FONDAZIONE BRUNO KESSLER

DH DIGITAL HUMANITIES

Cos'è l'Informatica Umanistica?
ovvero
Esplora i libri con le nuove tecnologie

Stefano Menini e Rachele Sprugnoli

Chi siamo

STEFANO

Liceo Scientifico

~~Corso di Laurea *Informatica* - Università di Verona~~

Corso di Laurea *Filosofia* Università di Trento

Fondazione Bruno Kessler + Scuola di Dottorato *Informatica* Università di Trento

RACHELE

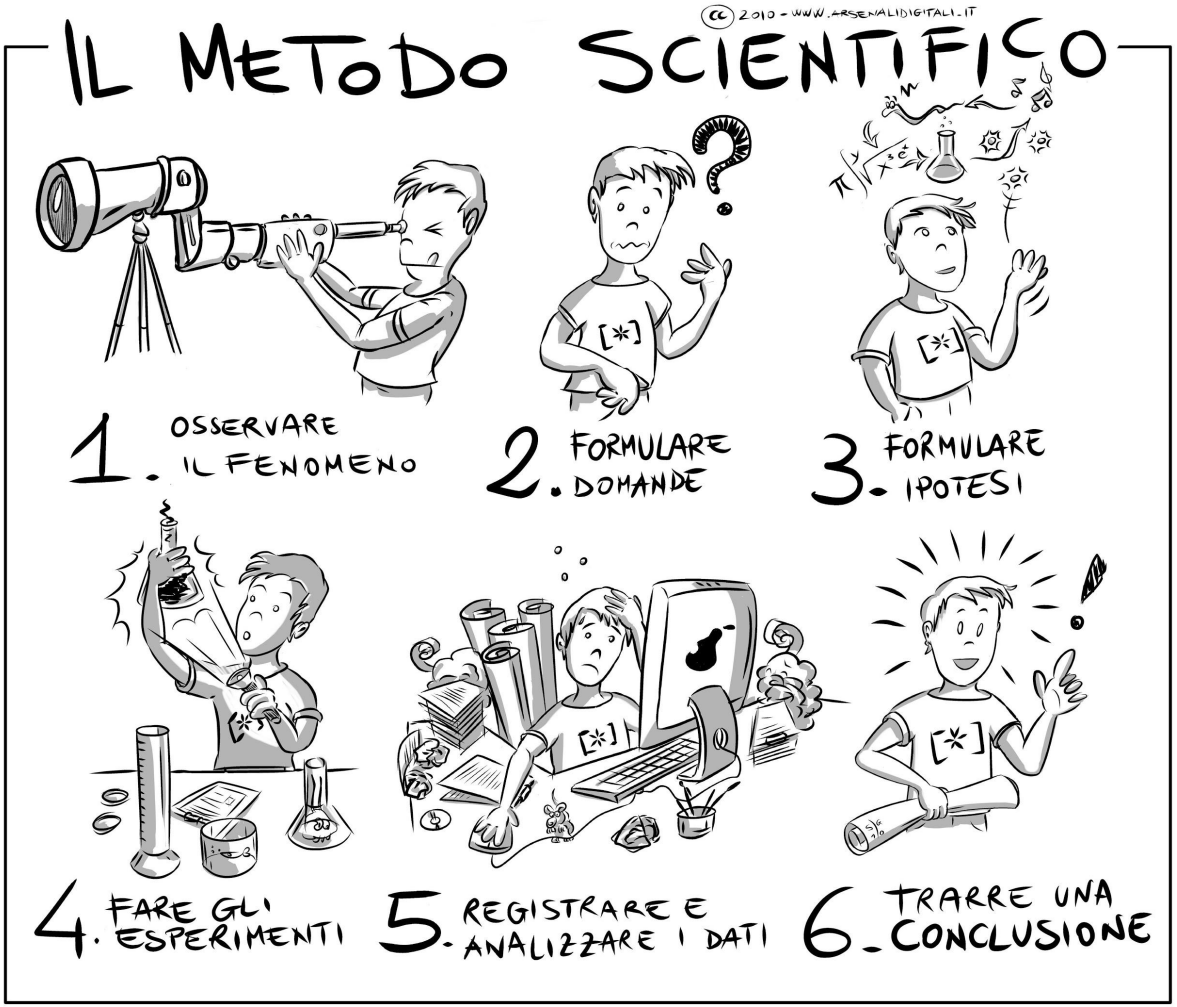
Liceo Classico

~~Corso di Laurea *Chimica* Università di Pisa~~

Corso di Laurea *Informatica Umanistica* Università di Pisa

Fondazione Bruno Kessler + Scuola di Dottorato *Informatica* Università di Trento

Perché vi definite ricercatori e scienziati?



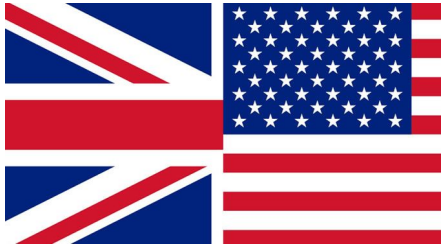
Dov'è il camice?

- Siamo ricercatori anche se non indossiamo il camice!
- Facciamo scienza anche se non usiamo il microscopio!

Cosa vuol dire fare ricerca?

- Scrivere articoli per conferenze e riviste

Tanto inglese!!



Digging in the Dirt: Extracting Keyphrases from Texts with KD

[1]Giovanni Moretti, [1-2]Rachele Sprugnoli, [1]Sara Tonelli
[1]Fondazione Bruno Kessler, Trento
[2] Università di Trento
{moretti,sprugnoli,satonelli}@fbk.eu

Abstract

English. In this paper we present a keyphrase extraction system called Keyphrase Digger (KD). The tool uses both statistical measures and linguistic information to detect a weighted list of n-grams representing the most important concepts of a text. KD is the reimplementation of an existing tool, which has been extended with new features, a high level of customizability, a shorter processing time and an extensive evaluation on different text genres in English and Italian (i.e. scientific articles and historical texts).

ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment

Giovanni Moretti, Sara Tonelli
Fondazione Bruno Kessler
Via Sommarive 18 Trento
moretti@fbk.eu
satonelli@fbk.eu

Stefano Menini, Rachele Sprugnoli
Fondazione Bruno Kessler and
University of Trento
menini@fbk.eu
sprugnoli@fbk.eu

Abstract

English. This work presents ALCIDE (*Analysis of Language and Content In a Digital Environment*), a new platform for Historical Content Analysis. Our aim is to improve Digital Humanities studies integrating methodologies taken from human language technology and an easily understandable data structure representation. ALCIDE provides a wide collection of tools that go beyond simple metadata indexing, implementing functions of textual analysis such as named entity recognition, key-concept extraction, lemma and string-based search and geo-tagging.

A SICK cure for the evaluation of compositional distributional semantic models

M. Marelli[1], S. Menini[1,2], M. Baroni[1], L. Bentivogli[2], R. Bernardi[1], R. Zamparelli[1]
[1]University of Trento, [2]Fondazione Bruno Kessler
marco.marelli@unitn.it, menini@fbk.eu, marco.baroni@unitn.it,
bentivo@fbk.eu, raffaella.bernardi@unitn.it, roberto.zamparelli@unitn.it

Abstract

Shared and internationally recognized benchmarks are fundamental for the development of any computational system. We aim to help the research community working on compositional distributional semantic models (CDSMs) by providing SICK (Sentences Involving Compositional Knowldedge), a large size English benchmark tailored for them. SICK consists of about 10,000 English sentence pairs that include many examples of the lexical, syntactic and semantic phenomena that CDSMs are expected to account for, but do not require dealing with other aspects of existing sentential data sets (idiomatic multiword expressions, named entities, telegraphic language) that are not within the scope of CDSMs. By means of crowdsourcing techniques, each pair was annotated for two crucial semantic tasks: relatedness in meaning (with a 5-point rating scale as gold score) and entailment relation between the two elements (with three possible gold labels: entailment, contradiction, and neutral). The SICK data set was used in SemEval-2014 Task 1, and it freely available for research purposes.

Keywords: Compositional Distributional Semantic Models, Data Sets, Semantic Relatedness, Textual Entailment

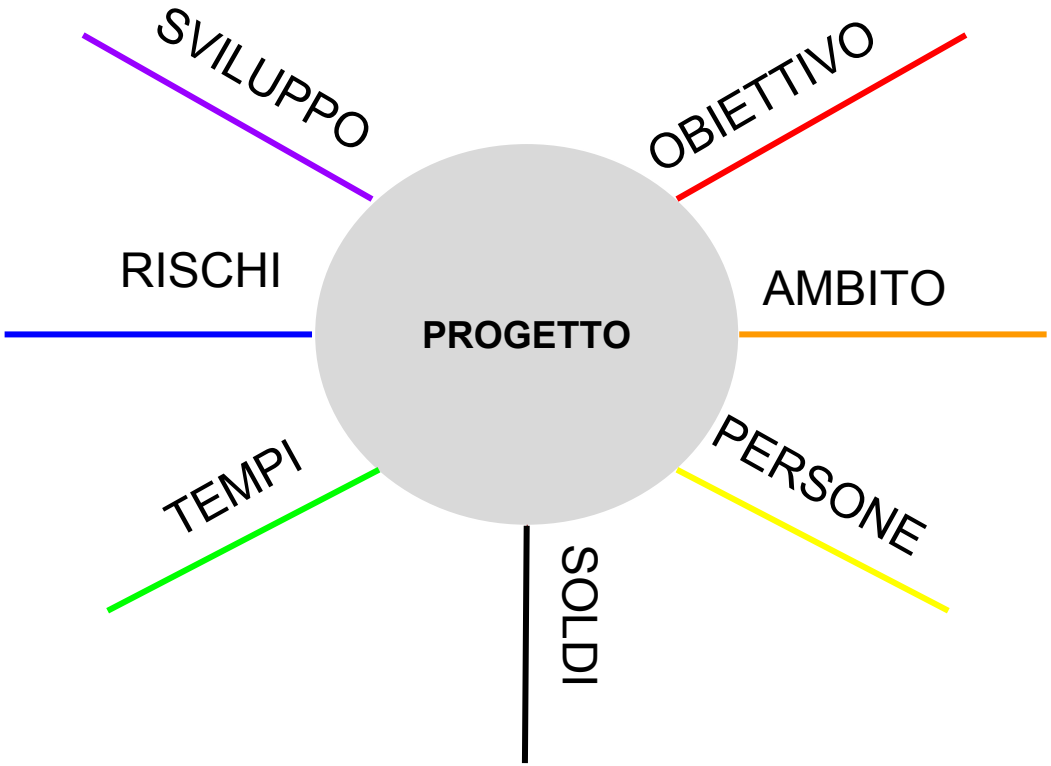
1. Introduction

Distributional Semantic Models (DSMs) approximate the meaning of words with vectors summarizing their patterns of co-occurrence in corpora. Recently, several compositional extensions of DSMs (Compositional DSMs, or CDSMs) have been proposed, with the purpose of representing the meaning of phrases and sentences by composing the distributional representations of the words they contain

(Sentences Involving Compositional Knowledge), a data set aimed at filling this void. SICK includes a large number of sentence pairs that are rich in the lexical, syntactic and semantic phenomena that CDSMs are expected to account for, but do not require dealing with other aspects of existing sentential data sets (multiword expressions, named entities, telegraphic language) that are not within the domain of compositional distributional semantics.

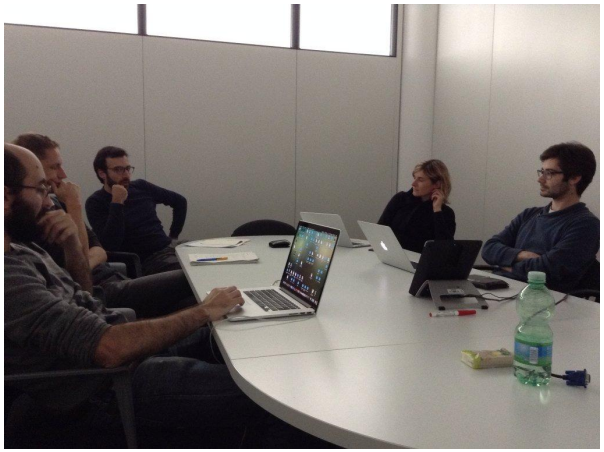
Cosa vuol dire fare ricerca?

- Partecipare a progetti di ricerca in ambito nazionale ed internazionale

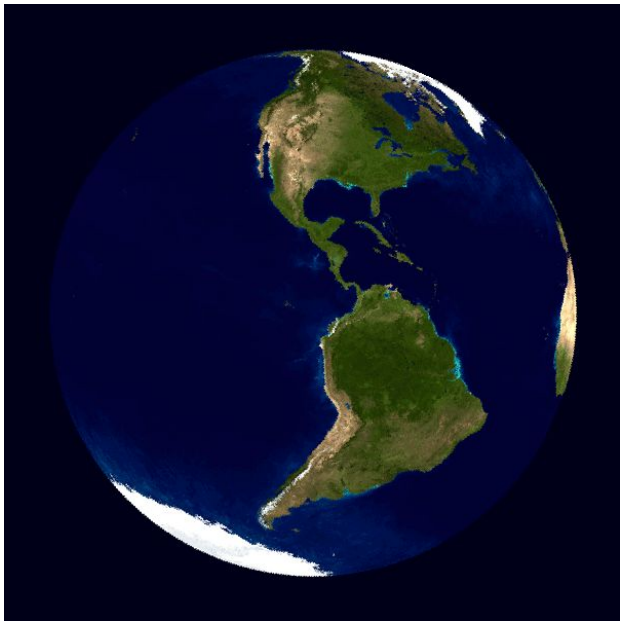


Cosa vuol dire fare ricerca?

- Fare tante riunioni



Cosa vuol dire fare ricerca?



- Viaggiare molto per partecipare a conferenze e workshop

- Parlare in pubblico

Cosa vuol dire fare ricerca?

- Portare il proprio lavoro anche fuori dall'università ed FBK
  - Scuole: proprio come oggi!
  - Musei
  - Festival

Cosa è l'Informatica Umanistica?

- Informatica Umanistica = Digital Humanities

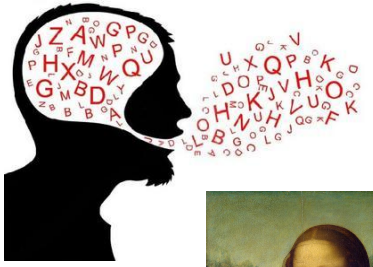
- Parole chiave scelte dai ricercatori:

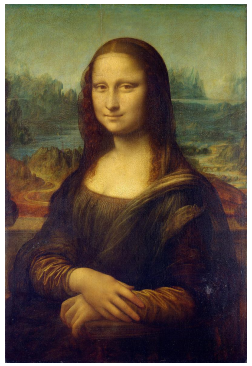
Un concetto fondamentale...

Interdisciplinarietà

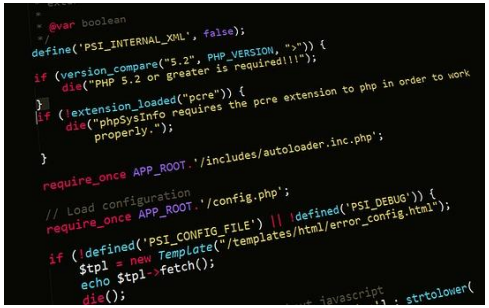
Scienze Umane + Scienze Esatte / Scienze Naturali / Scienze Applicate = Informatica Umanistica

