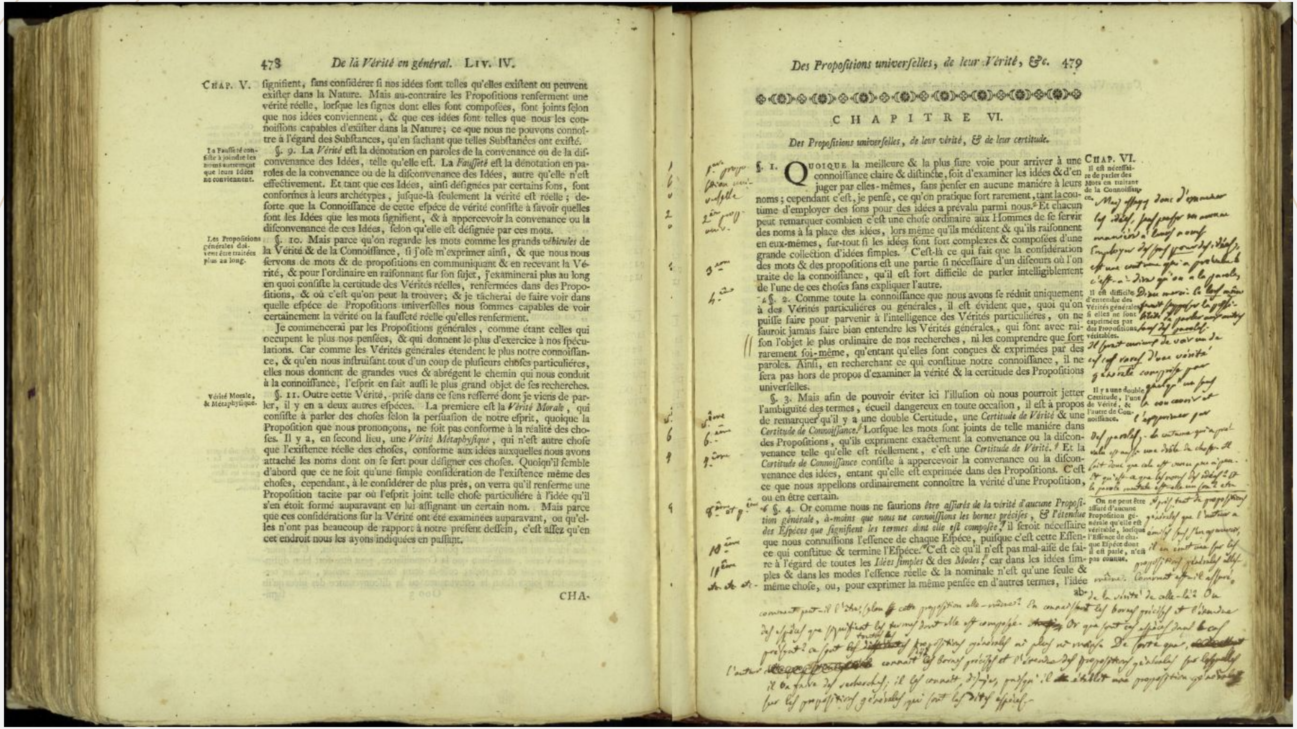
Esercitazione con Voyant

Rachele Sprugnoli (Università di Parma)

Trieste-online, 10 marzo 2022

1

I Concetti Fondamentali



Close Reading

http://www.alessandromanzoni.org/biblioteca/esemplari/3883/

Distant Reading

http://www.fedoabooks.unina.it/index.php/fedoapress/catalog/book/104

Scalable Reading

Non sto suggerendo di archiviare la lettura da vicino e "letture" di letteratura altamente interpretative. Al contrario, sto suggerendo un ***approccio misto****. […] È esattamente questo tipo di unificazione, della scala macro e micro, che promette una nuova, migliorata e migliore comprensione della letteratura.*
Le due scale di analisi ***lavorano in tandem e comunicano tra loro****. L'interpretazione umana dei "dati", sia che siano estratti su scala macro o micro, rimane essenziale. Sebbene i metodi di indagine, di raccolta delle prove, siano diversi, non sono antitetici e condividono lo stesso obiettivo finale di* ***rafforzare la comprensione della letteratura****, sia essa scritta in grande o in piccolo.*

Jockers, *Macroanalysis. Digital Methods and Literary History*, 2013

2

Panoramica su Voyant

Cosa è Voyant

- Voyant Tools è un ambiente web per la lettura e l'analisi di testi
    - vari formati di input: txt, pdf, html, xml
    - può essere integrato su altri siti
    - interattivo
    - permette una lettura scalabile
    - indipendente dalla lingua
    - analisi lessicale

