e l'opzione '3' (vedi sopra).

L'aggiunta di valori di variabili a singoli record va effettuata manualmente (vedi sotto) e in **un'unica sessione di lavoro**; questo perché il salvataggio dello schema non include le codifiche attribuite a ciascun record. Nel caso quindi l'utilizzatore si trovi a codificare manualmente un corpus che includa un numero considerevole di record e/o il lavoro richieda più di una sessione di lavoro, si raccomanda di procedere come segue:

1 - importare la quantità di file/record che si ritiene di poter codificare in un'unica sessione di lavoro;
2 - salvare il lavoro come un corpus (vedi opzione 'Save' del menu Corpus Builder).

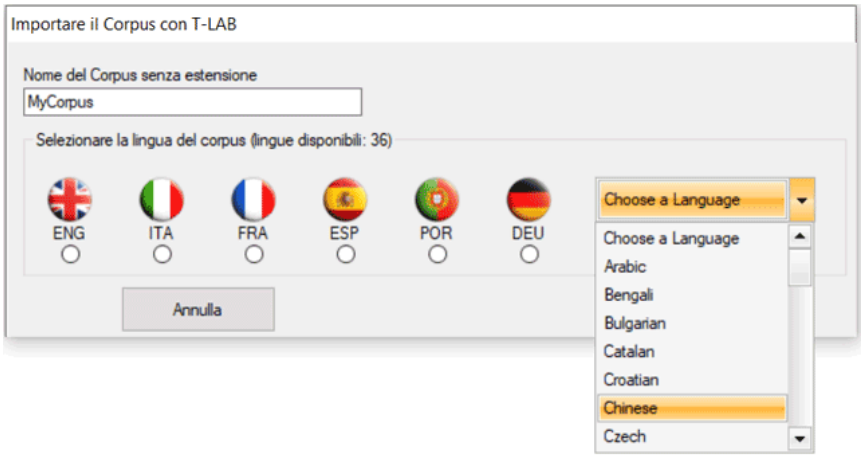
Quindi, nella successiva sessione, reimportare il corpus salvato in precedenza (vedi sopra, punto '2'), aggiungere altri record/file da codificare e continuare.



Quando l'utilizzatore ha completato le operazioni che ritiene opportune, l'opzione 'Check/Generate' consente di verificare la loro correttezza e, se tutto è ok, è possibile esportare (A) o salvare (B) un corpus pronto per essere importato da **T-LAB.**

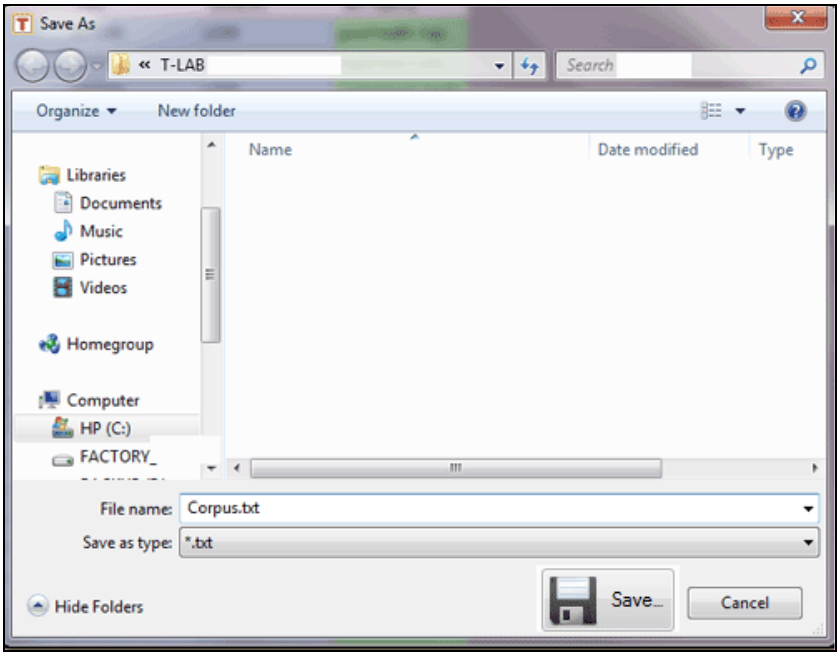
Nel primo caso (A – vedi sotto) Corpus Builder crea una nuova cartella nella directory '..\Miei Documenti\T-LAB PLUS\" e – automaticamente – avvia la procedura di importazione T-LAB.

N.B.: In questo caso, la nuova cartella che viene creata ha lo stesso nome del file corpus.



Nel secondo caso (B – vedi sotto) l'utilizzatore può salvare il corpus nella directory che preferisce e – successivamente – usare l'opzione 'Importa un corpus' del menu **T-LAB**.

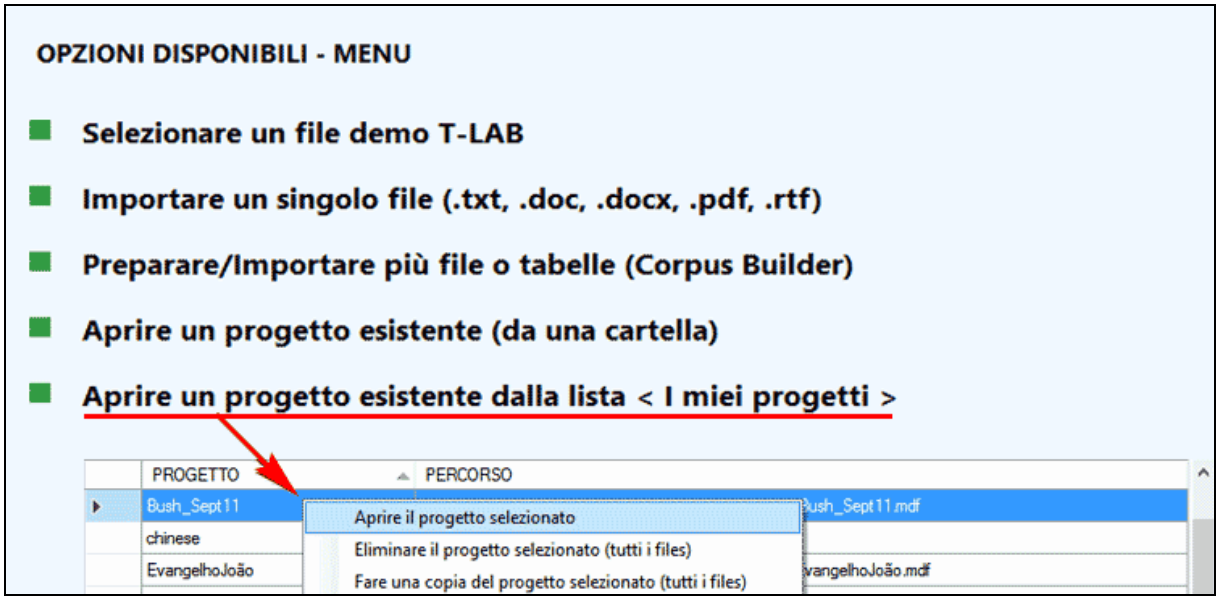
N.B.: In questo caso, si raccomanda di creare – ogni volta - una nuova cartella di lavoro con al suo interno il solo file corpus da importare.



Aprire un progetto esistente

Attraverso questa opzione l'utilizzatore può tornare a lavorare su un progetto già avviato, sia selezionando il file da una cartella esistente o dalla lista predisposta da T-LAB.

Inoltre, quando viene selezionato un item dalla lista predisposta da T-LAB, l'uso del tasto destro abilita l'utilizzatore a eliminare i relativi file o a farne un backup in un'altra cartella.



OPERAZIONI SUL LESSICO

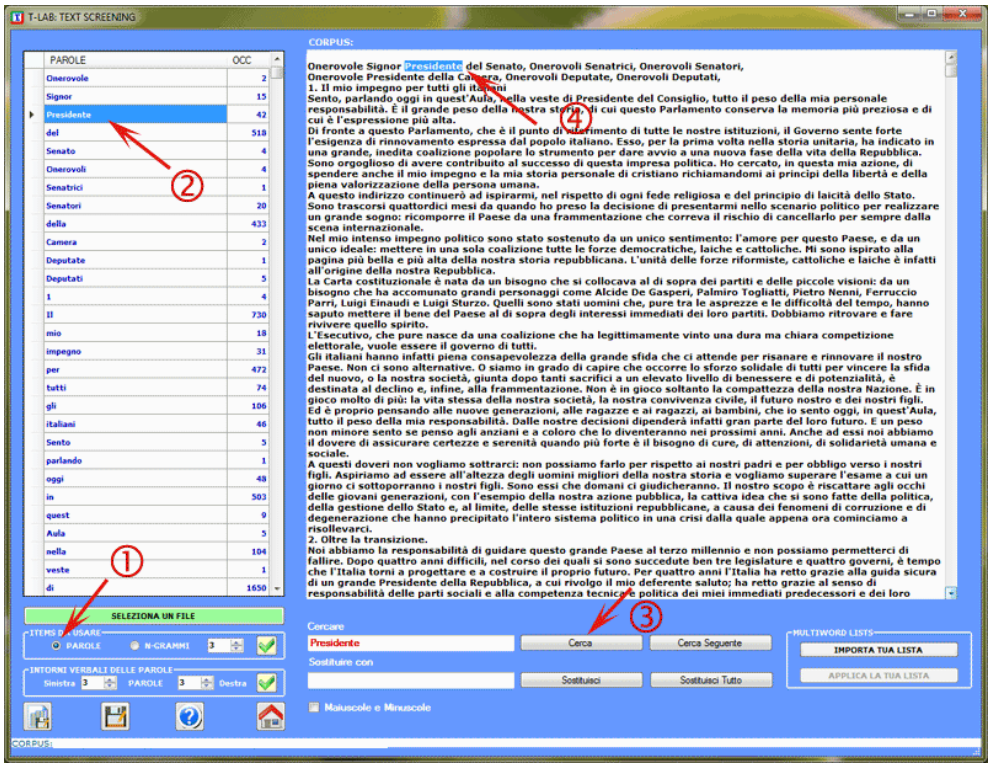
Text Screening / Disambiguazioni

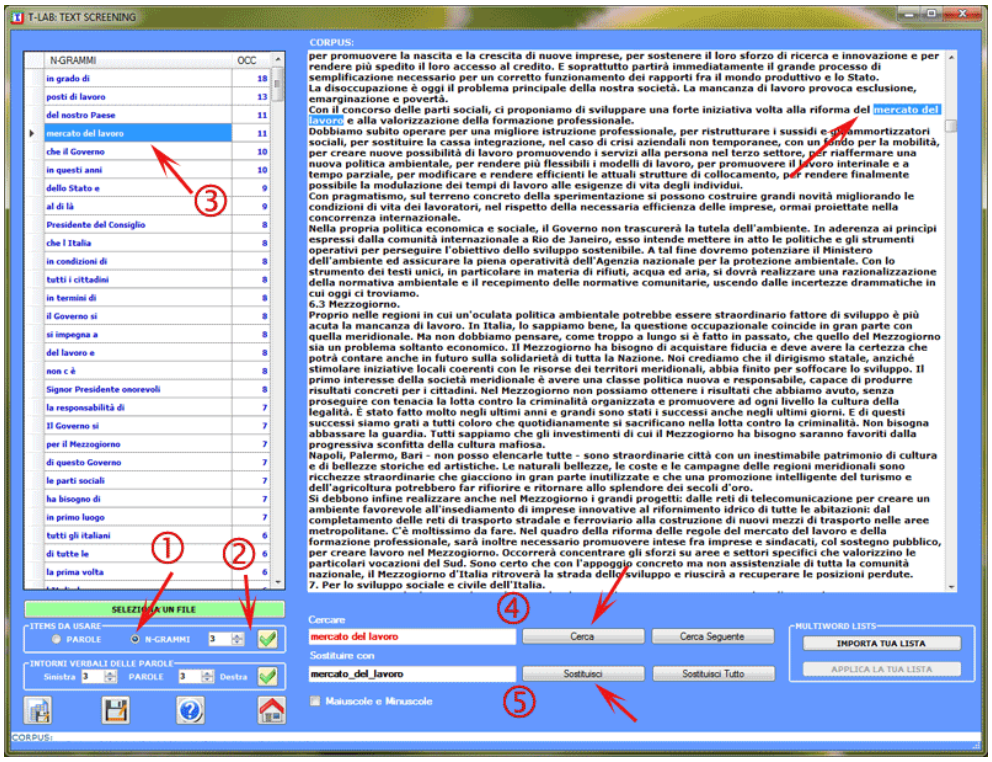
Questo strumento **T-LAB** consente di editare un qualunque file corpus (dimensione massima 30 Mb) e di effettuare **rapidamente** una serie di operazioni utili sia per una prima **esplorazione** dei suoi contenuti che per la **disambiguazione** di sue specifiche unità lessicali.

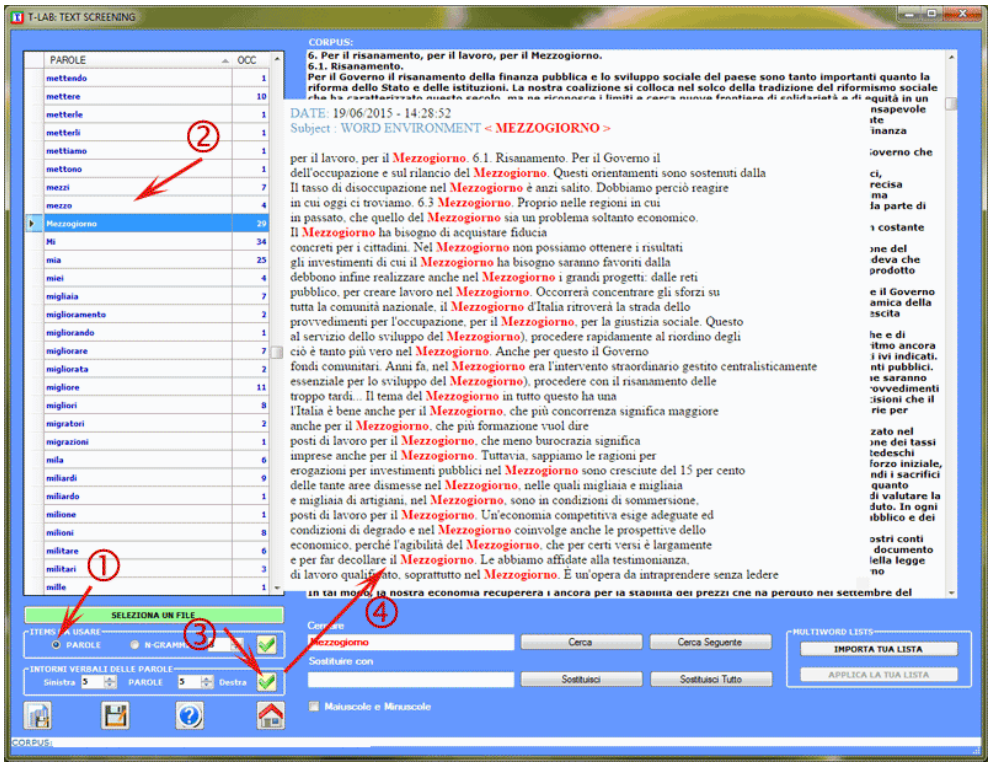
In particolare questo strumento produce rapidamente una serie di **liste** e consente operazioni del tipo **cerca**/**sostituisci**.

Le liste ottenibili sono le seguenti:
a- **parole** con le loro occorrenze;
b- **n-grammi** di parole con le loro occorrenze;
c- **intorni verbali** delle parole selezionate.

Le immagini seguenti illustrano le operazioni possibili nei tre casi (a-b-c).



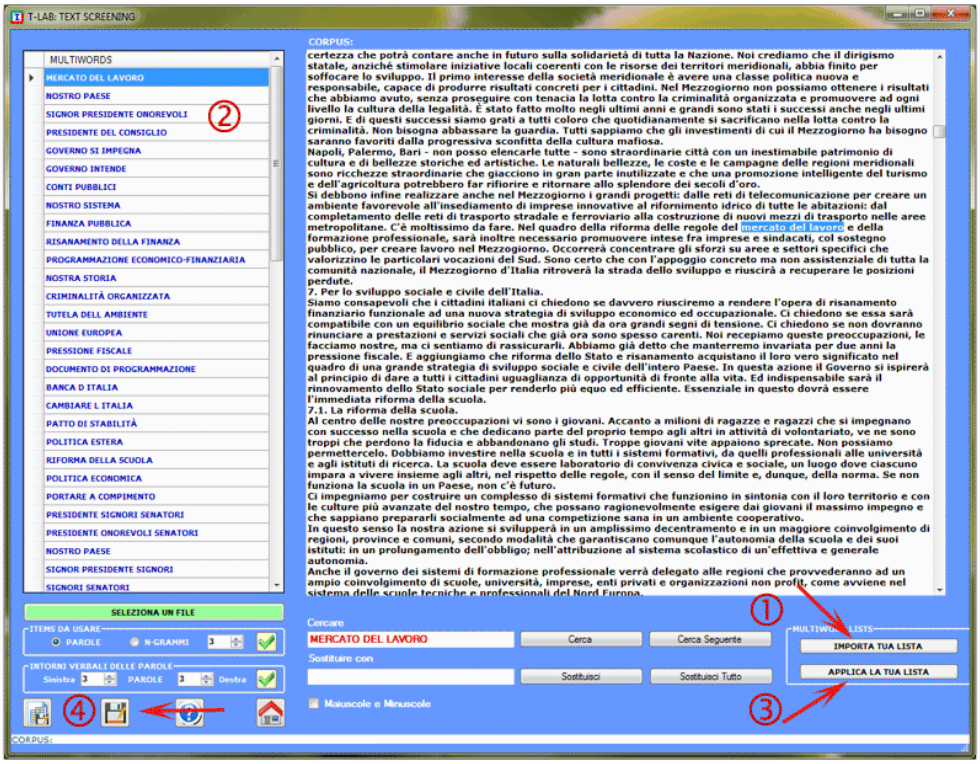




N.B.: Un click sul pulsante in basso a sinistra consente di esportate le liste a-b come file Excel. Diversamente, le liste 'c' vengono esportate automaticamente come file .html.

--

E' inoltre possibile importare liste personalizzate di **Multiword** (inserire link) ed, eventualmente, applicarle al corpus visualizzato (vedi immagine sotto)



Al termine delle operazioni, se l'utilizzatore ha modificato il testo e desidera salvarlo, **T-LAB** consente di creare un nuovo file (**corpus_dis.txt**) che, opportunamente rinominato, può essere importato e analizzato (vedi sopra '4').

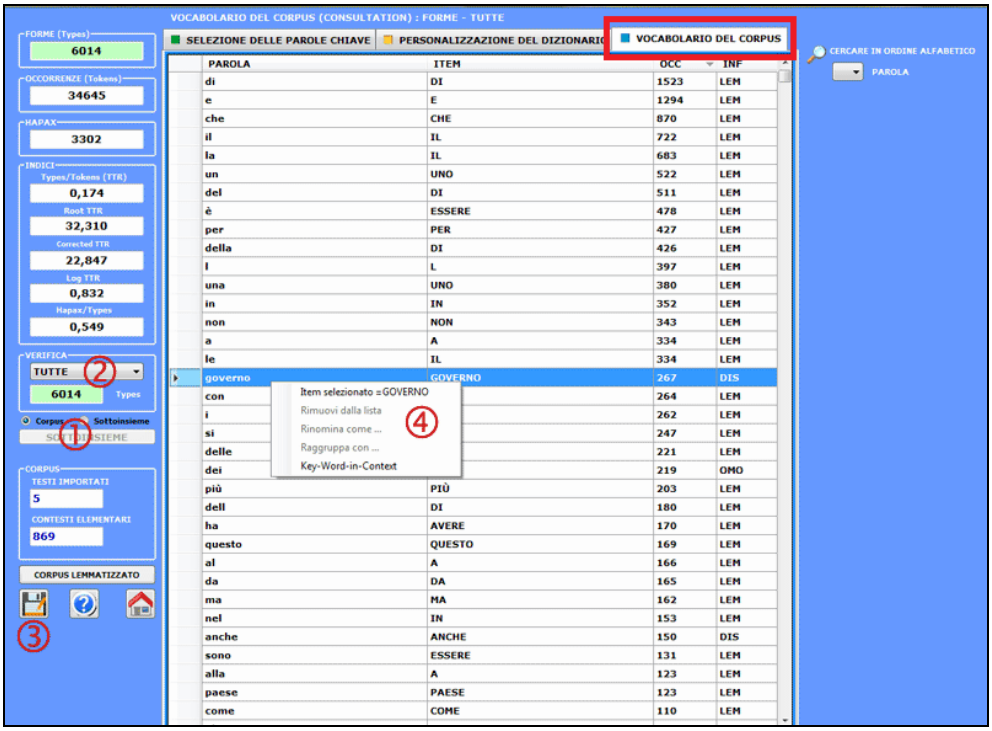
Vocabolario del Corpus

Questo strumento **T-LAB** consente di verificare il Vocabolario del corpus e dei suoi sottoinsiemi (vedi sotto opzione '1'). Inoltre fornisce alcune misure della **ricchezza lessicale**.