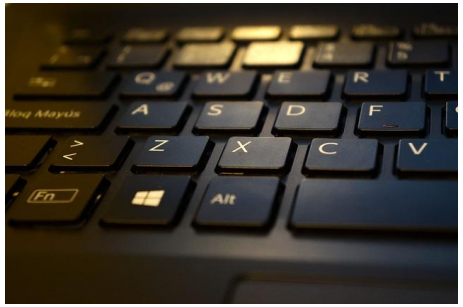
Scienze Umane





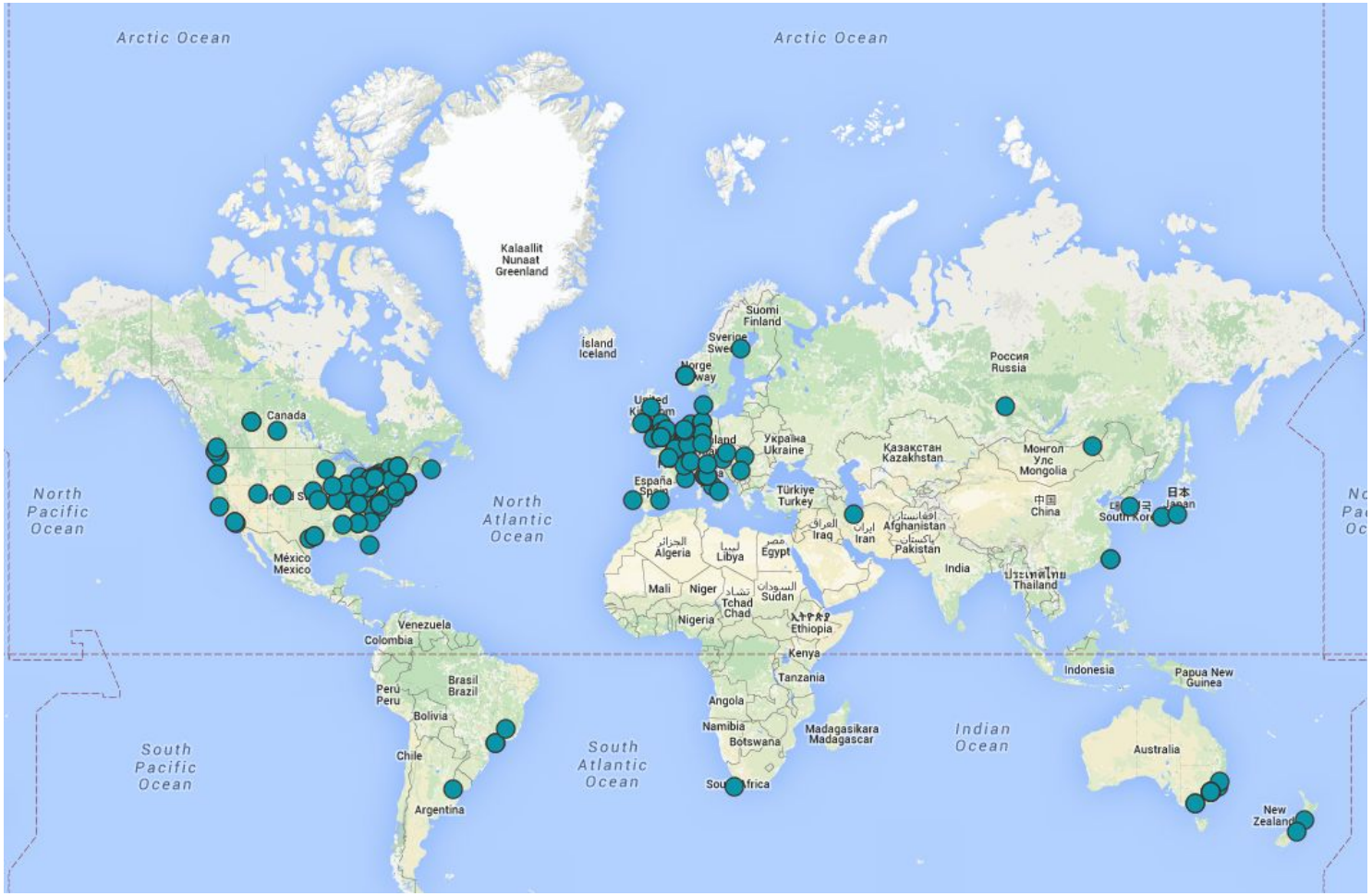
Altre Scienze





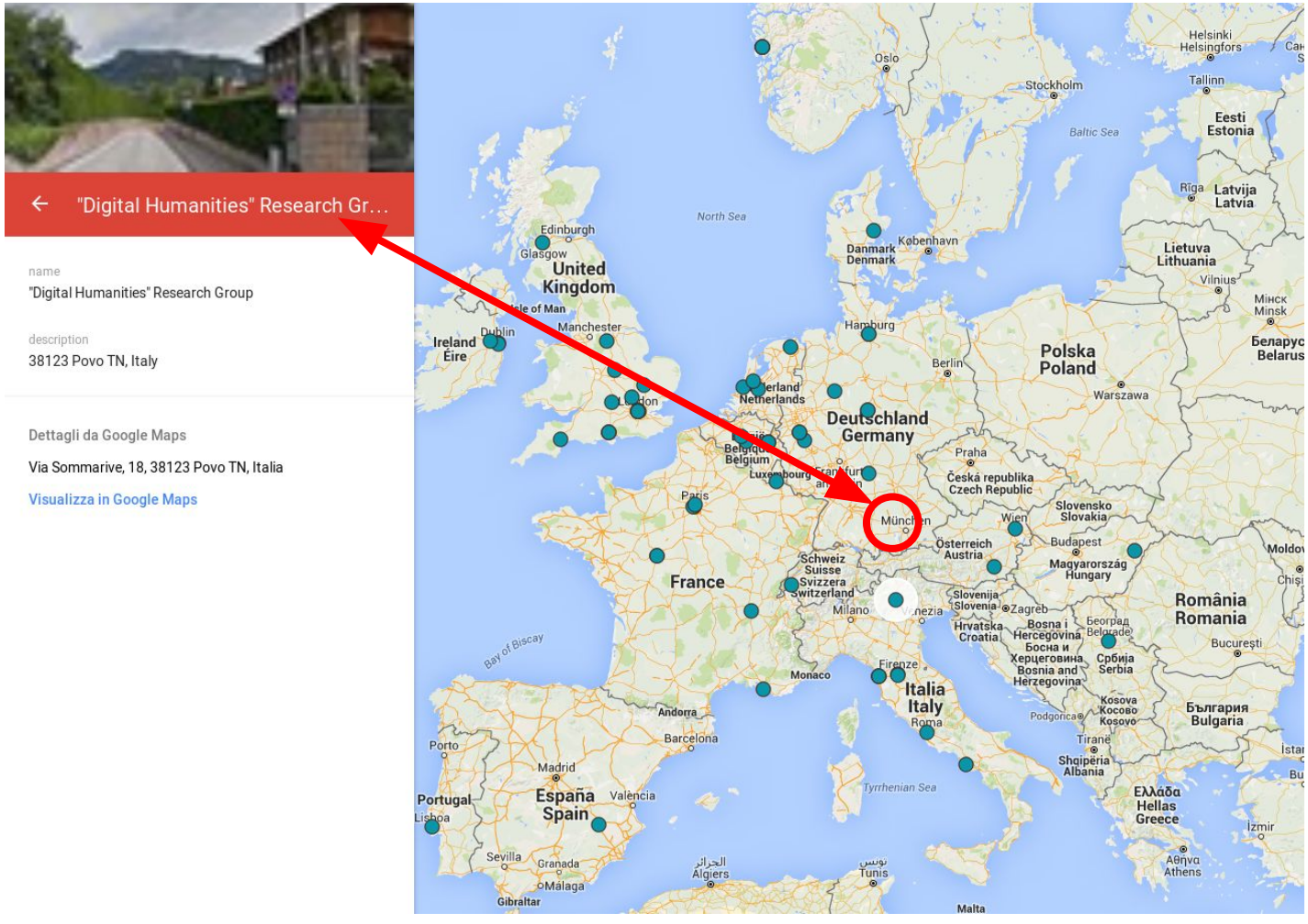
Dove si fa l'Informatica Umanistica?

Mappa dei centri che si occupano di Digital Humanities, da http://www.dhcenternet.org/



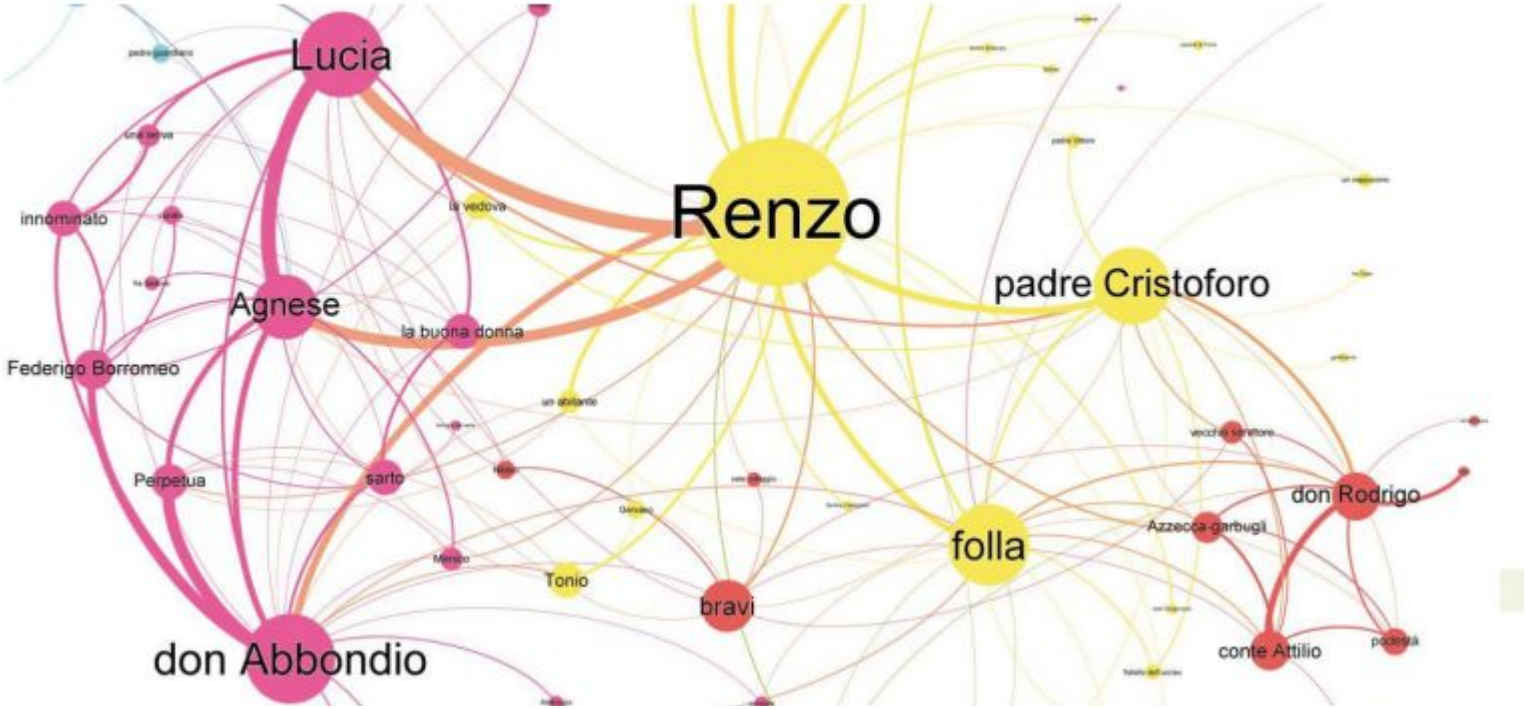
Dove si fa l'Informatica Umanistica?

Mappa dei centri che si occupano di Digital Humanities: l'Europa da http://www.dhcenternet.org/



Informatica Umanistica e… la letteratura

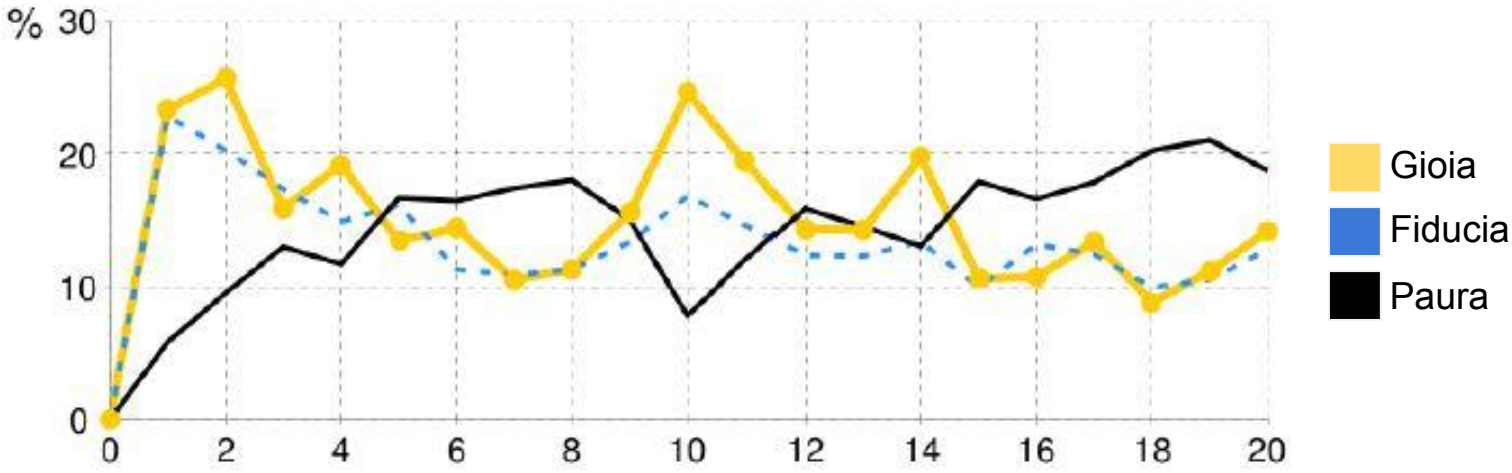
Interazione tra i personaggi de “IPromessi Sposi”



Visualizzazione dal progetto CBook di Cross Library (http://cbook.it/)

Informatica Umanistica e… la letteratura

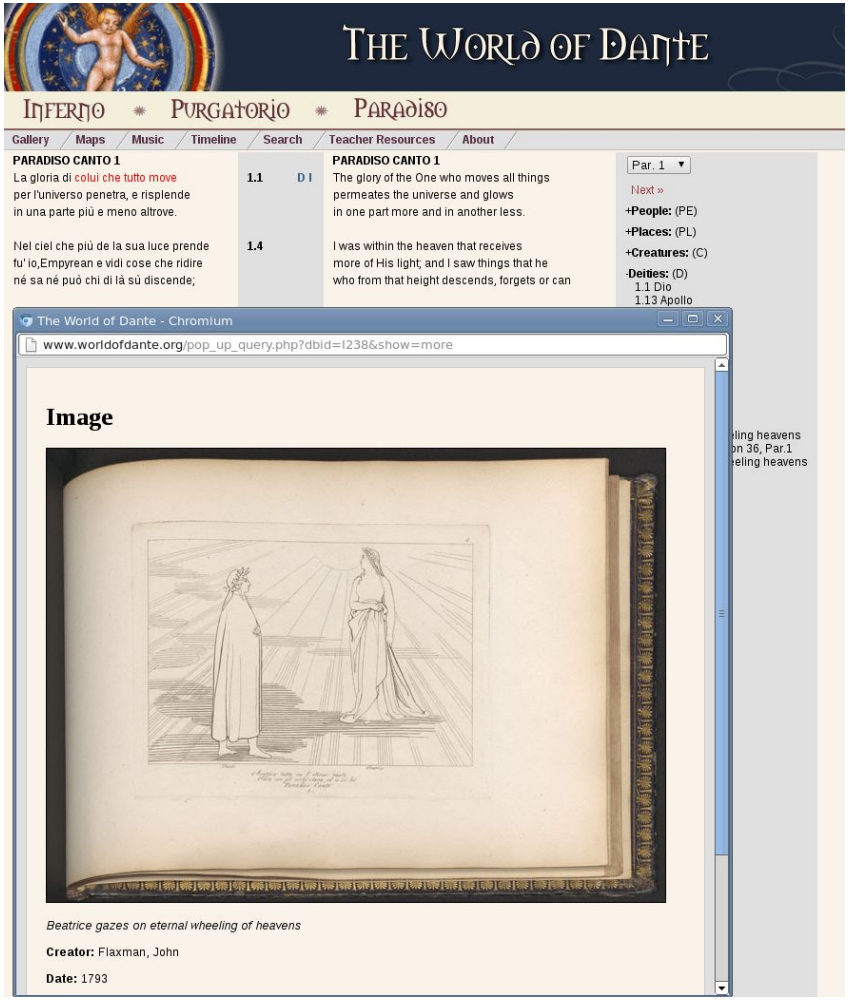
Emozioni nel romanzo “Frankenstein”



Visualizzazione dall’articolo scientifico “From once upon a time to happily ever after: tracking emotions in novels and fairy tales” di Mohammad S. (2011)

Informatica Umanistica e… la letteratura

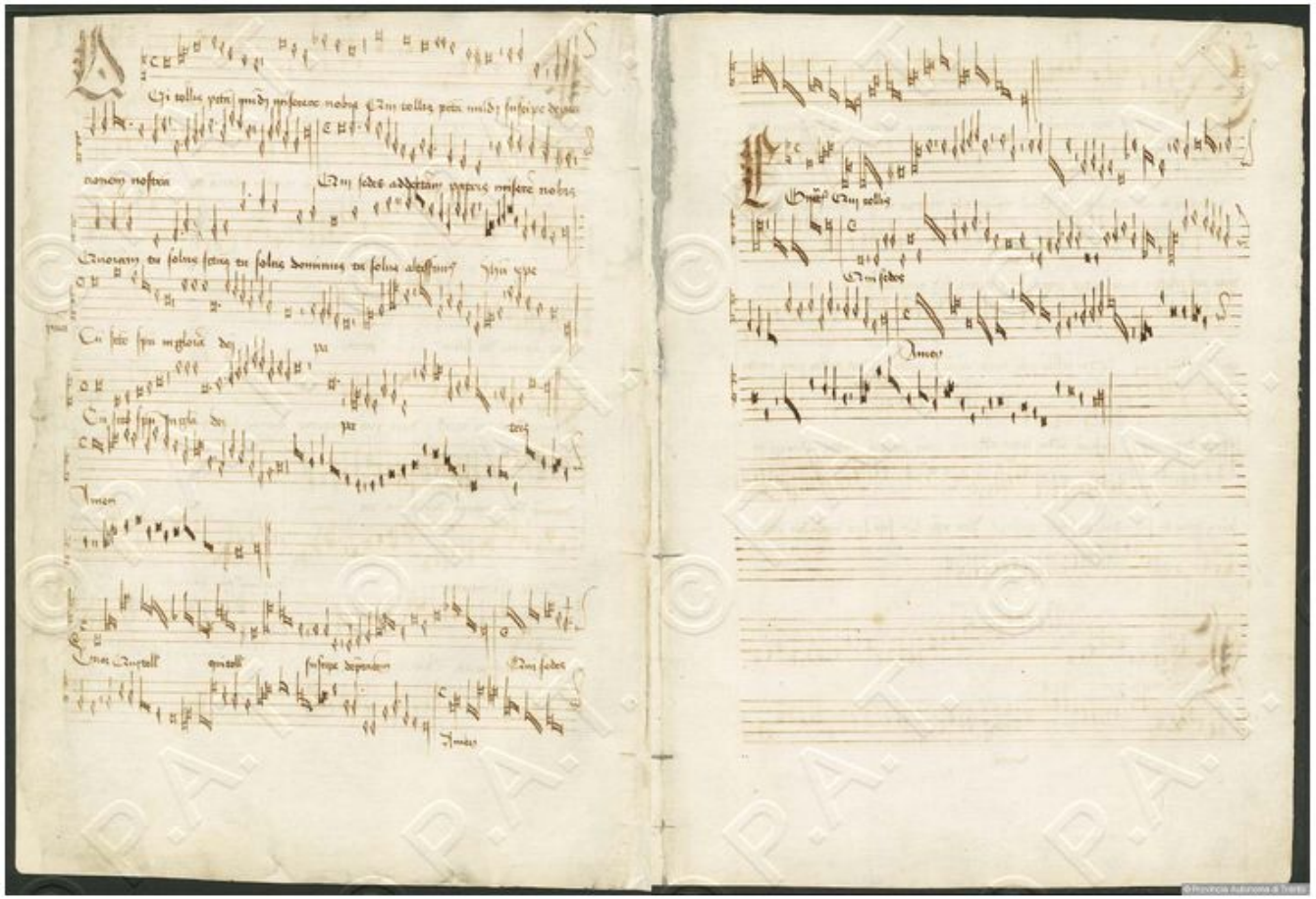
Edizione digitale della “Divina Commedia” di Dante



