



Metodi e tecnologie per l'analisi automatica di documenti del passato: il progetto A.L.C.I.D.E.

*Sara Tonelli
Fondazione Bruno Kessler, Trento
27 Novembre 2013*



Tecnologie dell'Informazione

Circa 20 unità di ricerca incluso il progetto esplorativo in
Digital Humanities (DH)

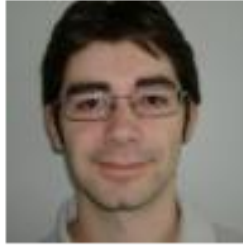
Il gruppo di Digital Humanities



**ALESSANDRO
MARCHETTI**

E-mail:
amarchetti@fbk.eu
PHONE:
0461314974

[VIEW PROFILE](#)



STEFANO MENINI

E-mail:
menini@celct.it
PHONE:
0461314942

[VIEW PROFILE](#)



**GIOVANNI
MORETTI**

E-mail:
moretti@fbk.eu
PHONE:
0461314807

[VIEW PROFILE](#)



**RACHELE
SPRUGNOLI**

E-mail:
sprugnoli@fbk.eu
PHONE:
0461314879

[VIEW PROFILE](#)



SARA TONELLI

E-mail:
satonelli@fbk.eu
PHONE:
0461314542

[VIEW PROFILE](#)



PAOLO POMBENI

[VIEW PROFILE](#)



MAURIZIO CAU

[VIEW PROFILE](#)

The Manifesto of the Digital Humanities (2011)

[...] For us, the digital humanities concern the totality of the **social sciences and humanities**. The digital humanities are not *tabula rasa*. On the contrary, they rely on all the paradigms, *savoir-faire* and knowledge specific to these disciplines, while mobilizing the tools and unique perspectives enabled by **digital technology**.

The digital humanities designate a “**transdiscipline**”, embodying all the methods, systems and heuristic perspectives linked to the digital within the fields of humanities and the social sciences.[...]

We call for the creation of scalable digital infrastructures responding to **real needs**. These digital infrastructures will be **built** iteratively, based upon methods and approaches that prove successful in research communities. [...]

Storia digitale

- Necessità di raccogliere, preservare e presentare il passato via web
- La digitalizzazione ci permette di sfruttare nell'ambito della ricerca storica caratteristiche come “*capacity, accessibility, flexibility, diversity, manipulability, interactivity and hypertextuality*” (Cohen and Rosenzweig “A Guide to gathering, preserving and presenting the past on the Web”, 2005).
- Alcuni potenziali problemi sono legati invece al mezzo digitale e non a quello cartaceo: “*quality, durability, readability, passivity and inaccessibility*” (Cohen and Rosenzweig, 2005).

Siti di storia digitale

- Siti di storia digitale hanno per lo più funzioni educative, offrono materiale da utilizzare in classe in sostituzione o come integrazione dei libri di testo: <http://www.digitalhistory.uh.edu/>

Digital History
using new technologies to enhance teaching and research

pre-1492 1490 1760 1770 1780 1790 1800

The First Americans Colonial Era American Revolution

TIMELINE

MEDIA
DOCUMENTS
TEXTBOOK
TEACHING

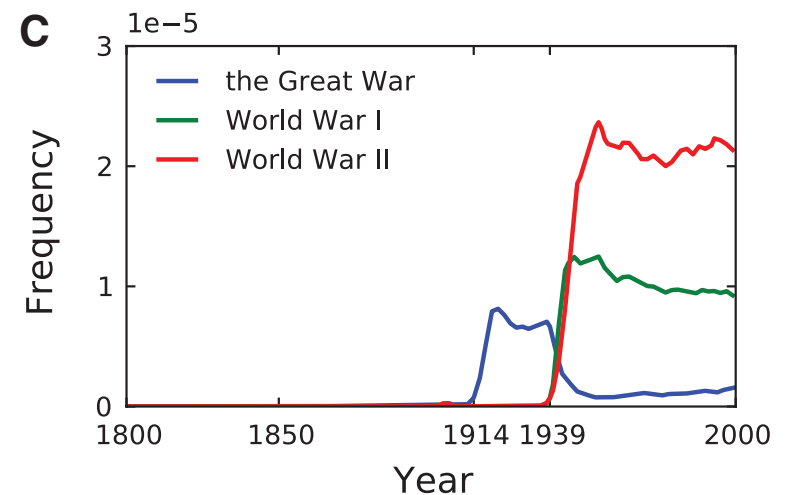
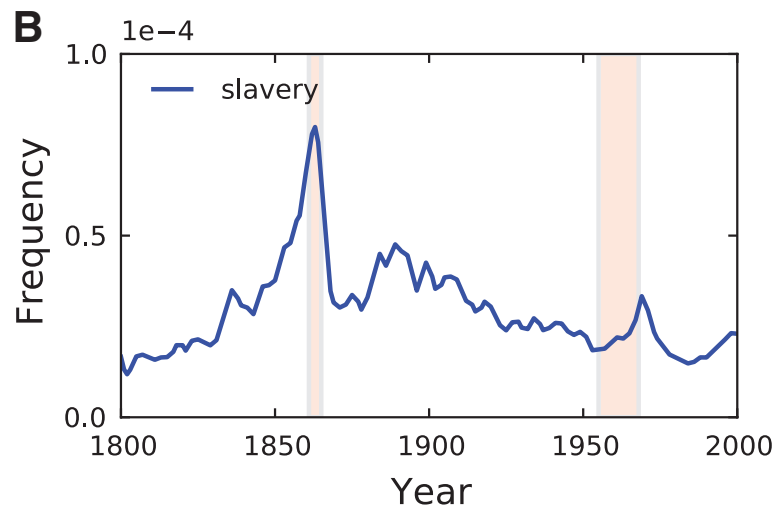
DRAG ME ROLL OVER AND CLICK ON AN AREA ABOVE FOR MORE INFORMATION

ERAS	TOPICS	RESOURCES	REFERENCES
HISTORY ERAS • The First Americans • Colonial Era • American Revolution • Early National Period • Pre-Civil War Era	DO HISTORY TOPICS • Advertisements • Architecture • Art • Beauty & Fashion • Cemeteries	PRIMARY SOURCES • Landmark Documents • Court Cases • Supreme Court Cases • Newspaper Articles • Obituaries	GENERAL REFERENCE • Timelines • Reference Room • Glossaries • History Profession • Image Sites

Analisi quantitativa di dati digitalizzati

Quantitative Analysis of Culture Using Millions of Digitized Books, Jean-Baptiste Michel et al., Science 331, 176 (2011)

- Analisi di una collezione di documenti contenente circa il 4% di tutti i libri pubblicati (GoogleBooks).
- *Culturomics*: studio di fenomeni linguistici e culturali veicolati dalla lingua inglese tra il 1800 e il 2000



La letteratura vista da lontano, Franco Moretti, 2005,
Einaudi editore.

