Distant Reading e Visualizzazione di Dati

Rachele Sprugnoli – rachele.sprugnoli@unicatt.it

Centro Interdisciplinare di Ricerche per la Computerizzazione
dei Segni dell'Espressione (CIRCSE)

COSA È IL DISTANT READING?

CLOSE READING

- Metodo tradizionale di analisi del testo
- Interpretazione del testo basata sulle interazioni tra un lettore umano e un testo

Unveil words, verbal images, elements of style, sentences, argument patterns
(Jasinski, 2001)

DISTANT READING

- Nuovo metodo di analisi introdotto nella critica letteraria
- Interpretazione del testo basata su caratteristiche generali e modelli astratti

A condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text.
(Moretti, 2000)

ALTRE DEFINIZIONI DI DISTANT READING

The construction of abstract models

Jasinski, “Sourcebook on Rhetoric”, 2001

A macroanalytic approach

Jockers, “On Distant Reading and Macroanalysis“, 2011

The idea of processing content in or information about a large number of textual items without engaging in the reading of the actual text.

Drucker, “Distant Reading and Cultural Analytics”, 2013

COSA È LA VISUALIZZAZIONE DEI DATI?

VISUALIZZAZIONE DEI DATI

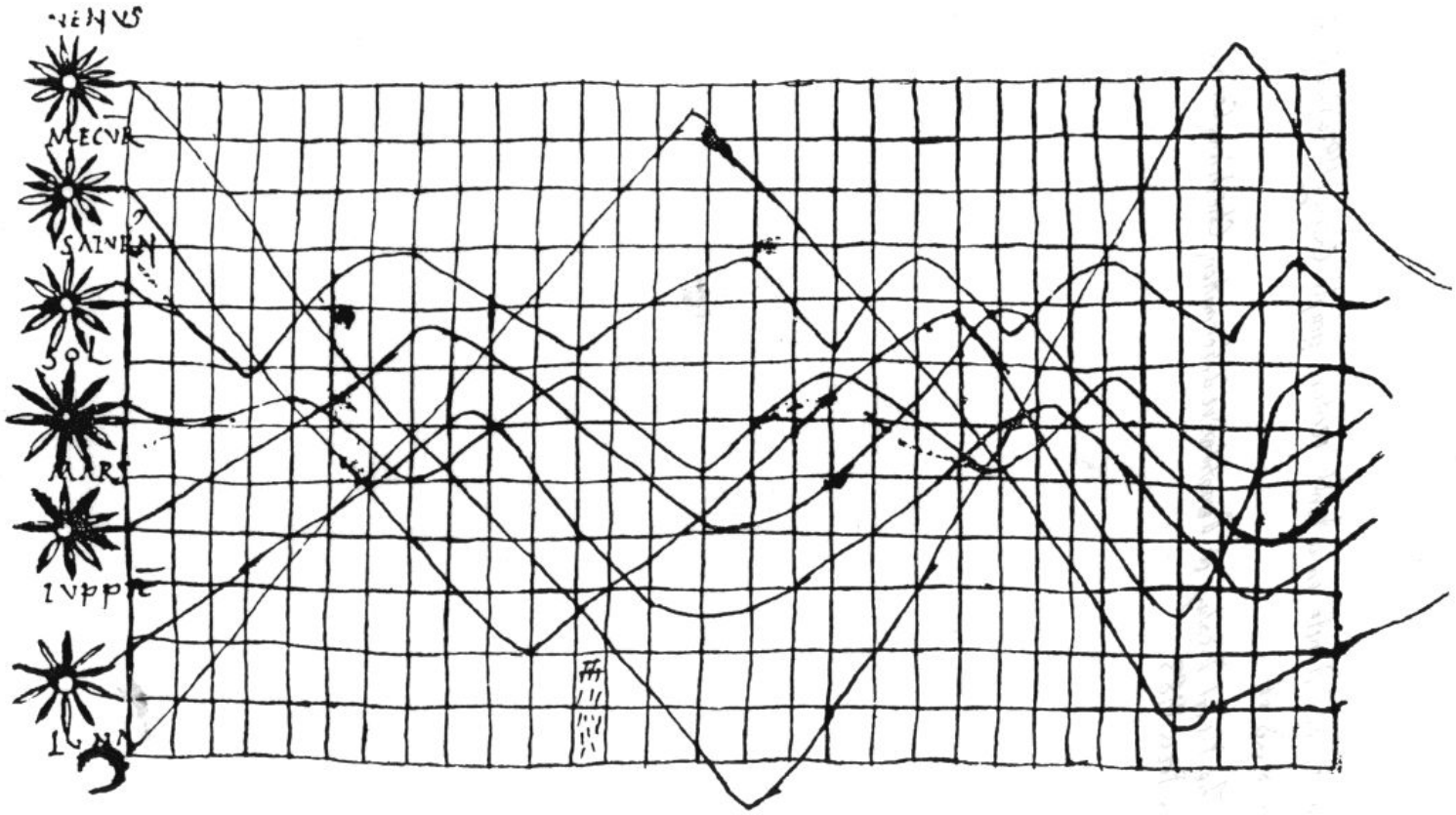
- Rappresentazione dei dati attraverso un linguaggio visivo

- Perché?
  - esplorare i dati: sia numeri che testi (singoli documenti, corpora e stream di dati)
  - trovare schemi ricorrenti
  - comprendere il contenuto
  - supervisionare la procedura di analisi
  - comunicare