http://www.worldofdante.org/

Informatica Umanistica e... la musica

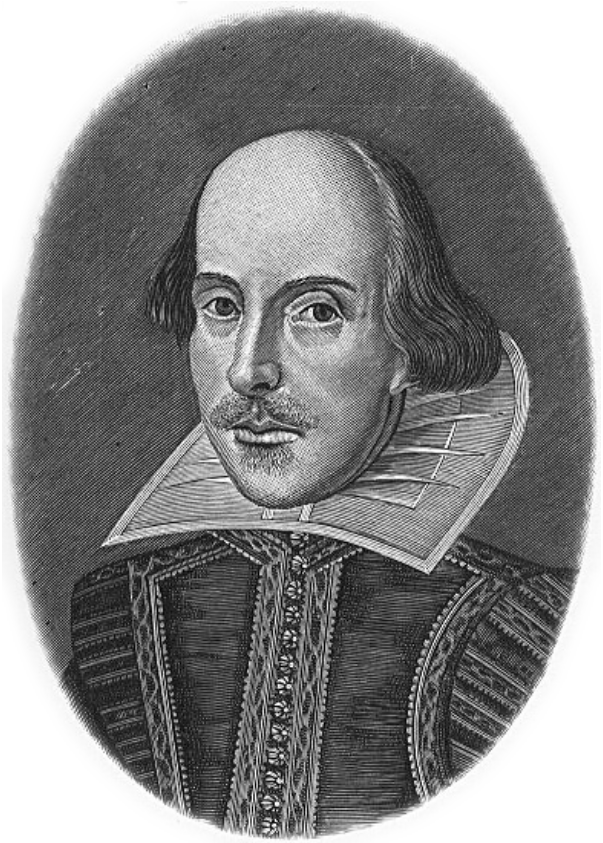
Archivi digitale: “I sette codici musicali trentini del Quattrocento”



Progetto della Soprintendenza per i beni culturali della Provincia autonoma di Trento, in collaborazione con il Ministero per i Beni e le Attività Culturali e con la Società Filarmonica di Trento.

Informatica Umanistica e… la musica

Analisi dei testi delle canzoni: Shakespeare contro l'Hip-Hop



http://poly-graph.co/vocabulary.html

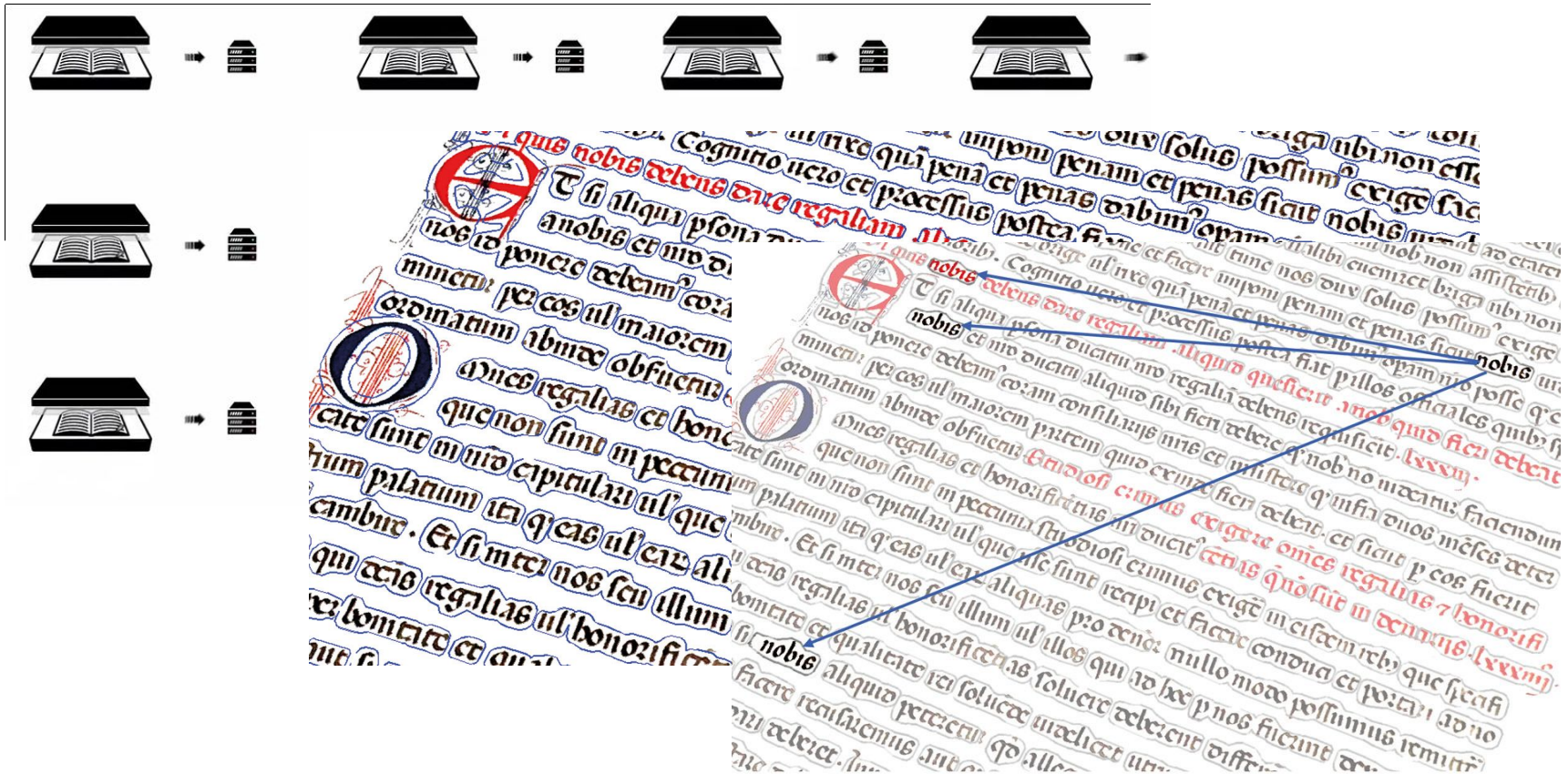
Informatica Umanistica e… l'arte

Giochi: il nostro progetto Verbo-Visuale-Virtuale con Mart di Rovereto e Museion di Bolzano!

https://youtu.be/PgiZl6noPns?t=5m29s

Informatica Umanistica e... la storia

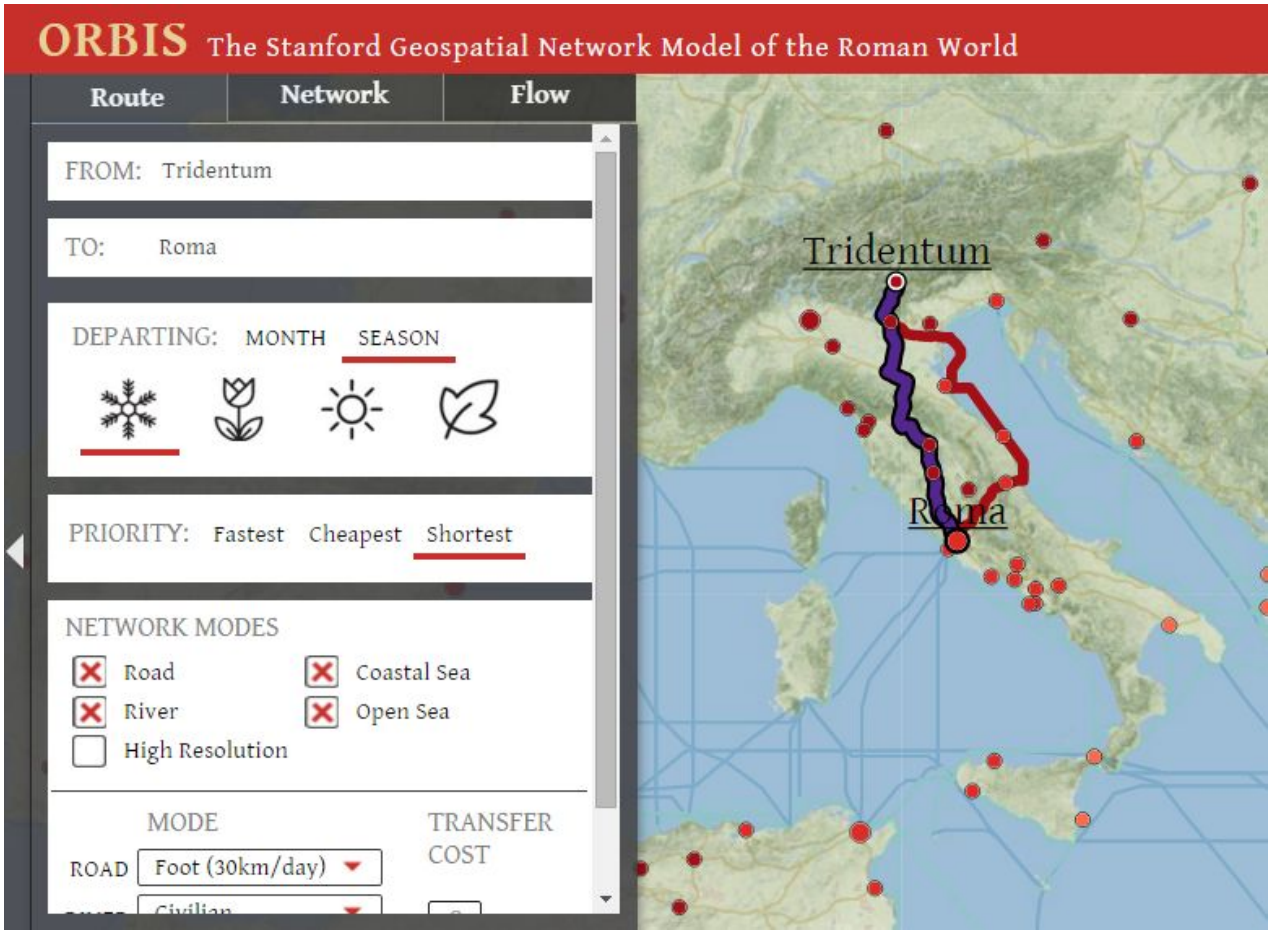
Digitalizzazione manoscritti: dalle biblioteche al computer.
Il caso di “Venice Time Machine”



http://vtm.epfl.ch/

Informatica Umanistica e… la storia

Il Google Map dell'antichità: progetto ORBIS



http://orbis.stanford.edu/

Informatica Umanistica e... la storia

Conservare la memoria con portali online

http://www.memoro.org/

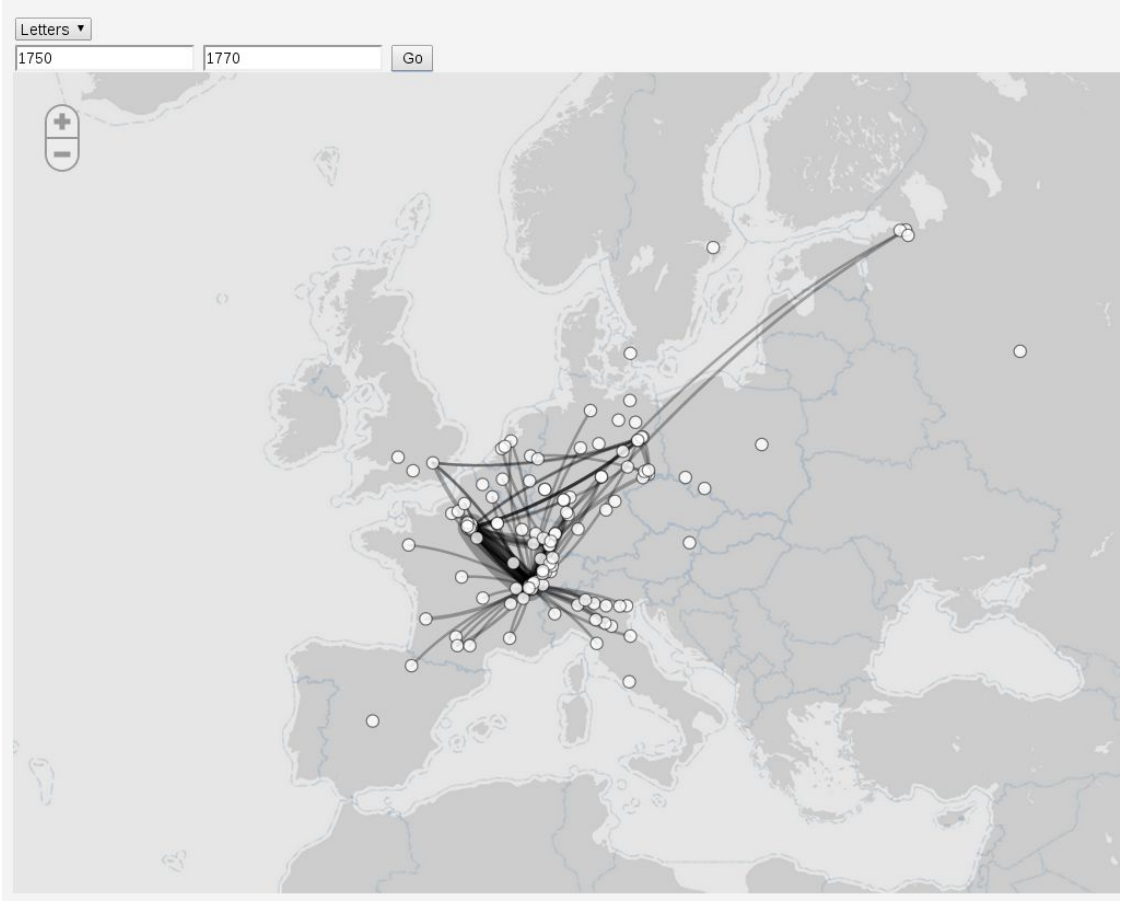
Informatica Umanistica e… la storia

Conservare la memoria con portali online

http://www.lagrandeguerrapiu100.it/

Informatica Umanistica e… la storia

Scoprire cose nuove sui personaggi storici: con chi erano in contatto quando non c'era Facebook?

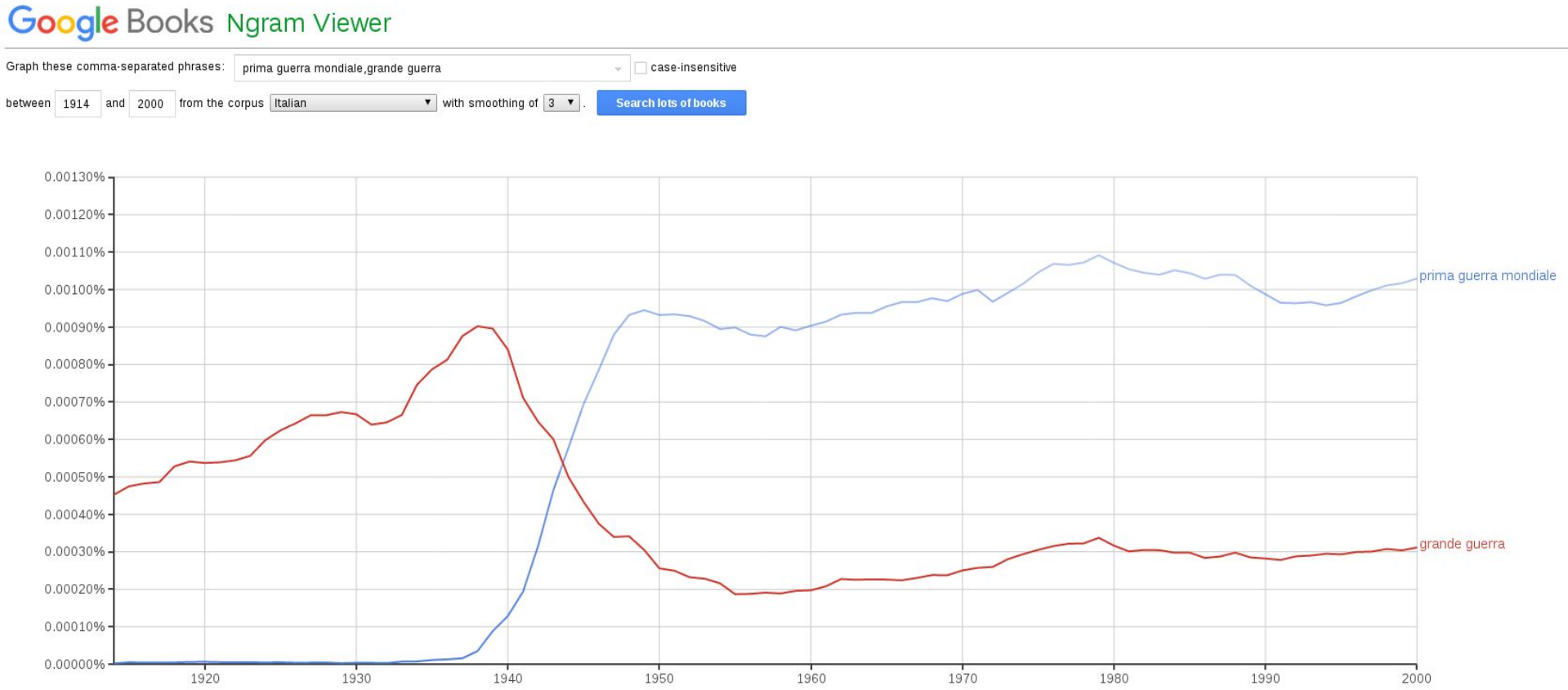


http://republicofletters.stanford.edu/

Informatica Umanistica e… la storia

Scoprire cose nuove sulla cultura del passato vicino e lontano usando milioni di libri in una volta sola