La tabella Vocabolario è una lista che include le "parole" (cioè i word types), le loro occorrenze (cioè i word tokens), i corrispondenti lemmi e alcune categorie utilizzate da **T-LAB** (vedi Glossario/Lemmatizzazione).

L'utilizzatore può agevolmente selezionare (vedi sotto opzione '2') le unità lessicali che appartengono a ciascuna categoria, consultare la relativa tabella ed esportarla in formato .xls (vedi sotto opzione '3').

Inoltre, usando il tasto destro del mouse, è possibile verificare le **concordanze** (Key-Word-in-Context) di ogni parola (vedi sotto opzione '4').

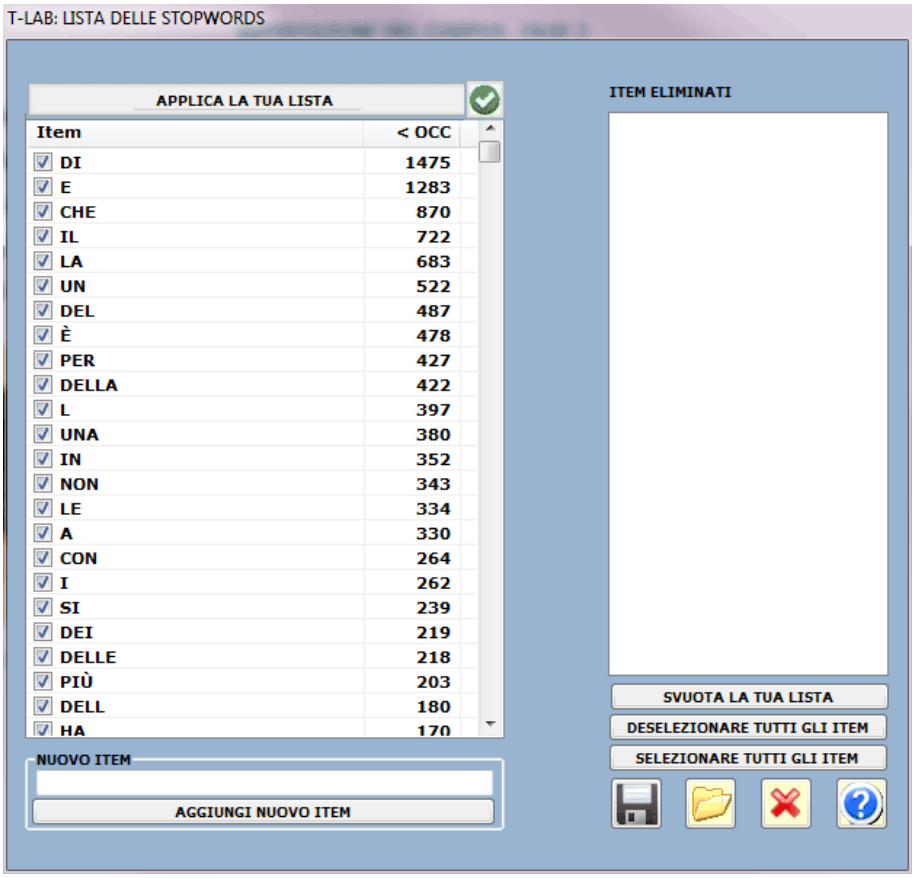


Le misure della ricchezza lessicale sono cinque:
Type/Token ratio (TTR);
Root TTR (Guiraud, 1960), ottenuta dividendo la quantità dei type per la radice quadrata dei token;
Corrected TTR (Carroll, 1964), ottenuta dividendo la quantità dei type per la radice quadrata di due volte la quantità dei token;
Log TTR (Herdan, 1960), ottenuta dividendo il logaritmo dei type per il logaritmo dei token;
Hapax/Types ratio.

N.B.:
- Hapax (i.e. Hapax Legomena) sono parole (type) che occorrono una sola volta nel corpus;
- quando vengono analizzati sottoinsiemi del corpus, tutte le misure della ricchezza lessicale non includono le stop words.

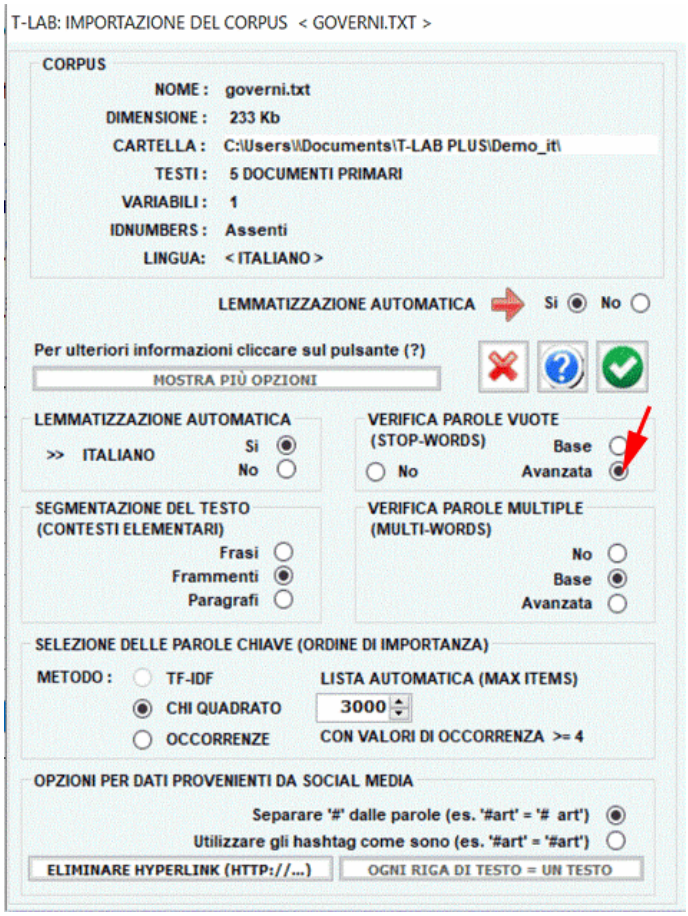
Stop-Words

Questa opzione consente di creare/modificare liste di **Stop-Words** (Parole Vuote) all'interno della finestra seguente.



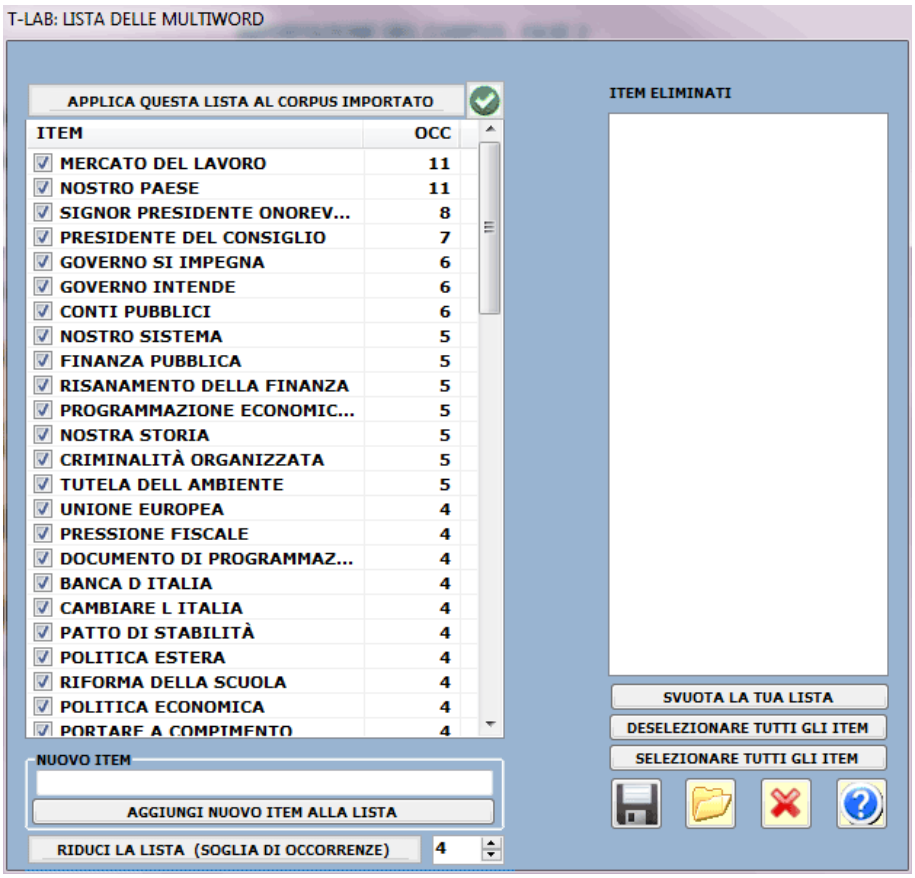
Ogni lista (file StopWords.txt) è costituita da "n" linee, ciascuna con una parola di max 50 caratteri, senza spazi vuoti e senza punteggiatura.

In ogni caso, per verificare/usare liste di StopWords durante la fase di importazione di un **nuovo corpus** è sufficiente selezionare l'opzione "**Avanzata**" nella finestra seguente:


T-LAB: IMPORTAZIONE DEL CORPUS < GOVERNI.TXT >
CORPUS
NOME : governi.txt
DIMENSIONE : 233 Kb
CARTELLA : C:\Users\\Documents\T-LAB PLUS\Demo_it\
TESTI : 5 DOCUMENTI PRIMARI
VARIABILI : 1
IDNUMBERS : Assenti
LINGUA: < ITALIANO >
LEMMATIZZAZIONE AUTOMATICA Si No
Per ulteriori informazioni cliccare sul pulsante (?)
MOSTRA PIÙ OPZIONI
LEMMATIZZAZIONE AUTOMATICA
>> ITALIANO Si No
VERIFICA PAROLE VUOTE (STOP-WORDS)
No Base Avanzata
SEGMENTAZIONE DEL TESTO (CONTESTI ELEMENTARI)
Frasi Frammenti Paragrafi
VERIFICA PAROLE MULTIPLE (MULTI-WORDS)
No Base Avanzata
SELEZIONE DELLE PAROLE CHIAVE (ORDINE DI IMPORTANZA)
METODO : TF-IDF CHI QUADRATO OCCORRENZE
LISTA AUTOMATICA (MAX ITEMS)
3000
CON VALORI DI OCCORRENZA >= 4
OPZIONI PER DATI PROVENIENTI DA SOCIAL MEDIA
Separare '#' dalle parole (es. '#art' = '# art')
Utilizzare gli hashtag come sono (es. '#art' = '#art')
ELIMINARE HYPERLINK (HTTP://...)
OGNI RIGA DI TESTO = UN TESTO

Multiwords

Questa opzione consente di creare/modificare liste di **Multiwords** (Locuzioni e Poliformi).



Ogni lista (file Multiwords.txt) è costituita da N linee (max 5.000), ciascuna con una sequenza di due o più parole (lunghezza massima: 50 caratteri, senza segni di punteggiatura).

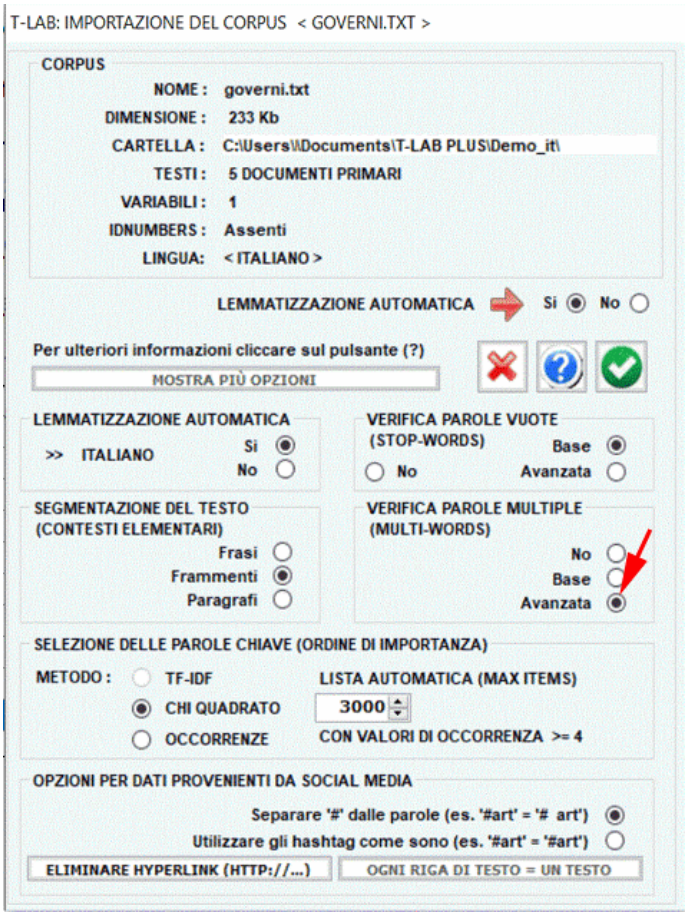
La struttura del file Multiwords.txt è quella di un semplice elenco, come l'esempio seguente:

ordine pubblico
servizio sanitario nazionale
val di fassa
forze dell'ordine
etc etc

Un click sul pulsante "**Applica questa lista …**" consente una rapida trasformazione delle parole multiple presenti in un corpus in altrettante stringhe che possono essere riconosciute e classificate da **T-LAB** (per es. "ministro dell'interno" viene trasformato in "ministro_dell_interno").

Al termine della trasformazione, è disponibile un nuovo file (**New_Corpus.txt**) che, opportunamente rinominato, può essere importato con **T-LAB**.

Per verificare/usare liste di Multiwords durante la fase di **importazione di un nuovo corpus** è sufficiente selezionare l'opzione "**Avanzata**" nella finestra seguente:

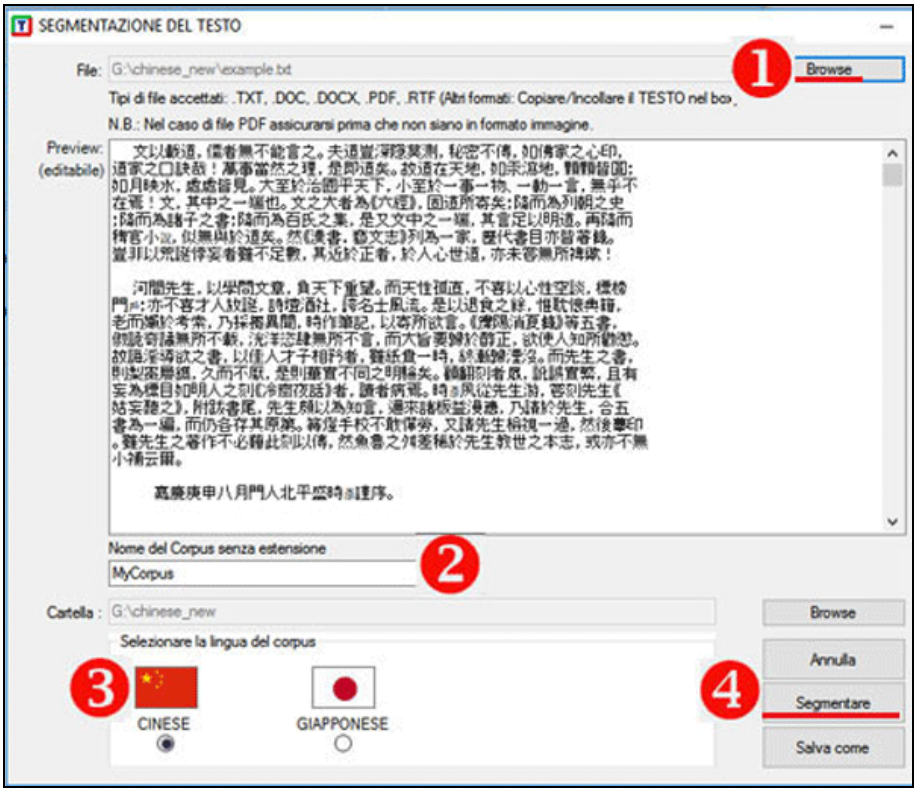


Segmentazione delle Parole

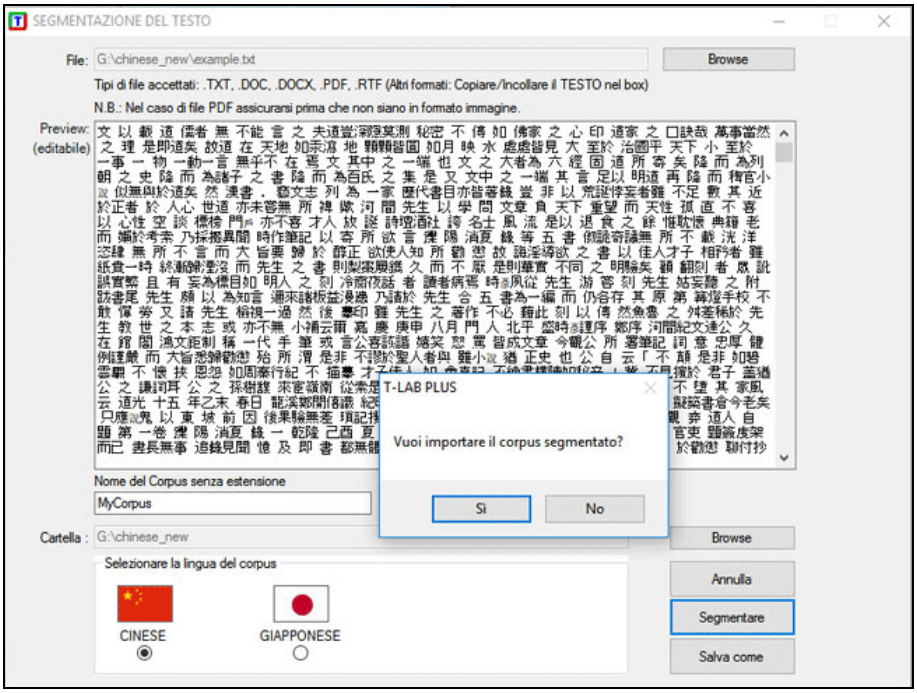
Questo strumento **T-LAB** può essere utilizzato prima di importare qualsiasi testo (*) cinese o giapponese che non abbia delimitatori (cioè spazi e / o segni di punteggiatura) tra le parole.
(*) Il testo da processare può essere costituito da un singolo documento o da una collezione di documenti che includono variabili categoriali.

Il suo utilizzo è molto semplice (vedi immagine seguente):

(1) selezionare un qualsiasi file;
(2) scegliere il nome del progetto;
(3) selezionare la lingua del testo;
(4) cliccare su 'Segmentare'.
Come risultato, verranno aggiunti spazi vuoti tra le parole.



Successivamente, se si vuole procedere con l' importazione, basta rispondere 'Sì' alla domanda "Vuoi importare il corpus segmentato?" (vedi immagine seguente).



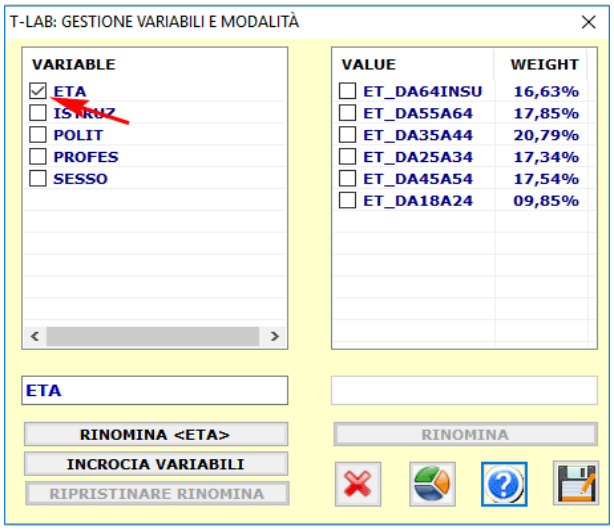
N.B.: Quando si desidera preparare un corpus costituito da vari testi che comprendono le linee di codifica (cioè variabili categoriali), si consiglia di procedere nel seguente modo:
1- 'Assemblare' i testi non segmentati (*) mediante lo strumento Corpus Builder e 'Salvare' il file corpus;
2 - Importare il corpus appena creato mediante lo strumento Segmentazione delle Parole; quindi procedere come spiegato in precedenza.
(*) Ciò significa che, quando si prepara il corpus, non è necessario segmentare ogni singolo file in anticipo.