PRINCIPI PER UNA BUONA VISUALIZZAZIONE DEI DATI

- Secondo Edward Tufte, una corretta visualizzazione dei dati dovrebbe:
  - mostrare i dati
  - indurre a pensare alla sostanza
  - evitare distorsioni
  - presentare molti dati in uno spazio piccolo
  - rendere coerenti anche grandi collezioni di dati
  - incoraggiare confronti
  - avere vari livelli di dettaglio
  - avere uno scopo chiaro
  - essere strettamente integrato con la parte statistica e descrittiva

MOLTO PRIMA DI TUFTE E MORETTI



MOLTO PRIMA DI TUFTE E MORETTI

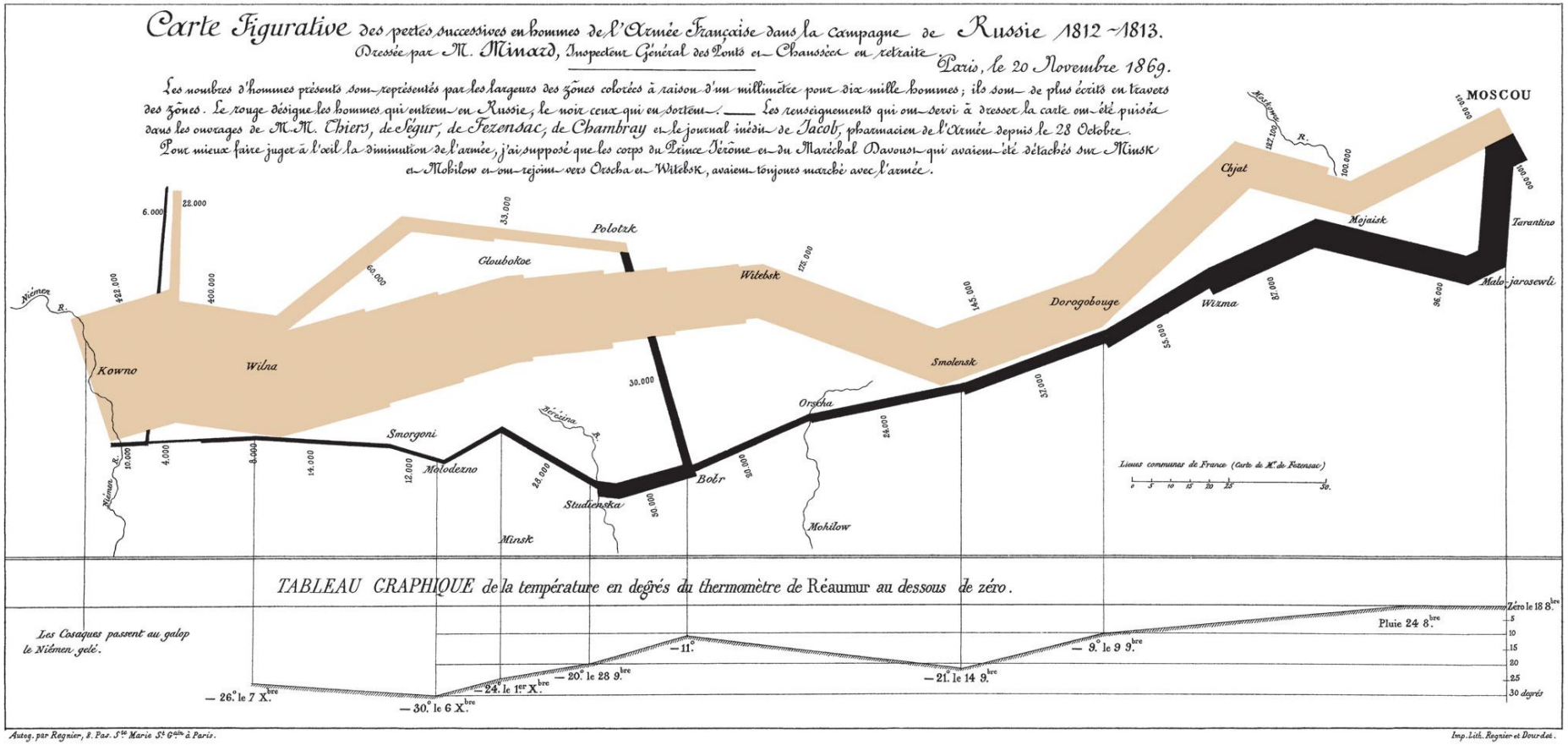
*"The great growth of **statistical research** in our times has made felt the need to record the results in forms **less dry**, **more useful**, and able to be explored more **rapidly** than numbers alone; thus, diverse representations have been imagined, among others my graphic tables and my figurative maps."*

Minard, "Graphic Tables and Figurative Maps" (1862)

translated by Edward Tufte

MOLTO PRIMA DI TUFTE E MORETTI

La campagna di Russia di Napoleone (1861)
“May well be the best statistical graphic ever drawn”, Tufte



QUAL È LA RELAZIONE TRA DISTANT READING E VISUALIZZAZIONE DEI DATI?