PRIMA GUERRA MONDIALE vs. GRANDE GUERRA

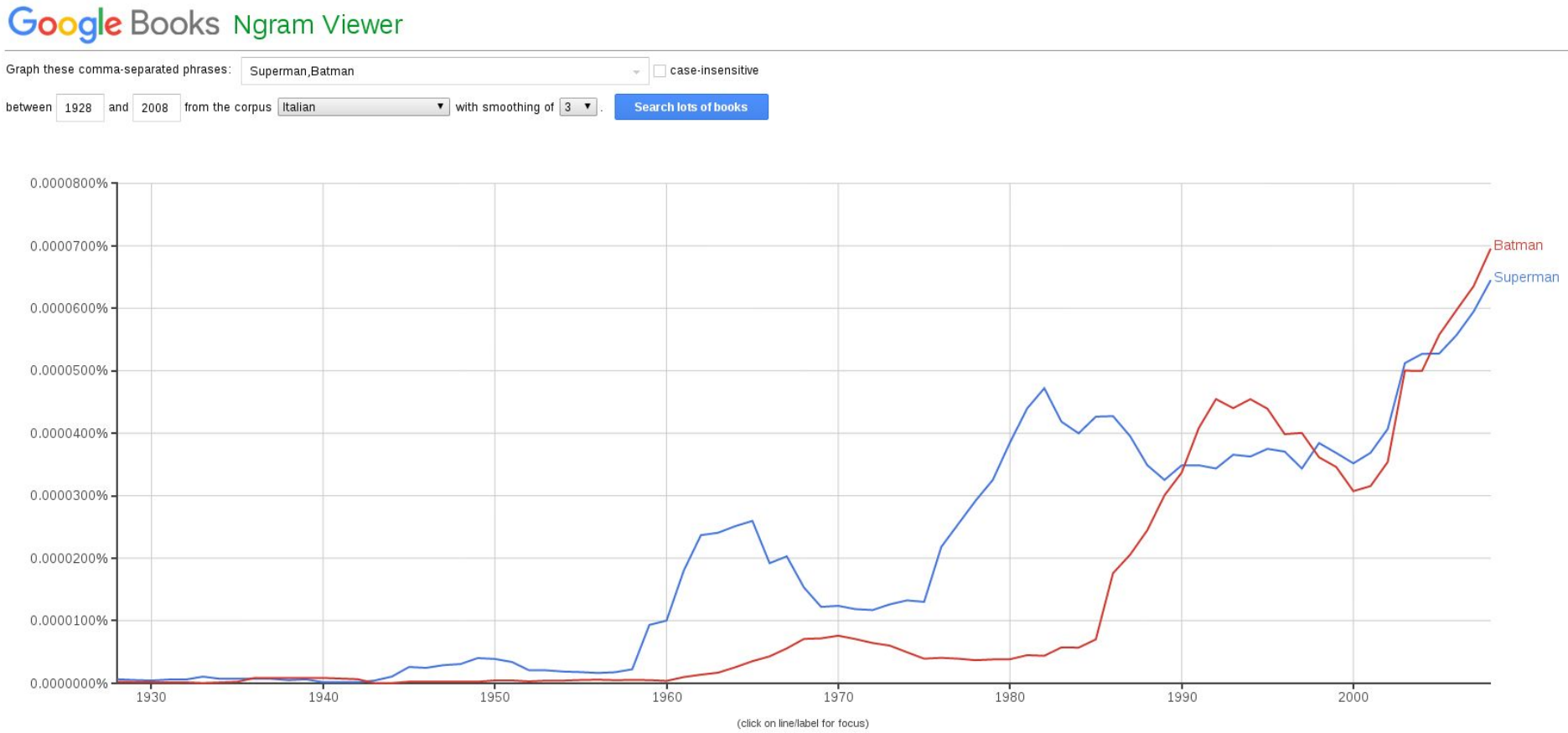


https://books.google.com/ngrams

Informatica Umanistica e...
la storia

Scoprire cose nuove sulla cultura del passato vicino e lontano usando milioni di libri in una volta sola

SUPERMAN vs. BATMAN



Informatica Umanistica e... la storia

Analisi linguistica dei testi storici: quello che facciamo noi!!

Visualizzazione dall'articolo scientifico "Towards sentiment analysis for historical texts" di Sprugnoli et al. (2015)

Esplora i libri con le nuove tecnologie

Cosa vuol dire fare analisi linguistica con un computer?

Elaborazione del linguaggio naturale:

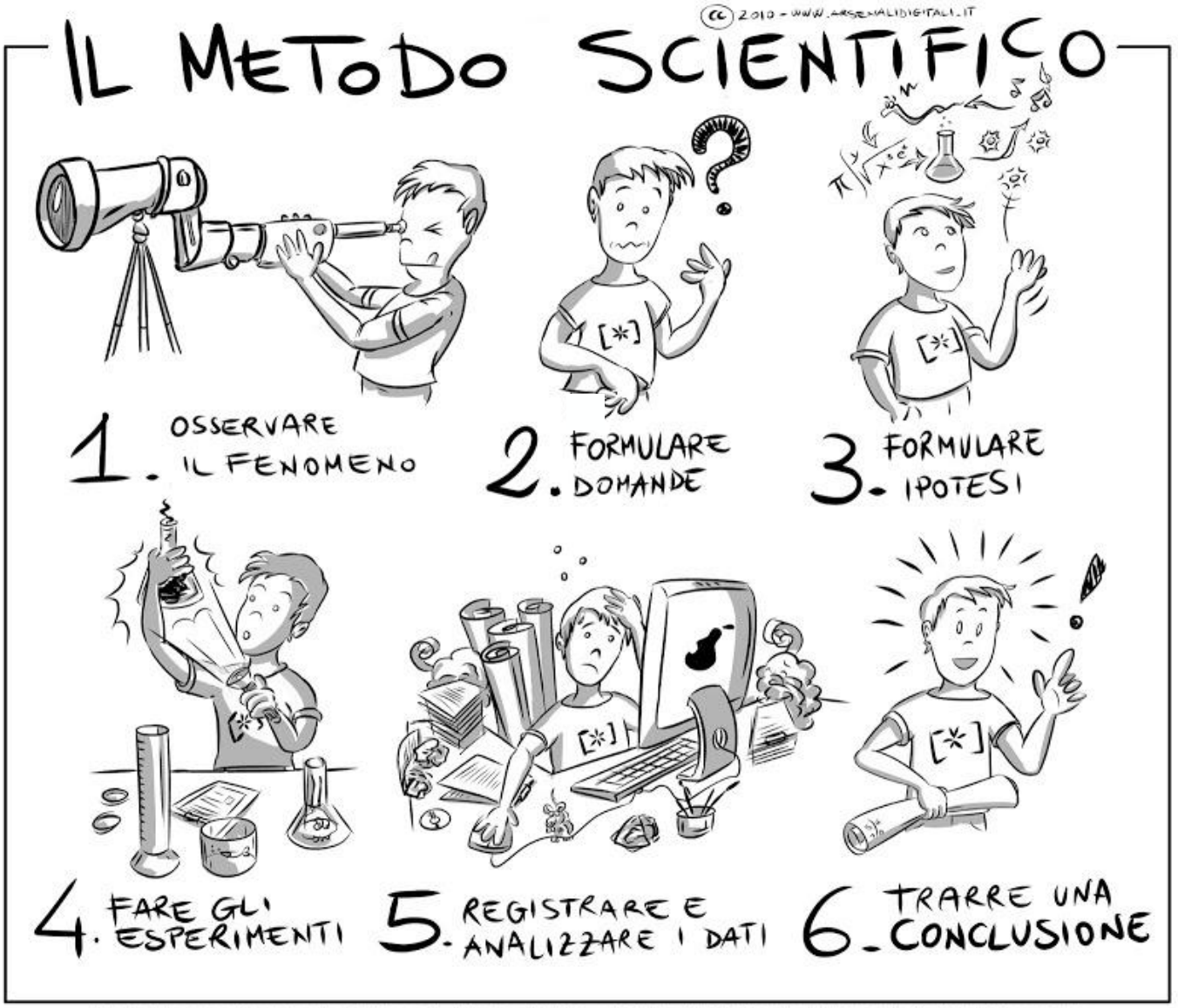
- Elaborazione automatica per fare analisi **lessicale**; **grammaticale**, **sintattica**, **semantica** (del significato)

Esplora i libri con le nuove tecnologie

A cosa serve fare analisi linguistica?

- Esplorare il contenuto dei testi con il computer.
- Leggere i libri in un modo diverso dal solito.

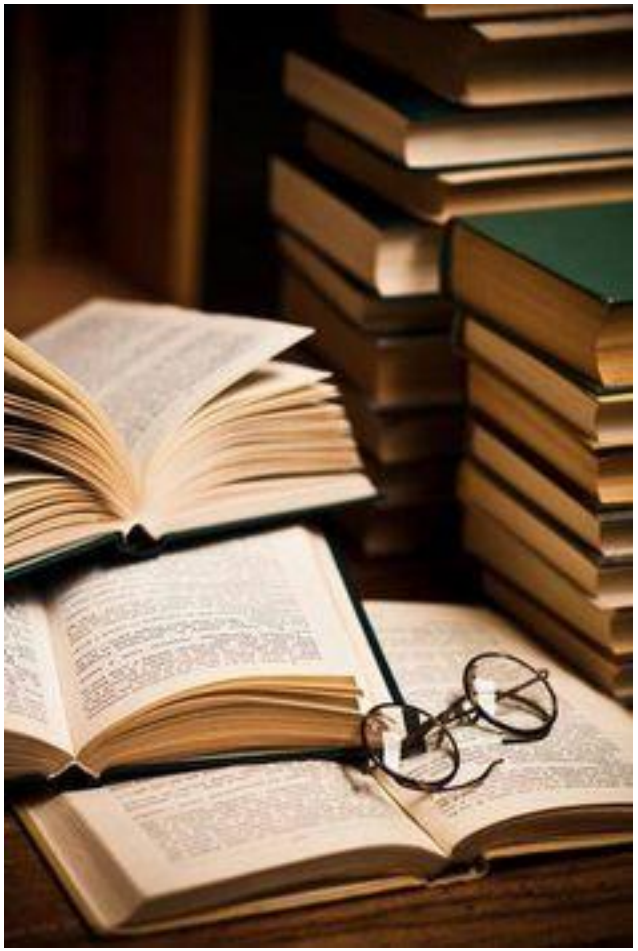
Come sviluppiamo un progetto?



Osserviamo

- Osserviamo il mondo della ricerca nelle **discipline umanistiche** (supratutto ricerca storica).
- Osserviamo il loro **metodo** di lavoro.
- Osserviamo il tipo di **dati** su cui lavorano.

Il loro metodo: Lettura “da vicino”

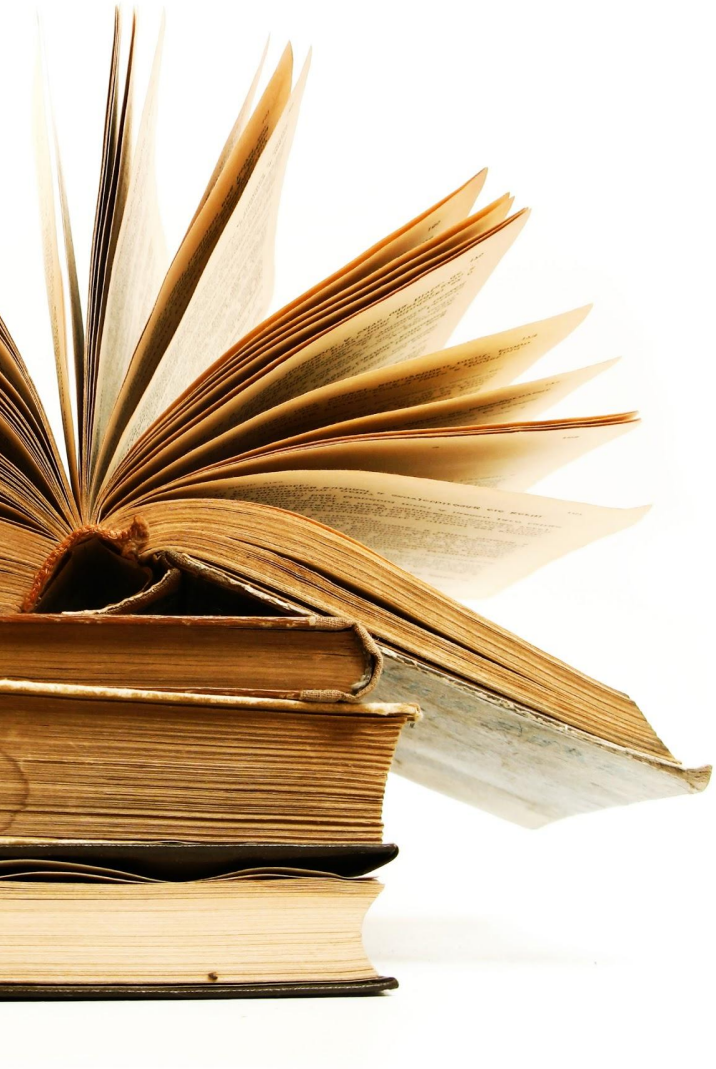


- Il testo viene letto passaggio per passaggio con attenzione.
- Viene fatta un’**analisi critica** dei testi (viene valutato il significato contenuto nel contesto in cui appare).

\+ Precisa

\- Richiede molto tempo.

Osserviamo i loro dati:



- Collezioni di centinaia o migliaia di testi storici.

Troviamo un Problema:

- Leggerli e studiarli tutti richiede troppo tempo.

- Serve uno strumento per esplorare il contenuto di questi testi.

Ci facciamo domande!

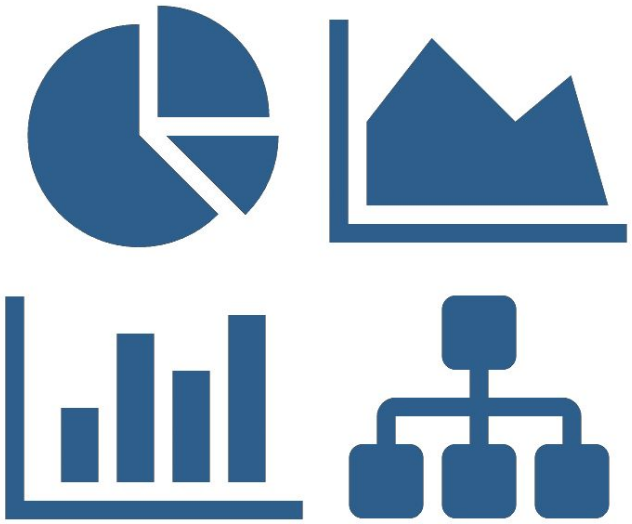
- Che tipo di **strumenti** servono agli storici?
- Come si può studiare il gran numero di documenti storici in modo **veloce e preciso**?
- L'**informatica** ci può aiutare?

e cerchiamo le risposte!

- Facciamo delle **ipotesi** su come si può **risolvere il problema** rendendo il lavoro degli storici più efficace.

- (Non sempre le ipotesi sono corrette, per questo **vanno verificate**)

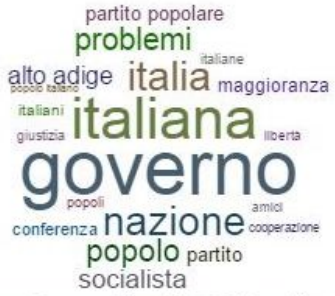
Lettura “da lontano”

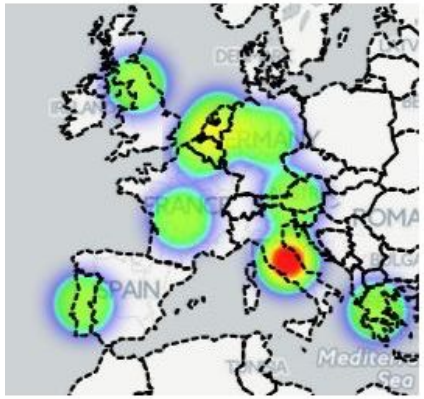


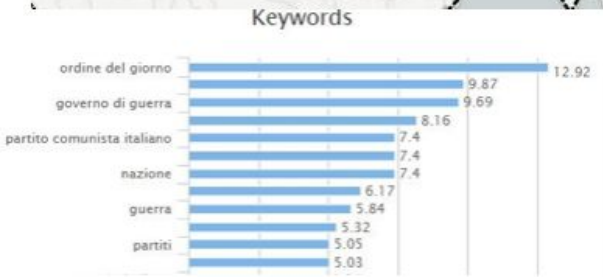
- Analisi dei contenuti e delle caratteristiche dei documenti su **larga scala**.
- Il contenuto dei documenti viene riassunto e reso disponibile attraverso **grafici e schemi**.