ALTRI STRUMENTI

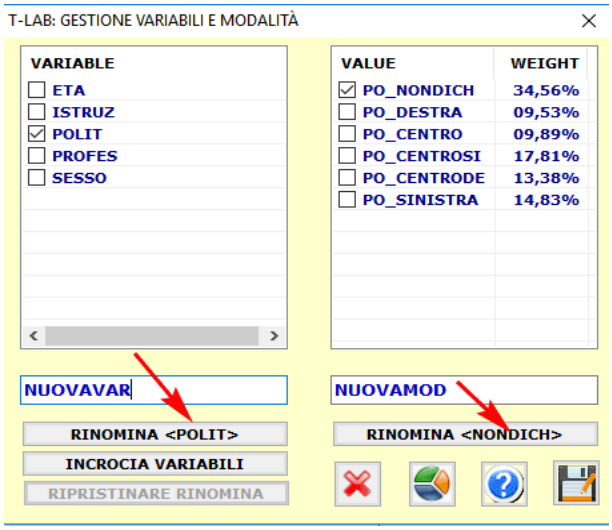
Gestione Variabili e Modalità

Questa funzione, attivata solo se il corpus include partizioni (variabili e modalità), consente di effettuare cinque tipi di operazioni:

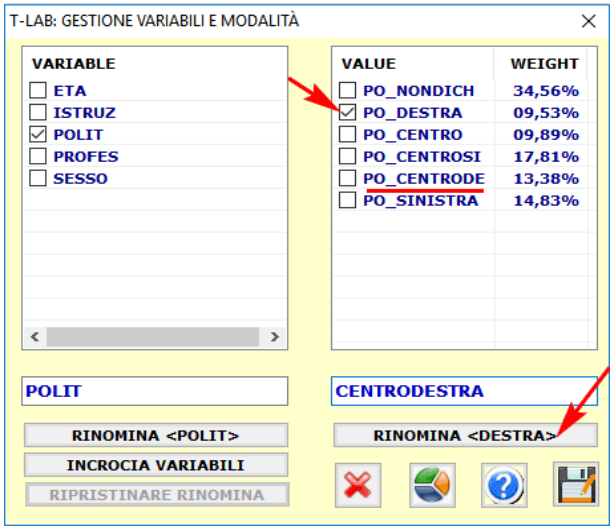
a) **verificare** le categorie di ogni variabile;



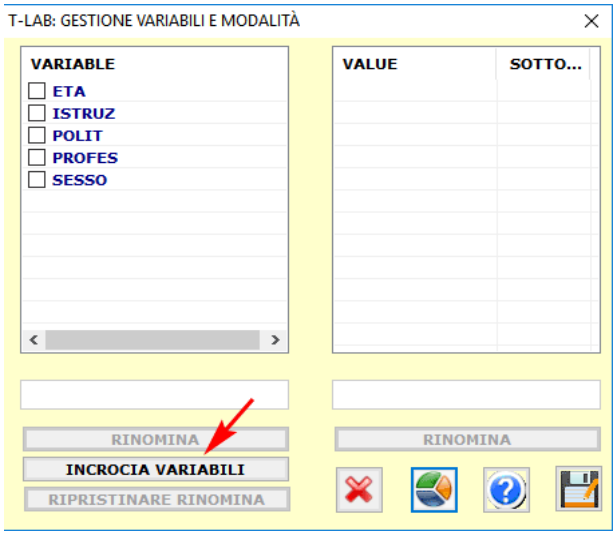
b) **rinominare** variabili e categorie;

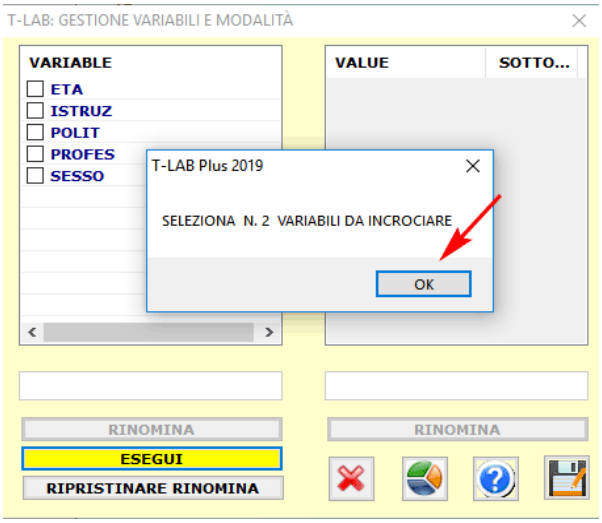


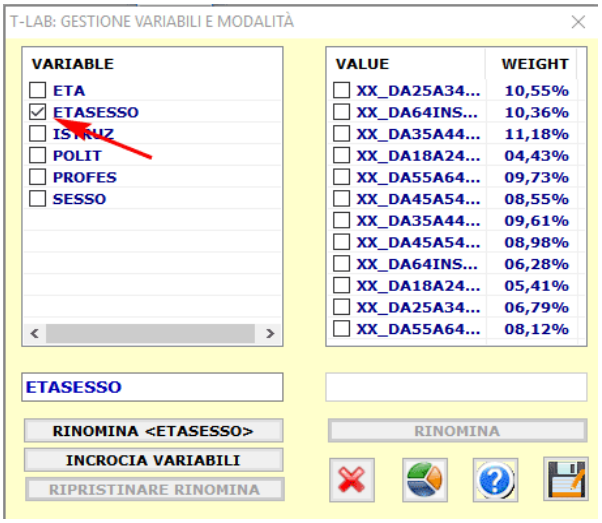
c) **raggruppare** due o più categorie tramite l'uso di label esistenti;



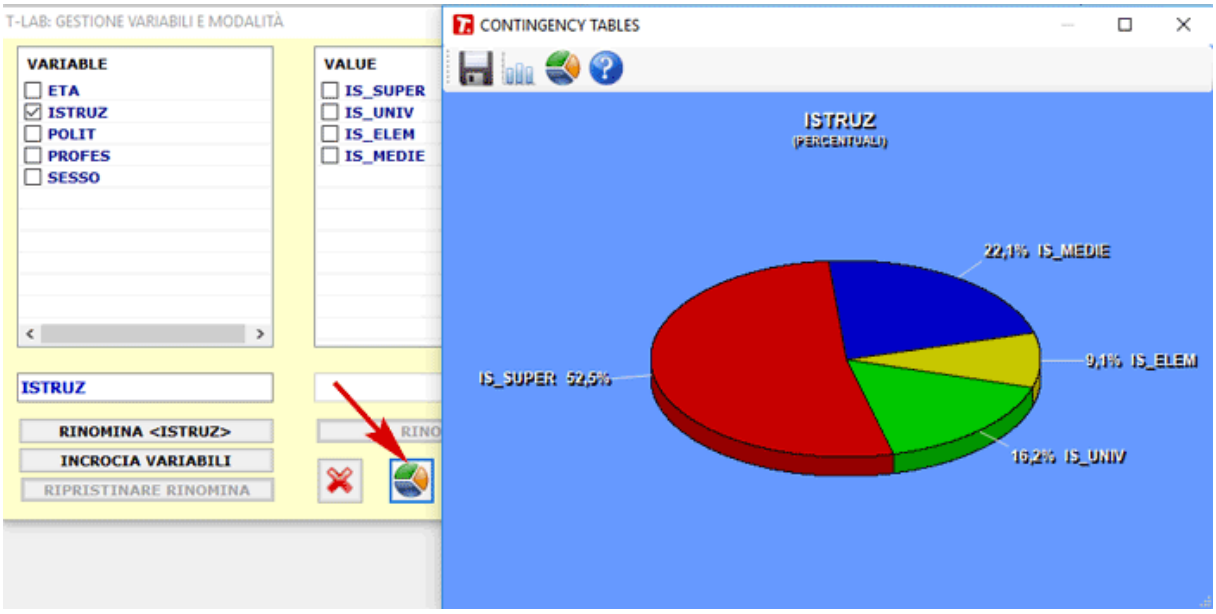
d) creare una **variabile incrocio** disponibile per ulteriori analisi.

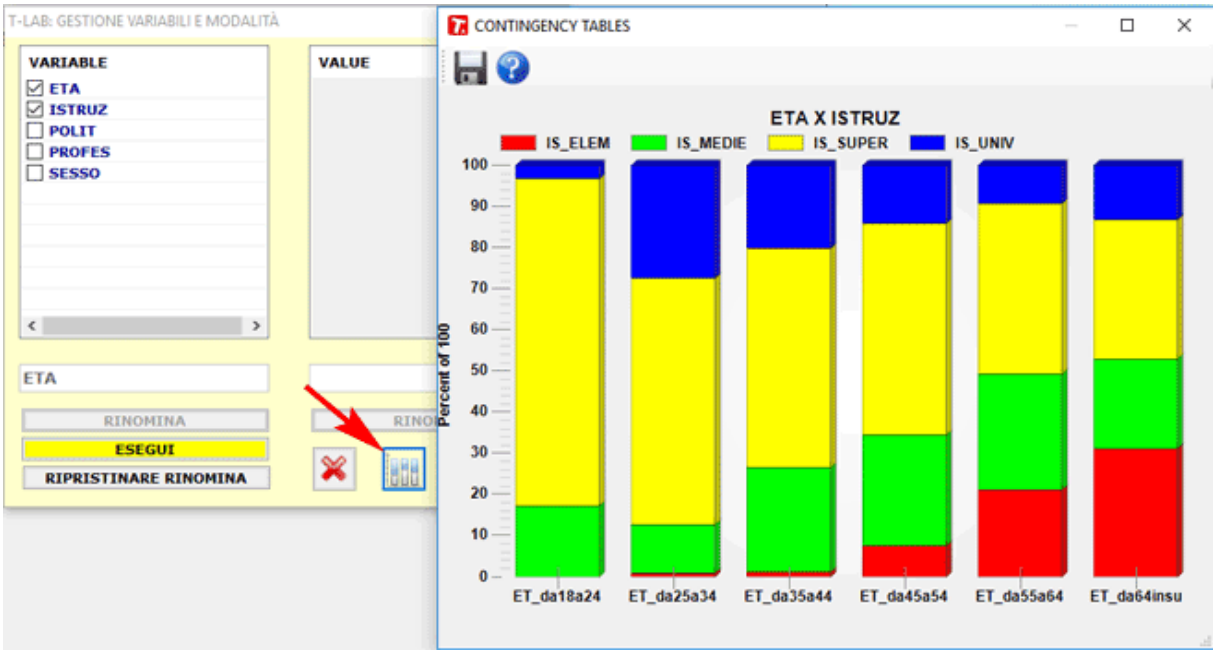






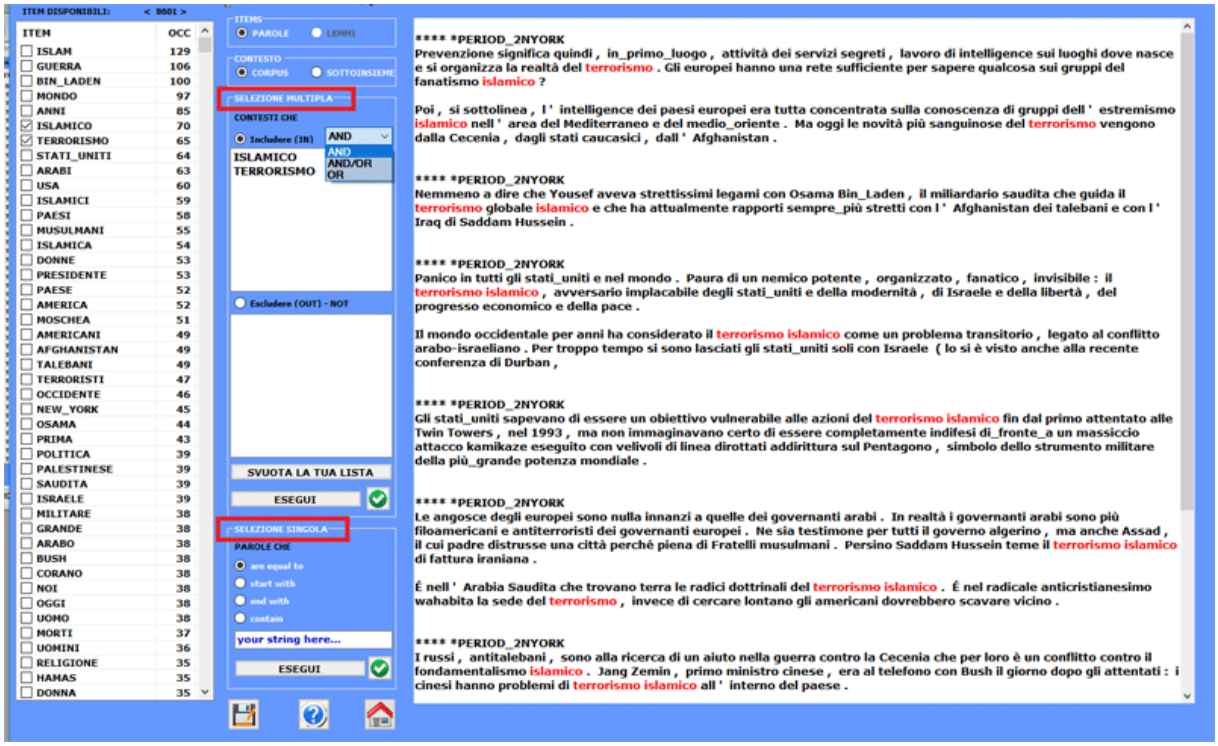
e) creare alcuni **grafici**.





Ricerca Avanzata nel Corpus

Questo strumento **T-LAB** ci consente di estrarre ed esportare tutti i frammenti di testo (cioè frasi o paragrafi) che corrispondono a query con parole singole o multiple, questo sia all'interno del corpus che dei suoi sottoinsiemi.



Il suo uso è estremamente intuitivo: basta selezionare le opzioni desiderate all'interno dei box corrispondenti (vedi sotto).

Nel caso delle selezioni ‘multiple’, le parole possono essere selezionate/aggiunte mediante click sugli item corrispondenti della tabella sulla sinistra.

Nel caso di selezioni ‘singole’, la stringa da cercare deve essere digitata nel box appropriato.

Dopo aver cliccato ‘esegui’, i risultati della ricerca sono mostrati nel box a destra e possono essere salvati in un file .rtf.

Tale file, che include tutte le codifiche **T-LAB**, può anche essere importato ed analizzato come un sub-corpus.

.

N.B.: L'uso di questa funzione è abilitato solo quando si lavora su un corpus già importato e sia stata selezionata una lista di parole chiave (vedi Impostazioni di Analisi).

Classificazione di Nuovi Documenti

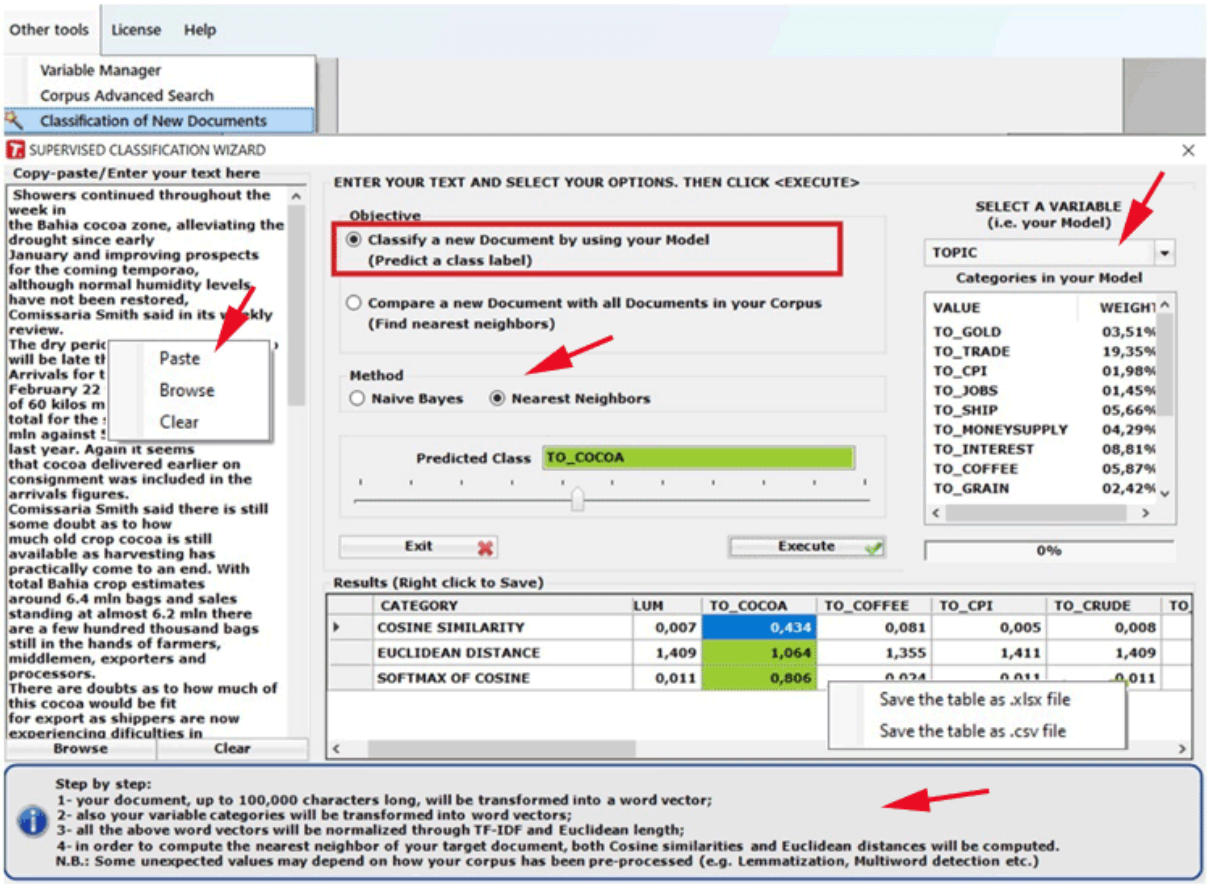
N.B .: Questa sezione dell'help è disponibile solo in inglese.

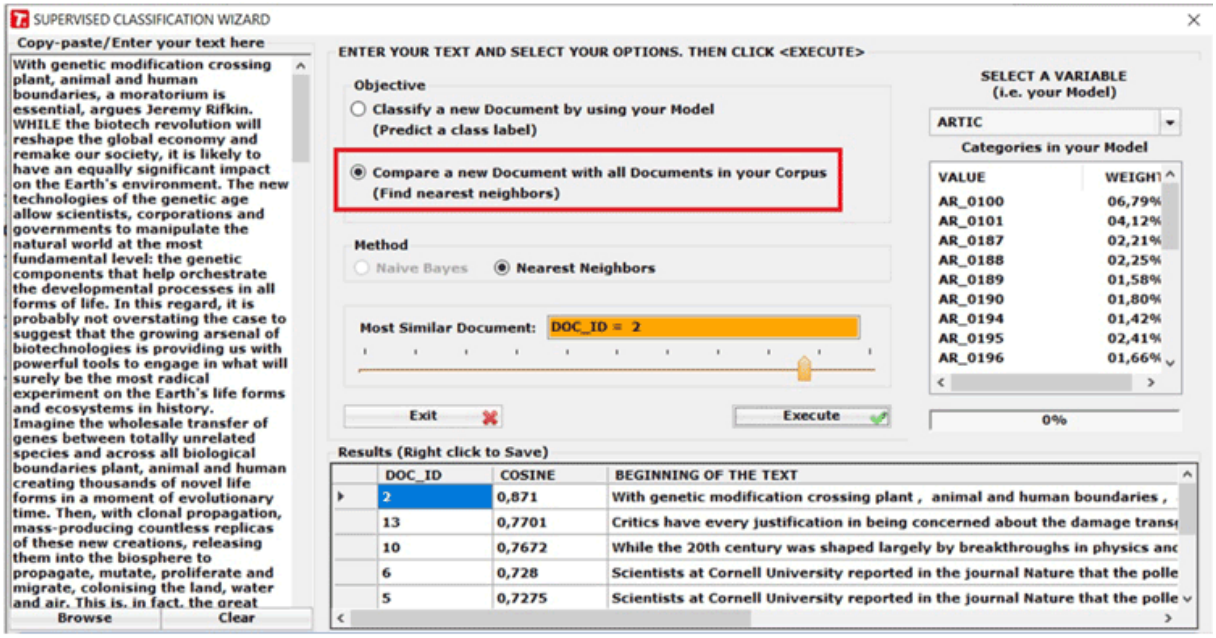
This tool, which is very easy to use, allows one to easily classify new documents according to a pre-existing model (i.e. any categorical variable) and also to compare any new document with all documents included in a corpus already analysed.