"*Le immagini vengono prima di tutto, nei nostri pamphlet, perché – visualizzando i dati empirici – costituiscono l'oggetto specifico di studio della critica computazionale*" Franco Moretti

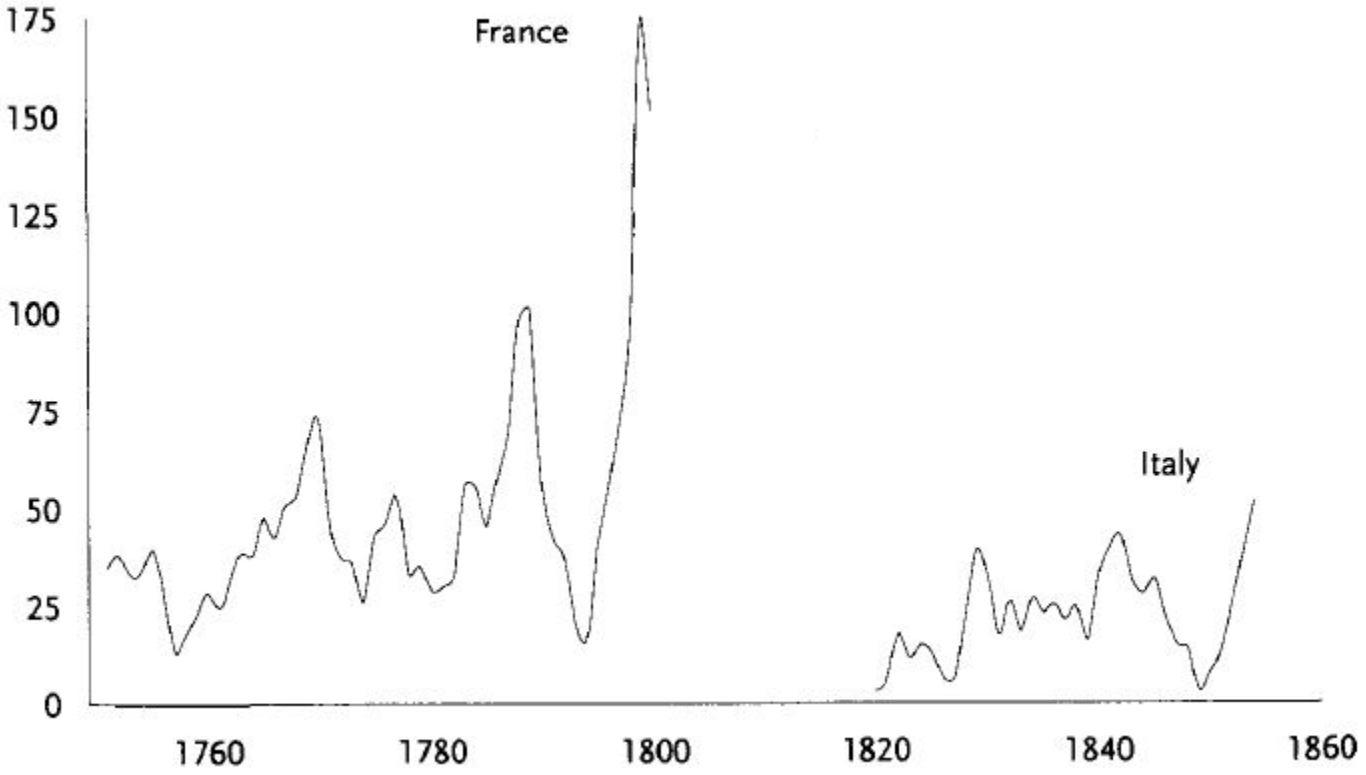
DAI TESTI AI MODELLI ALLE VISUALIZZAZIONI

Graphs, Maps, Trees: Abstract Models for a Literary History, Franco Moretti (2007)

- Il testo subisce un processo di deliberata riduzione e astrazione prendendo a prestito modelli da 3 discipline:
    1. Grafici → storia quantitativa
    2. Mappe → geografia
    3. Alberi → teoria dell'evoluzione

"Graphs, maps, and trees place humanities disciplines literally in front of our eyes-and show us how little we still know about it."

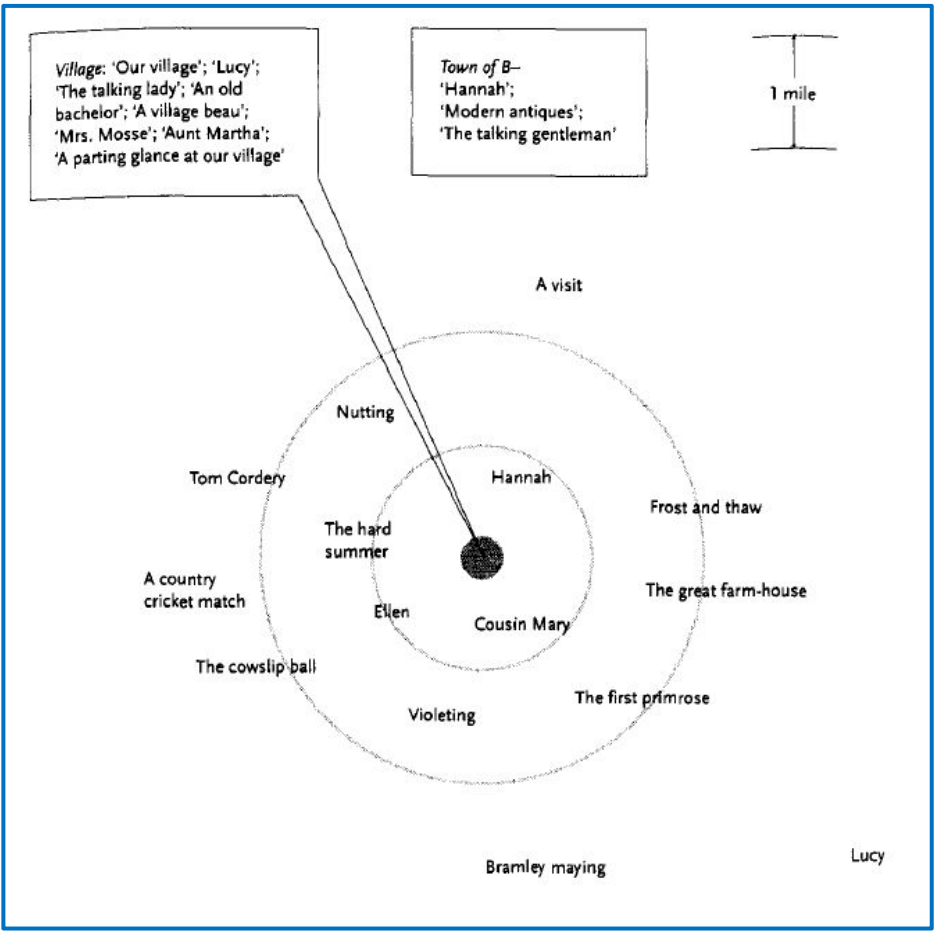
GRAFICI
L’ascesa e la caduta del romanzo