- Sostituire alla lettura ravvicinata del testo letterario (*close reading*) una lettura “da lontano”, costruendo grafi, carte e alberi partendo da un’analisi quantitativa dei testi letterari
- Processo di deliberata riduzione e astrazione, allontanamento rispetto alla concretezza del testo
- Lettura a distanza come modalità di analisi che riduce i dettagli ma fa emergere forme, regolarità e relazioni tra i testi

Progetto ALCIDE

Analysis of **L**anguage and **C**ontent In a **D**igital **E**nvironment:

- Sviluppo di una piattaforma web per l'analisi temporale, geografica, linguistica e semantica di documenti storici.
- Integrare tecnologie stato dell'arte per il trattamento automatico del linguaggio
- Fornire a studiosi in ambito storico (in particolare storia moderna e contemporanea) uno strumento che risponda realmente alle esigenze della ricerca storica
- Possibilità di analizzare, visualizzare ed esplorare grandi collezioni di documenti in tempo reale

Progetto ALCIDE: Requisiti dell'utente

- Possibilità di esplorare il *corpus* selezionando il periodo temporale di riferimento
- Il *periodo temporale* non deve essere predefinito
- Al termine di ogni ricerca deve essere possibile visualizzare il contenuto dei documenti trovati: evitare visualizzazioni statiche
- Possibilità di *annotare* e *esportare* i risultati della ricerca
- Creare un'infrastruttura non legata a un particolare corpus o periodo, renderla più flessibile possibile

Progetto ALCIDE: Requisiti tecnici

- Possibilità di aggiungere facilmente nuovi corpora, utilizzo di formati standard (XML) per consentire l'uso a utenti senza competenze tecniche
- Efficienza per consentire ricerche in tempo reale, velocità
- Eliminare lavoro manuale, architettura modulare automatizzata
- Consultabile tramite browser, evitare applicazioni da installare e plugin, utilizzare HTML5
- Evitare visualizzazioni troppo complesse

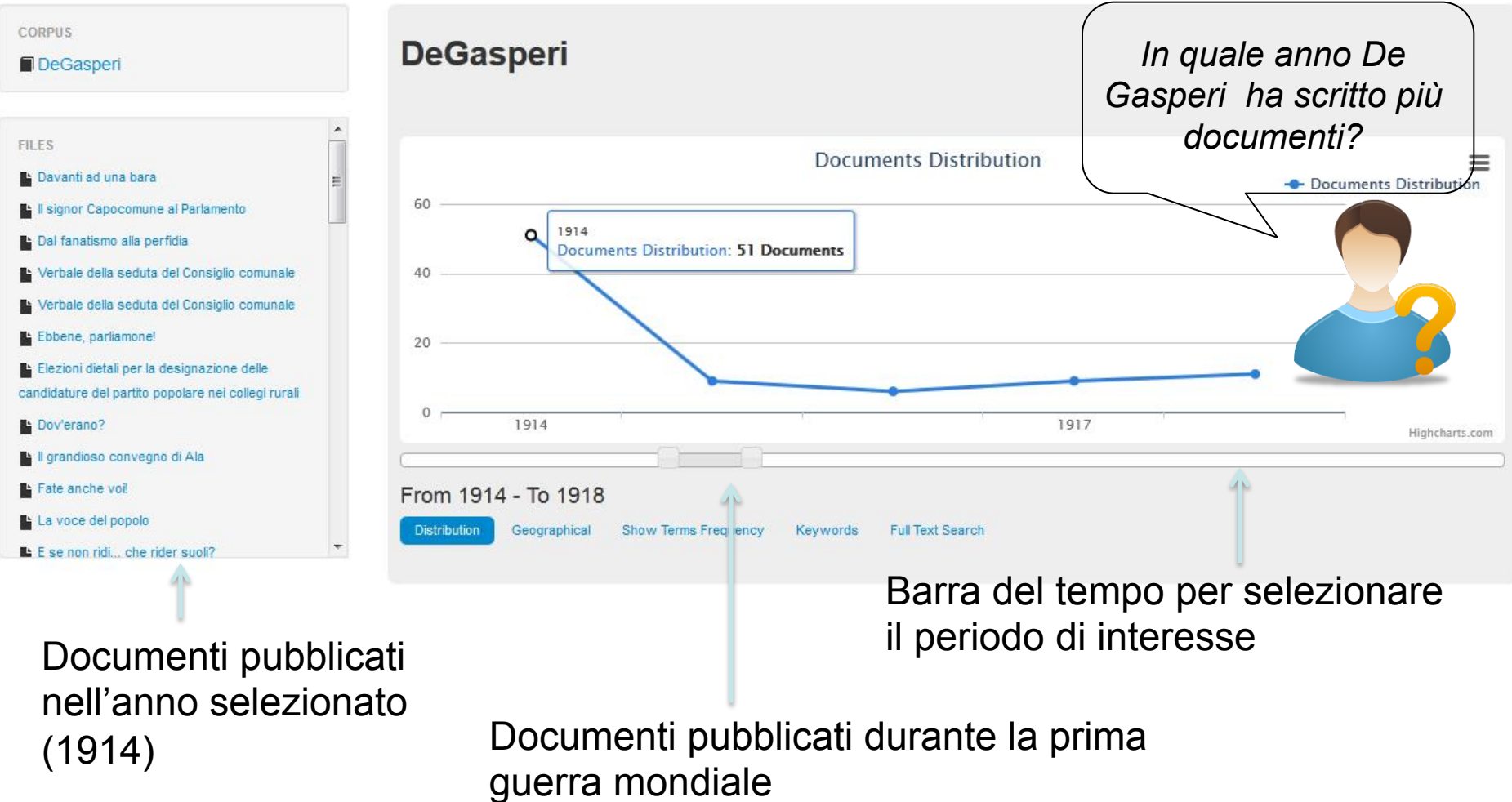
Progetto ALCIDE: Caso d'uso

- Collezione completa degli scritti pubblici di Alcide De Gasperi (1881 – 1954)
- Circa 3.000 documenti contenenti 3.000.000 di token (~70.000 lemmi)
- Comprende trascrizioni di discorsi, articoli di giornale, atti parlamentari, telegrammi, ecc.
- Versione cartacea raccolta in 7 tomi pubblicata da ISIG



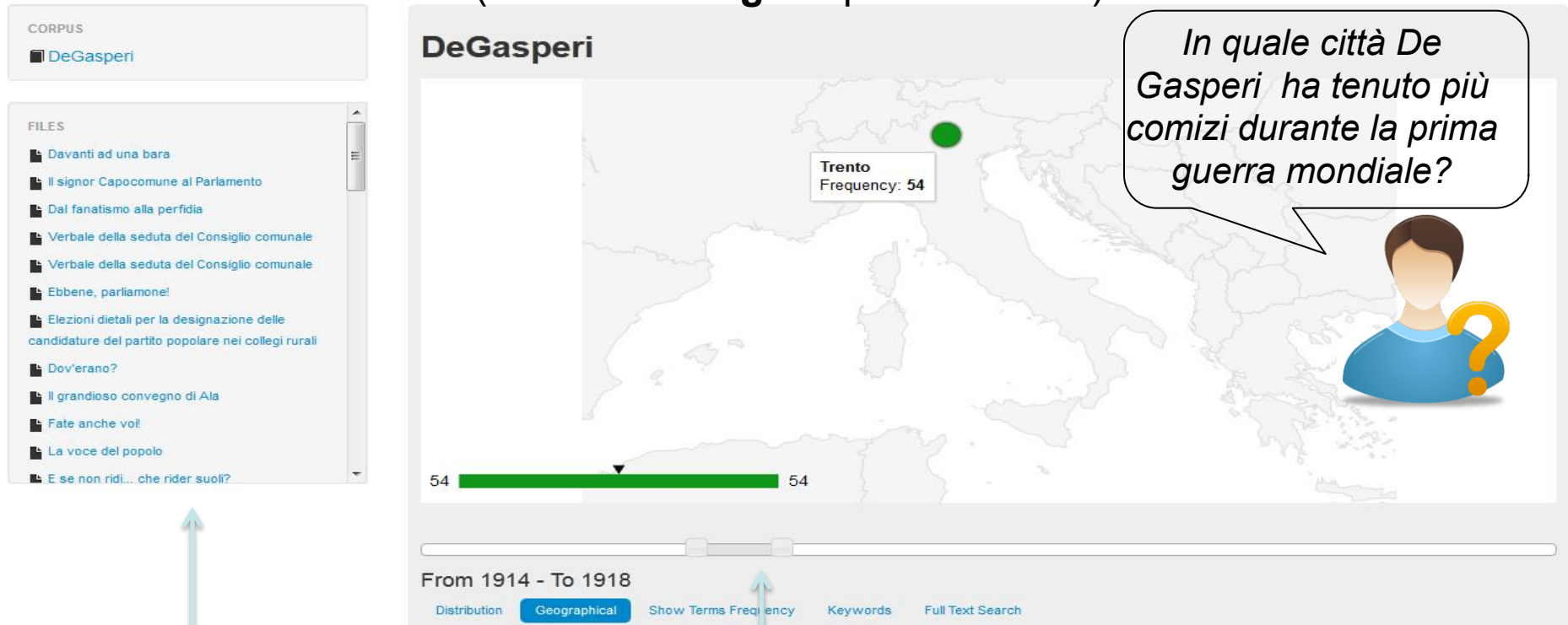
ALCIDE: Visualizzazione “da lontano”