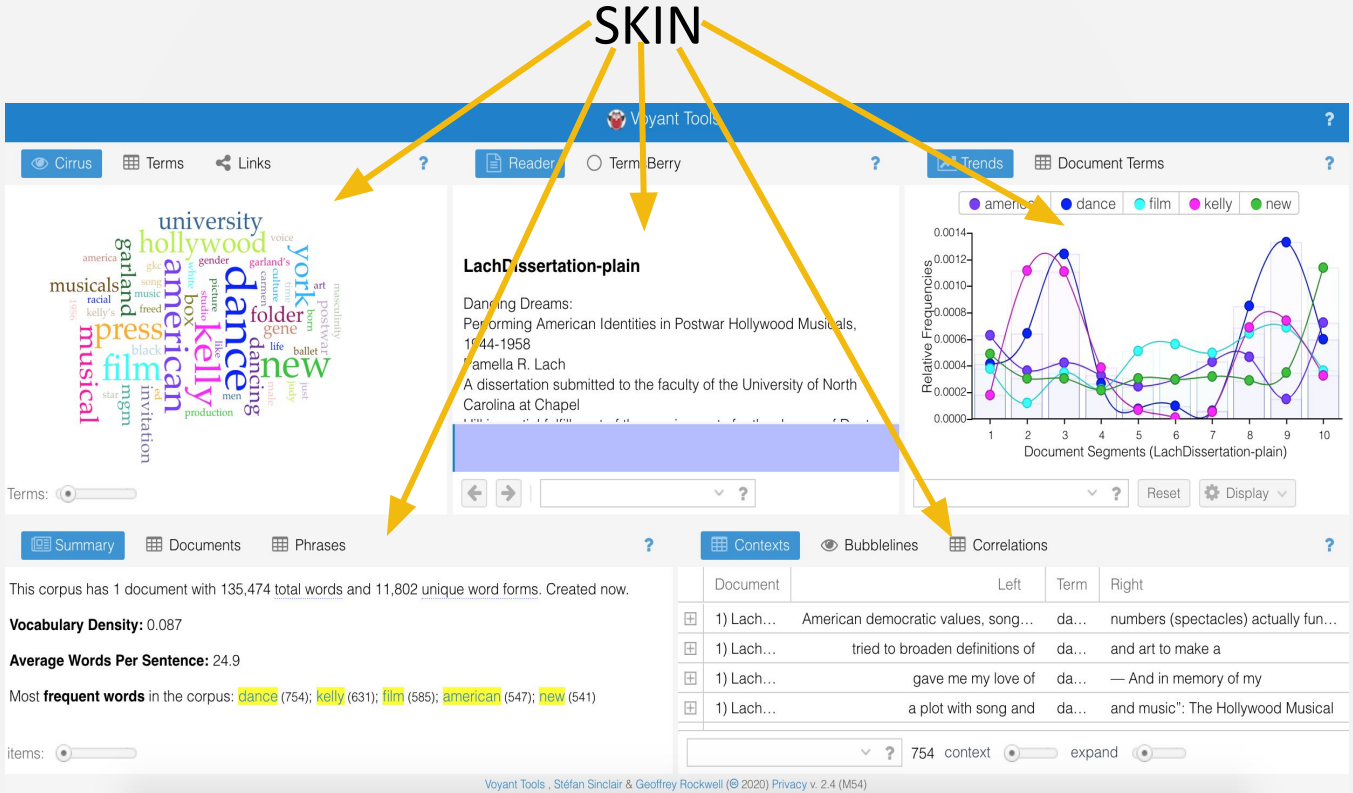
Dove trovare Voyant

- Sito ufficiale: https://voyant-tools.org/
- Mirror:
  - https://voyant-tools.huma-num.fr
  - https://voyant.lincsproject.ca
  - https://service.sadilar.org/voyant/
- Server installabile: https://github.com/voyanttools/VoyantServer