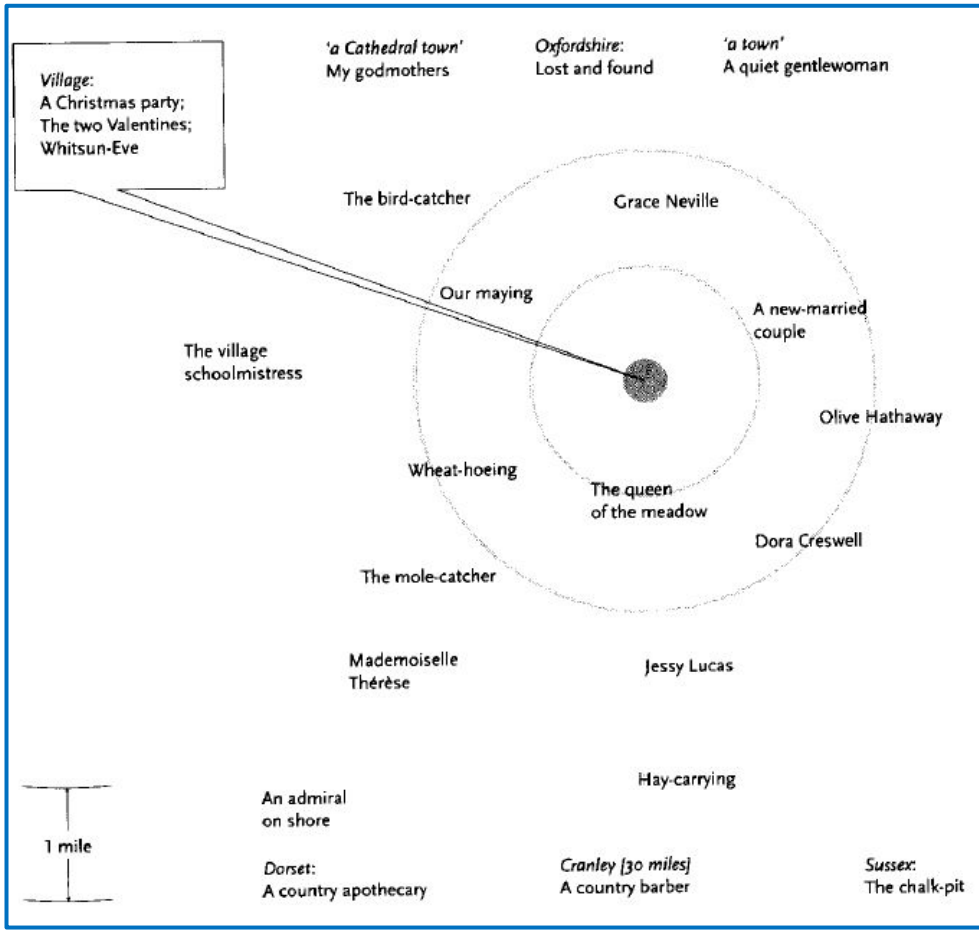
New novels per year

MAPPE
Mary Mitford, "Our Village"