1. Visualizzazione della distribuzione temporale dei documenti nel corpus (metadato: **anno** di pubblicazione)



ALCIDE: Visualizzazione “da lontano”

2. Visualizzazione della distribuzione geografica dei documenti nel corpus (metadato: **luogo** di pubblicazione)



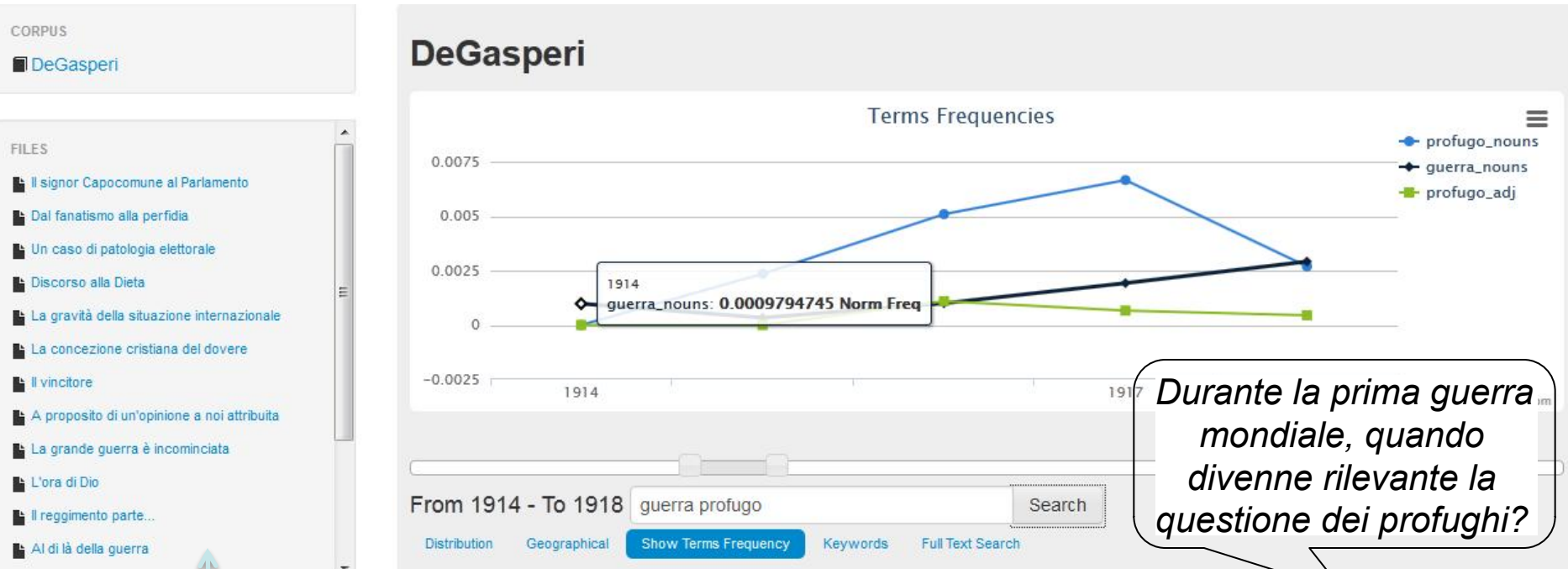
Lista dei documenti

Documenti di De Gasperi pubblicati a Trento durante la prima guerra mondiale

ALCIDE: Visualizzazione “da lontano”

3. Ricerca di termini con distribuzione temporale

I documenti sono precedentemente analizzati con un PoS-tagger e lemmatizzatore



Documenti pubblicati nel 1914
contenenti il termine “guerra”

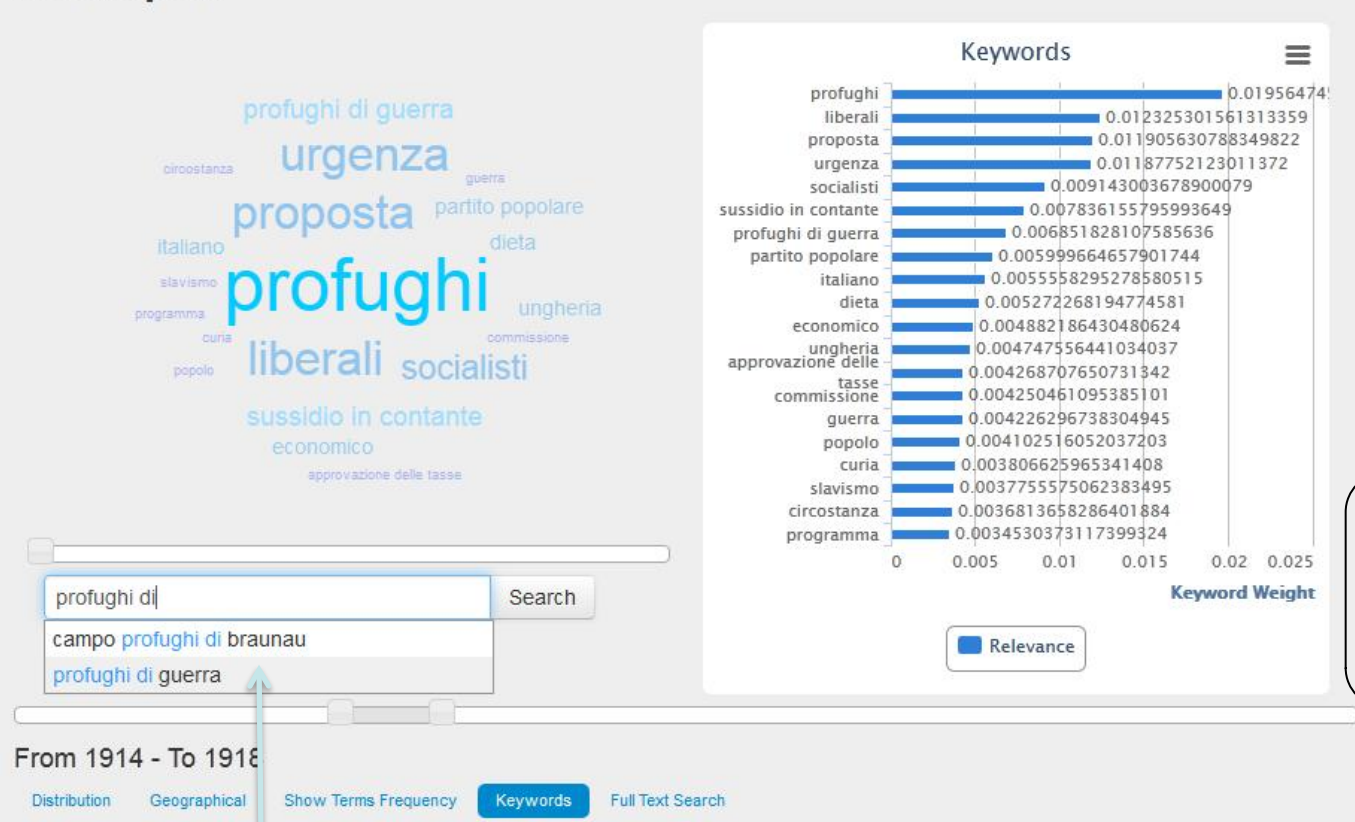
Frequenza **normalizzata**



ALCIDE: Visualizzazione “da lontano”