To this purpose, the following steps are required:

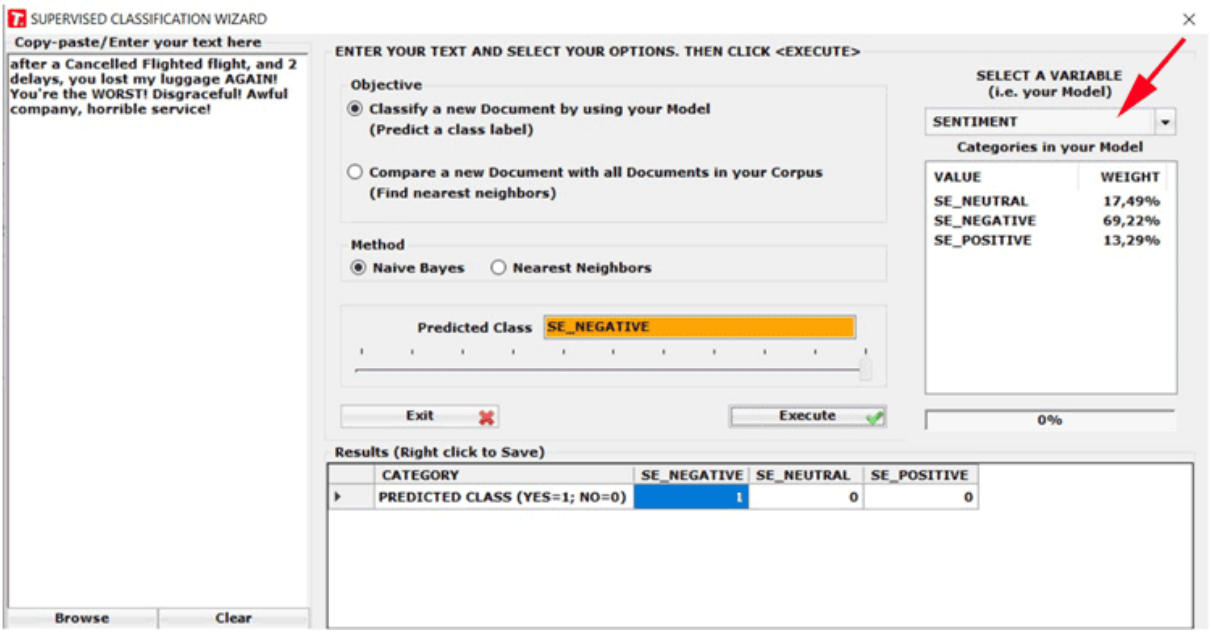
- enter a new document in the appropriate box;
- select a categorical variable to be used as a 'model';
- choose the desired 'objective' and a 'method';
- click 'execute'.

All results can be exported by using the right click options (see the below pictures).





When using this tool for sentiment analysis purpose, your corpus must include an appropriate categorical variable (see the below below).

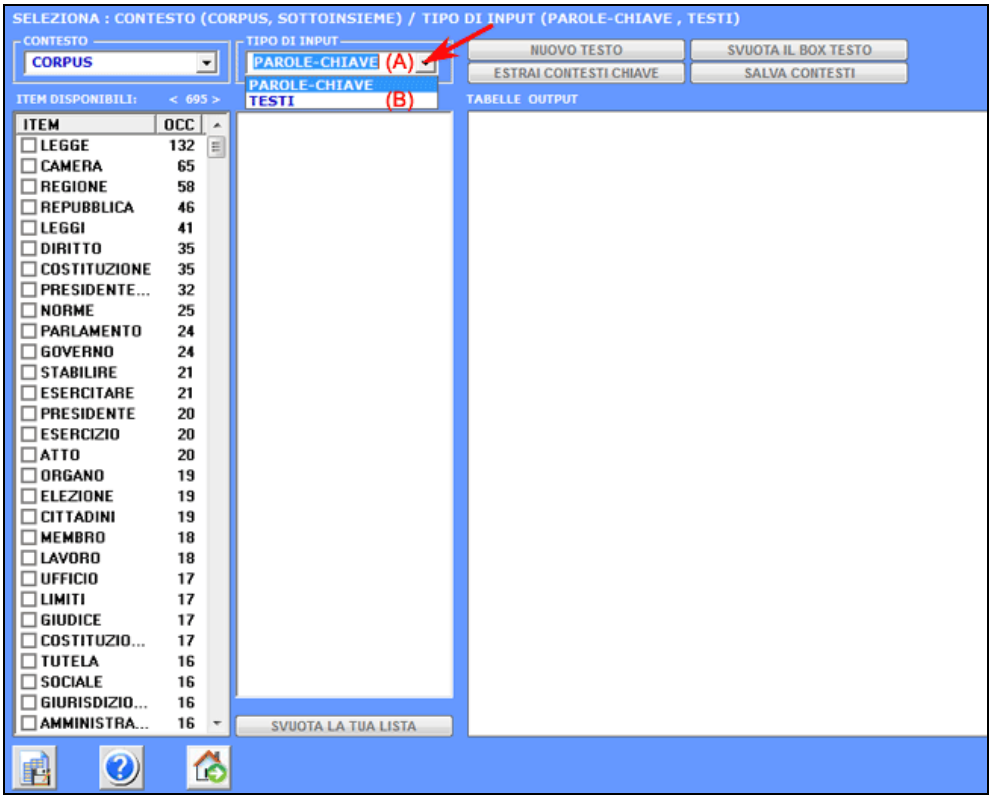


N.B.: When the user wishes to classify a dataset of new documents by using a supervised method, the dataset must be imported by T-LAB and then analysed by using a previously generated dictionary. To this purpose, the 'Thematic Document Classification' can be used, both for generating a dictionary of categories (i.e. unsupervised method) and for performing a supervised classification.

Contesti Chiave di Parole Tematiche

Questo strumento **T-LAB** può essere usato per due diversi scopi:

A) estrarre insiemi di unità di contesto che permettono di approfondire il valore tematico di specifiche **parole-chiave**;
B) estrarre le unità di contesto che risultano le più simili a **testi campione** proposti dall'utilizzatore.



Passo dopo passo, le rispettive procedure sono le seguenti:

Caso (A)

A differenza della funzione **Concordanze**, che consente di estrarre tutti i contesti elementari in cui sono presenti (occorrenze) le singole parole chiave selezionate , e della funzione **Associazioni di Parole**, che consente di estrarre i contesti elementari in cui le parole chiave selezionate sono "in coppia" con altre parole (co-occorrenze binarie), questo strumento consente di estrarre i contesti elementari in cui la parola selezionata è significativamente associata con un gruppo di altre parole (co-occorrenze multiple) che definiscono il suo ambito tematico.

Il suo funzionamento è il seguente:

1- l'utilizzatore sceglie una parola tematica "X" (vedi "lavoro" nell'immagine seguente);

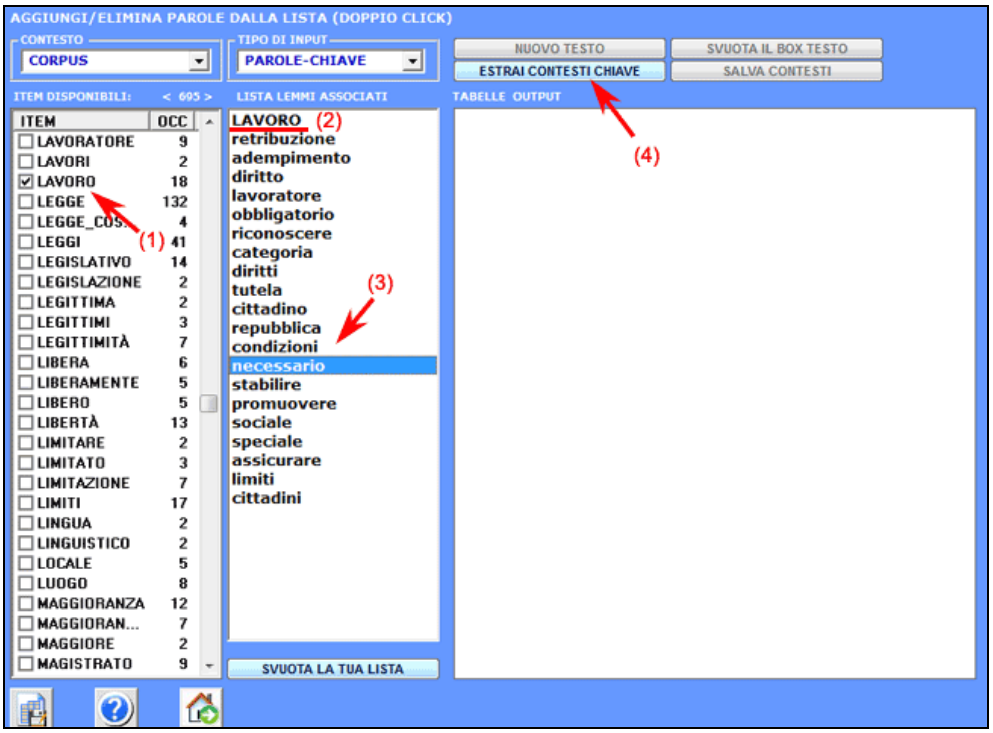
2- T-LAB propone una lista di parole (max. 50) i cui valori di co-occorrenza con "X" sono i più significativi;

3- l'utilizzatore può eliminare (doppio click) item irrilevanti dalla lista proposta;

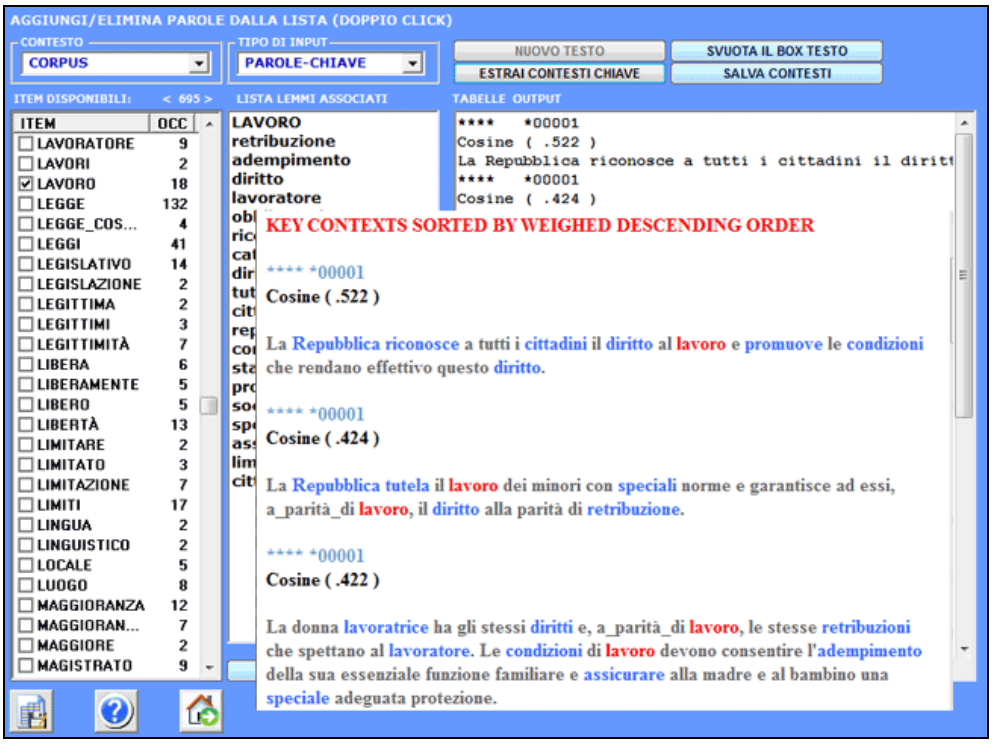
T-LAB assume la lista come un "query vector" e calcola i suoi **indici di associazione** (coefficiente del coseno) con tutti i contesti elementari del corpus o del **sottoinsieme** selezionato;

4- l'output fornito è una pagina **HTML** che contiene una lista dei più significativi contesti chiave di "X", ordinati per il valore decrescente dell'indice (vedi sotto);

I passi 1-4 possono essere ripetuti per "n" parole tematiche.



Gli output, sia in formato HTML e TXT, contengono un elenco dei più significativi contesti chiave di "X", elencati secondo l'ordine decrescente dei loro indici di associazione.

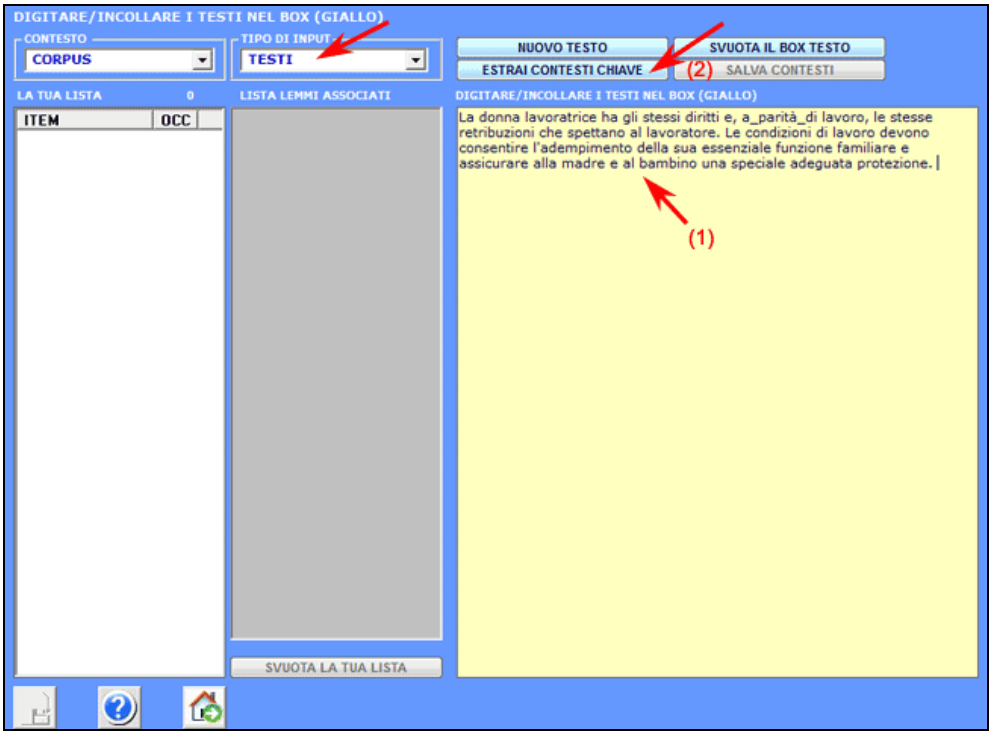


Caso (B)

Il suo funzionamento è modo seguente:

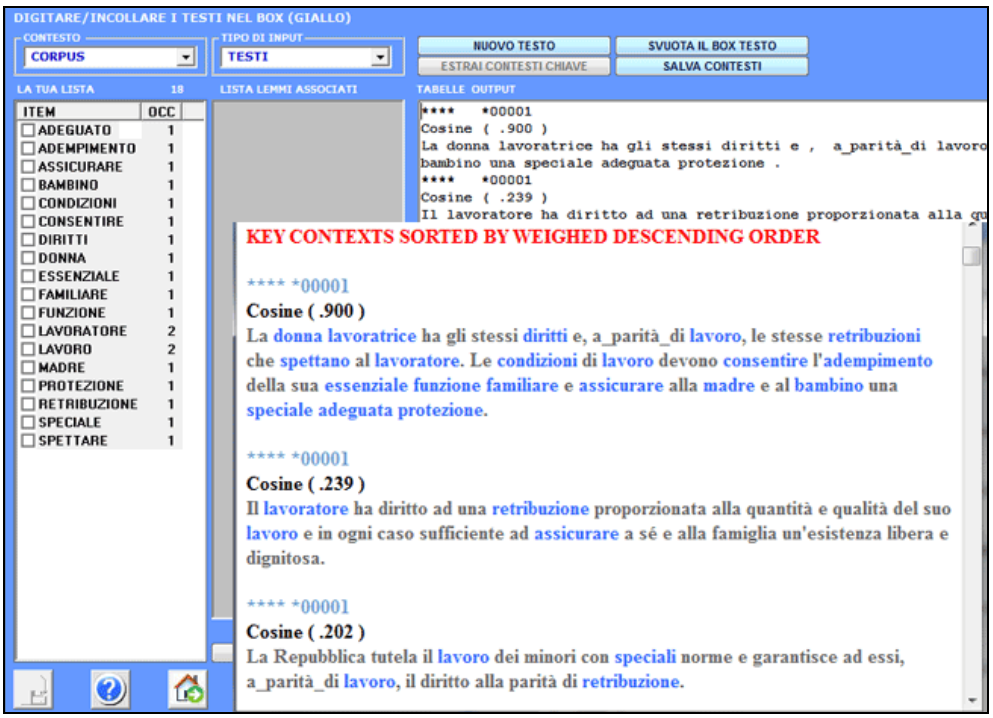
1 - l'utilizzatore copia / incolla un testo 'modello' (max 5000 caratteri) nella casella corrispondente;

2 - dopo aver cliccato l'opzione 'estrai contesti chiave', **T-LAB** trasforma il testo immesso in un vettore (query vector) e calcola i relativi indici di associazione (cioè i coefficienti coseno) con tutti i contesti elementari del corpus o del sottoinsieme selezionato.



Gli output, sia in formato HTML e TXT, contengono un elenco dei contesti chiave che sono più simili al testo in input.

N.B.: In questo caso la misura di similarità non tiene conto delle parole multiple le cui stringhe, con o senza il carattere underscore ('_'), non corrispondono al testo analizzato.



Esportare Tabelle Personalizzate

N.B.: Le immagini di questa sezione fanno riferimento a una versione precedente di **T-LAB 9**. In **T-LAB 10** l'aspetto è leggermente diverso, ma le funzionalità del software sono le medesime.

Questa opzione consente di creare, esplorare ed esportare tre tipi di tabelle:

a) quelle con le **occorrenze** delle varie unità lessicali entro i sottoinsiemi del corpus definiti da qualche variabile (matrici rettangolari);

b) quelle con le **co-occorrenze** delle varie unità lessicali (matrici quadrate) all'interno del corpus o di suoi sottoinsiemi;

c) quelle con le **occorrenze** delle varie unità lessicali all'interno di tutti i documenti (**matrici sparse** con gli indici dei vari elementi).

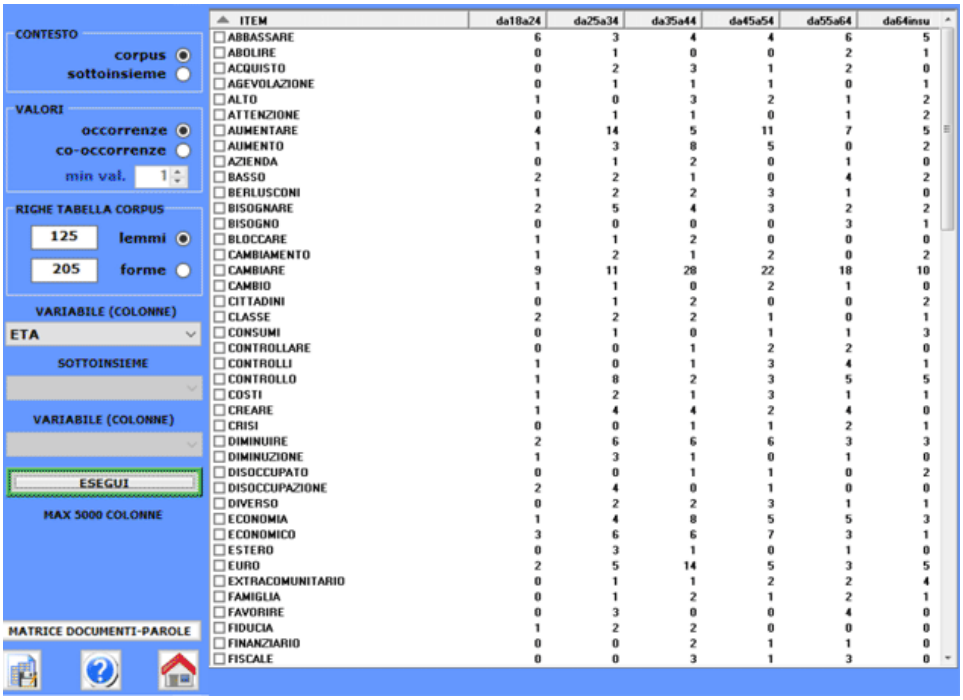
Le dimensione massime di tali tabelle sono rispettivamente: a) 10.000 righe per 150 colonne, b) 5.000 righe per 5.000 colonne; c) 30.000 documenti per 10.000 unità lessicali.