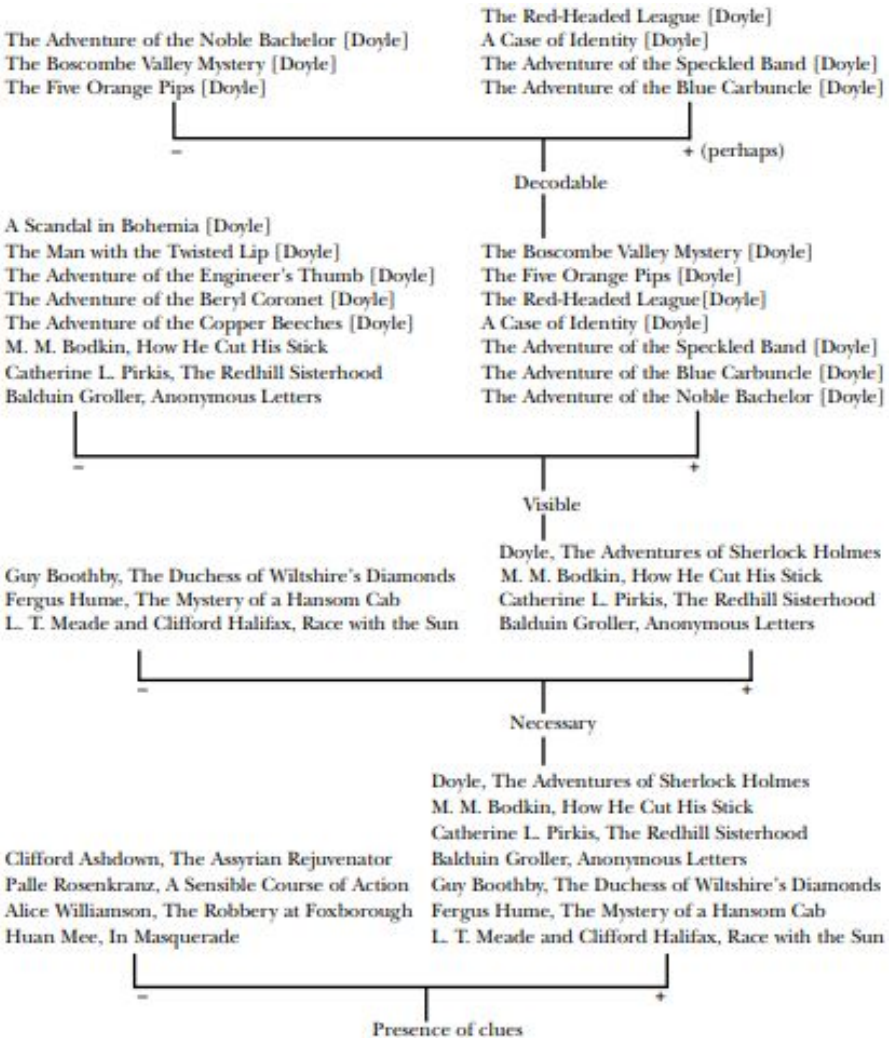
1824



1828



ALBERI
Il successo di Conan Doyle



PROCEDURA

- Raccolta di dati rilevante
  - accessibilità?
  - formato?
  - copyright?
- Formattazione e pulizia dei dati
- Elaborazione dei dati
- Formattazione dei risultati
- Produzione delle visualizzazioni
- Interpretazioni
- Comunicazione (Dissemination)

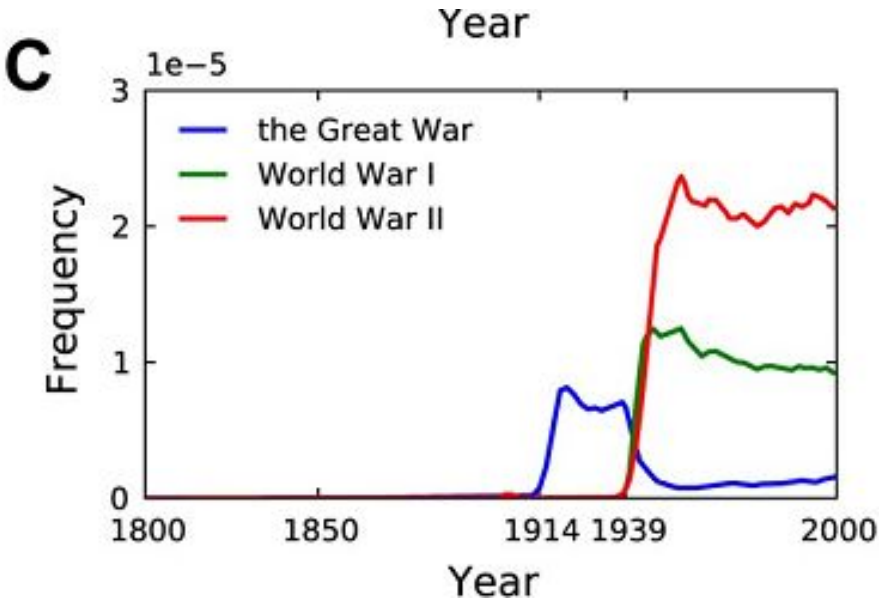
ALCUNI ESEMPI

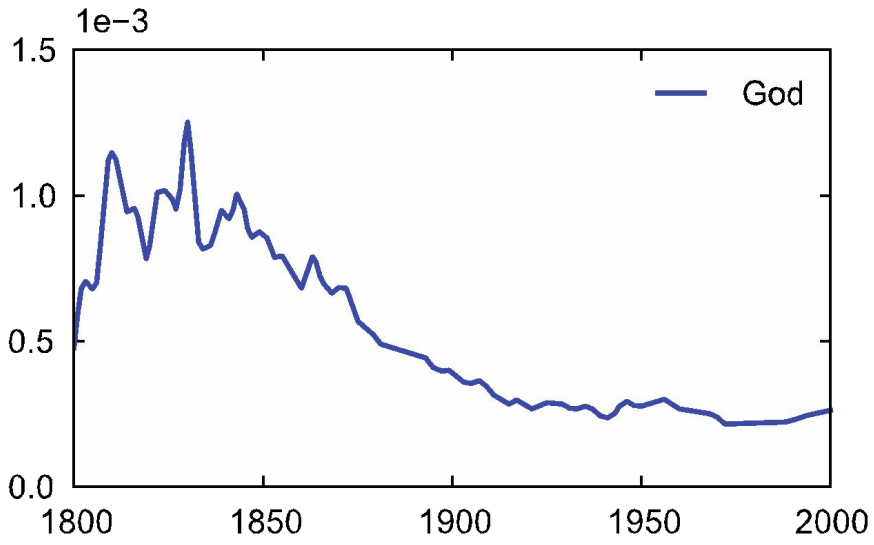
- Analisi linguistica
- Sentiment analysis
- Stilometria
- Topic modeling
- Georeferenziazione
- Reti di parole, concetti, personaggi
- Analisi degli n-grammi
- ...