4. Ricerca e visualizzazione di parole-chiave

DeGasperi



Parole-chiave estratte automaticamente dai documenti pubblicati nel periodo di interesse

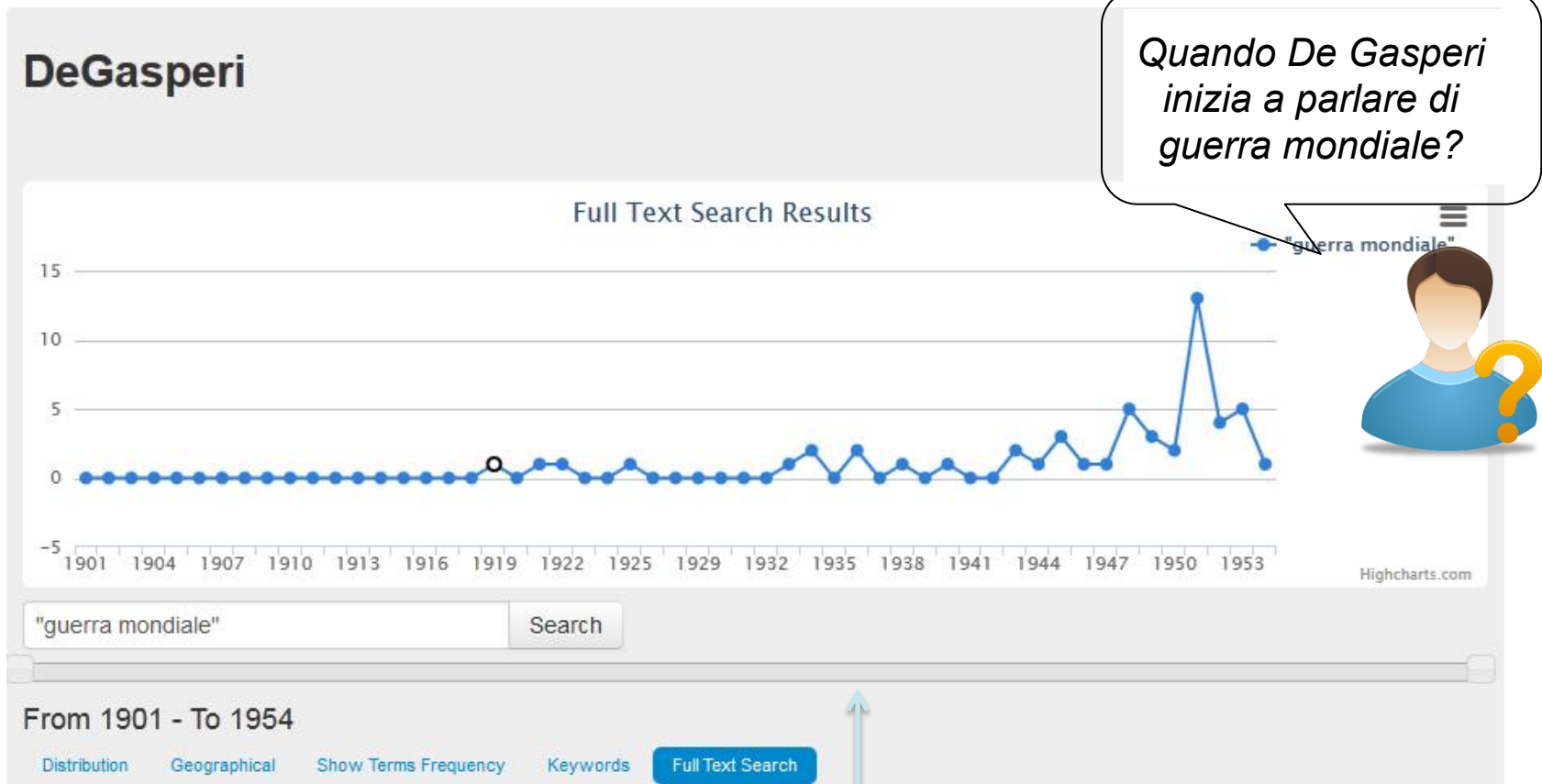
Quali erano le tematiche prevalenti nel discorso di De Gasperi tra il 1914 e il 1918?



Suggerimento di parole chiave nella maschera di ricerca

ALCIDE: Visualizzazione “da lontano”

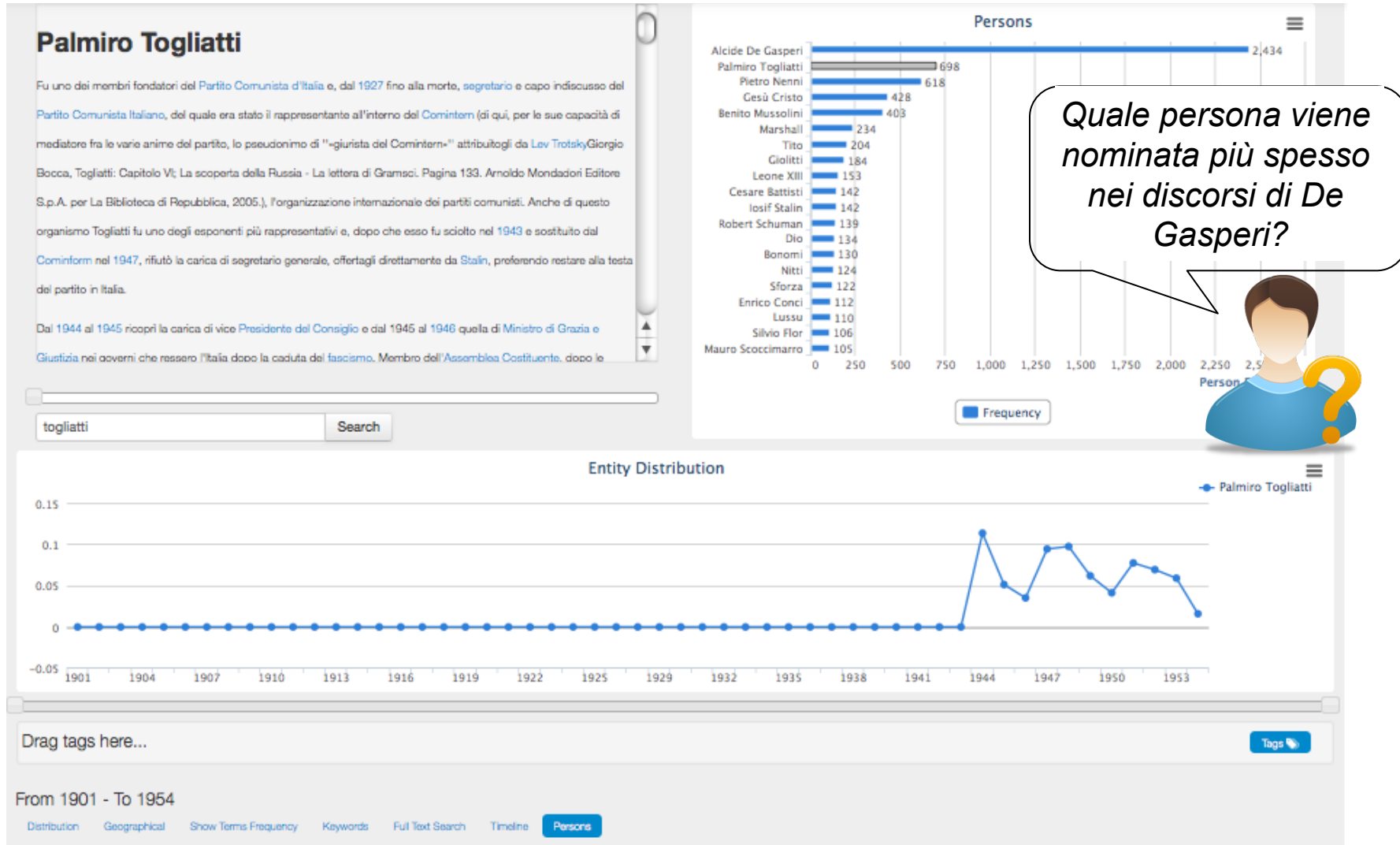
5. Ricerca di parole nel testo (forme flesse) senza normalizzazione