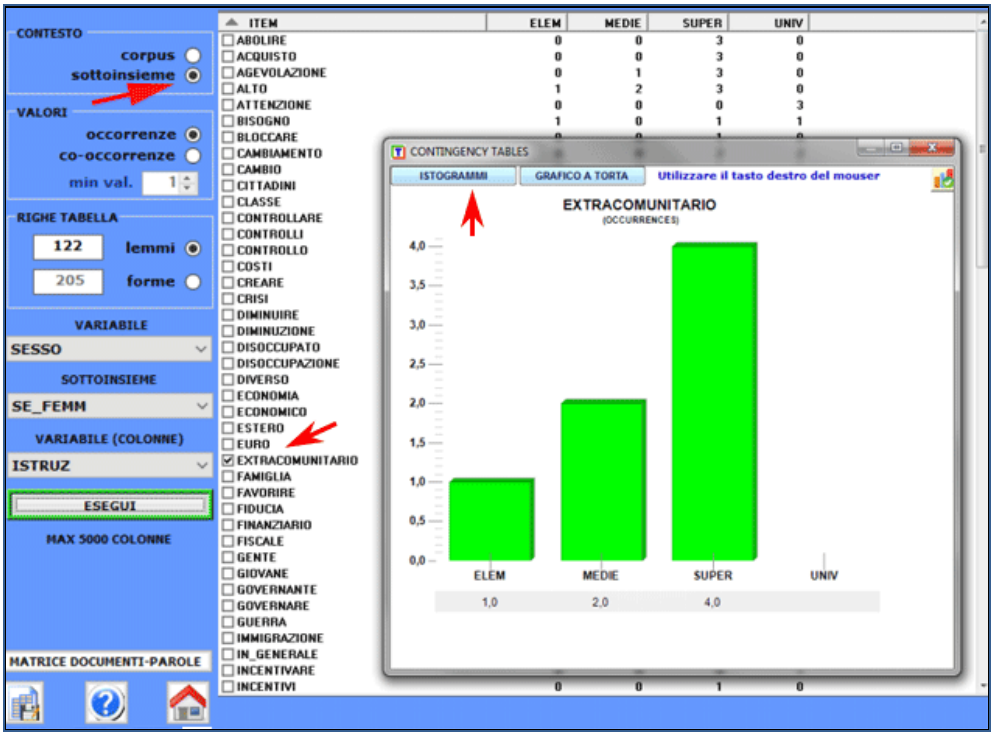
L'uso della funzione è molto intuitivo.

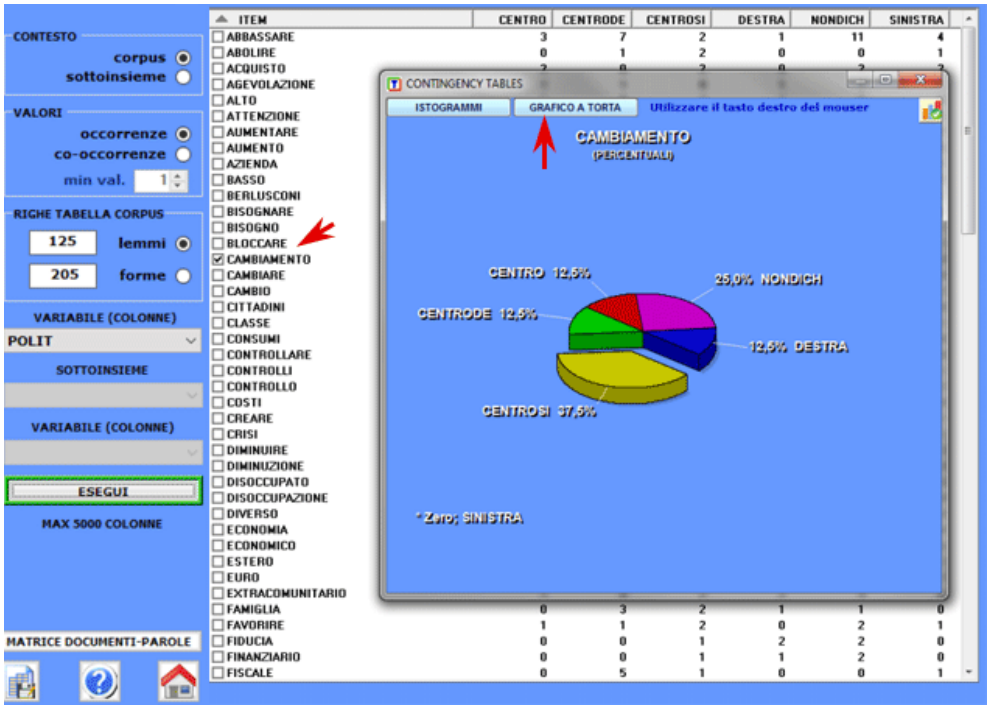
Nei casi più semplici, si richiede di selezionare la variabile le cui modalità costituiranno le colonne della tabella output.

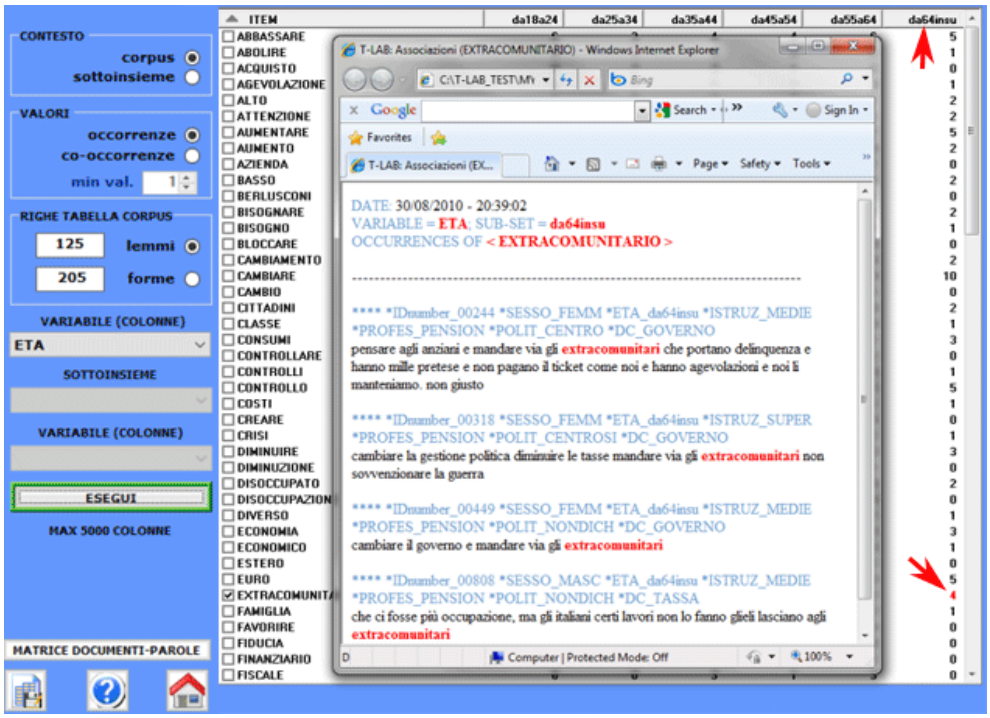
Nei casi più complessi, si richiede di selezionare una variabile e un sottoinsieme.

Tutte le tabelle di tipo 'a' e 'b' consentono di realizzare vari tipi **grafici**.
Inoltre, cliccando su specifiche celle, è possibile creare **file HTML** con i tutti i contesti elementari in cui la parola in riga è presente nel sottoinsieme in colonna (vedi sotto).


CONTESTO
corpus
sottoinsieme
VALORI
occorrenze
co-occorrenze
min val. 1
RIGHE TABELLA CORPUS
125 lemmi
205 forme
VARIABILE (COLONNE)
ETA
SOTTOINSIEME
VARIABILE (COLONNE)
ESEGUI
MAX 5000 COLONNE
MATRICE DOCUMENTI-PAROLE
ITEM da18a24 da25a34 da35a44 da45a54 da55a64 da64insu
ABBASSARE 6 3 4 4 6 5
ABOLIRE 0 1 0 0 2 1
ACQUISTO 0 2 3 1 2 0
AGEVOLAZIONE 0 1 1 1 0 1
ALTO 1 0 3 2 1 2
ATTENZIONE 0 1 1 0 1 2
AUMENTARE 4 14 5 11 7 5
AUMENTO 1 3 8 5 0 2
AZIENDA 0 1 2 0 1 0
BASSO 2 2 1 0 4 2
BERLUSCONI 1 2 2 3 1 0
BISOGNARE 2 5 4 3 2 2
BISOGNO 0 0 0 0 3 1
BLOCCARE 1 1 2 0 0 0
CAMBIAMENTO 1 2 1 2 0 2
CAMBIARE 9 11 28 22 18 10
CAMBIO 1 1 0 2 1 0
CITTADINI 0 1 2 0 0 2
CLASSE 2 2 2 1 0 1
CONSUMI 0 1 0 1 1 3
CONTROLLARE 0 0 1 2 2 0
CONTROLLI 1 0 1 3 4 1
CONTROLLO 1 8 2 3 5 5
COSTI 1 2 1 3 1 1
CREARE 1 4 4 2 4 0
CRISI 0 0 1 1 2 1
DIMINUIRE 2 6 6 6 3 3
DIMINUZIONE 1 3 1 0 1 0
DISOCCUPATO 0 0 1 1 0 2
DISOCCUPAZIONE 2 4 0 1 0 0
DIVERSO 0 2 2 3 1 1
ECONOMIA 1 4 8 5 5 3
ECONOMICO 3 6 6 7 3 1
ESTERO 0 3 1 0 1 0
EURO 2 5 14 5 3 5
EXTRACOMUNITARIO 0 1 1 2 2 4
FAMIGLIA 0 1 2 1 2 1
FAVORIRE 0 3 0 0 4 0
FIDUCIA 1 2 2 0 0 0
FINANZIARIO 0 0 2 1 1 0
FISCALE 0 0 3 1 3 0


CONTESTO
corpus
sottoinsieme
VALORI
occorrenze
co-occorrenze
min val. 1
RIGHE TABELLA
122 lemmi
205 forme
VARIABILE
SESSO
SOTTOINSIEME
SE_FEMM
VARIABILE (COLONNE)
ISTRUZ
ESEGUI
MAX 5000 COLONNE
MATRICE DOCUMENTI-PAROLE
ITEM ELEM MEDIE SUPER UNIV
CONTINGENCY TABLES
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouser
EXTRACOMUNITARIO
(OCCURRENCES)
4,0
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
ELEM
MEDIE
SUPER
UNIV
1,0
2,0
4,0


CONTESTO
corpus
sottoinsieme
VALORI
occorrenze
co-occorrenze
min val.
RIGHE TABELLA CORPUS
125
lemmi
205
forme
VARIABILE (COLONNE)
POLIT
SOTTOINSIEME
VARIABILE (COLONNE)
ESEGUI
MAX 5000 COLONNE
MATRICE DOCUMENTI-PAROLE
ITEM
CENTRO
CENTRODE
CENTROSI
DESTRA
NONDICH
SINISTRA
CONTINGENCY TABLES
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouse
CAMBIAMENTO
(PERCENTUALI)
CENTRO 12,5%
25,0% NONDICH
CENTRODE 12,5%
12,5% DESTRA
CENTROSI 37,5%
* Zero; SINISTRA


CONTESTO
corpus
sottoinsieme
VALORI
occorrenze
co-occorrenze
min val.
RIGHE TABELLA CORPUS
125
lemmi
205
forme
VARIABILE (COLONNE)
ETA
SOTTOINSIEME
VARIABILE (COLONNE)
ESEGUI
MAX 5000 COLONNE
MATRICE DOCUMENTI-PAROLE
ITEM
da18a24
da25a34
da35a44
da45a54
da55a64
da64insu
T-LAB: Associazioni (EXTRACOMUNITARIO) - Windows Internet Explorer
DATE: 30/08/2010 - 20:39:02
VARIABLE = ETA; SUB-SET = da64insu
OCCURRENCES OF < EXTRACOMUNITARIO >
**** *IDnumber_00244 *SESSO_FEMM *ETA_da64insu *ISTRUZ_MEDIE *PROFES_PENSION *POLIT_CENTRO *DC_GOVERNO
pensare agli anziani e mandare via gli extracomunitari che portano delinquenza e hanno mille pretese e non pagano il ticket come noi e hanno agevolazioni e noi li manteniamo. non giusto
**** *IDnumber_00318 *SESSO_FEMM *ETA_da64insu *ISTRUZ_SUPER *PROFES_PENSION *POLIT_CENTROSI *DC_GOVERNO
cambiare la gestione politica diminuire le tasse mandare via gli extracomunitari non sovvenzionare la guerra
**** *IDnumber_00449 *SESSO_FEMM *ETA_da64insu *ISTRUZ_MEDIE *PROFES_PENSION *POLIT_NONDICH *DC_GOVERNO
cambiare il governo e mandare via gli extracomunitari
**** *IDnumber_00808 *SESSO_MASC *ETA_da64insu *ISTRUZ_MEDIE *PROFES_PENSION *POLIT_NONDICH *DC_TASSA
che ci fosse più occupazione, ma gli italiani certi lavori non lo fanno gliell lasciano agli extracomunitari

Per esportare matrici sparse del tipo documenti per parole è sufficiente cliccare il pulsante ‘Matrice Documenti-Parole’.

In questo caso, i tipi di output sono due:

Il primo (Sparse_Matrix.csv) ha il formato seguente:

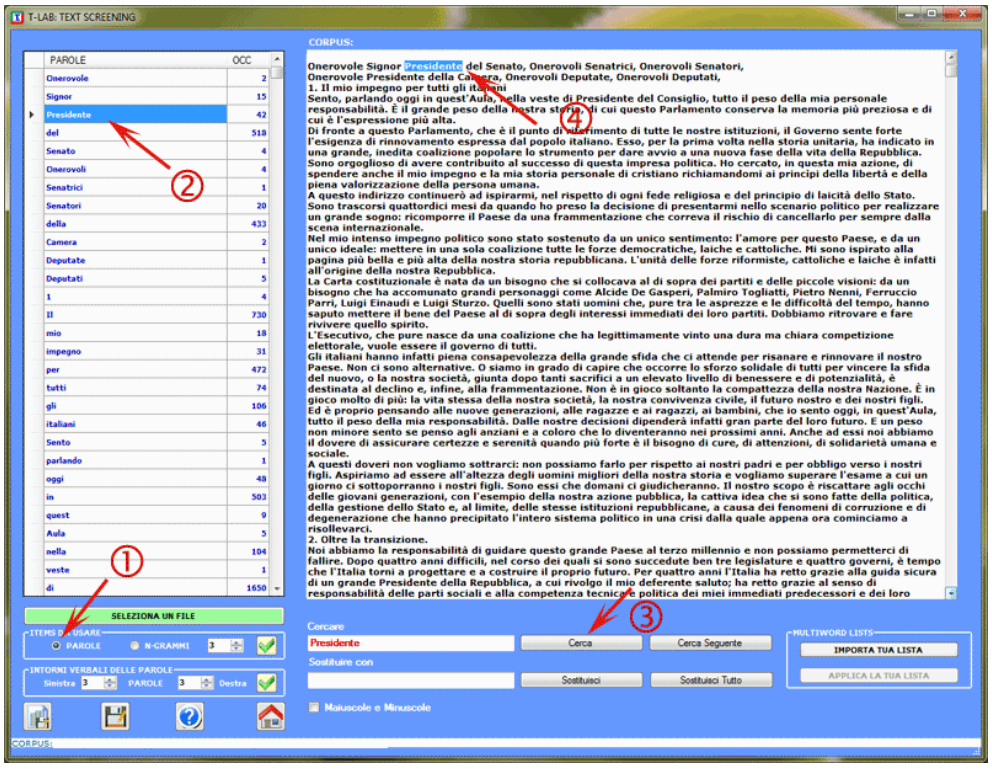
Doc_Index; Word_Index; Word_Occ
00001; 1; 12
00001; 2; 5
…..

Il primo (Word_Indexes.csv) ha il formato seguente:

Word_Index; Word_Label
1; abolire
2; acquisto
…..

Editor

N.B.: In **T-LAB 10** le funzioni per l'editing dei file in formato testo sono incluse nello strumento **Text Screening** (vedi sotto).



Importare-Esportare una Lista di Identificativi

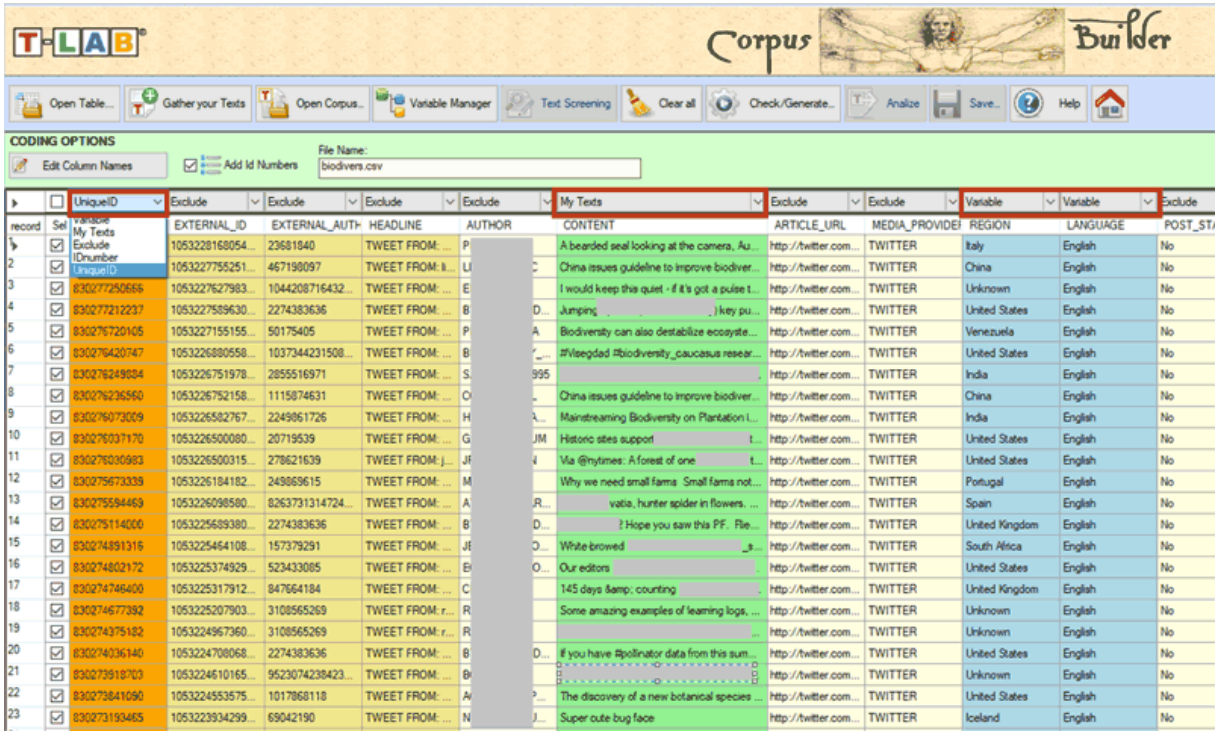
In **T-LAB** un identificatore univoco ('Unique identifier') è una variabile categoriale con un valore unico e diverso per ogni documento (o caso).

Un elenco di identificatori univoci può essere costituito da stringhe alfanumeriche di ogni tipo (ad es. ID degli intervistati, nomi propri, nomi geografici, nomi di libri, ecc.) di una lunghezza massima pari a 50 caratteri e senza spazi vuoti.

Poiché gli identificatori univoci sono singolari, è impossibile eseguire analisi dei dati che li usino come variabili. Essi vengono solo utilizzati per identificare i risultati negli output del software.