CULTUROMICS

Quantitative Analysis of Culture Using Millions of Digitized Books, Jean-Baptiste Michel et al., Science 331 (2010)

- *Culturomics*: application of data collection and analysis techniques to the study of human culture





"God is not dead but needs a new publicist"

ANALISI DEGLI N-GRAMMI - 1

- https://books.google.com/ngrams

Scegliere il corpus “Italian (2019) e provare:

- andare al cinema,andare a teatro
- andare_INF al cinema,andare_INF a teatro
- università di *
- bello_INF,saggio_INF
- bello_*
- bello_NOUN,bello_ADJ
- gustare=>pizza,gustare=>gelato
- gustare=>*_NOUN

(cliccare su ? per le istruzioni dettagliate)

ANALISI DEGLI N-GRAMMI - 2

- https://bookworm.htrc.illinois.edu/develop/

Confrontare:

- war, target audience: juvanile
- war, target audience: adult

- http://bookworm.library.yale.edu/

Confrontare:

- long skirt
- mini skirt
- miniskirt

PERCHÉ IL DISTANT READING È IMPORTANTE?

- È riproducibile e verificabile
- Fa affiorare schemi nascosti non visibili altrimenti
- Aiuta a gestire la quantità sempre crescente di dati digitali

MA NON DIMENTICHIAMO CHE...

La lettura da vicino (micro-analysis) rimane di importanza fondamentale!

SCALABLE READING

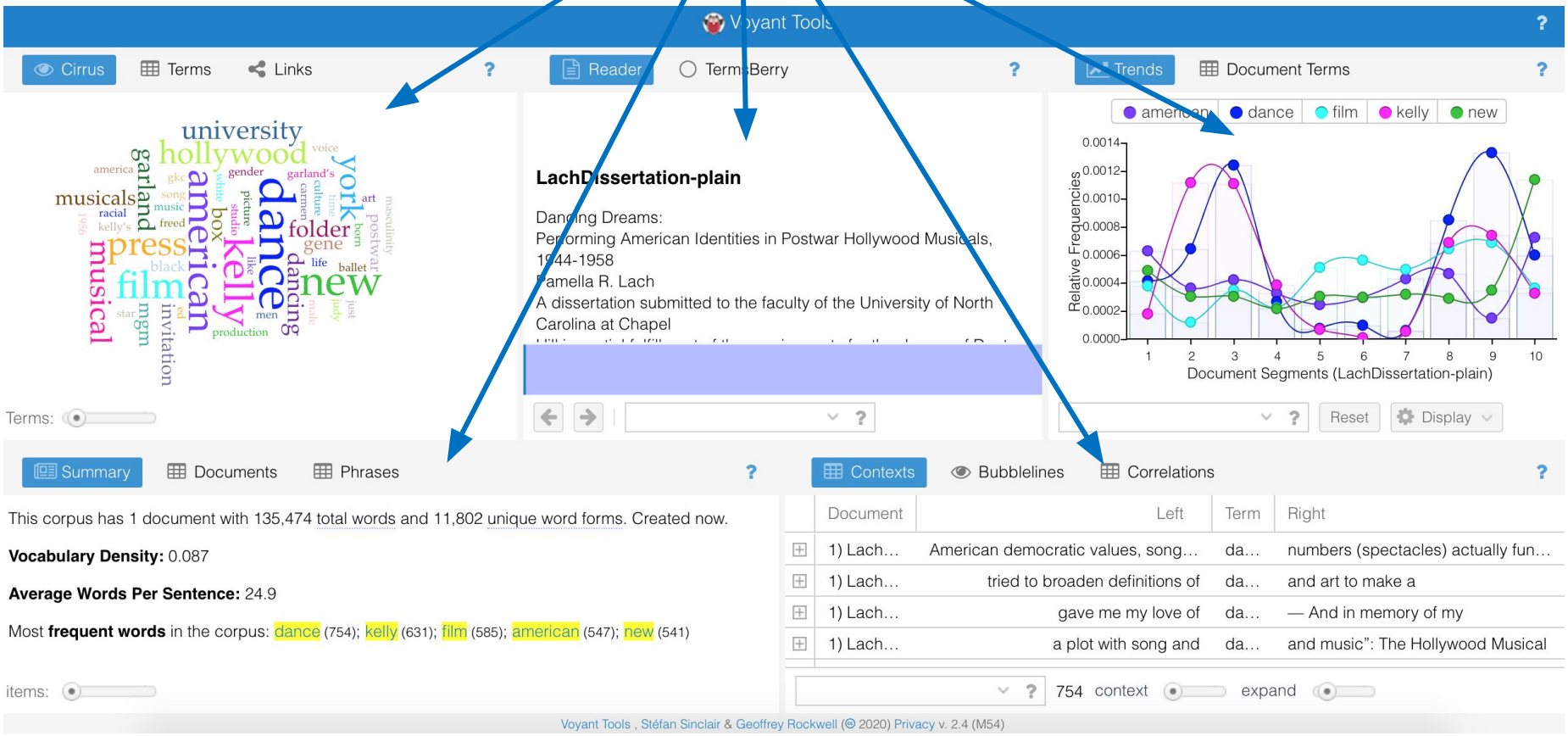
ESERCITAZIONE: VOYANT

- 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
  - scalable reading
  - indipendente dalla lingua
  - **https://voyant-tools.org/**

ESERCITAZIONE: VOYANT

SKIN



ESERCITAZIONE: VOYANT

- Caricare i dati in vari formati

HTML

- cercare su Google *us debate 2020 transcript* e scegliere una pagina web
- copiare la URL della pagina web, incollarla sotto Add Texts su Voyant e poi cliccare su Reveal

PDF

- andare su https://www.liberliber.it/online/autori/autori-m/niccol-machiavelli/, scegliere due opere, scegliere il formato PDF
- copiare la URL del PDF e incollarla sotto Add Texts su Voyant (una URL per riga) poi cliccare su Reveal