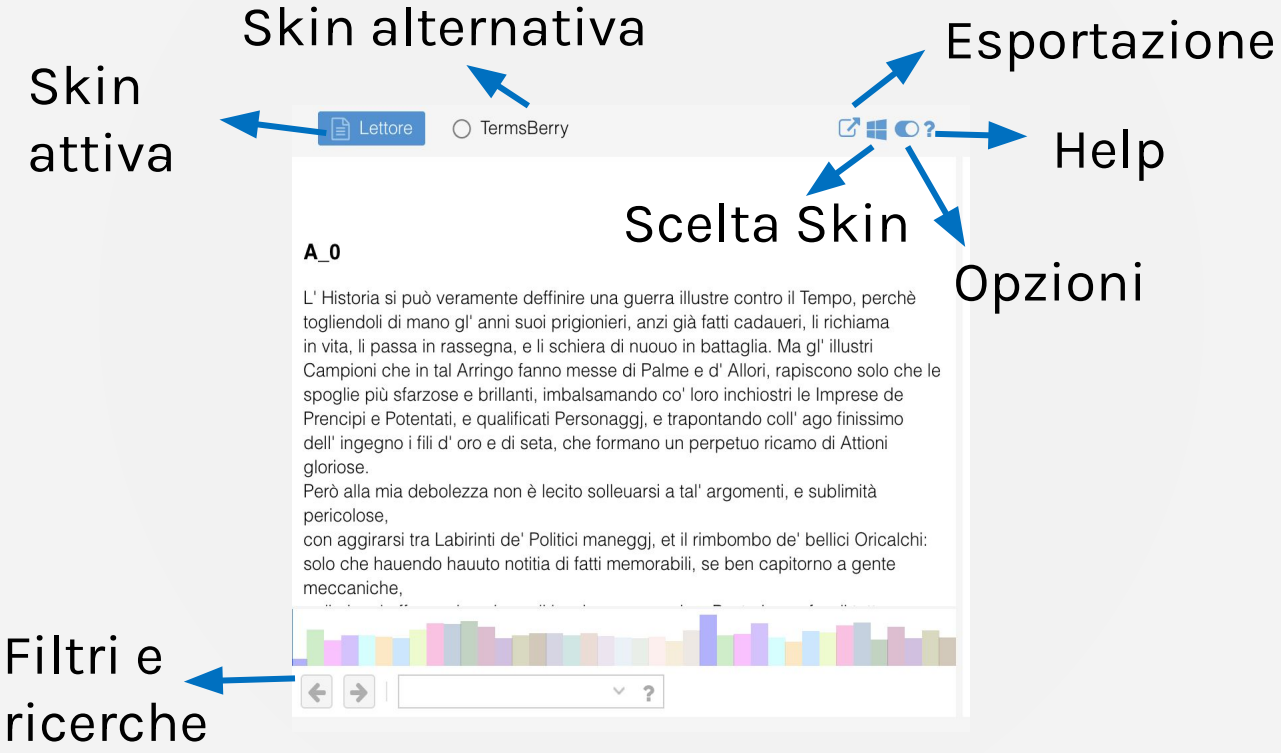
Che tipo di file caricare

- Formati accettati: TXT, HTML, XML, TEI, PDF, RTF, MS Word, JSON, tabelle in fogli di calcolo
- PDF: contiene un OCR, il risultato può variare
- **ATTENZIONE**: il modo in cui vengono caricati i file influenza il tipo di analisi successiva
  - eg.: testo unico *versus* divisione in capitoli

Struttura generale



Struttura della skin



Funzioni principali (1)

- Cambiare skin 

- READER: lettore del testo, permette il close reading
- CIRRUS: visualizzatore frequenza dei termini
- BUBBLES: visualizzatore frequenza dei termini
- TERMS: analisi della frequenza dei termini
- TRENDS: andamento delle frequenza dei termini
- BUBBLELINES: frequenza e distribuzione dei termini
- MICROSEARCH: frequenza e distribuzione dei termini
- CONTEXT: contesti di occorrenza dei termini

Funzioni principali (2)

- Cambiare skin 

- PHRASES: sequenze di parole che co-occorrono
- COLLOCATES: termini che appaiono vicino ad altri termini
- CORRELATIONS: termini la cui frequenza varia in sintonia
- MANDALA: relazioni tra termini e documenti
- SUMMARY: informazioni sul corpus
- DOCUMENTS: informazioni sui singoli documenti
- TOPICS: topic modeling

https://voyant-tools.org/docs/#!/guide/tools

Come e cosa esportare

- L’esportazione  si può applicare all'intero progetto Voyant o a una singola skin

- Puoi esportare una **URL**, uno strumento incorporabile (**interattivo**) o un riferimento **bibliografico**
- Puoi anche esportare un file **.png** statico nel caso delle visualizzazioni (uno screenshot potrebbe avere una migliore qualità dell'immagine)
- Puoi esportare i dati dagli skin a forma di **tabella** in vari formati

Come cercare

- Sintassi delle ricerche lessicali

- pestilenza: trova il termine esatto
- pestilen*: trova termini che iniziano con "pestilen"
- "marito e moglie": cerca l'intera espressione
- "opera misericordia"~5: "opera" e "misericordia" co-occorrono entro 5 termini
- @Personaggi: ricerca raggruppata di tutti i termini inclusi in una *categoria*
- ^@Personaggi: ricerca dei singoli termini inclusi in una *categoria*

3

Esercitazione Pratica

Come caricare file

- Andare su Voyant (sito ufficiale, mirror o versione in locale)

- cliccare su opzioni 
- sotto **Processing** scegliere **Simple Word Boundaries**

The following table summarizes tokenization for the string What's voyant-tools.org?:

Tokenization	Count	Tokens	Notes
Automatic	3	what's, voyant, tools.org	the hyphen is split but the tools.org is considered a URL token; tokens are lowercase
Word Boundaries	5	what, s, voyant, tools, org	any non-word character is a delimiter, tokens are lowercase
Whitespace Only	2	What's, voyant-tools.org?	punctuation is kept in tokens and case is unchanged