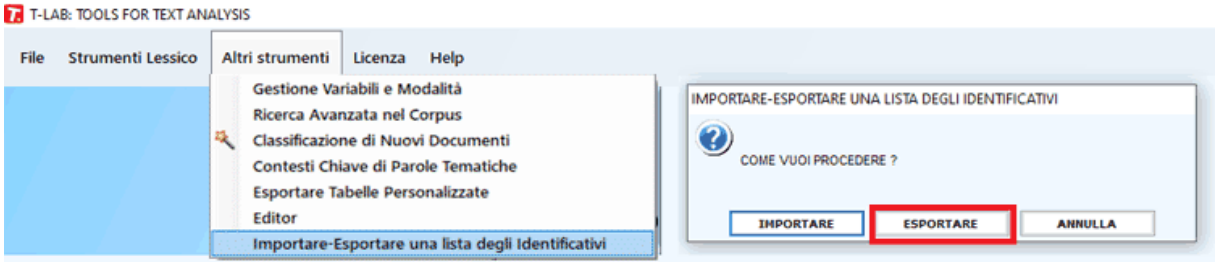
In **T-LAB**, tramite le opzioni di importazione / esportazione, qualsiasi elenco di identificatori univoci può essere modificato in qualsiasi momento.

Quando si importano dati in formato tabellare, gli identificatori univoci devono trovarsi nella prima colonna, come nell'esempio seguente relativo a messaggi Twitter.



Negli altri casi (ovvero raccolte di documenti che non sono in formato tabulare) la procedura raccomandata è la seguente:

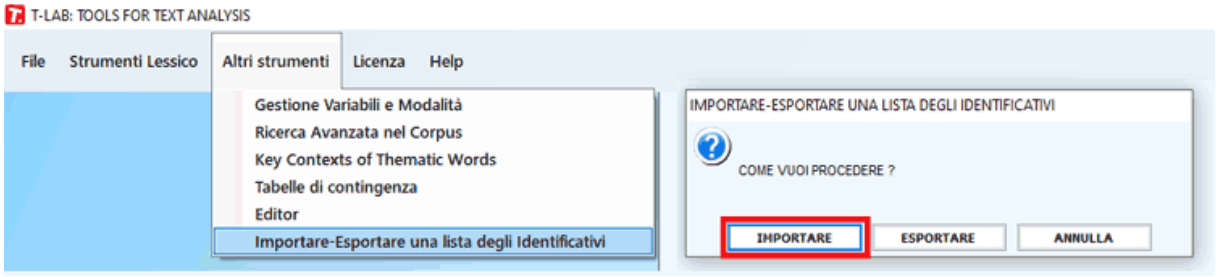
1- Importare prima il proprio corpus;
2- Esportare l'elenco degli identificatori creato automaticamente da **T-LAB.**



3- Modificare il file CSV creato da **T-LAB** (ovvero modificare i valori "MyIdentifier" in base alle proprie esigenze. Vedi l'immagine seguente).

MyID	MyIdentifier
1	TOBEREPLACED00001
2	TOBEREPLACED00002
3	TOBEREPLACED00003
4	TOBEREPLACED00004
5	TOBEREPLACED00005
6	TOBEREPLACED00006
7	TOBEREPLACED00007
8	TOBEREPLACED00008
9	TOBEREPLACED00009
10	TOBEREPLACED00010
...	...

4- Importare il file CSV che include gli identificatori univoci corretti.



GLOSSARIO

Analisi delle corrispondenze

Metodo di **analisi fattoriale** applicato allo studio di **tabelle dati** le cui "caselle" contengono valori di frequenza (numeri reali positivi) o di presenza-assenza ("1" e "0").

Come tutti i metodi di analisi fattoriale, l'analisi delle corrispondenze consente di estrarre nuove variabili - i **fattori** appunto - che hanno la proprietà di riassumere in modo ordinato l'informazione rilevante contenuta nelle innumerevoli caselle delle tabelle dati; inoltre, questo metodo di analisi consente di predisporre grafici atti a rappresentare - in uno o più spazi - i punti che individuano gli **oggetti** in riga e in colonna, cioè - nel nostro caso - le entità linguistiche (parole, lemmi, segmenti di testi e testi) con le rispettive caratteristiche di provenienza.

In termini geometrici, ciascun fattore organizza una dimensione spaziale - rappresentabile come una linea o asse - al cui centro (o baricentro) è il valore "0" e che si sviluppa in modo bi-polare verso le estremità "negativa" (-) e "positiva" (+), in modo tale che gli oggetti collocati su poli opposti sono quelli più diversi tra loro, un pò come la "sinistra" e la "destra" sull'asse della politica.

In **T-LAB** i risultati delle analisi vengono sintetizzati attraverso grafici bidimesionali (del tipo piani cartesiani) che consentono di apprezzare le relazioni di prossimità/distanza - ovvero di somiglianza/differenza - tra gli oggetti considerati.

Inoltre, in **T-LAB** vengono fornite delle misure - in particolare i **Contributi Assoluti** e i **Valori Test** - atte ad agevolare l'interpretazione delle **polarità fattoriali** che organizzano le somiglianze-differenze tra gli oggetti considerati.

Catene markoviane

Una catena markoviana (dal nome del matematico russo Andrei Andreiëvich Markov) è costituita da una **successione** (o sequenza) di eventi, generalmente indicati come **stati**, caratterizzata da due proprietà:

- l'insieme degli eventi e dei loro possibili esiti è finito;
- l'esito di ogni evento dipende solo (o al massimo) dall'evento immediatamente precedente.

Con la conseguenza che ad ogni transizione da un evento all'altro corrisponde un valore di probabilità.

In ambito scientifico, il modello delle catene markoviane è utilizzato per analizzare le successioni di eventi economici, biologici, fisici, ecc. Nell'ambito degli studi linguistici le sue applicazioni hanno come oggetto le possibili combinazioni delle varie unità di analisi sull'asse delle relazioni sintagmatiche (una dopo l'altra).

In **T-LAB** l'analisi delle catene markoviane concerne due tipi di **sequenze**:

- quelle concernenti le relazioni tra unità lessicali (parole, lemmi o categorie) presenti nel corpus in analisi;
- quelle presenti in file esterni predisposti dall'utilizzatore.

In entrambe i casi, in primo luogo vengono costruite tabelle quadrate in cui sono riportate le occorrenze delle transizioni, cioè quantità che indicano il numero di volte in cui una unità di analisi precede (o segue) l'altra. Successivamente, le occorrenze delle transizioni vengono trasformate in valori di probabilità (vedi immagini seguenti).

	s_1	s_2	s_3	s_4	s_5	s_6	TOT
s_1	0	8	7	11	2	1	29
s_2	6	0	24	5	10	8	53
s_3	9	24	0	3	28	16	80
s_4	3	7	5	0	6	14	35
s_5	4	5	26	11	0	7	53
s_6	7	9	18	5	7	0	46

	s_1	s_2	s_3	s_4	s_5	s_6	TOT
s_1	0,00	0,28	0,24	0,38	0,07	0,03	1
s_2	0,11	0,00	0,45	0,09	0,19	0,15	1
s_3	0,11	0,30	0,00	0,04	0,35	0,20	1
s_4	0,09	0,20	0,14	0,00	0,17	0,40	1
s_5	0,08	0,09	0,49	0,21	0,00	0,13	1
s_6	0,15	0,20	0,39	0,11	0,15	0,00	1

Per ulteriori informazioni vedi **Analisi delle Sequenze**.

Chi quadro

Si tratta di un test statistico atto a verificare se i valori di frequenza ottenuti tramite rilevazione, e registrati in una qualche tabella a doppia entrata, sono significativamente diversi da quelli "teorici".

Generalmente in **T-LAB** questo test è applicato a tabelle 2 x 2; quindi il valore di soglia è fissato a 3.84 (df =1; p. 0.05) o 6.64 (df =1; p. 0.01).

Ad esempio, quando si tratta di stabilire la significatività delle occorrenze di una parola ("x") entro una unità di contesto ("A") il test viene applicato a una tabella di questo tipo

	Context "A"	Other Contexts		
Word "x"	15	198	213	N_j
Other Words	572	2420	2992	
	587	2618	3205	N_{ij}
	N_i			

La formula del CHI quadro, nella sua versione semplificata, è la seguente:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

dove "O" ed "E" stanno rispettivamente per le frequenze osservate e per quelle teoriche.

Per ogni cella, le occorrenze attese (E) sono calcolate nel modo seguente: (Ni x Nj)/Nij.

Ne risulta che nell'esempio considerato il valore del CHI quadro è pari a 19.38.

Poiché è maggiore del valore critico, l'ipotesi nulla (assenza di differenza significativa) può essere respinta.

Cluster Analysis

Insieme di tecniche statistiche il cui obiettivo è costituito dall'individuare **raggruppamenti di oggetti** che abbiano due caratteristiche complementari:

A - al loro interno, la massima somiglianza tra gli elementi che li costituiscono (gli oggetti appartenenti a ciascun cluster);

B - tra di loro, la massima differenza.

Nel linguaggio della statistica, le caratteristiche “A” e “B” corrispondono rispettivamente alla varianza interna (within cluster variance) e a quella esterna (between cluster variance).

In generale, i metodi della Cluster Analysis vengono distinti in due tipi:

- **Hierarchical methods**, i cui algoritmi ricostruiscono l'intera gerarchia degli oggetti in analisi (il cosiddetto "albero"), vuoi in senso ascendente, vuoi in senso discendente;

- **Partitioning methods**, i cui algoritmi prevedono che l'utilizzatore abbia preventivamente definito il numero di cluster in cui l'insieme degli oggetti in analisi va diviso.

In **T-LAB** sono utilizzati algoritmi di entrambi i tipi.

In particolare:

- la funzione **Co-Word Analyi**s e Mappe Concettuali utilizza un metodo gerarchico;
- la funzione **Cluster Analysis** consente di utilizzare tre diversi metodi: due gerarchici e uno a partizioni;
- le funzioni **Analisi Tematica dei Contesti Elementari** e **Classificazione Tematica dei Documenti** utilizzano un algoritmo del tipo bisecting K-means.

Alcune delle pubblicazioni citate in **Bibliografia** consentono di approfondire sia aspetti generali dei vari metodi (Bolasco S., 1999; Lebart L., A. Morineau, M. Piron, 1995), sia aspetti specifici concernenti Hdbscan (Campello R. J. G. B., Moulavi D., Zimek A. & Sander J. , 2015) e il metodo bisecting K-means (Steinbach, M., G. Karypis, V. Kumar, 2000; Savaresi S.M., D.L. Boley, 2001).

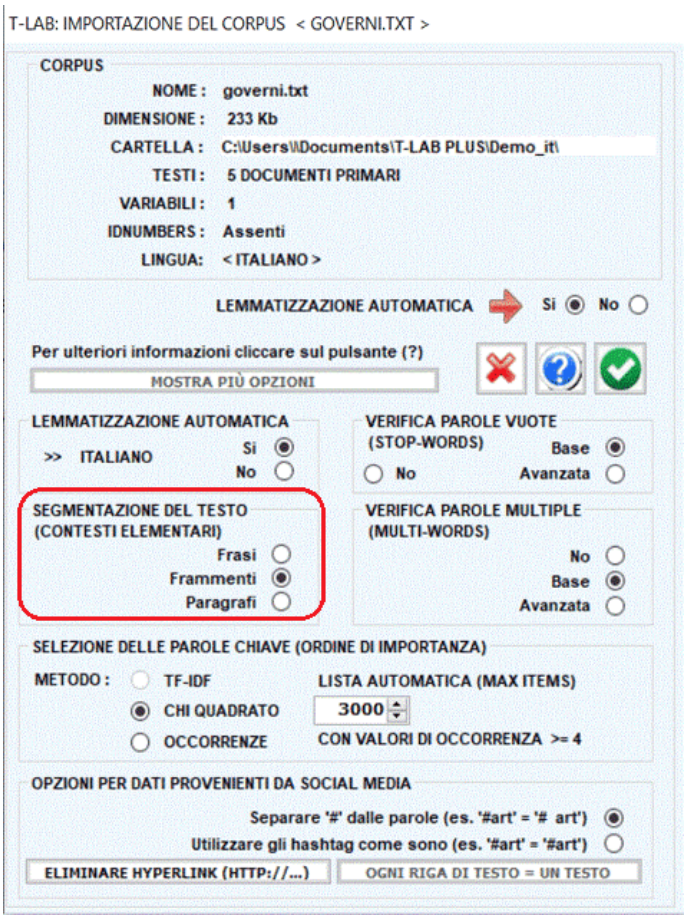
Codifica

Prima dell'importazione del corpus, l'utilizzatore può inserire delle righe di codifica all'inizio di ogni **unità di contesto** che desidera classificare mediante l'uso di una o più **variabili**.

Normalmente, le unità di contesto codificate corrispondono ai **documenti primari**.

Contesto elementare

Nella fase di importazione, **T-LAB** effettua una **segmentazione** del **corpus** in **contesti elementar**i: ciò per facilitarne l'esplorazione da parte dell'utilizzatore e, soprattutto, per effettuare analisi che richiedono il calcolo delle **co-occorrenze**.



Ne risulta che, a seconda delle scelte dell'utilizzatore, i contesti elementari possono essere di quattro tipi:

1 - **Frasi**

Contesti elementari marcati dalla punteggiatura forte (.? !), con lunghezza minima di 50 caratteri (Max. 1000 caratteri).

2 - **Frammenti**

Contesti elementari di lunghezza comparabile costituiti da uno o più enunciati.

In questo caso, l'algoritmo di segmentazione rispetta le seguenti regole:

- considerare come contesto elementare ogni sequenza di parole interrotta dal "punto e capo" (ritorno di carrello) e le cui dimensioni siano inferiori 400 caratteri;

- nel caso in cui, entro la lunghezza massima, non sia presente alcun punto e a capo, cercare, nell'ordine, altri segni di punteggiatura (? ! ; : ,). Se non vengono trovati, segmentare in base a un criterio statistico, ma senza troncare le unità lessicali.

3 - **Paragrafi**

Contesti elementari marcati dalla punteggiatura forte (.? !) e dal ritorno di carrello, con lunghezza massima di 2000 caratteri.

4 - **Testi Brevi**

Questa opzione è abilitata solo quando il corpus è costituito da testi con dimensione massima di 2000 caratteri (es. risposte a domande aperte).

N.B.:

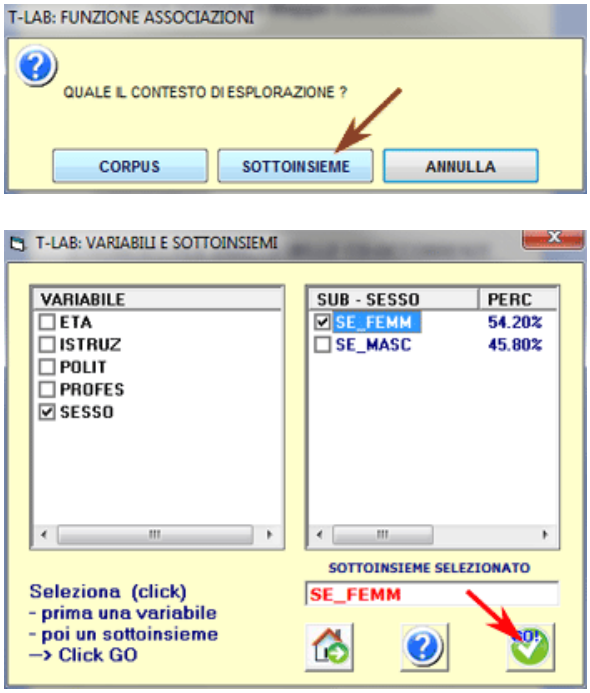
- il file **corpus_segments.dat** contiene il risultato della segmentazione del corpus;
- la funzione **Concordanze** consente la verifica dei contesti elementari in cui ogni parola (forma grafica o lemma) è presente.

Corpus e Sottoinsiemi

Corpus: collezione di uno o più testi selezionati per un lavoro di analisi.

Sottoinsieme: una parte del corpus definita tramite l'uso di **variabili** e **modalità**.

T-LAB consente - in modo automatico - di esplorare e di analizzare le relazioni tra le unità di analisi di tutto il **corpus** o di suoi **sottoinsiemi**.



Qualche esempio di **corpus**:

- un singolo testo o documento che tratti un qualunque argomento;
- un insieme di articoli tratti dalla stampa e che affrontano lo stesso argomento;
- una o più interviste realizzate entro un progetto di ricerca;
- uno o più libri dello stesso autore o che affrontano temi simili;
- una mailing-list scaricata da internet;
- un insieme di risposte a una "domanda aperta" di un questionario;
- una o più trascrizioni di focus group.

Qualche esempio di **sottoinsieme**:

- nel caso di un corpus costituito da articoli pubblicati in vari anni (es ANNO = variabile usata), tutti gli articoli di un determinato anno (es 2001 = modalità della variabile anno);
- nel caso di risposte a domande aperte, tutte le risposte di una determinata categoria di persone (es FEM = modalità della variabile SESSO);
- nel caso di un corpus suddiviso per aree tematiche (es TEMA = variabile), tutte le parti che si riferiscono allo stesso tema (es SCUOLA = modalità della variabile TEMA).

N.B.: Sottoinsiemi del corpus sono anche i "**cluster tematici**" di documenti o di contesti elementari ottenuti utilizzando i corrispondenti strumenti **T-LAB**.

Nel caso di un corpus costituito da più testi, perché questo sia un **insieme utilmente analizzabile**, si richiede che le sue parti abbiano due caratteristiche che li rendano comparabili:

a) una qualche omogeneità tematica e/o del contesto in cui sono stati prodotti, in modo da ottenere dati tra loro confrontabili;

b) un equilibrato rapporto tra le loro dimensioni, sia in termini di occorrenze sia in termini di Kbytes, per non incorrere in "anomalie" di tipo statistico.

Entro la logica di **T-LAB**, il corpus è un **database** organizzato in **record** e **campi**. Più precisamente, i record sono costituiti dalle entità archiviate (testi, frammenti di testi, parole) e i campi sono costituiti dalle variabili utilizzate per classificare le varie entità (gli autori dei testi, i contesti di riferimento, i tipi di temi, etc.).

Vedi **Preparazione del Corpus.**

Disambiguazione

Operazione attraverso la quale si cerca di risolvere i casi di **ambiguità** semantica, in particolare quelli attribuibili agli **omografi**, cioè alle parole (forme o lemmi) con **forma grafica** equivalente ma di diverso significato.

 N.B.: In **T-LAB 10** specifiche funzioni per la disambiguazione sono implementate nello strumento **Text Screening**; inoltre, nella fase di importazione, **T-LAB** riconosce e "distingue" tre tipi di oggetti linguistici:

- nomi propri (di persone e di luoghi);
- locuzioni e **multiwords**;
- i tempi composti dei verbi.

In tutti e tre i casi vengono utilizzate liste presenti nel database, costruite e testate per limitare i casi più frequenti di ambiguità (criterio di **efficacia**) e per contenere i tempi di elaborazione (criterio di **efficienza**).

Dizionario

I dizionari **T-LAB** sono tabelle o file che contengono schemi di classificazione delle **unità lessicali**.

Gli schemi di classificazione, e quindi i dizionari, possono essere di due tipi: (a) basati su **caratteristiche linguistiche** o (b) basati su **categorie tematiche**. Entrambi possono essere esportati e personalizzati.

Nel primo caso ('a'), cioè rinominare o raggruppare item presenti nell'elenco di parole chiave, l'utente può fare riferimento allo strumento **Personalizzazione del Dizionario.**

Nel secondo caso ('b'), cioè creare / usare un dizionario per una classificazione supervisionata, l'utente può fare riferimento a qualsiasi strumento **T-LAB** per l'analisi tematica (ad es. **Classificazione basata su Dizionari**, **Classificazione Tematica di Documenti**, etc.).

Documento primario

I documenti primari sono testi (o parti del corpus) che corrispondono alle unità di contesto precedute da una riga di **codifica**.

A seconda dei casi, essi possono essere: libri o capitoli di libri, articoli di quotidiani, trascrizioni di interviste, risposte a domande aperte etc.

Forma e Lemma

I software per l'analisi dei testi, in primo luogo, riconoscono le cosiddette **forme grafiche**, ovvero le stringhe di caratteri separati da spazi vuoti. Poi, a seconda degli algoritmi implementati o a seconda delle categorie utilizzate dagli studiosi, si passa ai **lessemi**, ai **lemmi**, alle **parole chiave**, etc.

Le tabelle **T-LAB**, per tutte le unità lessicali presenti nel database del corpus, riportano due informazioni:

- la prima, denominata **forma**, contiene la trascrizione delle unità lessicali (singole parole, lessie o multiword) come "stringhe" riconosciute dal software;

- la seconda, denominata **lemma**, contiene le label (o tag) con la quali sono state raggruppate e classificate le unità lessicali.

A seconda dei casi, il lemma può essere:

- il risultato del processo di lemmatizzazione automatica;
- una voce di un "dizionario personalizzato";
- una categoria che indica un gruppo di sinonimi;
- una categoria di analisi del contenuto;
- etc.