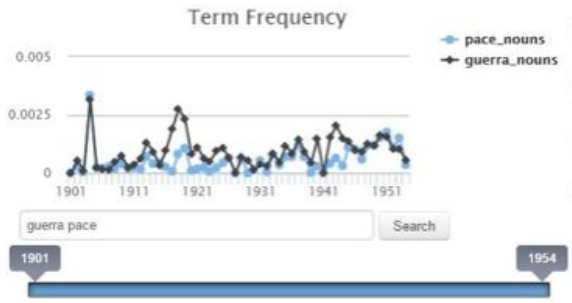
Lettura “da lontano”

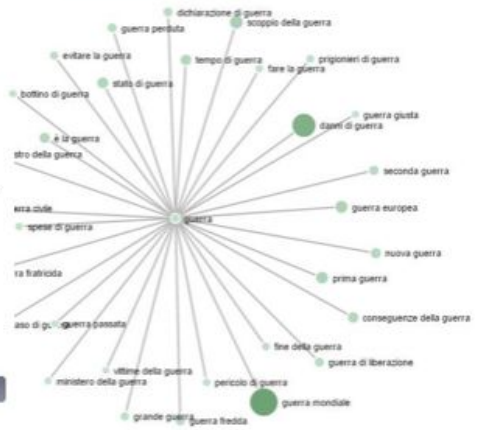
3000 Documenti:



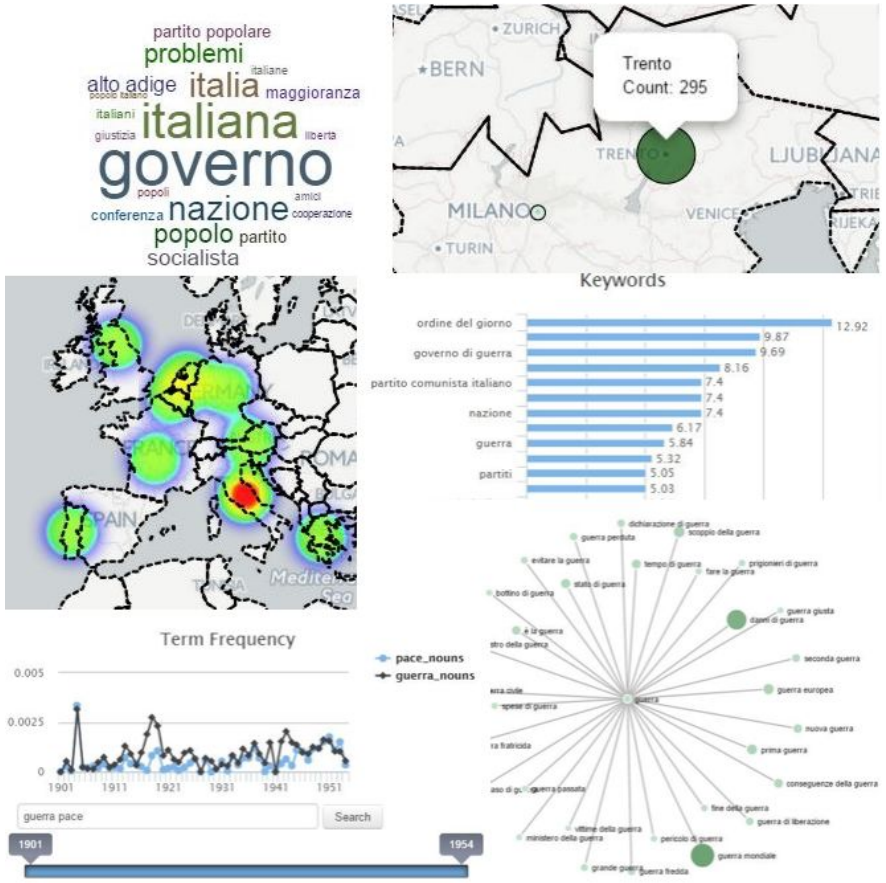








Lettura “da lontano”



+ **Veloce**: in pochi secondi si possono “leggere” centinaia di documenti.

+ **Potente**: si possono ottenere informazioni che sono difficili da avere leggendo il testo normalmente.

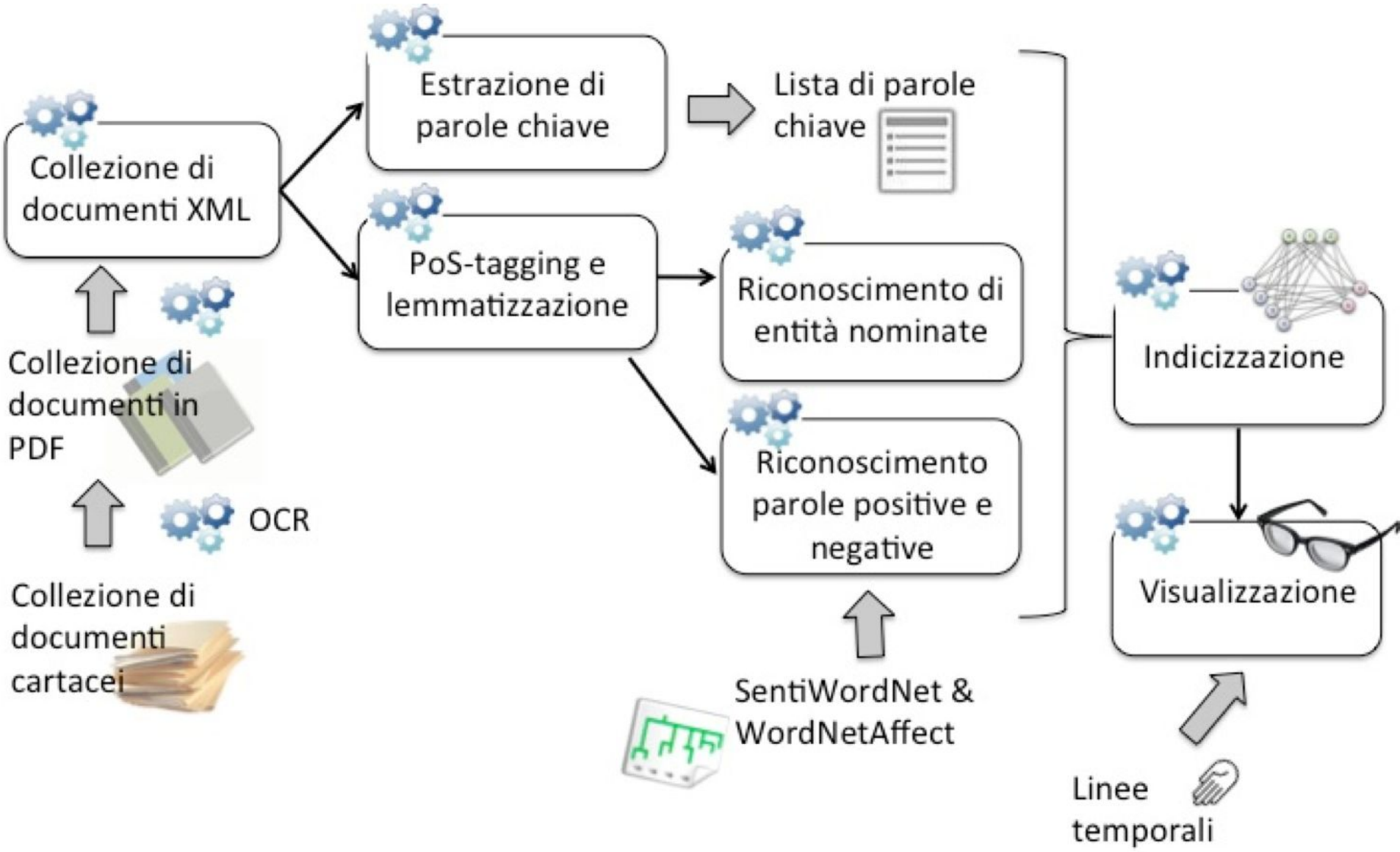
+ Colpo d’occhio.

- Più lontana dal testo.

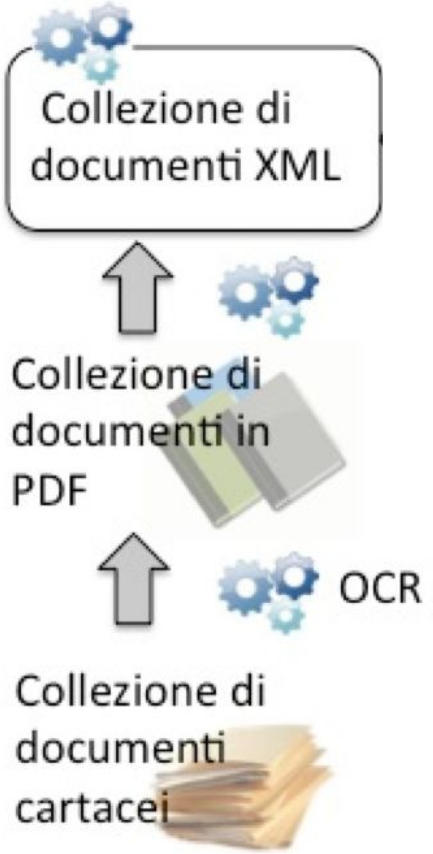
Esperimento


```
response.setContentType("text/html");
response.setCharacterEncoding("UTF-8");
PrintWriter out = response.getWriter();

Map<Integer, Person> mapOfPerson = (Map<Integer, Person>) getServletContext()
KX_configuration configuration = (KX_configuration) getServletContext()

Language lang = Language.ITALIAN;
KX_core kxc = new KX_core(Threads.FOUR);

Calendar yesterday = Calendar.getInstance();
yesterday.add(Calendar.DATE, -1);

Calendar two_days_ago = Calendar.getInstance();
two_days_ago.add(Calendar.DATE, -2);

JSONObject root = new JSONObject();
try {
    root.put("name", "");
    JSONArray rootChilds = new JSONArray();
    for (Person p : mapOfPerson.values()) {
        JSONObject person = new JSONObject();
        person.put("name", p.surname);
        JSONArray personChilds = new JSONArray();
        for (String s : p.sources) {
            JSONObject source = new JSONObject();
            source.put("name", s);

            JSONArray sourceChilds = new JSONArray();
            StringBuffer text_of_source = new StringBuffer();
            int size_of_source_by_year = 0;
            for (Calendar c : p.post_by_source.get(s).keySet()) {
                if (c.get(Calendar.YEAR) == yesterday.get(Calendar.YEAR)
                    text_of_source.append(p.post_by_source.get(s).get(c).
                    size_of_source_by_year++;
                }
            }

            //source.put("size", p.post_by_source.get(s).size());
            source.put("size", size_of_source_by_year);
```

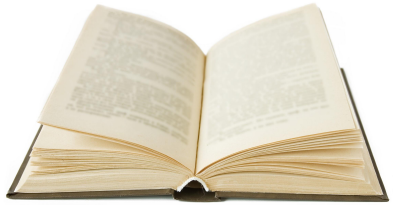
Estrarre Informazioni



Estrarre Informazioni

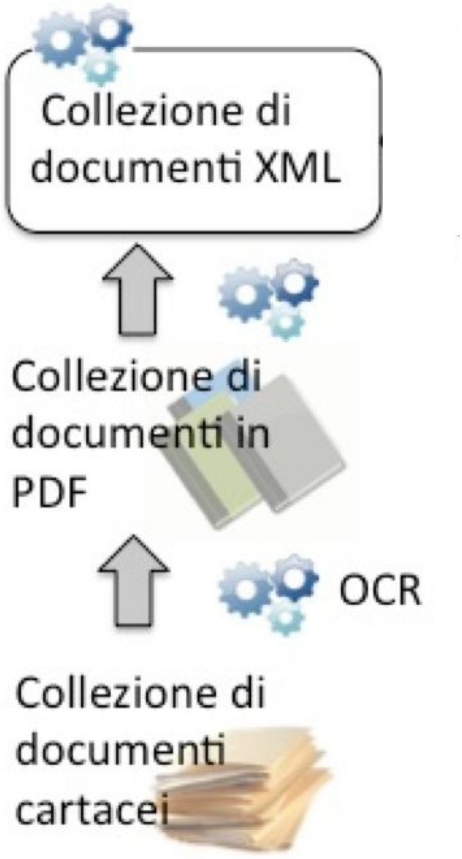


Dobbiamo fare leggere i documenti al computer, quindi devono essere **scritti in modo comprensibile** per il computer



```
<file id="1">
    <head>
        <url>I.pdf_doc_number_1</url>
        <date>1901-01-01</date>
        <title>Title of the document</title>
        <description>Description of the Document</description>
        <location>Geographic location</location>
        <abstract>Content of the abstract</abstract>
    </head>
    <content>Text of the document in Italian</content>
    <original_text>Text of the document not in Italian - if present</original_text>
</file>
```

Estrarre Informazioni



De Gasperi nacque a Pieve Tesino in una povera famiglia tirolese : infatti i suoi genitori dovettero chiedere un sussidio allo Stato austriaco per farlo studiare . Era il primo dei quattro figli di Maria Morandini , nata a Predazzo , e Amedeo De Gasperi , nato a Sardagna.

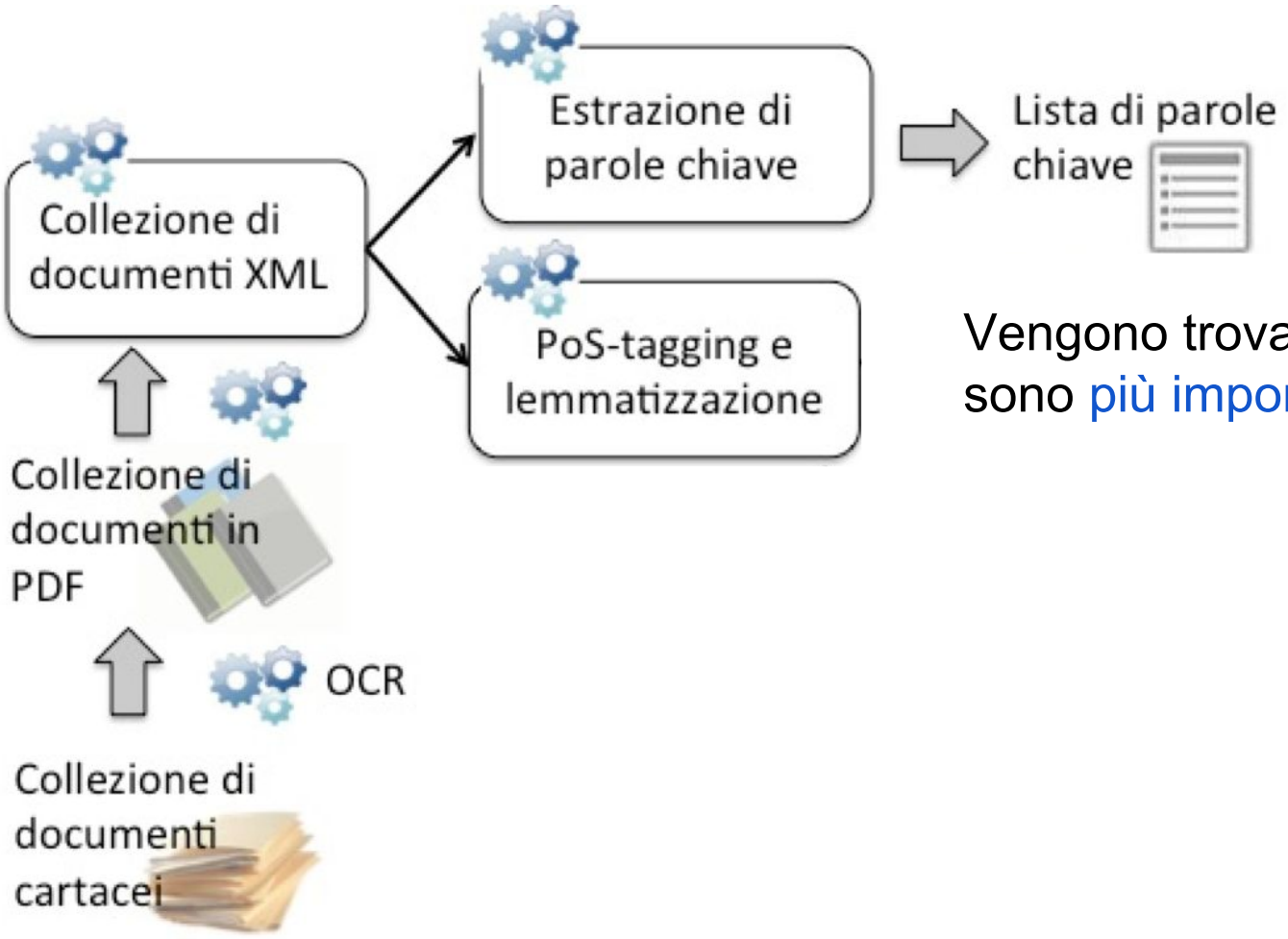
Un computer **capisce** questa frase**?**

Estrarre Informazioni

- Token (parole)
- Parti del discorso (la categoria grammaticali delle parole)
- Lemmi (la forma che trovate sul vocabolario)

token	pos	lemma
De	SPN	__NULL__
Gasperi	SPN	__NULL__
nacque	VI	nascere
a	E	a
Pieve	SPN	__NULL__
Tesino	SPN	__NULL__
in	E	in
una	RS	indet
povera	AS	povero
famiglia	SS	famiglia
tirolese	AS	tirolese

Estrarre Informazioni



Vengono trovate le parole che sono più importanti nel testo.