- cliccare su **Upload**, aprire la cartella **capitoli**, selezionare (ctrl+a) tutti i file e cliccare su **Apri:** sono i capitoli de I Promessi Sposi

Aggiungere le stopword (1)

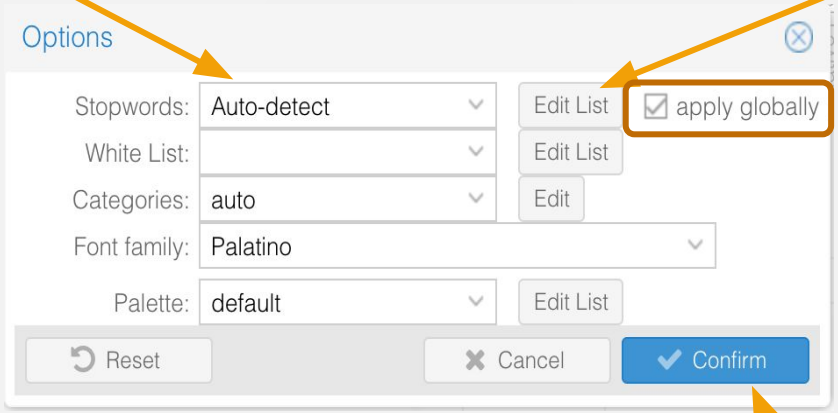
- Parole funzione (*versus* parole contenuto) da ignorare: congiunzioni, preposizioni, articoli…
- Dove trovare le liste di stopword?

  - Lingue moderne: https://www.ranks.nl/stopwords
  - Latino e greco antico: https://github.com/aurelberra/stopwords/tree/master/ancientstopwords/data
  - Lista di stopword per l'italiano nella cartella Voyant: aprire il file **stopword-it.txt** con un editor di testo, selezionare e copiare la lista

Aggiungere le stopword (2)

- Su Voyant, cliccare sulle opzioni della skin Cirrus 

1) Selezionare “Italian”



2) Cliccare su **Edit List,** incollare la lista del file **stopwords-it.txt**

3) Cliccare su **Confirm**

Aggiungere le stopword (3)

- L’effetto su Cirrus

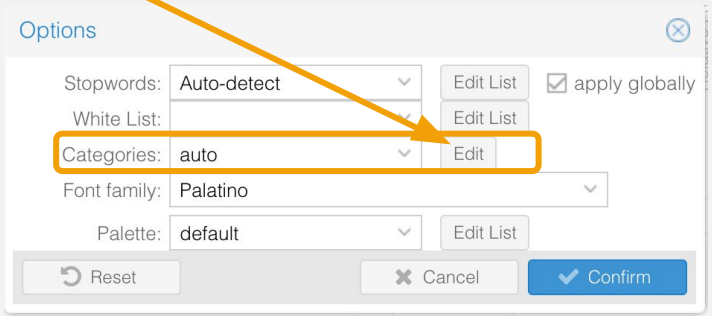
PRIMA

DOPO

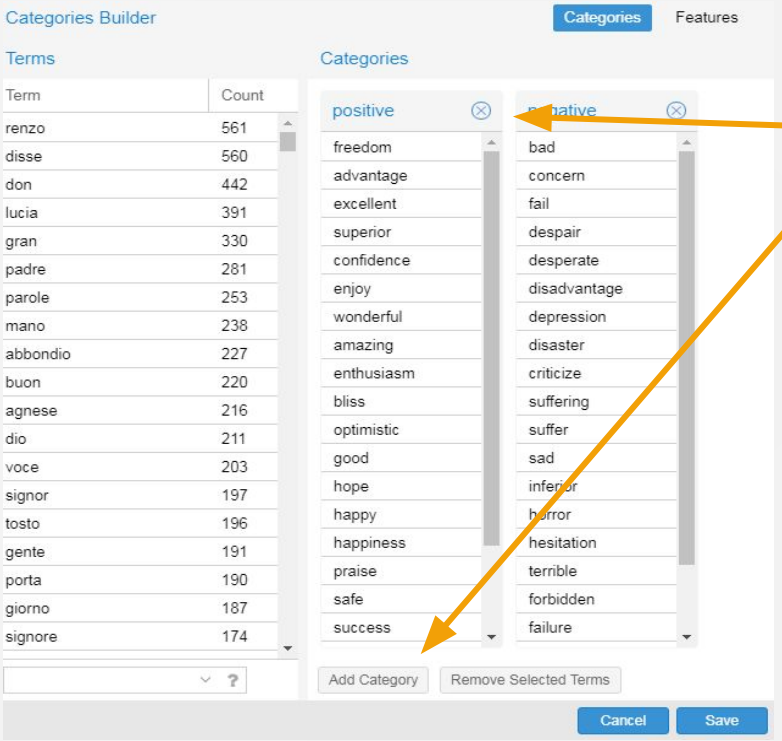
Aggiungere le categorie (1)