Barra temporale per selezionare il periodo d'interesse

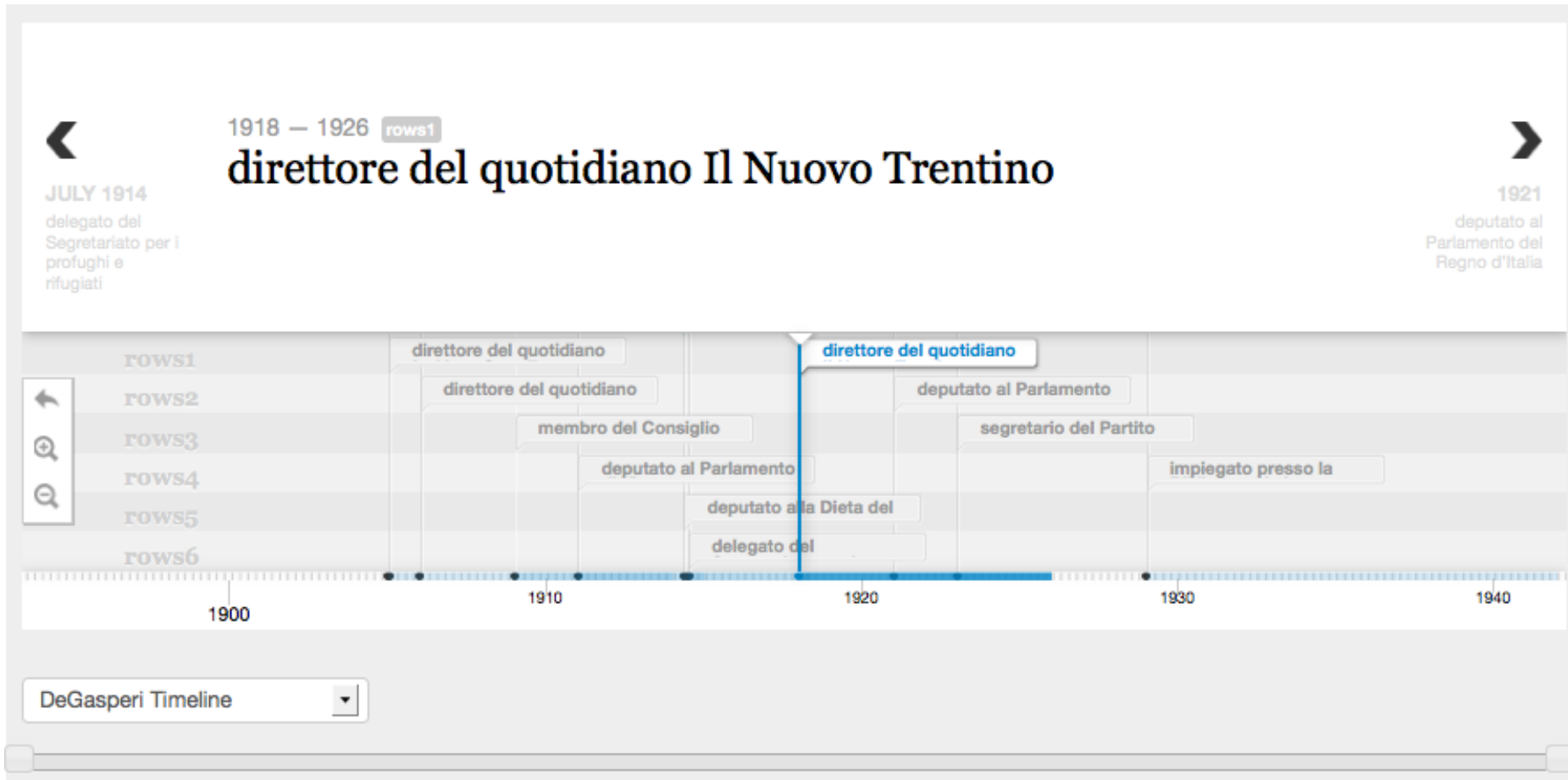
ALCIDE: Visualizzazione “da lontano”

6. Ricerca e visualizzazione di persone nominate nei documenti



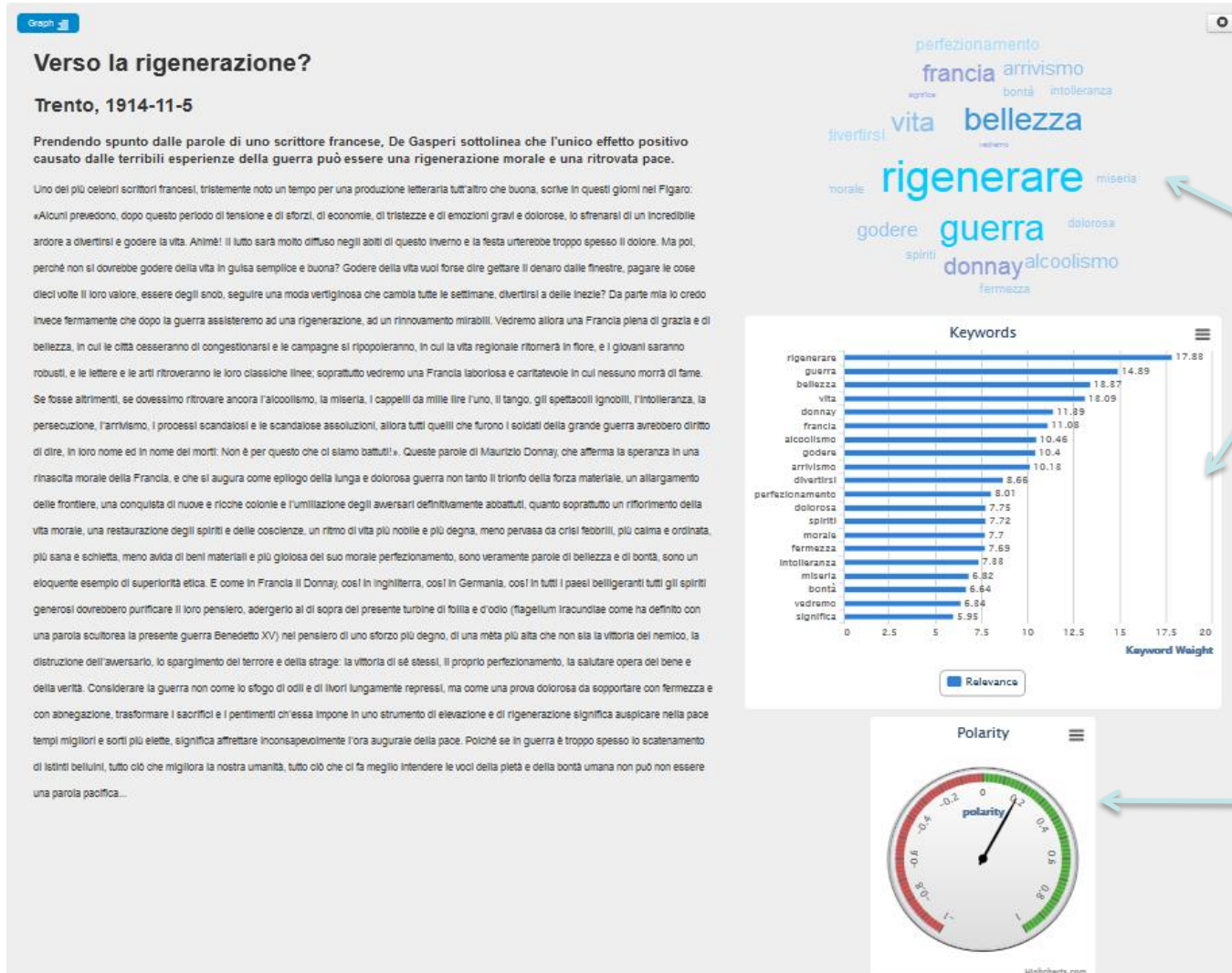
ALCIDE: Visualizzazione “da lontano”