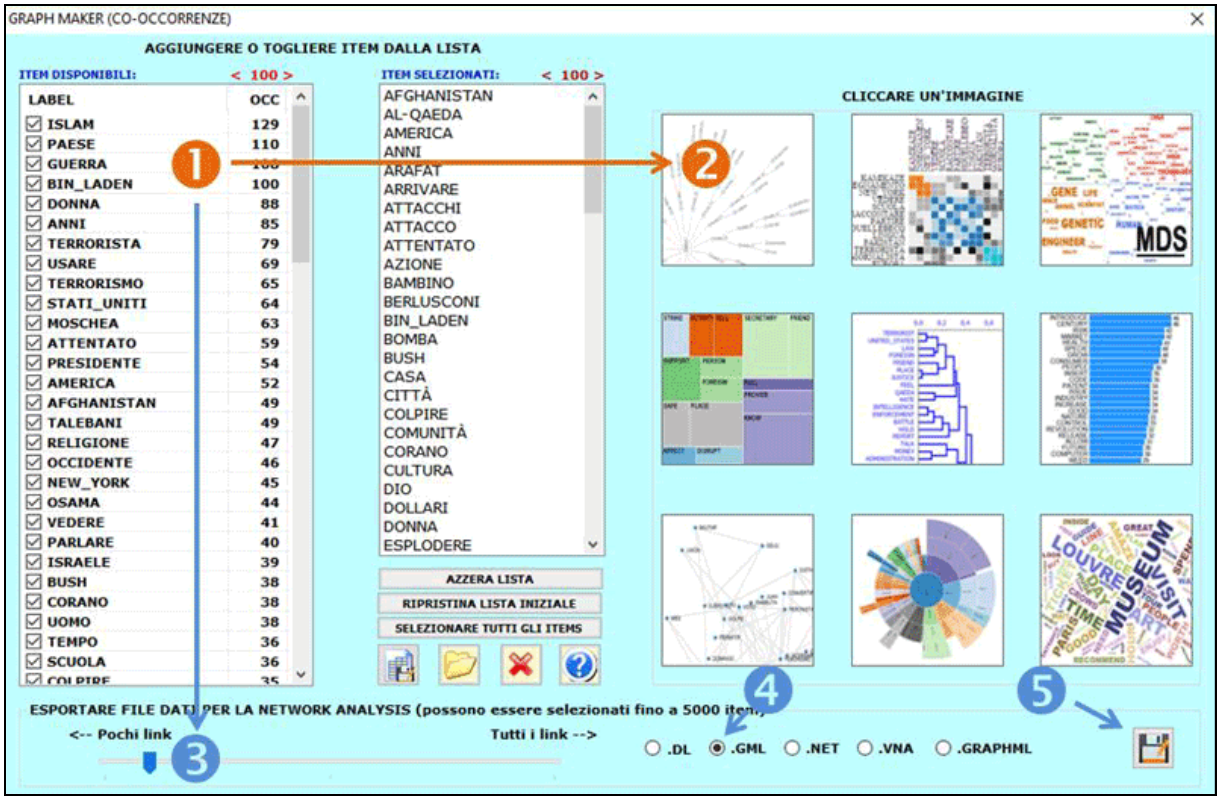
Graph Maker

Lo strumento **Graph Maker** permette all'utilizzatore di creare ed esportare vari tipi di grafici dinamici che possono essere utilizzati per due obiettivi:

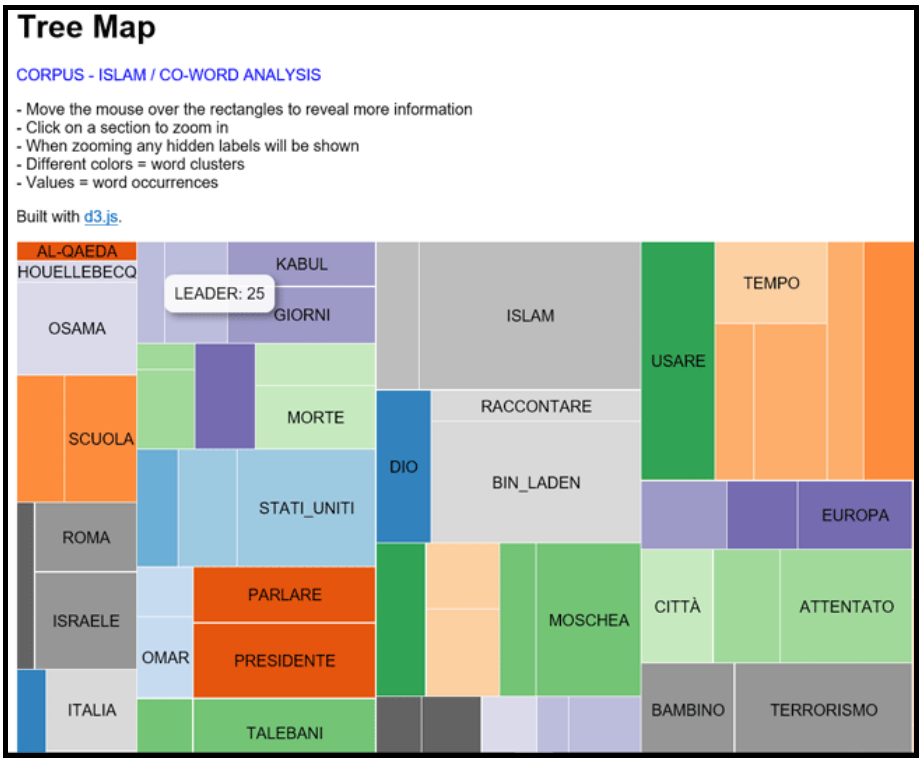
(a) esplorare le **relazioni di co-occorrenza** tra parole;
(b) effettuare un qualche tipo di **network analysis**.

Nel caso (a) si richiedono solo due passaggi (vedi immagine a seguire):
1- Selezionare gli item (cioè le parole-chiave) da utilizzare;
2- Cliccare una qualsiasi immagine per visualizzare il grafico corrispondente.

Nel caso (b), dopo la selezione delle parole-chiave (vedi sotto punto '1'), l'utilizzatore può filtrare i link da utilizzare (vedi sotto punto '3'), quindi può scegliere il formato dell'output (vedi sotto punto '4') e cliccare sul pulsante 'salva' (vedi sotto punto '5').



N.B.: Ogni output in formato HTML include alcune semplici istruzioni che ne aiutano l'esplorazione (vedi immagine a seguire).


Tree Map
CORPUS - ISLAM / CO-WORD ANALYSIS
- Move the mouse over the rectangles to reveal more information
- Click on a section to zoom in
- When zooming any hidden labels will be shown
- Different colors = word clusters
- Values = word occurrences
Built with d3.js.
AL-QAEDA
HOUELLEBECQ
OSAMA
KABUL
LEADER: 25
GIORNI
ISLAM
TEMPO
USARE
SCUOLA
MORTE
RACCONTARE
DIO
BIN_LADEN
EUROPA
ROMA
STATI_UNITI
ISRAELE
PARLARE
OMAR
PRESIDENTE
MOSCHEA
CITTÀ
ATTENTATO
ITALIA
TALEBANI
BAMBINO
TERRORISMO

IDnumber

IDnumber è una label che può essere inserita nelle righe di codifica come identificativo dei soggetti (es nel caso di risposte a domande aperte) o delle unità di contesto in cui è suddiviso il corpus da importare (vedi **Preparazione del corpus**).

In **T-LAB**, ogni volta che viene utilizzata, la label "IDnumber" deve essere seguita da un trattino basso ("_") e da un numero progressivo di max 5 cifre (vedi esempio seguente).

**** *IDnumber_0001 *ETA_ADUL *SES_FEM *PROF_OPER
Segue il testo di una risposta o di un documento.
……

**** *IDnumber_0002 *ETA_GIOV *SES_MAS *PROF_IMP
Segue il testo di una risposta o di un documento.
……

**** *IDnumber_0003 *ETA_ADUL *SES_MAS *PROF_OPER
Segue il testo di una risposta o di un documento.
……

Ogni corpus può includere numerazioni progressive (IDnumber) di max 30.000 soggetti o unità di contesto.

N.B.:
Il primo valore dell'IDnumber deve essere "1" (es. IDnumber_00001).
Nel caso in cui i testi raccolti dall'utilizzatore siano in un formato MS Excel, nel pacchetto di installazione T-LAB è disponibile una macro che in modo automatico li trasforma in un corpus codificato e pronto per l'importazione.

Indici di associazione

In **T-LAB** gli indici di associazione (o di similarità) sono utilizzati per analizzare le **co-occorrenze** delle **unità lessicali** (**LU**, lexical units) all'interno dei **contesti elementari (EC,** elementary context**s)**, cioè dati binari del tipo presenza/assenza.

Ad esempio, dati due **LU** e dieci **EC**, possiamo costruire il seguente esempio:

	EC_1	EC_2	EC_3	EC_4	EC_5	EC_6	EC_7	EC_8	EC_9	EC_10
LU_1	1	0	1	1	1	0	1	0	1	1
LU_2	0	1	0	1	0	0	1	1	0	1

Gli stessi dati possono essere rappresentati nel modo seguente:

	LU_2		
LU_1	*Present*	*Absent*	*Total*
Present	*3*	*4*	*7*
Absent	*2*	*1*	*3*
Total	*5*	*5*	*10*

Generalizzando e utilizzando le lettere dell'alfabeto:

	LU_2		
LU_1	*Present*	*Absent*	*Total*
Present	*a*	*b*	*a + b*
Absent	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*n*

Le formule corrispondenti ai sei indici di associazioni usati da **T-LAB** sono le seguenti:

Jaccard

$$\frac{a}{a+b+c}$$

Dice

$$\frac{2a}{2a+b+c}$$

Coseno

$$\frac{a}{\sqrt{(a+b)} \times \sqrt{(a+c)}}$$

Equivalenza

$$\frac{a^2}{(a+b)\times(a+c)}$$

Inclusione

$$\frac{a}{\text{Min}((a+b),(a+c))}$$

Mutua Informazione

$$\text{Log}\frac{a/N}{(a+b)\times(a+c)}$$

Ipotizzando di aver ottenuto indici di associazione delle relazioni tra dieci **LU**, possiamo costruire una tabella come la seguente:

	LU_1	LU_2	LU_3	LU_4	LU_5	LU_6	LU_7	LU_8	LU_9	LU_10
LU_1		0,067	0,048	0,286	0,154	0,077	0,060	0,309	0,231	0,077
LU_2	0,067		0,269	0,134	0,000	0,072	0,056	0,072	0,072	0,072
LU_3	0,048	0,269		0,048	0,156	0,104	0,040	0,052	0,052	0,156
LU_4	0,286	0,134	0,048		0,077	0,000	0,060	0,154	0,000	0,077
LU_5	0,154	0,000	0,156	0,077		0,667	0,000	0,000	0,000	0,333
LU_6	0,077	0,072	0,104	0,000	0,667		0,000	0,000	0,000	0,417
LU_7	0,060	0,056	0,040	0,060	0,000	0,000		0,129	0,129	0,000
LU_8	0,309	0,072	0,052	0,154	0,000	0,000	0,129		0,167	0,083
LU_9	0,231	0,072	0,052	0,000	0,000	0,000	0,129	0,167		0,000
LU_10	0,077	0,072	0,156	0,077	0,333	0,417	0,000	0,083	0,000	

Di fatto, **T-LAB** costruisce ed analizza tabelle analoghe di dimensioni N x N (dove N può corrispondere a varie centinaia di colonne), sia mediante **Multidimensional Scaling** che mediante **Cluster Analysis**.

Tabelle simili sono anche utilizzate per calcolare indici di similarità del secondo ordine tra coppie di parole chiave (vedi lo strumento Associazioni di Parole).

Isotopia

La nozione di isotopia (iso = uguale, topos = luogo) rinvia a una concezione del significato come "effetto del contesto", cioè come qualcosa che non appartiene alle parole prese singolarmente, bensì che risulta dai loro rapporti all'interno dei testi.

La funzione delle isotopie è quella di facilitare l'interpretazione dei discorsi (o dei testi); in effetti, ciascuna di esse individua un contesto di riferimento "condiviso" da più parole, che non deriva però dai loro specifici significati. Ciò nella logica che l'insieme è qualcosa di diverso dalla sommatoria dei suoi elementi.

Il riconoscimento di un'isotopia, quindi, non è la mera constatazione di un "dato", bensì il risultato di un processo interpretativo (F. Rastier 1987).

La nozione di isotopia è stata inizialmente proposta dal semiologo A.J. Greimas (1966) per definire la ricorrenza, all'interno delle unità sintagmatiche (frasi e/o testi), di più parole con tratti semantici in comune tra loro.

Nella logica di **T-LAB**, la rilevazione delle isotopie deriva dall'analisi delle occorrenze e delle co-occorrenze.

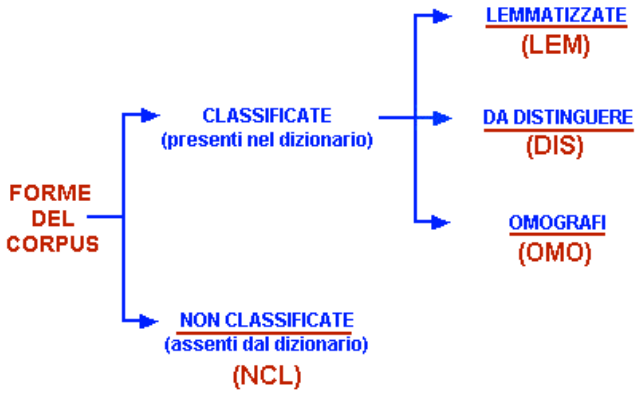
Lemmatizzazione

Nei dizionari linguistici che consultiamo, ogni voce corrisponde a un lemma che - generalmente - definisce un insieme di parole con la stessa radice lessicale (o lessema) e che appartengono alla stessa categoria grammaticale (verbo, aggettivo, etc.).

Di norma, la **lemmatizzazione** comporta che le forme dei verbi vengono ricondotte all'infinito presente, quelle dei sostantivi e degli aggettivi al maschile singolare, quelle delle preposizioni articolate alla loro forma senza articolo, e così via.

Ad esempio, le **forme flesse** "parliamo" e "parlato", risultanti dalla combinazione di un'unica **radice** (<parl->) con due diversi suffissi (<-iamo>, <-ato>), sono ricondotte allo stesso lemma "parlare".
Si danno tuttavia dei casi in cui la lemmatizzazione non segue la regola della radice comune; in particolare, nella categoria dei verbi irregolari. Ad esempio, "vado" e "andremo" sono entrambe forme del lemma "andare".

Nella fase di importazione del **corpus**, **T-LAB** consente di effettuare un particolare tipo di lemmatizzazione automatica che segue la logica del seguente "albero".



Ovviamente, il dizionario di riferimento è quello implementato in **T-LAB**.

Le sigle delle quattro categorie sono utilizzate in molte tabelle, sempre nella colonna (o campo) "INF".

N.B.:
- la categoria "DIS" ("da distinguere") è costituita dai casi in cui **T-LAB** riconosce parole - in generale, nomi e aggettivi - per le quali è opportuno non applicare la lemmatizzazione standard; ciò per evitare che vengano appiattite le differenze tra i diversi significati delle forme singolari e plurali (ad es. "beni" e "bene", "culture" e "cultura"), oppure delle forme femminili e maschili ("singola" e "singolo", "tecnica" e "tecnico");
- a volte, per marcare casi di omografia, **T-LAB** aggiunge il carattere ('_') a uno dei lemmi corrispondenti.

Lessia e Lessicalizzazione

Secondo Pottier (vedi **Bibliografia**), la **lessia** è una espressione costituita da una o più parole che si comportano come una unità lessicale con significato autonomo.

I suoi tipi fondamentali sono tre: *semplice*, corrispondente alla parola nel senso comune del termine (es. "*cavallo*", "*mangiava*"); *composta*, costituita da due o più parole integrate in un'unica forma (es. "*biotecnologie*", "*mangianastri*"); *complessa*, costituita da una sequenza in via di lessicalizzazione (es. "*a mio giudizio*", "*complesso industriale*").

La **lessicalizzazione** è il processo linguistico attraverso il quale un sintagma o un raggruppamento di parole diventa una unità lessicale o come tale si comporta.

In **T-LAB** la funzione **Locuzioni e Multiwords** consente di costruire una lista delle lessie complesse presenti nel corpus e di procedere alla loro trasformazione in stringhe unitarie (lessicalizzazione).

MDS (Muldimensional Scaling)

Insieme di tecniche statistiche che consentono di analizzare matrici di similarità e di rappresentare le relazioni tra i dati entro uno spazio di dimensioni ridotte.

In **T-LAB** un tipo di **MDS** (metodo Sammon) è usato per rappresentare le relazioni tra unità lessicali o tra nuclei tematici (vedi **Co-Word Analysis** e **Modellizzazione dei Temi Emergenti**).

I dati in analisi sono costituiti da matrici quadrate in cui sono riportati valori di prossimità (dissimilarità) derivati dal calcolo di un indice di associazione.

I risultati ottenuti, analogamente a quelli dell'analisi delle corrispondenze, consento di interpretare sia le relazioni tra gli "oggetti" (vicinanza/distanza), sia le dimensioni che organizzano lo spazio in cui essi sono rappresentati.

La bontà dell'adattamento, cioè il grado di corrispondenza tra le distanze risultanti dalla mappa MDS e quelle della matrice input, è misurata dalla funzione di Stress. Minore è il valore dello stress (es. < 0.10), maggiore è la bontà dell'adattamento.

La formula dello stress è la seguente:

$$S = \sum_{i \neq j} \frac{(d_{ij}^{*} - d_{ij})^2}{d_{ij}^{*}}$$

dove d_{ij}^{*} indica le distanze tra i punti (ij) nella matrice input e d_{ij} indica le distanze tra gli stessi punti nella mappa MDS.

Multiwords (Parole Multiple)

Sequenza di due o più parole che, al livello del significato, costituiscono una **unità lessicale**.

La categoria delle multiwords, i cui confini dipendono dal modello analitico adottato, include sottoinsiemi quali i **nomi composti** (ad es. "mezzi pubblici", "livello di occupazione") e le **locuzioni** usate come modi di dire (es. "nella misura in cui", "a buon fine", "a onor del vero").

La lista delle multiwords implementata in **T-LAB**, ovviamente, non è esaustiva.
E' stata costruita e testata con due criteri:

a) limitare i casi più frequenti di ambiguità (criterio di **efficacia**);

b) contenere i tempi di esecuzione della procedura **normalizzazione** (criterio di **efficienza**).

In **T-LAB** è anche possibile utilizzare una **lista personalizzata di Multiwords**.

N-Grammi

In **T-LAB** un n-gramma è una sequenza di due (bi-gramma) o più parole chiave presenti all'interno dello stesso **contesto elementare**.

Il suo uso è riservato al calcolo delle **co-occorrenze** e, all'interno dello stesso contesto elementare, la contiguità delle parole considerate non tiene conto né delle 'parole vuote' (cioè stop-word) né della punteggiatura.

Prendiamo, ad esempio, il seguente contesto elementare:

L'Italia è una Repubblica democratica, fondata sul lavoro.

Supponendo che i cinque item in rosso siano inclusi nella nostra lista di parole chiave, la suddivisione in bi-grammi produce i seguenti contesti di co-occorrenza:

Italia & Repubblica
Repubblica & democratica
democratica & fondata
fondata & lavoro.

Diversamente, nel caso di tri-grammi il risultato sarebbe il seguente:

Italia & Repubblica & democratica
Repubblica & democratica & fondata
democratica & fondata & lavoro
fondata & lavoro.

E' importante sottolineare che, nel caso dei contesti elementari, le co-occorrenze sono basate sulla presenza delle parole nello stesso 'luogo' (es. frase, paragrafo etc.); diversamente, nel caso degli n-grammi, le co-occorrenze sono basate su una relazione di contiguità.

In **T-LAB** l'analisi delle co-occorrenze basate su n-grammi può essere realizzata con lo strumento **Associazioni di Parole**. Inoltre, l'analisi markoviana dei bi-grammi può essere effettuata con lo strumento **Analisi delle Sequenze**.

Naïve Bayes Classifier

La formula del Naïve Bayes Classifier (NB) usata in **T-LAB** è la seguente:

$$v_{\mathbf{NB}} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

Dove:

argmax – si riferisce al massimo valore della probabilità a posteriori;
$v_j \in V$ - si riferisce al j-cluster (v_j) della partizione (V);
$P(v_j)$ - si riferisce alla probabilità a priori di ogni j-cluster;
$\prod_i P(a_i \mid v_j)$ - è il prodotto delle probabilità di ogni (a_i) parola per ogni (v_j) cluster, dove ogni $P(a_i \mid v_j)$ è un elemento di un vettore normalizzato di frequenze relative e ogni (a_i) parola è presente all'interno della i-unità di contesto da riclassificare.