- CATEGORIE: gruppi di parole semanticamente connesse, ad esempio lista di personaggi, lista di luoghi, lista di emozioni da usare per ricerche mirate

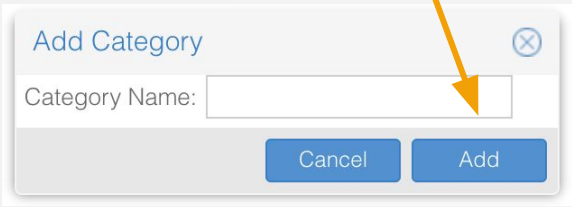
  - Cliccare sulle opzioni 
  - Cliccare su **Edit** vicino a **Categories**



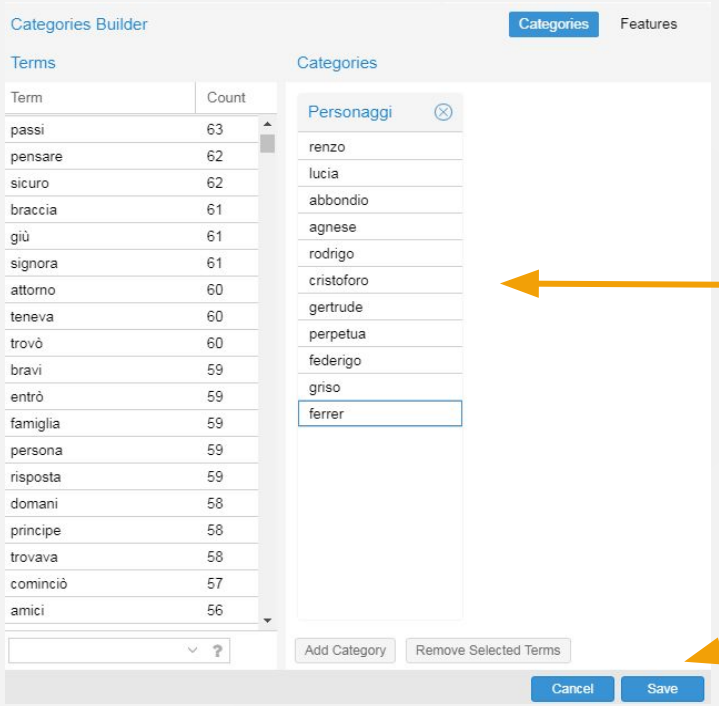
Aggiungere le categorie (2)



1) Rimuovere le categorie esistenti
2) Cliccare su **Add Category**
3) Scrivere un nome per la categoria, e.g. **Personaggi** e cliccare su **Add**



Aggiungere le categorie (3)



4) Trascinare i nomi dei personaggi dalla lista **Terms** alla lista **Personaggi**

5) Salvare cliccando su **Save**

Esempi di utilizzo

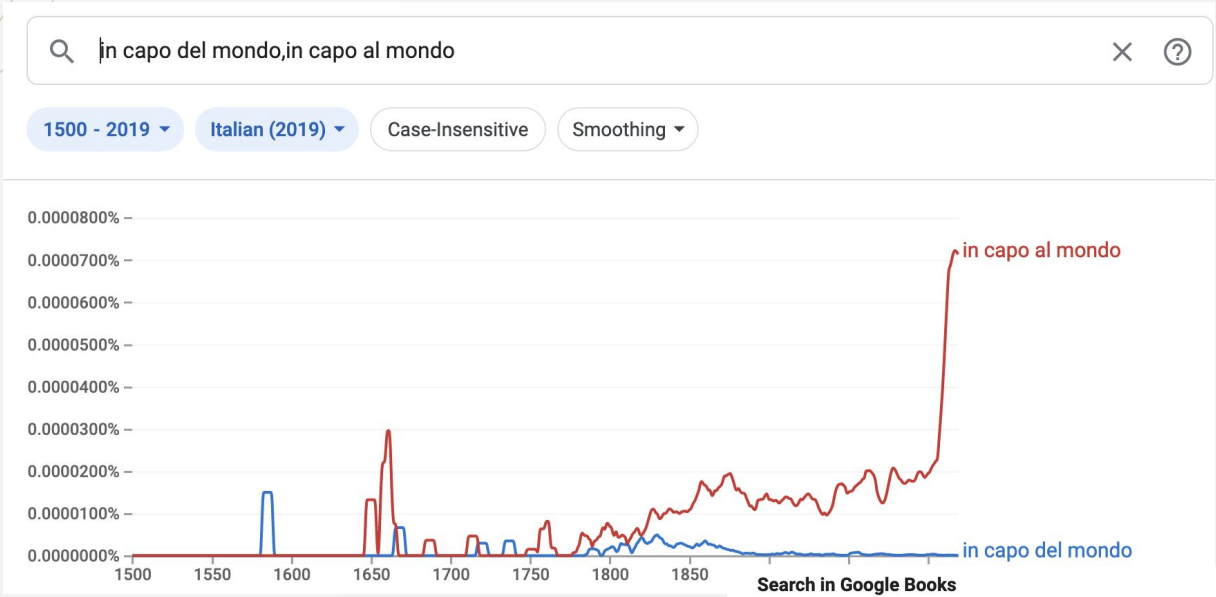
1. Come individuare gli hapax legomena? → TERMS
2. Quali sono le parole più frequenti del cap 34? → CIRRUS + SCALE, DISTINCTIVE WORDS (nella skin DOCUMENTS)
3. Quali sono i capitoli più connessi alla pestilenza? → MANDALA
4. Quale personaggio viene menzionato di più e in che parti? → usare categoria su TRENDS, BUBBLELINES, MICROSEARCH
5. Qual è il capitolo con più densità lessicale? → DOCUMENTS
6. Si possono intuire delle caratteristiche psicologiche di Renzo e Lucia? → COLLOCATES
7. Come individuare gli usi metaforici di una parola? CONTEXTS
8. Come individuare forme desuete? → CONTEXTS (es. is*)