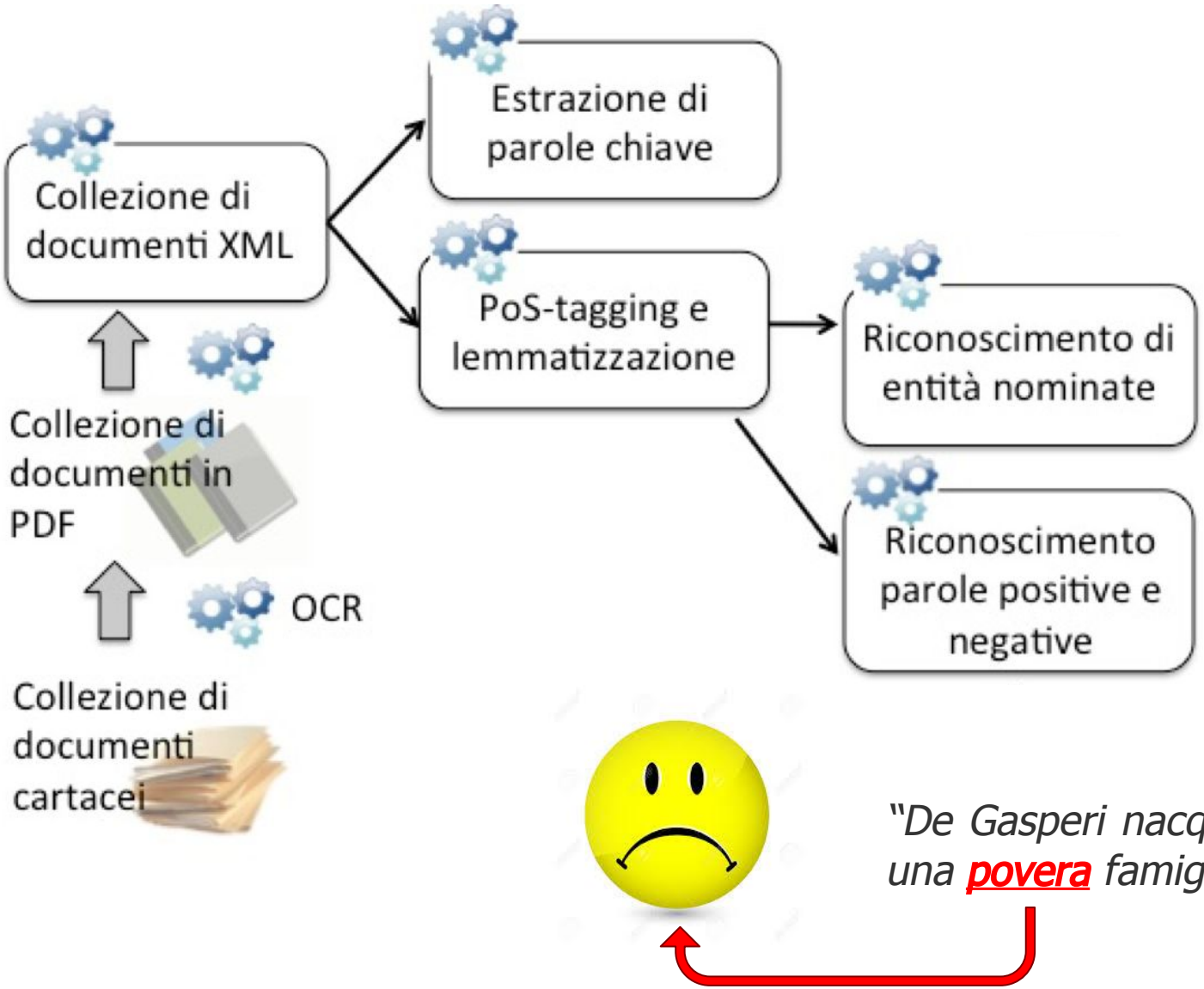
Estrarre Informazioni

Entità nominate: cose così importanti da avere un loro nome.

☑ Person ☐ Organization ☑ Location

De Gasperi nacque a Pieve Tesino in una povera famiglia tirolese : infatti i suoi genitori dovettero chiedere un sussidio allo Stato austriaco per farlo studiare .
Era il primo dei quattro figli di Maria Morandini , nata a Predazzo , e Amedeo Degasperi , nato a Sardagna .
Dopo di lui nacquero Mario , che fu sacerdote , Marcella e Augusto .

Estrarre Informazioni



*"De Gasperi nacque a Pieve Tesino in una **povera** famiglia tirolese"*

Estrarre Informazioni

ALCIDE

Analisi del linguaggio e del contenuto in un ambiente digitale

- Alcide è una piattaforma web.

Esplorare Documenti e Libri
(senza avere competenze informatiche)

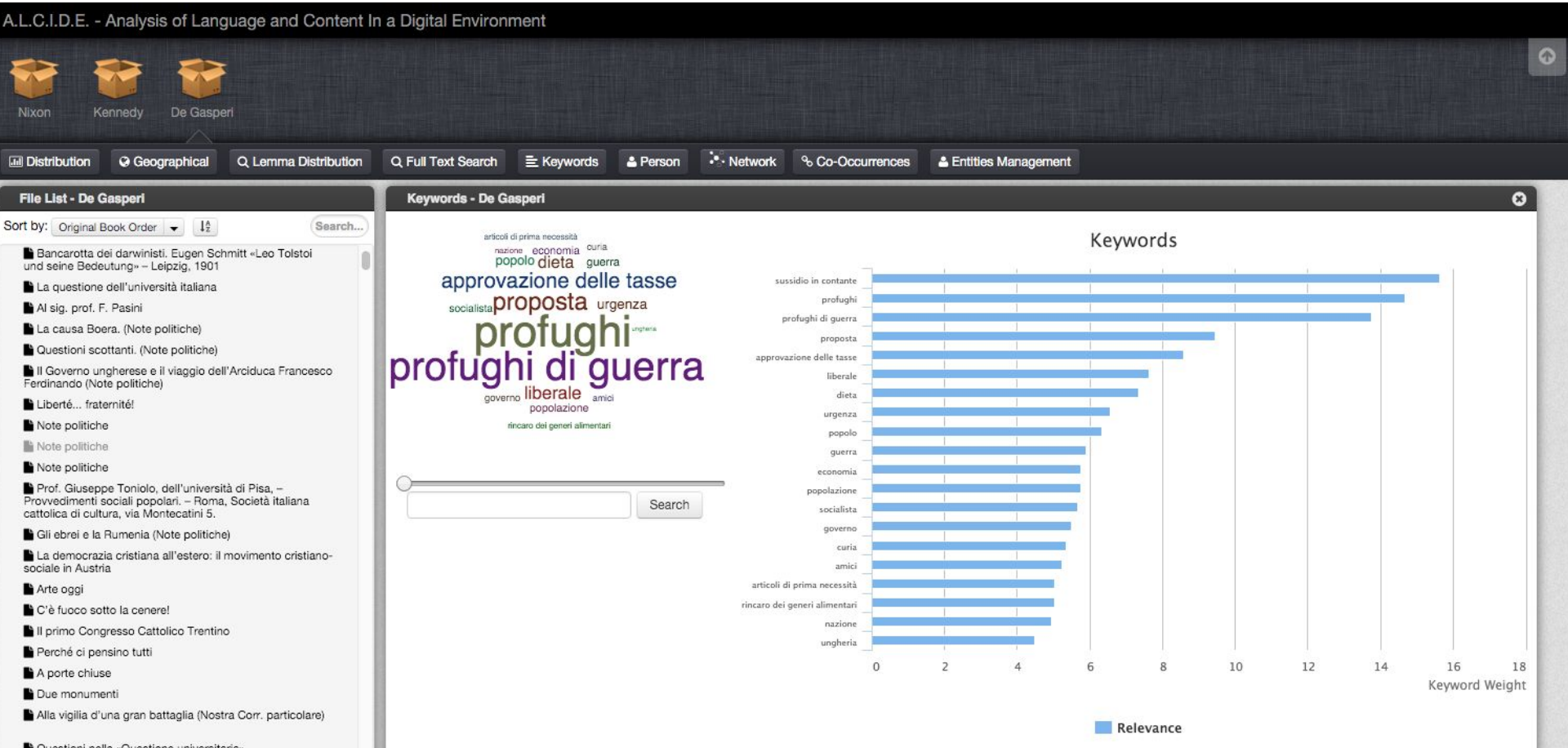
Raccoglie le informazioni estratte

Visualizza i Dati

Rappresentazioni intuitive e facili da capire

ALCIDE
Analisi del linguaggio e del contenuto in un ambiente digitale

http://dh.fbk.eu/projects/alcide-analysis-language-and-content-digital-environment



Leggere non serve più?

Serve ancora!

Lettura da vicino

Lettura da lontano

I computer a volte sbagliano…

… e non riescono a capire tutto.

Scriviamo cosa abbiamo imparato!

ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment

Giovanni Moretti, Sara Tonelli
Fondazione Bruno Kessler
Via Sommarive 18 Trento
moretti@fbk.eu
satonelli@fbk.eu

Stefano Menini, Rachele Sprugnoli
Fondazione Bruno Kessler and
University of Trento
menini@fbk.eu
sprugnoli@fbk.eu

Abstract

English. This work prese
(*Analysis of Language and*
Digital Environment), a new
Historical Content Analysis.
to improve Digital Humaniti
tegrating methodologies tak
man language technology a
understandable data structur
tion. ALCIDE provides a wi
of tools that go beyond sim
indexing, implementing fun
tual analysis such as named
nition, key-concept extractio
string-based search and geo-

Italiano. *Questo articolo*
CIDE (Analysis of Langua
tent In a Digital Enviro
nuova piattaforma per l'an
umenti storici. Il nostro
quello di migliorare la ricerc
dell' Informatica Umanistic
metodologie mutuate dalle t
linguaggio con la rappresen
itiva di strutture dati comple
offre una vasta gamma di
l'analisi testuale che vanno
plice indicizzazione dei met
empio, il riconoscimento di
di entità, estrazione di con
basata su lemmi e stringhe,

1 Introduction

In this paper we present ALC
Language and Content In a Digi
a new platform for Historical C
Our aim is to improve Digital
ies implementing both methodol

Figure 2: A simplified db graph of the structure

time span. To display data about the locations, the platform uses the Google GeoChart library[5].

4.2 Named Entity Recognition

The automatic extraction of person, location and organization names rely on the EntityPro (Pianta et al., 2008) module of TextPro. The module was originally trained on contemporary newspaper stories, on which it reached a performance of 92.12 F1 for Persons and 85.54 for GPEs. However, since the same tool obtained respectively 75.75 and 86.23 F1 on historical data (a set of Alcide De Gasperi's writings) a domain-specific adaptation was necessary. This was carried out by compiling black and white lists of common proper names for the period of interest and exploiting the tool in-built filtering functionality.

The data obtained is displayed together with the documents to highlight the most relevant persons in the text. It is also possible to query the system in order to obtain all the documents related to a specific entity or visualize in a graph the relevance of an entity over time.

4.3 Keyword Extraction

Keyword extraction is provided by the KX module embedded into the TextPro Suite. KX is a system for key-phrase extraction (both single and multi-word expressions) which exploits basic linguistic annotation combined with simple statistical measures to select a list of weighted keywords from a document (Pianta and Tonelli, 2010). KX was initially developed to work on news, patent documents and scientific articles. However, since ALCIDE is typically meant to deal with historical corpora, we tailored key-words extraction to the historians' requirement giving a higher rank to abstract concepts. This is done by boosting the relevance of concepts with a specific ending (e.g. '-ism', '-ty' in English and '-ismo', '-itidine' in Italian) usually expressing an abstract meaning. We also gave higher priority to generic key-concepts by boosting those expressed by single words.

Similarly to Named Entities, documents are displayed together with their most relevant keywords. Moreover, the portal allows the user to query the keywords characterizing a selected time span, the documents related to a specific keyword and the relevance of a keyword over the time.

4.4 Advanced Search Functions

One of the features we are interested in is to perform an efficient search of words or group of words in the whole collection of documents. The platform offers two main text search options. The first one is a full text search that gives the possibility to search for the match of one or more specific strings in a text. The second function performs a lemma based search, that looks for documents containing a specific verb, noun, or adjective in all its forms giving a lemma in input (e.g. searching for the verb *fight* the engine retrieves all the document containing *fight*, *fighting*, *fought*, etc).

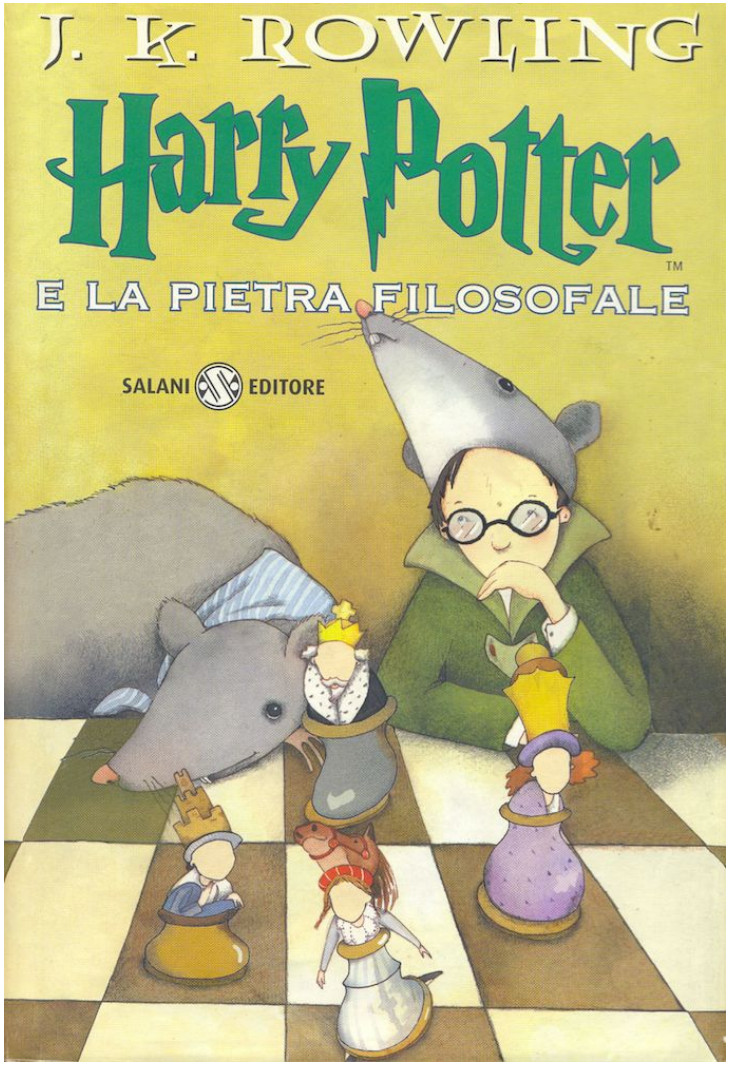
Both the search functions give the possibility to perform the query in documents issued in a specific time span and to display in a graph the trend of the target term usage over time.

5 Graphical Interface

The graphical interface was developed to represent all previously mentioned data in an intuitive visualization framework. The interface provides the

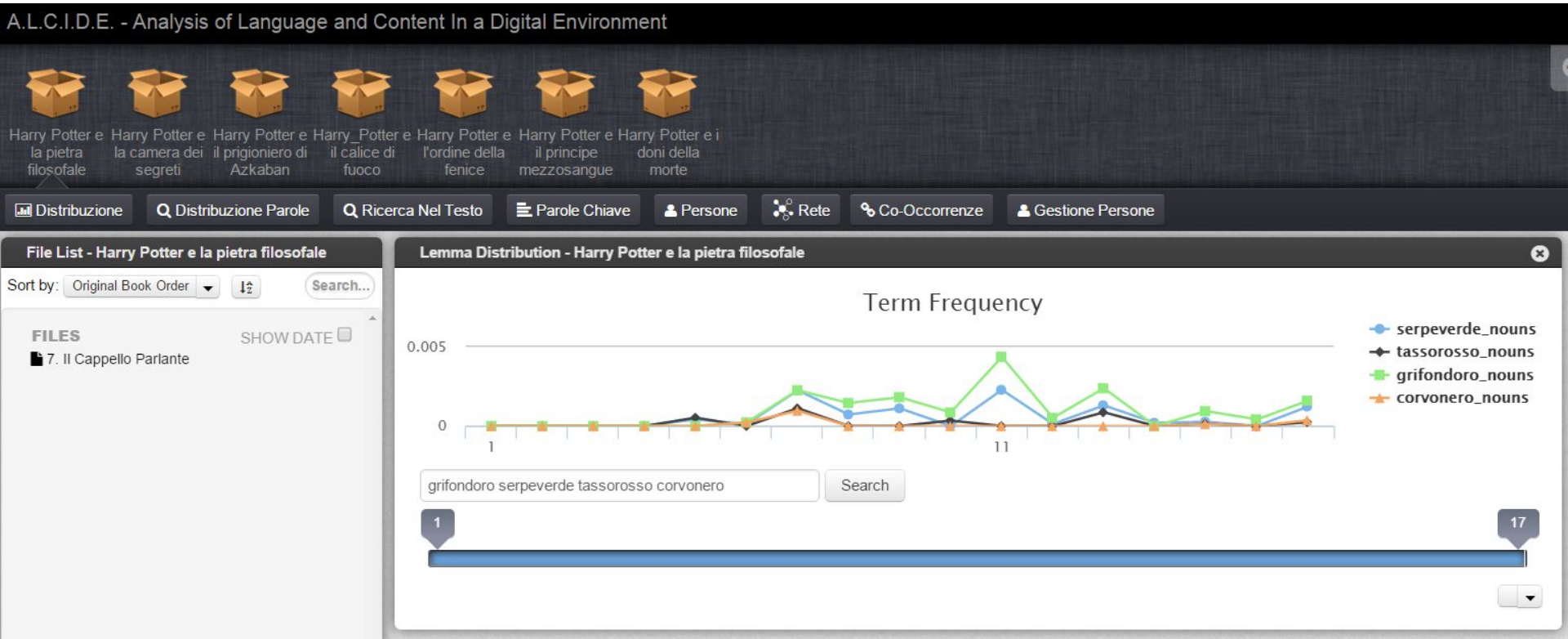
[5] https://developers.google.com/chart/interactive/docs/gallery/geochart

Vediamo come funziona (Demo)