ESERCITAZIONE: VOYANT

- Caricare i dati in vari formati

XML + ZIP

- cliccare su Upload in Voyant
- cercare la cartella in cui sono stati salvati i materiali del corso, entrare nella sotto-cartella voyant e fare doppio click sulla cartella compressa Aristotele.zip

TXT

- cliccare su opzioni 
- cliccare su Upload in Voyant
- aprire la sotto-cartella capitoli, selezionare (ctrl+a) tutti i file e cliccare su Apri
- sono i capitoli de I Promessi Sposi (le prossime slide si riferiscono a questo caso d'uso)

ESERCITAZIONE: VOYANT

- Caricare i dati in vari formati

TXT

- 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 in Voyant
- aprire la sotto-cartella capitoli, selezionare (ctrl+a) tutti i file e cliccare su Apri
- sono i capitoli de I Promessi Sposi (le prossime slide si riferiscono a questo caso d’uso)

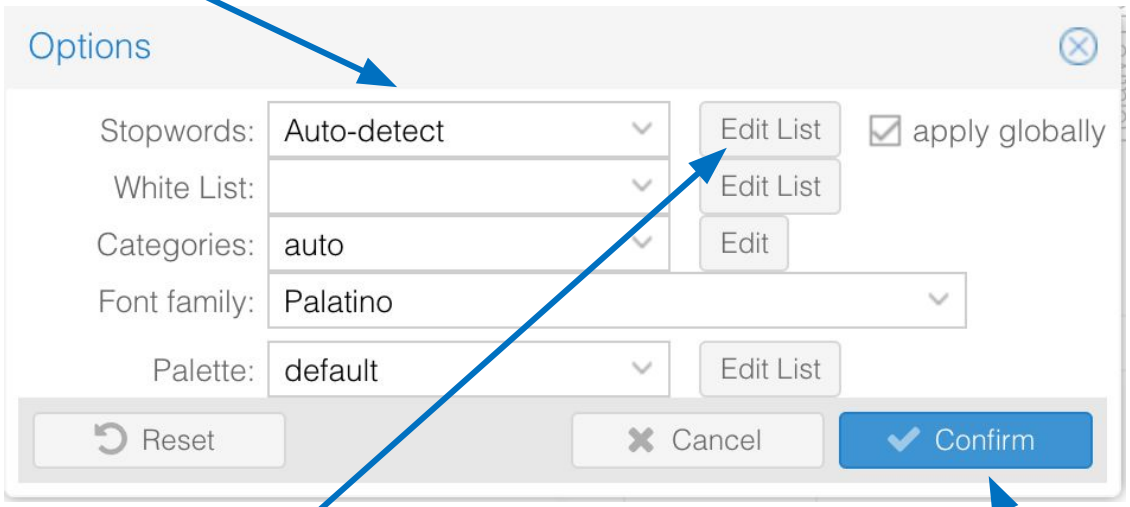
ESERCITAZIONE: VOYANT

- Selezionare la giusta lista delle stopword (parole da ignorare)
  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 stopwords-it.txt con un editor di testo

- Per modificare le stopword, cliccare sulle opzioni 

ESERCITAZIONE: VOYANT

1) Selezionare “Italian”



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

3) Cliccare su “Confirm”

ESERCITAZIONE: VOYANT

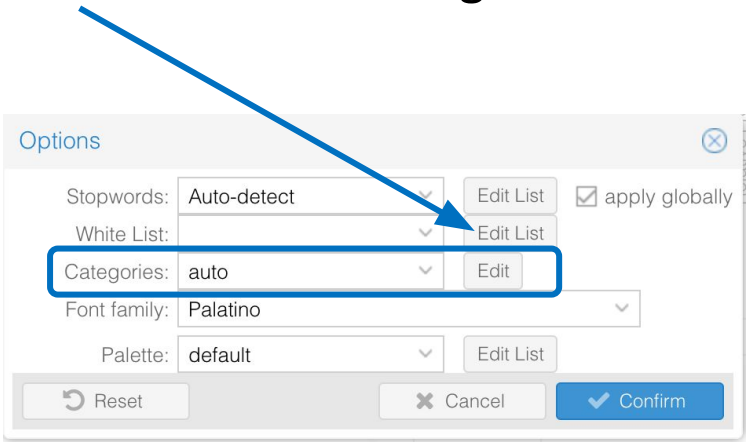
- CIRRUS: l'effetto dell'applicazione delle stopword