Un’analisi più dettagliata (1)

- “In capo del mondo” oppure “in capo al mondo”?

- Cercare in CONTEXTS "capo mondo"~1: cosa osserviamo?
- TLIO: http://tlio.ovi.cnr.it/TLIO/ (cercare *capo*)
- Vocabolario della Crusca del 1826 appartenuto a Manzoni: https://www.alessandromanzoni.org/biblioteca/esemplari/4155/reader#page/335/mode/1up
- Google Books Ngram Viewer: https://books.google.com/ngrams (vedi slide successiva)

Un’analisi più dettagliata (2)



Search in Google Books

in capo del mondo	1500 - 1584	1585	1586 - 1829	1830 - 1907	1908 - 2019	Italian (2019)
in capo al mondo	1500 - 1759	1760 - 1981	1982 - 2003	2004 - 2013	2014 - 2019	Italian (2019)

Analisi di una locuzione (1)

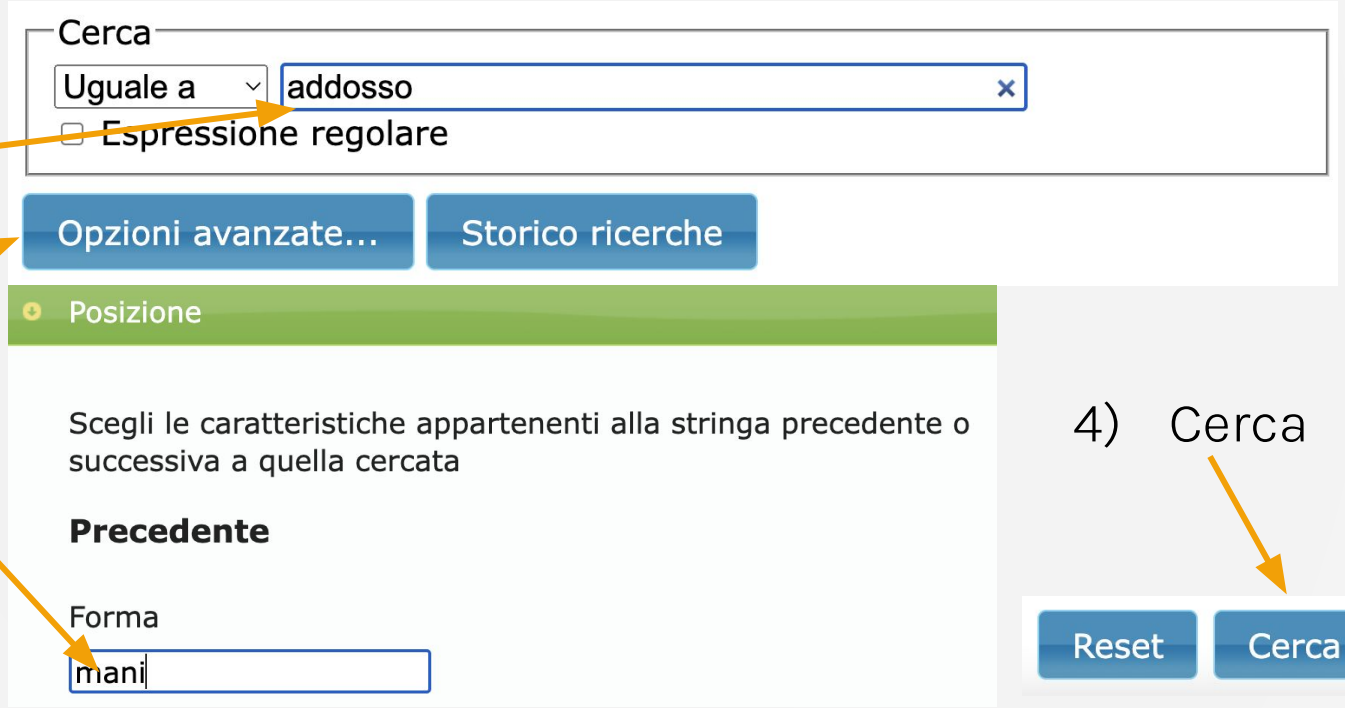
- “Porre le mani addosso” oppure “mettere le mani addosso”?