8. Visualizzazione di eventi su linee del tempo predefinite o personalizzate



ALCIDE: Visualizzazione “da vicino”

9. Visualizzazione del contenuto del documento



Visualizzazione
delle parole
chiave

Polarità generica
(positiva o
negativa) del
documento

ALCIDE: Visualizzazione “da vicino”

Graph

L'Unione Europea vuole la pace

Roma, 1949-02-16

Si sta facendo il tentativo di unire i paesi d'Europa in una rappresentanza comune di governi e Parlamentari. Il «Comitato dei Ministri» dovrebbe accordarsi su provvedimenti d'intesa e di collaborazione e il «Consiglio dei Delegati dei Parlamentari» sarebbe chiamato a dare parere sulle proposte del Comitato. Per intanto si comincia cogli Stati che sono più affini tanto politicamente che civilmente, poi si vedrà. A costituire il nucleo iniziale sarà chiamata anche l'Italia e presto ci sarà una riunione costitutiva che dovrà elaborare lo Statuto di questa società di popoli. Pare che si preveda un Comitato di sette ministri degli Esteri e un Consiglio di un centinaio di delegati circa. In Italia, governo e opinione pubblica hanno accolto con soddisfazione l'invito. Finalmente usciamo dalla situazione umiliante del dopoguerra e rientriamo, come tutti gli altri, nella famiglia europea nella quale potremo far valere le nostre tendenze conciliative e universalistiche che mirano al consolidamento della pace e alla sicurezza delle nazioni minori: e potremo dire una parola ragionevole sulle questioni del dopoguerra, rimaste ancora insolute, quali ad esempio la questione germanica e la questione delle colonie. Il Trattato di pace prevedeva che in riconoscimento del contributo dato dall'Italia alla guerra di liberazione, avremmo potuto partecipare al consesso mondiale dell'ONU: ma la Russia ce l'ha impedito. Ecco ora l'Unione Europea come un parziale surrogato, perché badate, essa non vuol essere che uno di quegli organismi supplementari e integrativi che sono previsti dalla stessa Carta di San Francisco. Non era logico supporre che «tutta la nazione», senza differenza di partito, si sarebbe pronunciata favorevole a questo primo concreto tentativo di unità o di federazione europea, tanto invocata da scrittori, politici e pensatori di tutte le fedi, lungo tutto un periodo di cento anni? Invece comunisti e paracomunisti mobilitano di nuovo le folle ignare in nome della pace, ed entrano in lizza le donne dell'UDI e i giovani del Fronte, e di nuovo in Parlamento e sulle piazze si griderà contro il tradimento del governo e della borghesia capitalista che vuole la guerra. Borghesi, capitalisti e imperialisti? Ma gli uomini coi quali noi assoceremo i nostri sforzi di collaborazione europea sono laburisti, socialisti o cristiani democratici, rappresentanti di tendenze di pace, venuti su dai sindacati operai e dalle classi del popolo. E questi uomini governano gli Stati che durante le passate guerre furono tutti vittime di aggressione e che fanno e devono fare tutti gli sforzi per evitare un altro disastro che si rovescerebbe sui loro territori annientandoli. A chi la volete dare ad intendere che...

Persons Keywords Polarity Tags

popoli pace italia
unione europea
passate guerre
guerra economica
parlamento

Text Classification

Polarity

Positivo

Graph

La marcia di...Sfruz

Trento, 1923-03-30

Quando Mussolini o i suoi generali richiamano alla disciplina e condannano con gradita precisione le violenze degli eroi della sesta giornata, nella stampa fascista e nelle riunioni di partito si risponde con un atto di sottomissione: qui e là però gli avvenimenti ed il linguaggio che riprende poco dopo ci dicono che la sottomissione è formale, ma non nello spirito. Tuttavia che ciò accada nelle province per le quali la marcia di Roma fu o doveva essere la fine di tutto un lungo periodo di conflitti sanguinosi, è ancora spiegabile, perché l'uso ripetuto e prolungato della violenza produce una perturbazione delle coscienze così profonda che gli stessi esaltatori d'un metodo di eccezione da loro ritenuto necessario ma transitorio, né d'anzi prevedevano né oggi sanno dominare. Ma tale spiegazione e il compatimento che ne deriva non si confanno a regioni che del fascismo hanno ignorata la epoca cruenta. Nella Venezia Tridentina, se si eccettuano le due imprese di Bolzano e di Trento, le quali ebbero stile e sviluppo militare e carattere antistatale, ma non furono rivolte contro l'incolumità delle persone, o di un gruppo di persone, nessun episodio grave si ebbe a deplorare, né per difesa né per offesa. Le risse domenicali non insanguinarono né i centri né i villaggi della provincia, cosicché la tradizione d'ordine, di quiete tolleranza, di rispetto alla legge, non subì interruzioni notevoli. Lo spirito di violenza non è quindi nel Trentino né endemico né epidemico. Se tuttavia qui e là fa capolino anche fra noi, vuol dire che vi viene importato di proposito e forzando il naturale sviluppo delle cose. Vestigia terrenti e, per conto nostro, ci preoccupiamo che anche le faville più leggere non suscitino qualche incendio. Onde ci chiediamo se sia lecito che un giornale locale che fa professione di fascismo pubblici minacce e intimidazioni che (in una corrispondenza da Tione) polemizzando con un o.d.g. di una assemblea di capifamiglia stampi bravate consimili: «Illustri capi di famiglia, fate il piacere di tacere anche e soprattutto perché oltre al Partito fascista che vostro malgrado oggi governa energicamente e sagacemente potrebbe svegliarsi quel santo manganello che vi farebbe ingoiare, ecc... Come? Oltre il Governo, del quale vantate l'energia e la sagacia, invocate ancora il manganello? Oltre la forza della legge minacciate la violenza del