PRIMA

DOPO

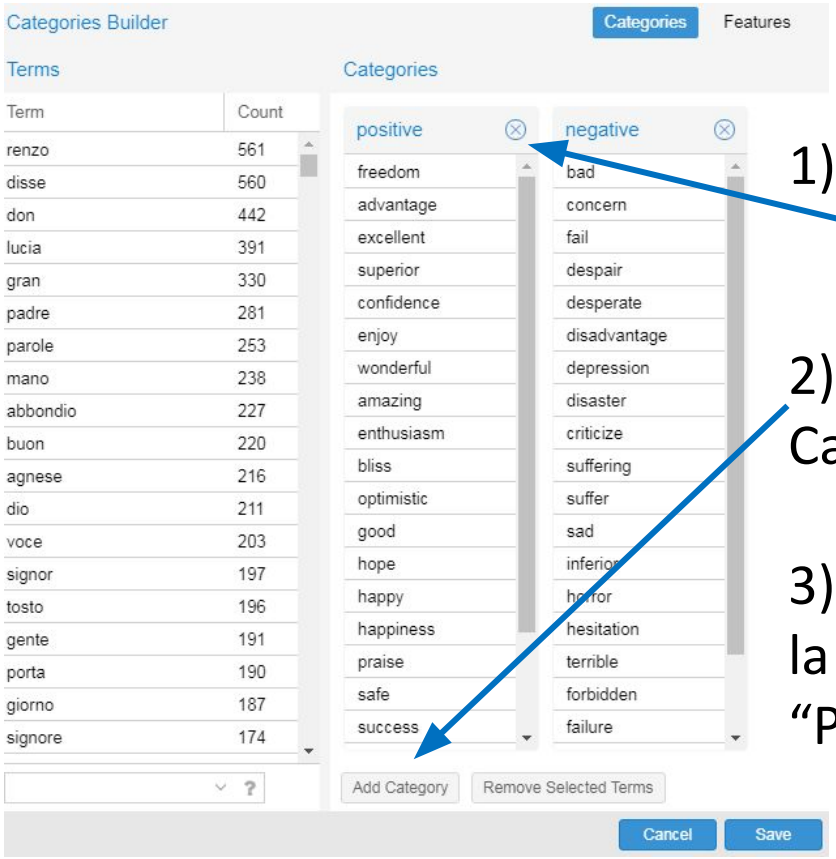
ESERCITAZIONE: VOYANT

- 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



ESERCITAZIONE: VOYANT

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



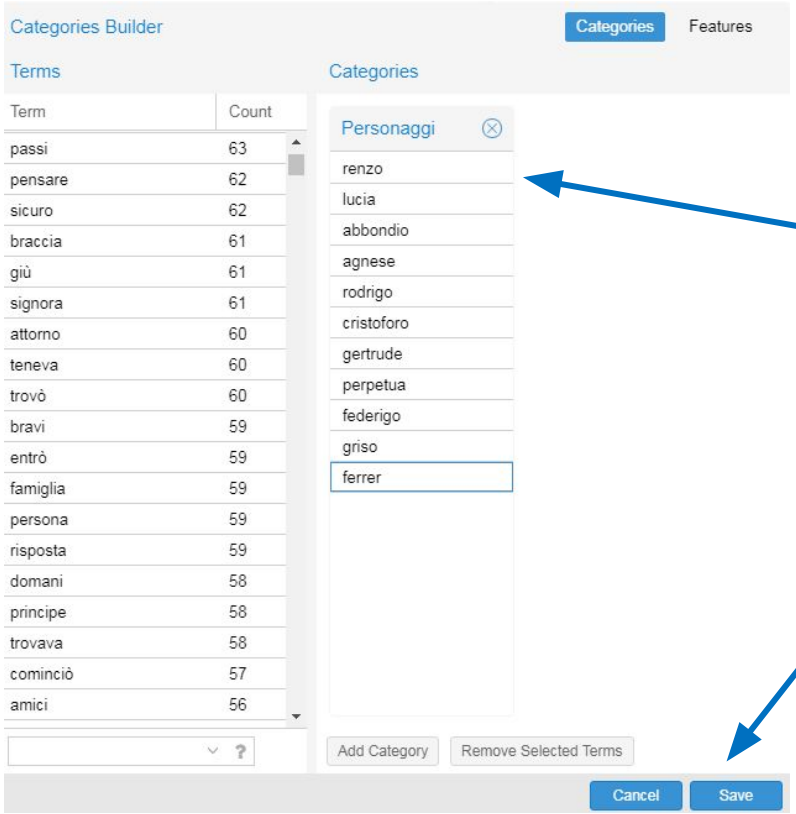
1) Rimuovere le categorie esistenti

2) Cliccare su “Add Category”

3) Scrivere un nome per la categoria, e.g. “Personaggi”

ESERCITAZIONE: VOYANT

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



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

5) Salvare cliccando su “Save”

ESERCITAZIONE: VOYANT

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

- 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
- 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 che formano il corpus

ESERCITAZIONE: VOYANT

- ESPORTAZIONE 

- Puoi esportare una URL, uno strumento incorporabile (interattivo) o un riferimento bibliografico
- Si applica all'intero progetto Voyant o a un singolo strumento particolare ("skin")
- 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

ESERCITAZIONE: VOYANT

- SINTASSI DELLE RICERCHE POSSIBILI

- 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 nella categoria
- ^@Personaggi: ricerca dei singoli termini inclusi nella categoria

ESERCITAZIONE: VOYANT

- ESEMPI

- Come individuare gli hapax legomena? → TERMS
- Quali sono le parole più frequenti del cap 34? → CIRRUS + SCALE
- Quali sono i capitoli più connessi alla pestilenza? → MANDALA
- Quale personaggio viene menzionato di più e in che parti? → TRENDS, BUBBLELINES, MICROSEARCH
- Qual è il capitolo con più diversità/ricchezza lessicale? → DOCUMENTS
- Quali termini appaiono più frequentemente accanto alla parola “dio”? → COLLOCATES

ESERCITAZIONE: VOYANT

- Caricare i 3 file nella sotto-cartella versioni: Fermo_e_Lucia.txt, Promessi_Sposi_27.txt, Promessi_Sposi_40.txt e poi caricare la lista di stopword

- Ci sono differenze notevoli tra i 3 testi dal punto di vista numerico? Lunghezza, densità del vocabolario?
- Ci sono sintagmi ricorrenti? Sintagmi con piccole variazioni?
- Il personaggio di Geltrude appare in “Fermo e Lucia” similmente a come appare Gertrude nei “Promessi Sposi”?
- Quante parole contengono la lettera “j” nelle varie versioni?

GRAZIE!

Email: rachele.sprugnoli@unicatt.it
Twitter: @RSprugnoli