- Cercare in CONTEXTS "mani addosso": cosa osserviamo?
- Confrontiamo con due corpora diacronici:
  - https://www.corpusmidia.unito.it/index.php
  - https://corpora.ficlit.unibo.it/DiaCORIS/
- Confrontiamo con un corpus d’italiano contemporaneo:
  - https://www.corpusitaliano.it/it/access/simple_interface.php

Analisi di una locuzione (2)

- MIDIA

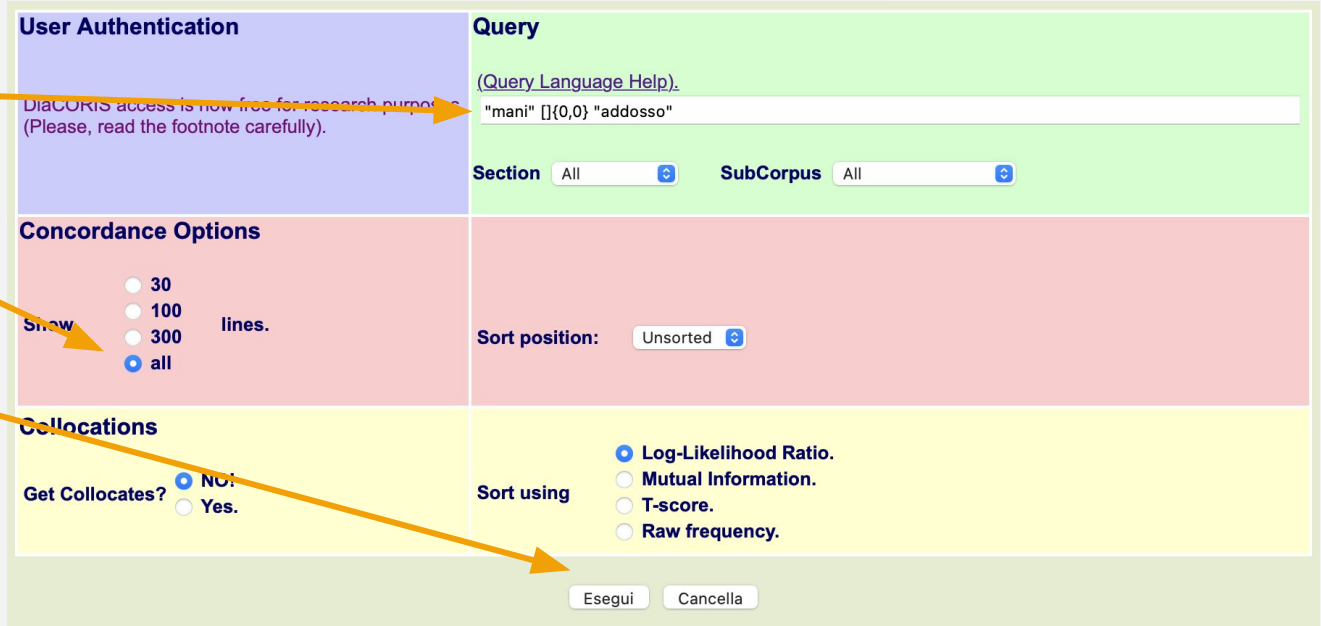
1) Ricerca forma
2) Opzioni avanzate
3) Ricerca forma precedente



Analisi di una locuzione (3)