Normalizzazione del corpus

In **T-LAB** , la normalizzazione del corpus ha un duplice obiettivo:

a) consentire una corretto riconoscimento delle parole come forme grafiche.

b) risolvere preliminarmente alcuni casi di ambiguità.

Ciò comporta che **T-LAB**, in primo luogo, effettua una serie di trasformazioni del file in analisi: eliminazione di spazi vuoti in eccesso, marcatura degli apostrofi, aggiunta di spazi dopo i segni di interpunzione, riduzione delle maiuscole, etc.

In secondo luogo, **T-LAB** marca una serie di stringhe riconosciute come nomi **propri** (di persone e luoghi); quindi trasforma le sequenze di forme grafiche riconosciute come locuzioni o **multiwords** in stringhe unitarie da utilizzare come tali nel processo di analisi ("nella misura in cui" e "il punto di vista" diventano quindi rispettivamente "nella_misura_in_cui" e "il_punto_di_vista).

I parametri di queste operazioni non sono modificabili dall'utilizzatore.

*Nella fase di normalizzazione, per il corretto riconoscimento delle forme grafiche, in **T-LAB** viene utilizzata la seguente lista di **separatori**:*

, ; : . ! ? ' " ( ) < > + / = [ ] { }

Nuclei tematici

In **T-LAB**, la dizione **nuclei tematici** viene utilizzata in alcune funzioni che producono mappe delle **parole chiave**.

Essa sta ad indicare i piccoli cluster di parole, **co-occorrenti** nei contesti elementari del corpus, che - sulle mappe - vengono rappresentati con etichette che possono essere definite e cambiate dall'utilizzatore.

Occorrenze e co-occorrenze

In analisi dei testi, queste due nozioni sono di fondamentale importanza.

Le **occorrenze** sono quantità risultanti dal conteggio del numero di volte (frequenze) in cui una unità lessicale (**LU**, lexical unit) ricorre all'interno del **corpus** o delle unità di contesto (**CU**, context unit) in cui è suddiviso.

La loro distribuzione può essere rappresentata in tabelle di contingenza come la seguente:

	CU_1	CU_2	CU_3	CU_4
LU_1	19	1	12	14
LU_2	17	0	1	8
LU_3	8	4	2	9
LU_4	101	0	13	0
LU_5	32	1	29	11
LU_6	4	3	0	30
LU_7	10	1	3	21
LU_8	5	1	1	34
LU_9	25	5	0	54

Le **co-occorrenze** sono quantità risultanti dal conteggio del numero di volte in cui due o più

unità lessicali sono contemporaneamente presenti - all'interno degli stessi contesti elementari (**EC**, elementary contexts).

La loro distribuzione può essere rappresentata in tabelle del tipo presenza/assenza come quella seguente:

(A)

	LU_1	LU_2	LU_3	...	LU_n
EC_1	0	1	0	...	1
EC_2	1	0	0	...	0
EC_3	0	1	1	...	0
EC_4	0	0	0	...	0
EC_5	1	1	0	...	1
EC_6	0	0	0	...	0
EC_7	0	0	1	...	0
EC_8	1	0	0	...	0
EC_9	0	0	0	...	0
EC_10	0	1	0	...	0
EC_11	1	0	1	...	0
EC_12	0	0	0	...	1
EC_13	1	1	0	...	0
EC_14	0	0	1	...	0
EC_15	0	0	0	...	0
EC_16	0	1	0	...	1
EC_17	0	0	1	...	0
EC_18	0	0	0	...	0
EC_19	1	0	0	...	0
EC_20	0	0	0	...	1

Con una semplice trasformazione, le tabelle del tipo “A” (rettangolare) possono essere trasformate in tabelle del tipo “B” (quadrate e simmetriche) in cui per ogni coppia di unità lessicale è indicata la quantità delle loro co-occorrenze, cioè il totale di contesti elementari in cui sono contemporaneamente presenti.

(B)

	LU_1	LU_2	LU_3	...	LU_n
LU_1		2	1	...	1
LU_2	2		1	...	3
LU_3	1	1		...	0
...	...	...	...		...
LU_n	1	3	0	...	

In gran parte - in T-LAB - l'analisi dei testi si realizza attraverso lo studio delle relazioni tra occorrenze e tra co-occorrenze: ciò, sia attraverso particolari **indici di associazione**, sia attraverso l'uso di tecniche statistiche di tipo multidimensionale quali la **cluster analysis** e **l'analisi delle corrispondenze**.

Omografia

Due o più parole (forme o lemmi) sono **omografe** quando hanno la stessa forma grafica (sono scritte allo stesso modo), ma hanno significati diversi.

Nella lingua italiana, casi di questo tipo sono migliaia.

In **T-LAB** sono implementate delle procedure di **disambiguazione** che ne riducono l'incidenza; in particolare, la normalizzazione di locuzioni, **multiwords** e tempi composti dei verbi.

In questo modo - ad esempio - la normalizzazione della sequenza "il punto di vista" (trasformata in "il_punto_di_vista"), consente di distinguere le specifiche occorrenze di "punto" e "vista" (due classici omografi); così come la normalizzazione della sequenza "sono stato" ("sono_stato") consente di distinguere i casi in cui si parla dello "stato" come condizione o come forma politica.

Parole chiave

All'interno della logica **T-LAB** sono Parole Chiave tutte le **unità lessicali** (parole, lemmi, lessie, categorie) che, di volta in volta, vengono incluse nelle tabelle da analizzare.

Operativamente, la selezione delle parole chiave può essere effettuata secondo due modalità: **automatica** e **personalizzata**.

Delle due, solo la seconda consente di modificare le liste delle unità lessicali e di usare **dizionari personalizzati**.

Polarità fattoriali

In **analisi delle corrispondenze** ciascun fattore organizza una dimensione spaziale - rappresentabile come una linea o un asse - al cui centro (o baricentro) è il valore "0" e che si sviluppa in modo bi-polare verso le estremità "negativa" (-) e "positiva" (+); in modo tale che gli oggetti collocati sui poli opposti sono quelli più diversi tra loro, un pò come la "sinistra" e la "destra" sull'asse della politica.

A questo proposito, val la pena di ricordare quanto ha scritto J.P. Benzecri, uno dei matematici che più ha contribuito a definire questo specifico modello di analisi:

"Interpretare un asse fattoriale significa trovare ciò che vi è di analogo, da una parte tra tutto ciò che è situato a destra dell'origine (o baricentro), dall'altra tra tutto ciò che è alla sinistra di questo, ed esprimere poi con concisione ed esattezza l'opposizione tra i due estremi" (1984, p. 302). (Vedi **Bibliografia**)

N.B.: *Quando i grafici fattoriali sono di tipo bidimensionale (o tridimensionale), le opposizioni sono più di due: oltre alla sinistra e alla destra, c'è l'alto e il basso. Tuttavia i criteri di interpretazione restano invariati.*

Profilo

In **T-LAB** il profilo di una **unità di analisi** (unità lessicale o unità di contesto) corrisponde al vettore (riga o colonna) della tabella dati che contiene i suoi valori di **occorrenza** o di **co-occorrenza**.

N.B.:

Nell'**Analisi delle Corrispondenze** sono **attivi** i profili che corrispondono a righe o colonne delle tabelle analizzate e che intervengono nella costruzione degli assi fattoriali; mentre sono **illustrativi** (o supplementari) i profili di righe o colonne i cui valori sono calcolati a posteriori.

Soglia di Frequenza

Nella fase di importazione **T-LAB** calcola una soglia di frequenza minima per selezionare le parole (forme o lemmi) da inserire nelle analisi del menu **configurazioni automatiche** e, in particolare, per costruire l'elenco delle **Parole-Chiave**.

In ogni caso, per garantire l'affidabilità di alcuni calcoli statistici, la soglia minima **T-LAB** è fissata a 4.

Per questo calcolo viene utilizzato un algoritmo documentato in uno dei volumi in ***bibliografia*** *(Bolasco S., 1999), e che prevede i seguenti passi:*

a) *individuazione del range delle frequenze basse, che - a partire dalla frequenza minima ("1") - è definito dal primo "salto" nei valori crescenti delle occorrenze;*

b) *scelta del valore di soglia che, a seconda delle dimensioni del corpus, viene fatto corrispondere al valore minimo nel primo o nel secondo decile (10% o 20%) del range.*

Specificità

Analisi delle Specificità è il nome di uno strumento **T-LAB** che permette di verificare le unità lessicali (vale a dire: parole, lemmi o categorie) e i contesti elementari (vale a dire: frasi o paragrafi) che sono tipici (o 'caratteristici') di un testo o un sottoinsieme del corpus definito da una variabile categoriale.

Le **tipiche unità lessicali**, definite dalla proporzione delle rispettive occorrenze (vale a dire dal loro sovra/sotto utilizzo), sono individuate tramite il calcolo **chi-quadrato** o del **valore test**.

I **tipici contesti elementari** vengono individuati calcolando e sommando i valori **TF-IDF normalizzati** assegnati alle parole di cui ogni frase o paragrafo è costituito.

Stop word list

Nella pratica dell'analisi dei testi, molte parole vengono definite "vuote" in quanto - da sole - non veicolano contenuti specifici e/o rilevanti.

Per costruire un loro elenco (**Stop Word List**) non esiste un criterio standard.

In **T-LAB** l'elenco è tratto dalle seguenti categorie:

- aggettivi indefiniti
- articoli
- avverbi
- esclamazioni
- interiezioni
- preposizioni
- pronomi (dimostrativi, indefiniti e relativi)
- verbi ausiliari (essere, avere, andare, venire)
- verbi modali (dovere, parere, potere, sapere, sembrare, solere, volere).

In ogni caso, l'utilizzatore può importare **liste personalizzate di StopWords.**

TF-IDF

Questa misura, proposta da Salton (1989) nell'ambito dell'Information Retrieval, consente di valutare l'importanza di un termine (unità lessicale) all'interno di un documento (unità di contesto).

La sua formula è la seguente

:

w *i,j* = ***tf*** *i,j* *x* ***idf*** *i* (*Term Frequency* x *Inverse Document Frequency*)

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Dove:

tf *i,j* = numero delle occorrenze di ***i*** (un termine) all'interno di ***j*** (un documento)
df *i* = numero dei documenti che contengono ***i***
N = totale dei documenti che costituiscono il corpus in analisi.

Il valore ***tf*** *i,j* (Term Frequency) può essere normalizzato nel modo seguente:

tf *i,j* = ***tf*** *i,j* / Max (***f*** *i,j*)

dove Max (***f*** *i,j*) è la frequenza massima di **i** (un qualunque termine) all'interno di **j** (documento).

Tabelle dati

Le tabelle (o **matrici**) dati sono costituite da righe e colonne, e da valori registrati nelle rispettive "caselle". Esse consentono di sintetizzare - in modo ordinato - sia le osservazioni da sottoporre ad analisi statistiche (input), sia i risultati ottenuti attraverso la loro applicazione (output).

Per più di una ragione, gli statistici affermano che la buona riuscita di un'analisi è affidata alla costruzione di una "buona tabella".

In **T-LAB,** a seconda del tipo di analisi, le tabelle possono essere di tre tipi, corrispondenti ad altrettanti modi di costruire incroci tra righe e colonne:

- lemmi (o forme) in riga e testi (o **variabili**) in colonna;
- testi (o frammenti di testi) in riga e lemmi (o forme) in colonna;
- lemmi (o forme) sia in riga che in colonna.

A seconda dei casi, i valori riportati nelle celle sono **occorrenze o co-occorrenze**.

Unità di analisi

Le **unità di analisi** archiviate nel database **T-LAB** sono di due tipi: **unità lessicali** e **unità di contesto**.

A - le **unità lessicali** (**LU**, lexical units) sono **parole**, singole o "multiple", archiviate e classificate in base a un qualche criterio. Più precisamente, nel database **T-LAB** ogni unità lessicale costituisce un record classificato con due campi: **forma** e **lemma**. Nel primo campo, denominato "forma", sono elencate le parole così come compaiono nel corpus, mentre nel secondo, denominato "lemma", sono elencate le label attribuite a gruppi di unità lessicali classificate secondo criteri linguistici (es. **lemmatizzazione**) o tramite **dizionari** e **griglie semantiche** definite dall'utilizzatore.

B - le **unità di contesto** (**CU,** context units) sono porzioni di testo in cui può essere suddiviso il corpus. Più esattamente, nella logica **T-LAB**, le unità di contesto possono essere di tre tipi:

- **B.1** **documenti primari**, corrispondenti alla suddivisione "naturale" del corpus (es. interviste, articoli, risposte a domande aperte, etc.), ovvero ai **contesti iniziali** definiti dall'utilizzatore;
- **B.2** **contesti elementari**, corrispondenti a unità sintagmatiche di una o più frasi e definiti in modo automatico (o semi-automatico) da **T-LAB**. Quindi, nel database **T-LAB** ogni documento primario risulta costituito da uno o più contesti elementari;
- **B.3** **sottoinsiemi del corpus**, corrispondenti a gruppi di documenti primari riconducibili alla stessa "categoria" (es. interviste di "uomini" o di "donne", articoli di un particolare anno o di una particolare testata, e così via);

Unità di Contesto

Vedi **unità di analisi**.

Unità Lessicale

Vedi **unità di analisi**.

Valore Test

Si tratta di una misura statistica che **T-LAB** utilizza per misurare e caratterizzare due tipi di relazioni:

a) quelle tra una qualsiasi unità lessicale e una qualsiasi categoria di variabile, i cui rispettivi valori di occorrenza siano riportati in una tabella di contingenza;

b) quelle riguardanti qualsiasi riga o colonna di una tabella di contingenza con i fattori estratti tramite un'analisi delle corrispondenze della stessa tabella.

A seconda delle relazioni analizzate, le formule del valore test, tratte da uno dei volumi in bibliografia (Lebart L. Morineau A. Piron M., 1995, pp 181-184), sono le seguenti:

a)

$$t_k(j) = \frac{n_{jk} - n_k \cdot \frac{n_j}{n}}{\sqrt{n_k \cdot \frac{n - n_k}{n - 1} \cdot \frac{n_j}{n} \cdot \left(1 - \frac{n_j}{n}\right)}}$$

dove 'njk' indica le occorrenze all'interno di una cella, mentre 'nj' e 'nk' corrispondono rispettivamente ai totali marginali di riga e colonna;

b)

$$t\alpha(j) = \sqrt{\mathrm{nj}\,\frac{\mathrm{n}\text{ - }1}{\mathrm{n}\text{ -nj}}}\,\varphi\alpha\mathrm{j}$$

- dove 'nj' and ' j' indicano rispettivamente le occorrenze del j-esimo oggetto e la sua coordinata sul -esimo fattore.

Il valore test ha due proprietà importanti: un valore di soglia (1.96), corrispondente alla significatività statistica più comunemente utilizzata (p. 0.05), e un segno (- / +).

Ciò significa che, ordinando i valori in modo ascendente o discendente, è possibile individuare rapidamente la rilevanza di ogni elemento analizzato.

In **T-LAB**, la consultazione delle tabelle con i valori test è di tipo interattivo.

Variabili e modalità

In **T-LAB** le **variabili** sono le etichette utilizzate per identificare e classificare le varie parti del corpus: nomi di caratteristiche che identificano tipi di soggetti, di testi e di contesti.

Ogni variabile si declina attraverso due o più **modalità**, ciascuna delle quali - in modo univoco - corrisponde a un valore di codifica: ad esempio, la variabile "sesso" prevede due modalità (femminile e maschile).

In **T-LAB**, ogni testo può essere identificato attraverso **un massimo di 50 variabili**. Ovviamente, per ciascuna di esse, vanno indicate le rispettiva modalità (Max 150): ciò secondo le indicazioni contenute nella **Preparazione del corpus**.

BIBLIOGRAFIA ESSENZIALE

Bardin L. (1977):*L'analyse de contenu, Paris, P.U.F.*

Benzécri J.P & F. (1984):*Pratique de l'analyse des données. Analyse des correspondances & Classification, Paris, Dunod*

Blei D.M. (2012):*Introduction to Probabilistic Topic Models, Communications of the ACM, Volume 55 Issue 4, April 2012 Pages 77-84*

Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E. (2008): *Fast unfolding of communities in large networks*. *Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)*

Bolasco S. (1999):*Analisi Multidimensionale dei dati. Metodi, strategie e criteri di interpretazione, Roma, Carocci*

Boley D.L. (1998):*Principal direction divisive partitioning, Data Mining and Knowledge Discovery, 2(4), 325-344*

Campello R. J. G. B., Moulavi D., Zimek A. & Sander J. (2015): *Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection*. ACM Trans. Knowl. Discov. Data 10, 1, Article 5 (July 2015)

Carroll J.B. (1964): *Language and Thought, Englewood Cliff NJ, Prentice Hall*

De Mauro T. Mancini F. Vedovelli M. Voghera M. (1993):*Lessico di frequenza dell'italiano parlato (Fondazione IBM), Milano, Etas Libri*

Fernández A., Gómez S. (2008): *Solving Non-uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms, Journal of Classification, 25: 43-65*

Greenacre M.J. (1984):*Theory and Applications of Correspondance Analysis, New York, Academic Press*

Greimas A.J. (1966):*Sémantique structurale, Paris, Larousse*

Guiraud P. (1960): *Problèmes et méthodes de la statistique linguistique. Dordrecht, Reidel*

Herdan, G. (1960): *Quantitative Linguistics. London, Butterworth*

Kohonen T. (1989):*Self-Organization and Associative Memory, Berlin, Springer-Verlag*

Krippendorff K. (1980):*Content Analysis. An Introduction to its Methodology, London, Sage inc.*

Lancia F. (2004) :*Strumenti per l'analisi dei testi. Introduzione all'uso di T-LAB, Milano, FrancoAngeli*

Lancia F. (2005), *Word co-occurrence and Similarity in Meaning www.tlab.it*

Lancia F. (2012) : *The Logic of the T-LAB Tools Explained, www.tlab.it*

Lebart L., Morineau A., Piron M. (1995):*Statistique exploratoire multidimensionnelle, Paris, Dunod*

Lebart L., Salem A. (1994):*Statistique textuelle, Paris, Dunod*

Maranda P. (1990):*DisCan: User's Manual, Québec, Nadeau Caron Informatique*

Marwan N., Romano M., Thiel M. & Kurths J. (2007): *Recurrence Plots for the Analysis of Complex Systems*, Phys. Rep. 438, 240-329.

Michelet B. (1988):*L'analyse des associations, Thèse de doctorat, Université Paris VII, Paris*

Miller M.M. Riechert B.P. (1994):*Identifying Themes via Concept Mapping: A New Method of Content Analysis, Paper presented at the Communication Theory and Methology Division of the Association for Education in Journalism and Mass Communication Annual Meeting,*

Atlanta
Pottier B.(1974) : *Linguistique générale, théorie et description, Paris, Klincksieck*
Rastier F. (1987):*Sémantique interprétative, Paris, PUF*
Rastier F., Cavazza M., Abeillé A. (2002):*Semantics for Descriptions, Stanford, CSLI*
Salton G. (1989):*Automatic text processing: the transformation, analysis, and retrieval of Information by Computer, Addison-Wesley, Reading, Massachussets*
Saussure (de) F. (1916), *Cours de Linguistique générale, Lusanne-Paris, Payot,*
Savaresi S.M., D.L. Boley (2001): *On the performance of bisecting K-means and PDDP, 1st SIAM Conference on DATA MINING, Chicago, IL, USA, April 5-7, paper n.5, pp.1-14*
Savaresi S.M., Boley D.L. (2004): *A Comparative Analysis on the Bisecting k-means and the PDDP Clustering Algorithms, International Journal on Intelligent Data Analysis, 8(4): 345-362*
Steinbach M., Karypis G., Kumar V. (2000): *A comparison of Document Clustering Techniques. Proceedings of World Text Mining Conference, KDD2000, Boston*
Steyvers M., Griffiths T. (2007). *Probabilistic Topic Models. In Landauer, T.; McNamara, D; Dennis, S.; et al. Handbook of Latent Semantic Analysis, Mahwak, NJ, Lawrence Erlbaum*
van der Maaten L.J.P., & G.E. Hinton (2008): *Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research* 9(Nov):2579-2605, 2008
Webber C. L., & Zbilut J. P. (2005) : *Recurrence Quantification Analysis of Nonlinear Dynamical Systems.* In M. Riley, & G. Van Orden (Eds.), *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences* (pp. 26-94)