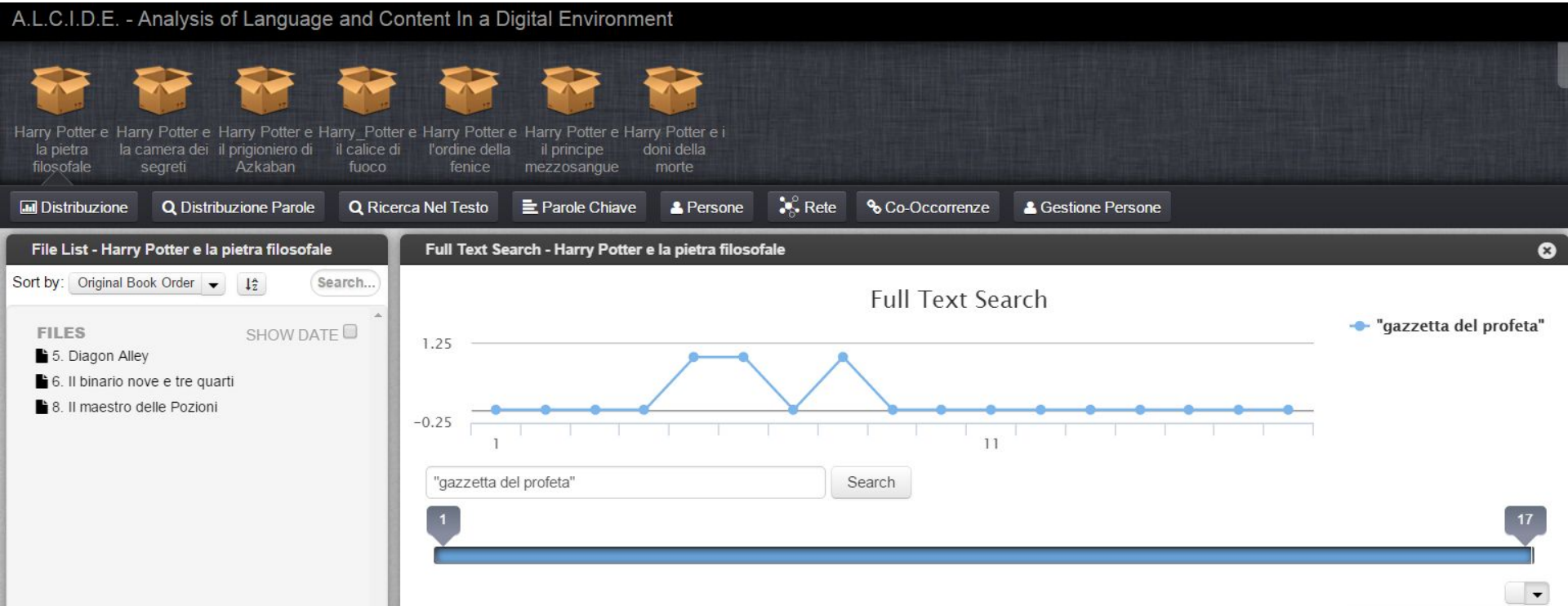
Quando appaiono insieme i nomi delle Case?

- Il primo picco è al capitolo 7: Il Cappello Parlante
- Il secondo picco è al capitolo 11: Il Quiddich
- Quando viene nominata "Grifondoro" viene di solito nominata anche "Serpeverde"



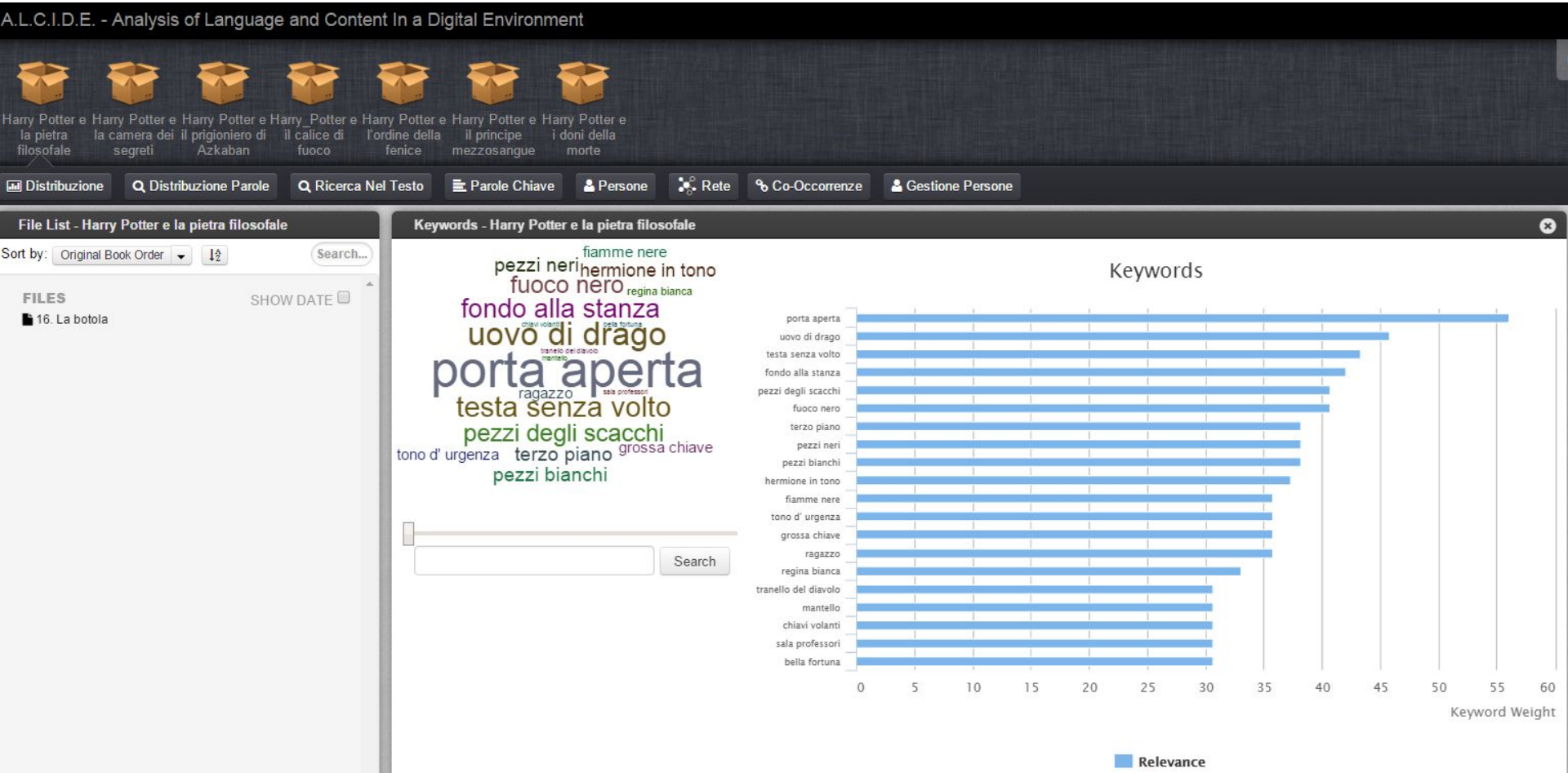
In che capitolo viene nominata per la prima volta la Gazzetta del Profeta?

- Nel capitolo 5: Diagon Alley



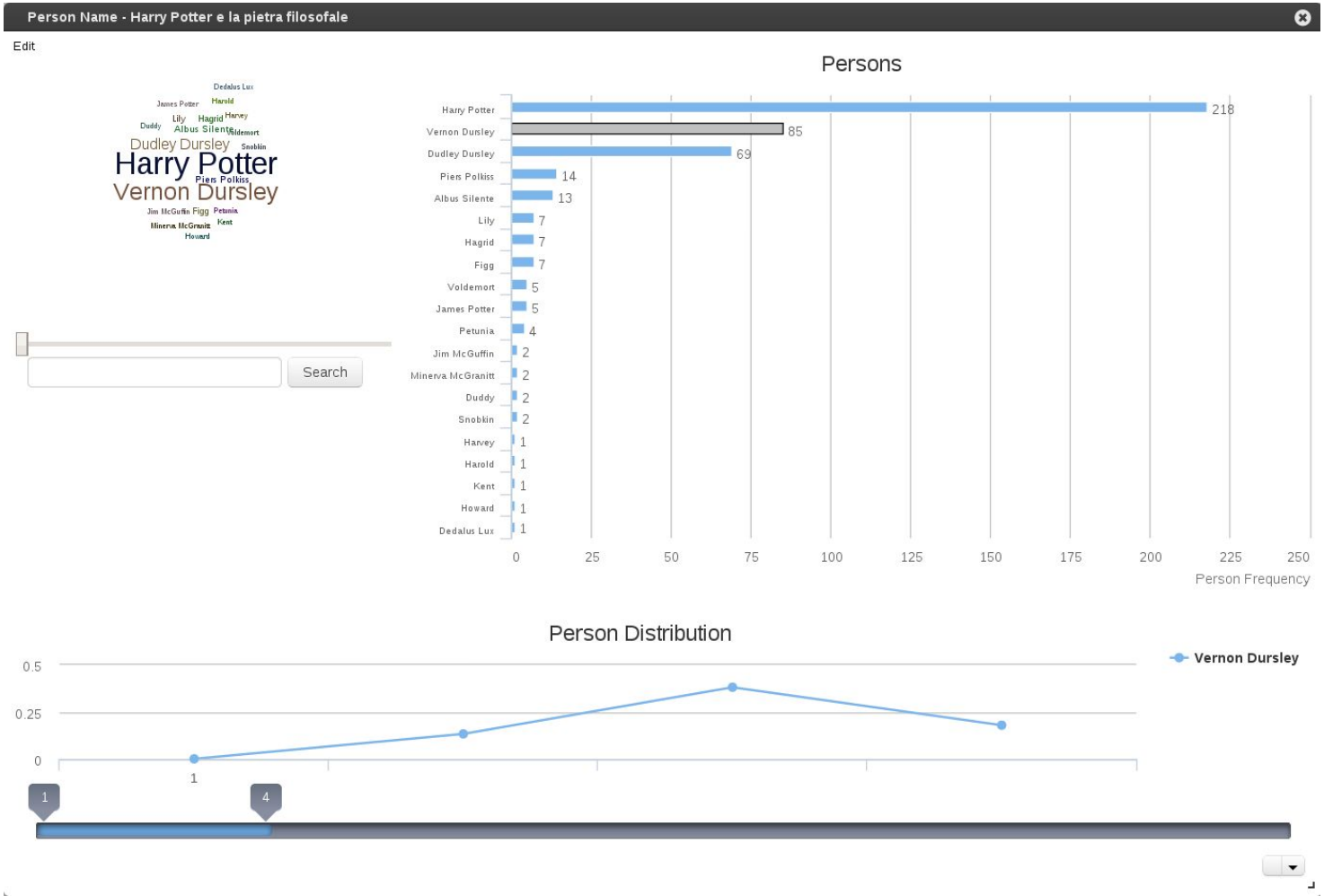
Quali sono le parole chiave del capitolo 16?

- Le parole chiave sono relative agli ostacoli che Harry, Ron ed Hermione devono superare come: la *porta aperta* della stanza di Fuffi, il *fuoco nero* dell'indovinello, la *grossa chiave* tra le chiavi alate ed i vari pezzi degli scacchi



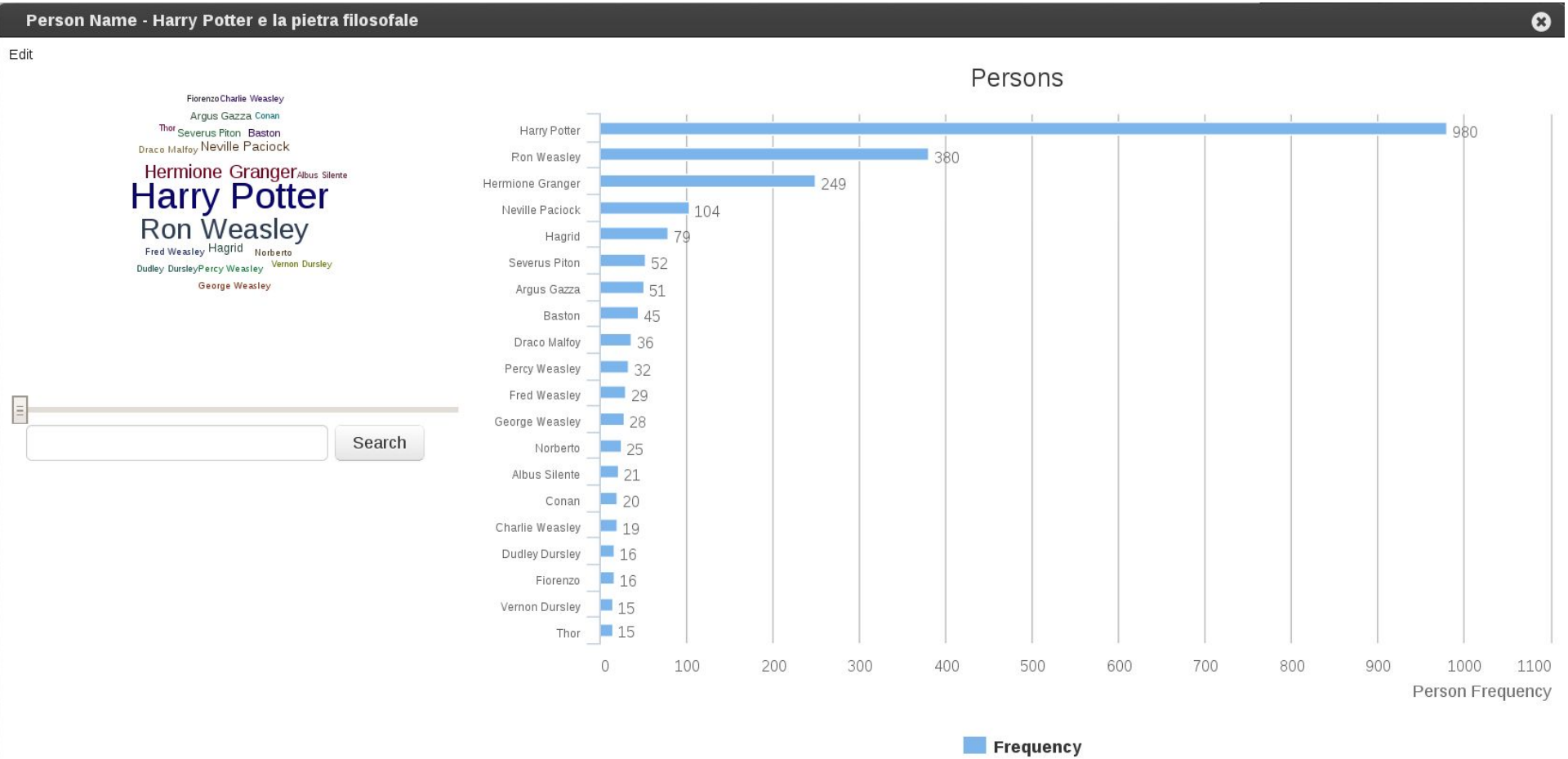
Come cambiano i personaggi nel corso del libro?

In basso si possono selezionare i capitoli usando la barra di scorrimento e si può vedere la distribuzione delle menzioni.
Nei primi capitoli ci sono i membri della famiglia e gli amici di Dudley.



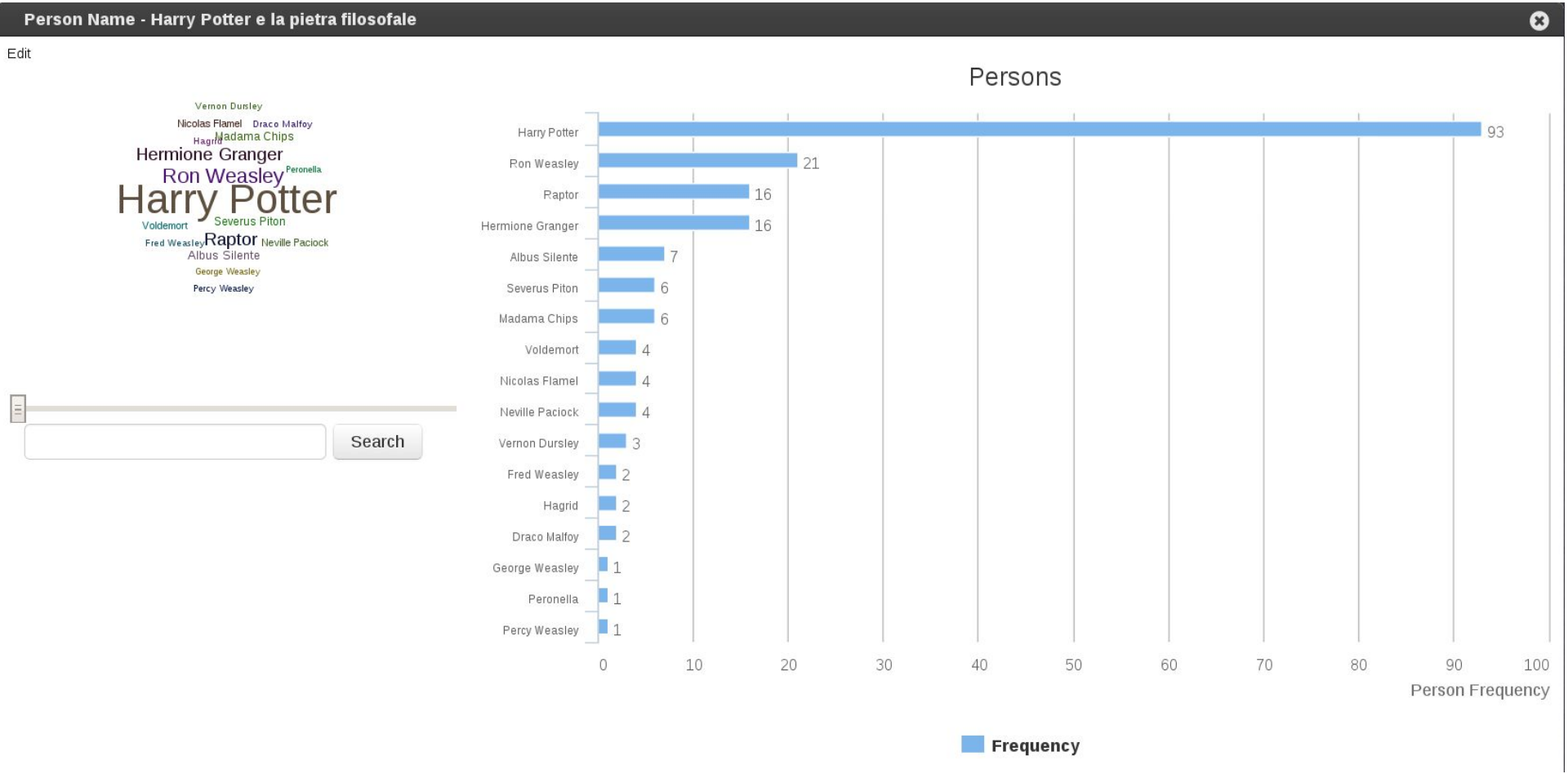
Come cambiano i personaggi nel corso del libro?