- DiaCORIS

1) Ricerca
2) Mostra tutti risultati
3) Esegui

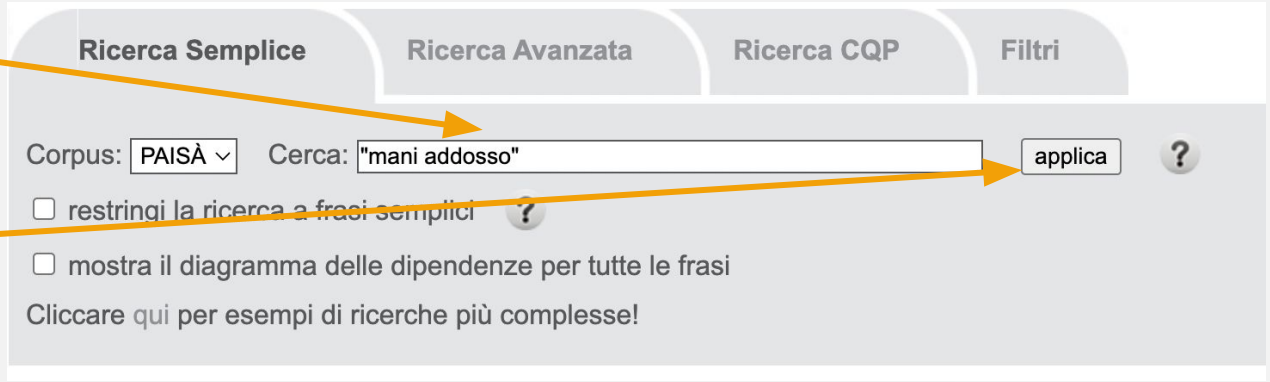


Analisi di una locuzione (4)

- Paisà (ricerca semplice)

1) Ricerca stringa

2) Applica

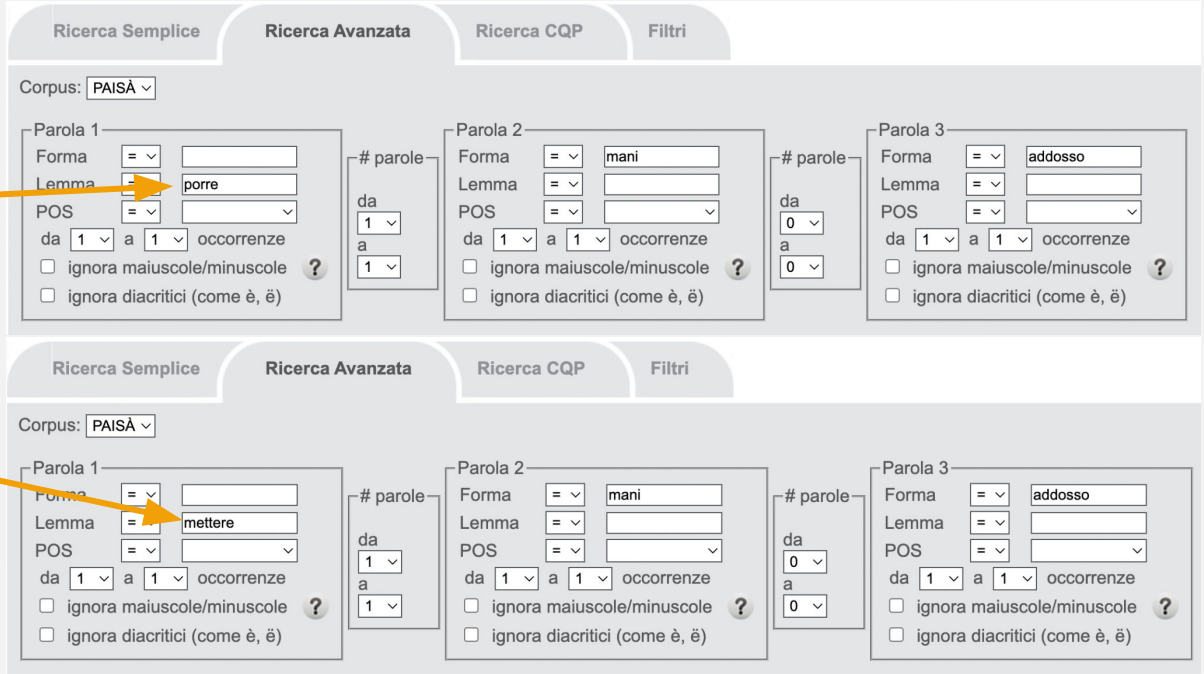


Analisi di una locuzione (5)

- Paisà (ricerca avanzata)

1) Ricerca con lemma “porre”

2) Ricerca con lemma “mettere”



Provate voi!

- Caricate i 3 file della cartella **versioni**, caricate la lista di stopword e provate a rispondere alle seguenti domande:
    1. Ci sono differenze evidenti tra i 3 testi dal punto di vista quantitativo?
    2. Ci sono sintagmi ricorrenti? Sintagmi con piccole variazioni?
    3. La presenza di Lucia è simile nelle 3 versioni?
    4. Il personaggio di Geltrude appare in “Fermo e Lucia” similmente a come appare Gertrude ne “I Promessi Sposi”?
    5. Quante parole contengono la lettera “j” nelle varie versioni?
    6. Cosa notate nella frequenza di “egli”? !!ATTENZIONE!!

Grazie!

Domande?

Mi trovate a rachele.sprugnoli@unipr.it.

Su Twitter: @RSprugnoli

Per scoprire di più sulle mie ricerche https://personale.unipr.it/it/ugovdocenti/person/236480.