Persons Keywords Polarity Tags

giustizia marcia
mussolini legge
olio di ricino
violenza fascismo
manganello corrispondenza
criterio soggettivo

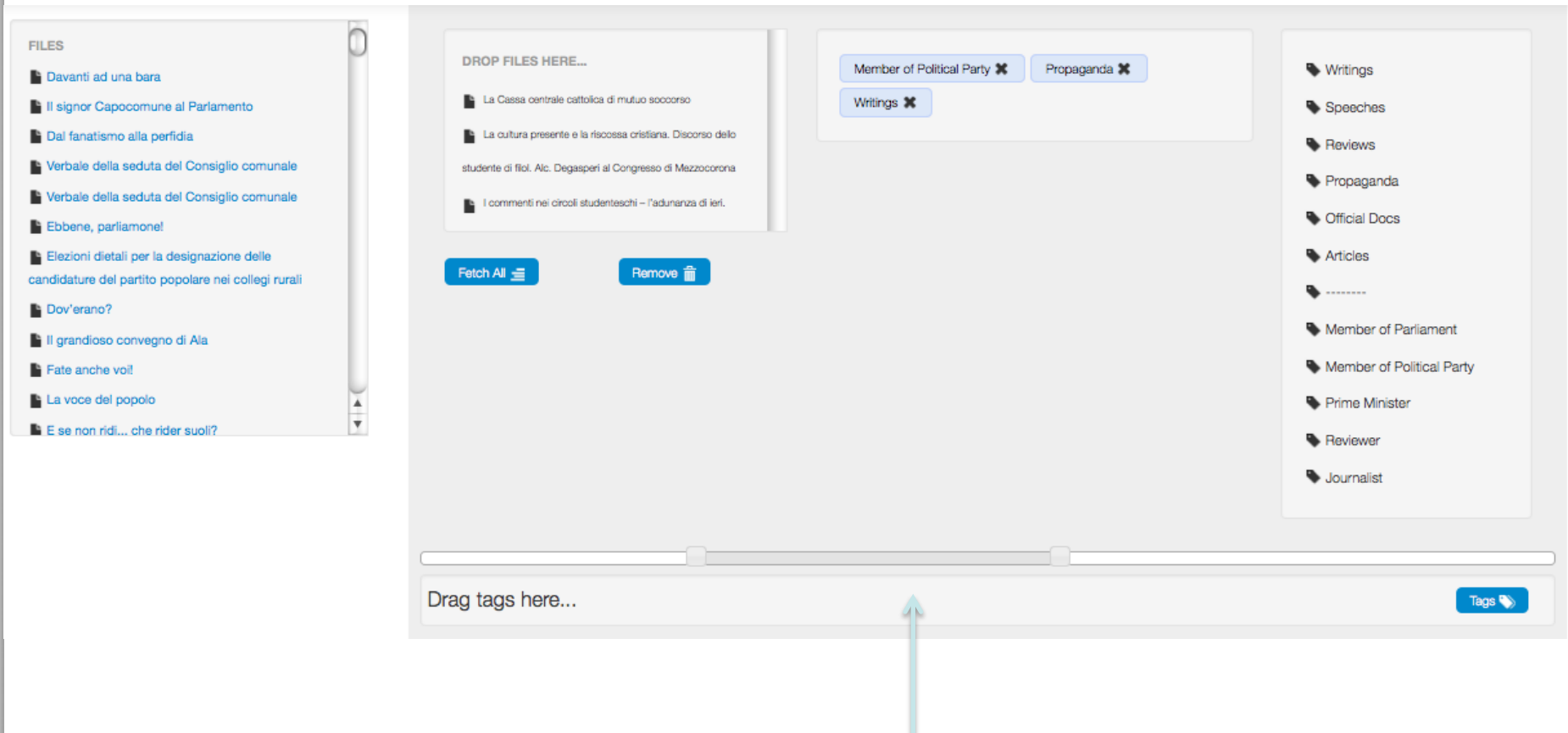
Text Classification

Polarity

Negativo

ALCIDE: Annotazione di documenti

10. Assegnazione di etichette personalizzate a documenti singoli o raggruppati per periodo

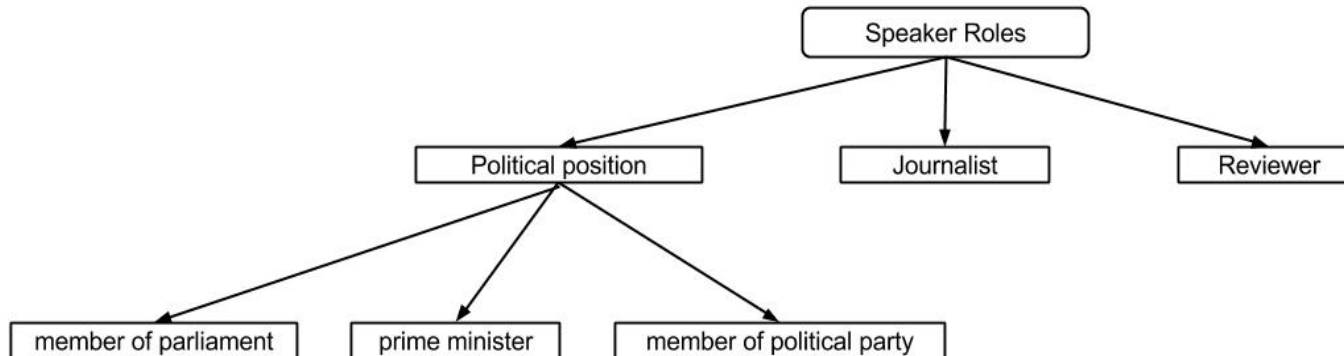
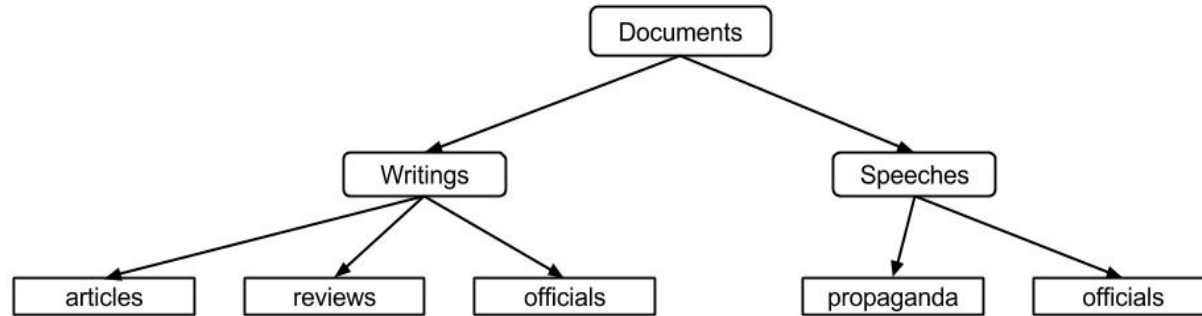


The interface displays a list of files on the left, a central area for dropping files and fetching/removing them, and a right sidebar for assigning tags. The tags are categorized into various roles and topics. A timeline slider at the bottom allows for selecting a group of documents for tagging.