Nei capitoli centrali ci sono i personaggi collegati al mondo della magia, gli amici di Hogwarts sono i primi



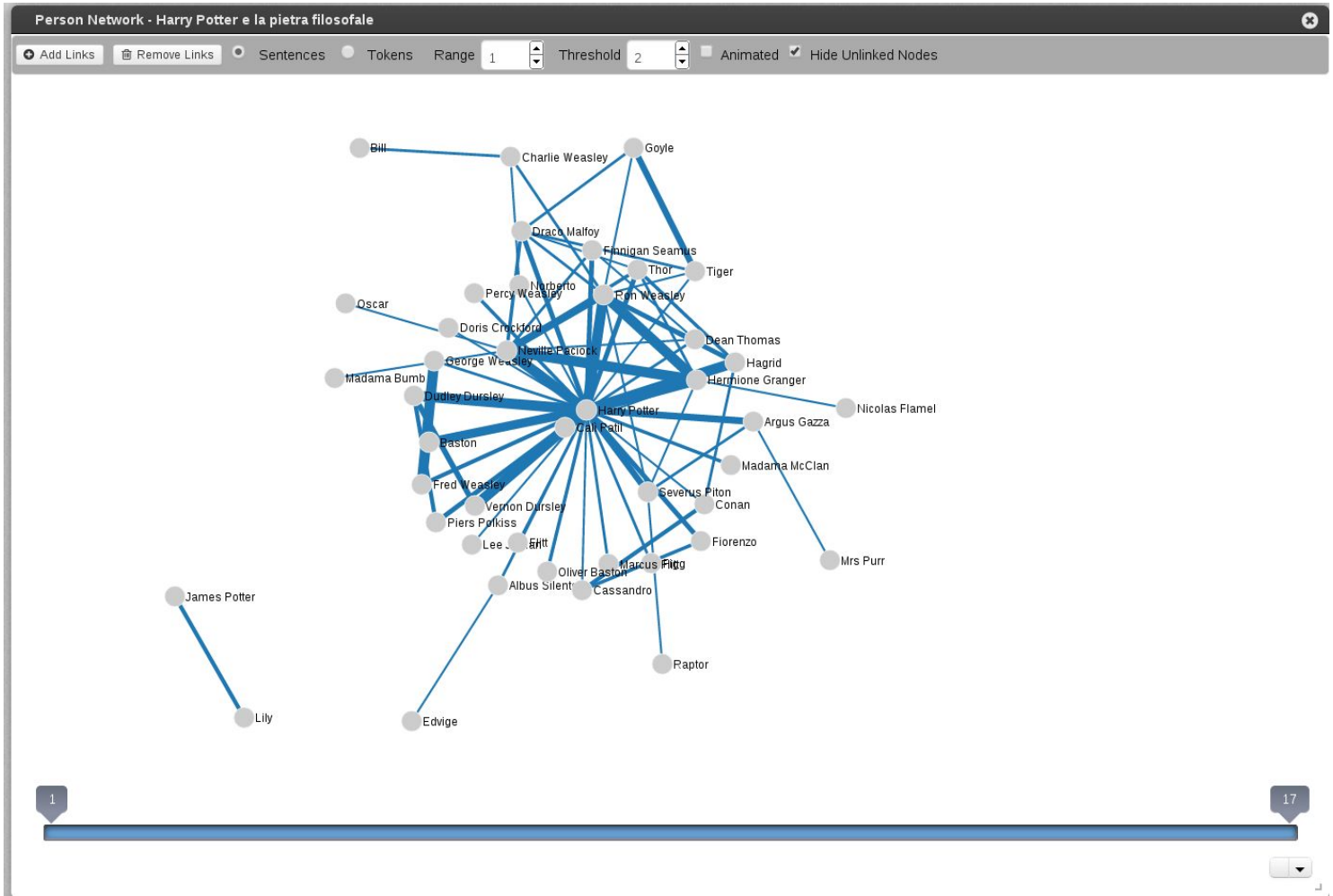
Come cambiano i personaggi nel corso del libro?

Nel capitolo finale il cattivo, Raptor, ha una frequenza alta mai avuta nei capitoli precedenti



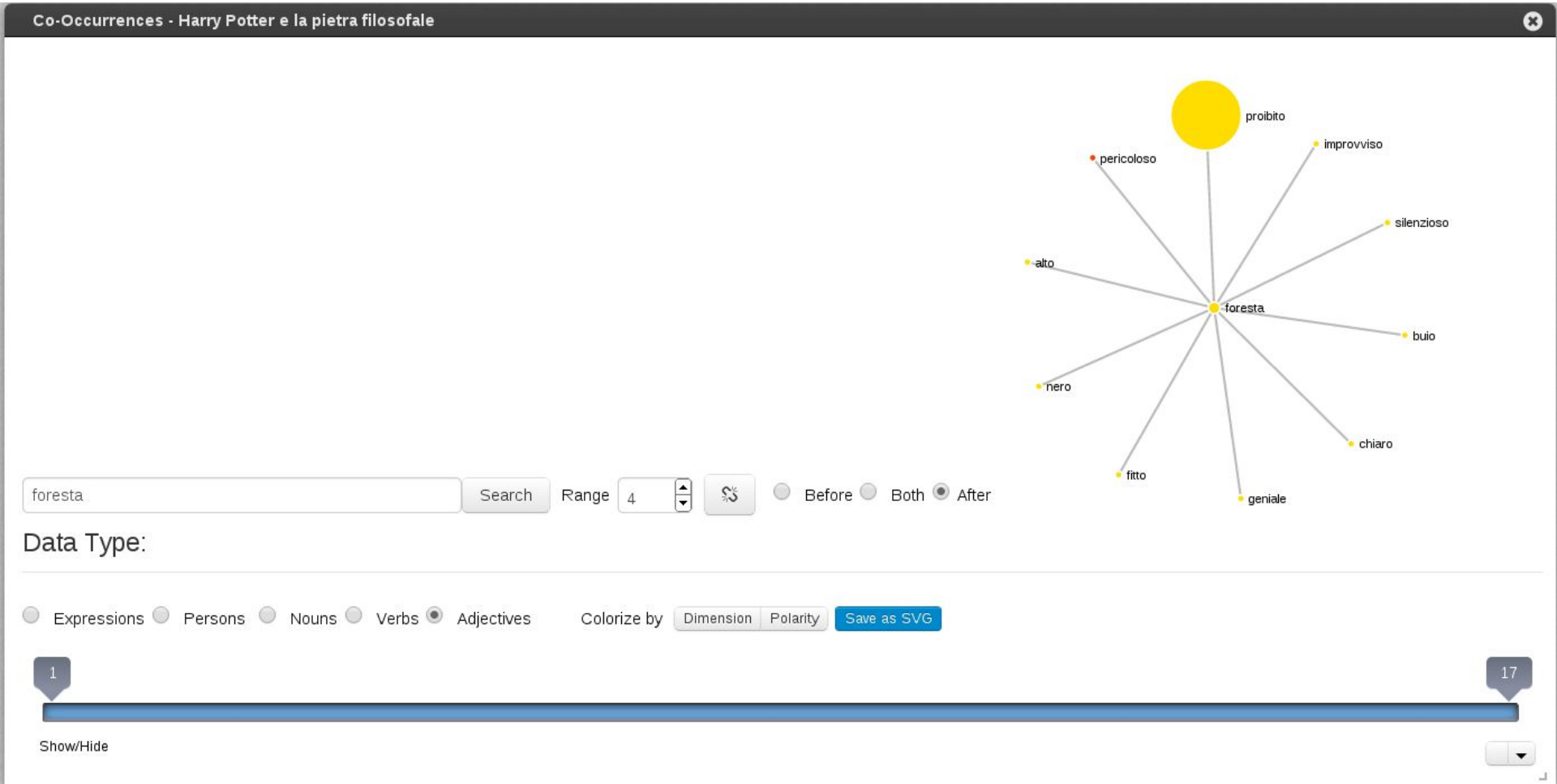
Come interagiscono tra di loro i personaggi?

Harry è al centro della rete di interazioni. Più è spesso il segmento che collega i nodi e più è frequente l'interazione tra i personaggi collegati. Goyle, Tiger e Draco formano un triangolo. James e Lily Potter appaiono insieme.



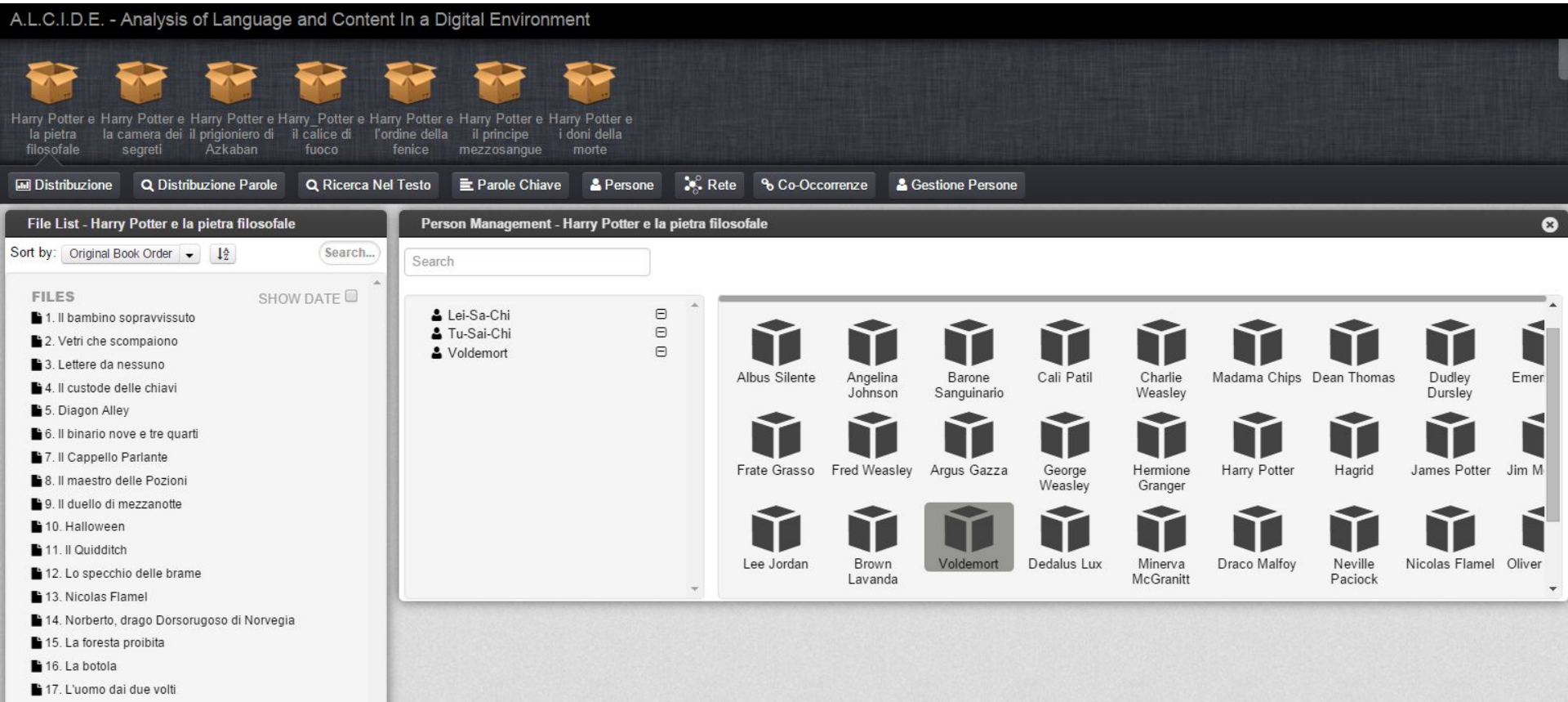
Quali sono gli aggettivi che descrivono la foresta?

Co-occorrenze dei soli aggettivi che appaiono entro 4 parole dopo il lemma “foresta”: in rosso gli aggettivi che esprimono un sentimento negativo.

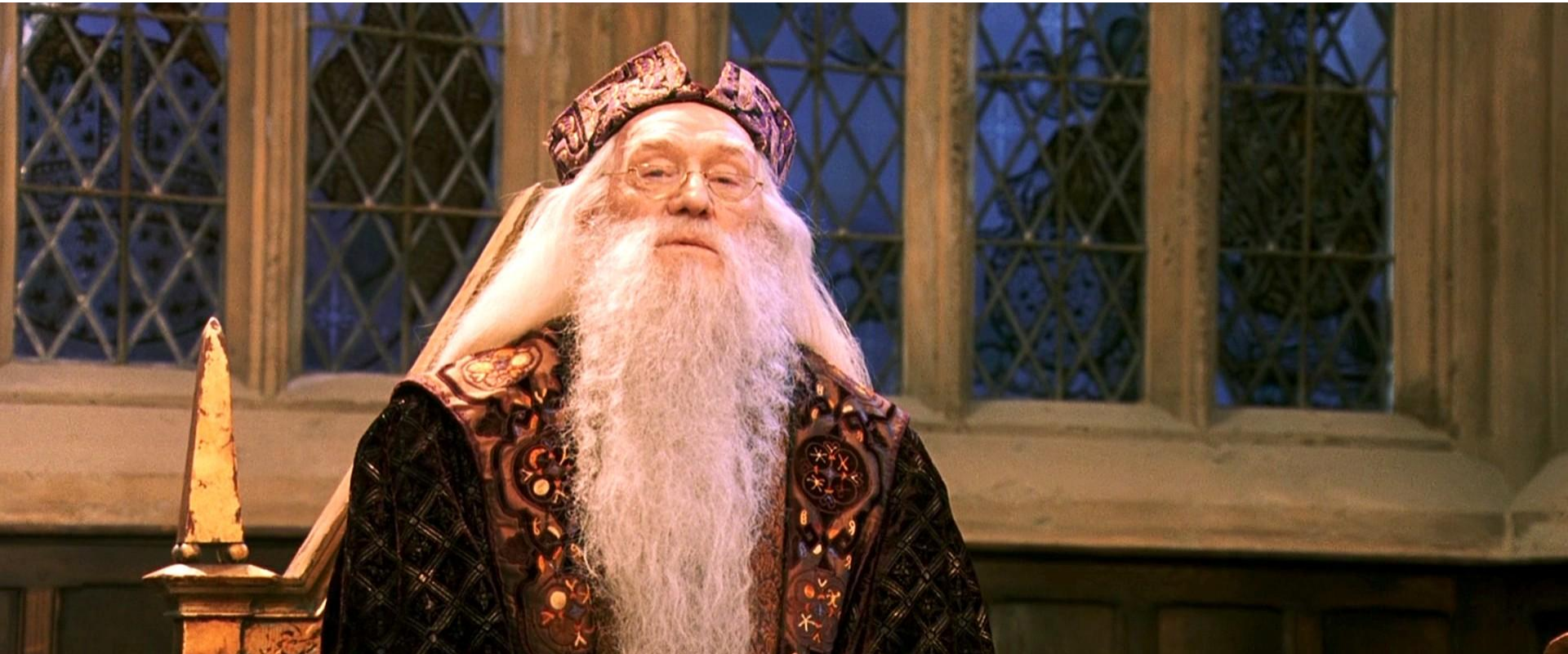


Quali sono i nomi con i quali è chiamato Voldemort?

Voldemort viene nominato usando 3 espressioni; l'espressione "Signore Oscuro" non appare mai nel primo libro della saga



Tocca a voi!



Tocca a voi!

RISPONDETE A QUESTE DOMANDE USANDO LA PIATTAFORMA

- In quale libro e in quale capitolo appare per la prima volta l' espressione "Signore Oscuro" ?
- Quali sono le parole chiave che descrivono la seconda prova in "Harry Potter e il Calice di Fuoco" ?
- In quale dei 7 libri Hermione è più nominata di Ron?
- In quali libri Voldemort ha un ruolo più attivo?
- Quale personaggio "sussurra" di più nel settimo libro? E chi "borbotta"?
- In quali parti di "Harry Potter e il Calice di Fuoco" appare più spesso Voldemort?

Tocca a voi!

- Nel quinto libro si parla di “vita” e di “morte”, com’è il loro andamento?
- In che modo vi aspettate che cambi la presenza di Voldemort nell’ultimo libro?
- Quali sono i personaggi nuovi che compaiono nel secondo libro?
- Come vengono descritti i mezzosangue in “Harry Potter e i Doni Della Morte”?
- Nell’ultimo libro cosa si deve fare con gli Horcrux?
- Quali tipi di incantesimi sono usati in “Harry Potter e l’Ordine della Fenice”?