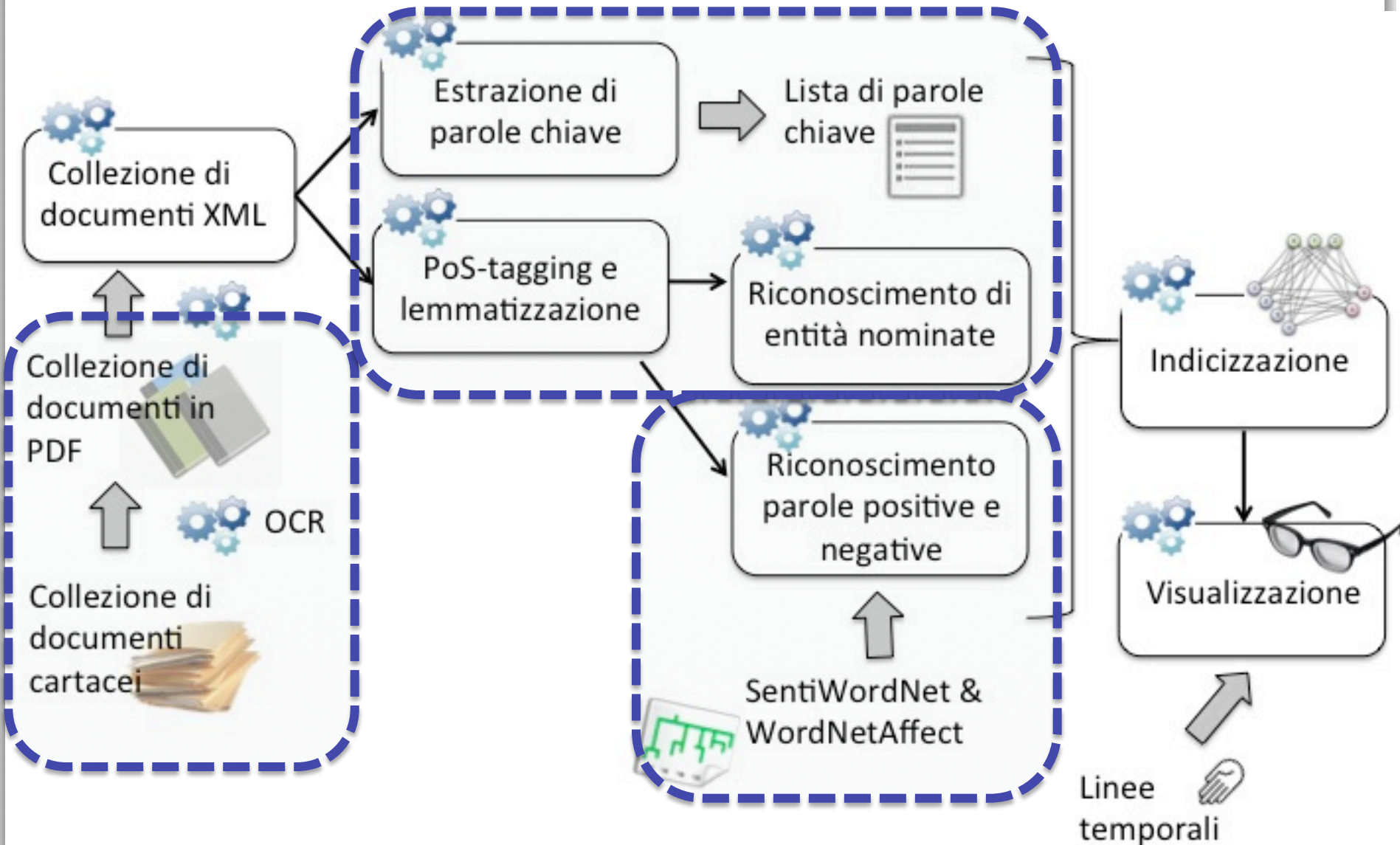
Barra temporale per selezionare un gruppo di documenti

ALCIDE: Annotazione di documenti

Intervento degli esperti
di dominio



Dal documento cartaceo alla piattaforma web

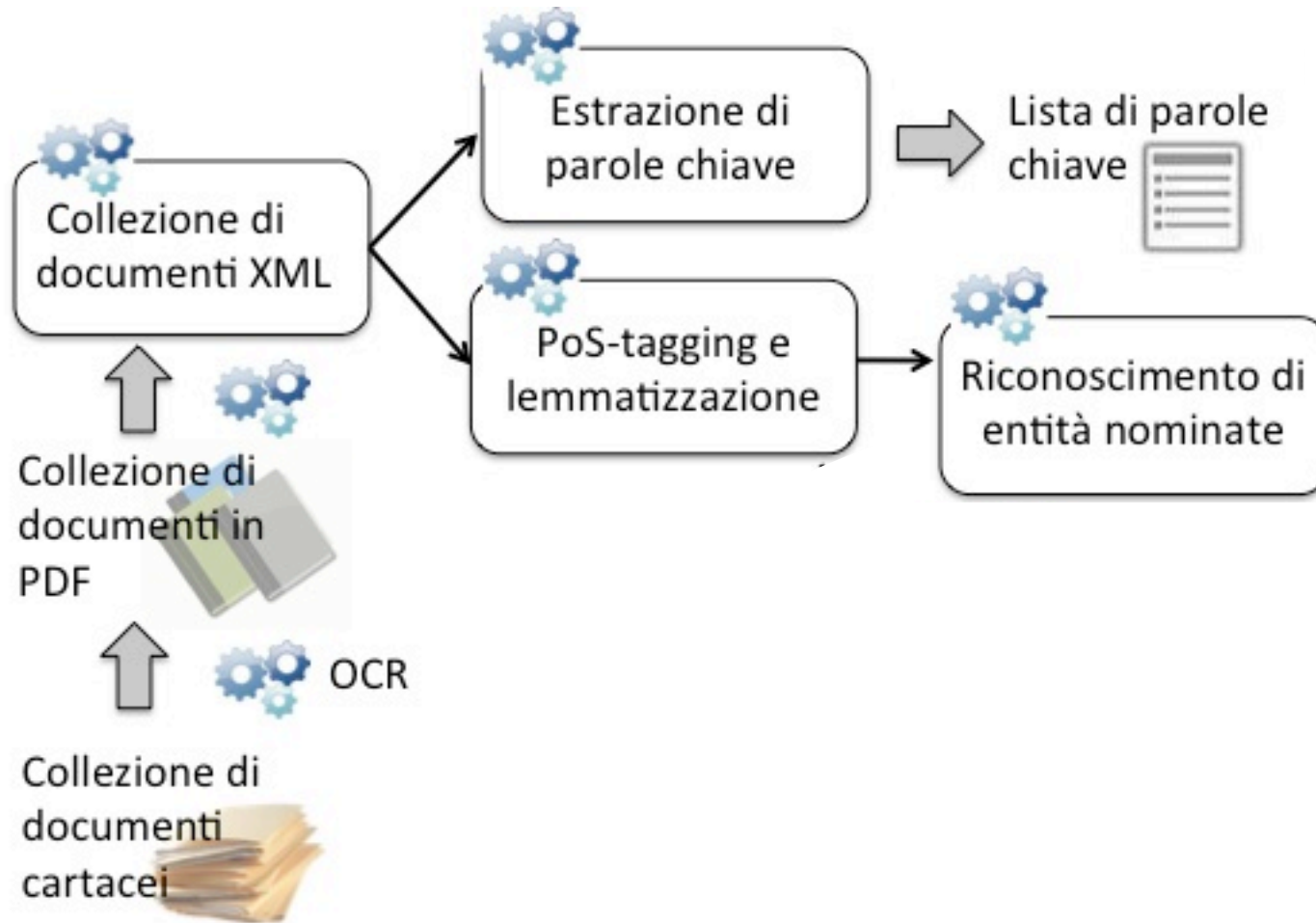


Dal documento cartaceo alla piattaforma web

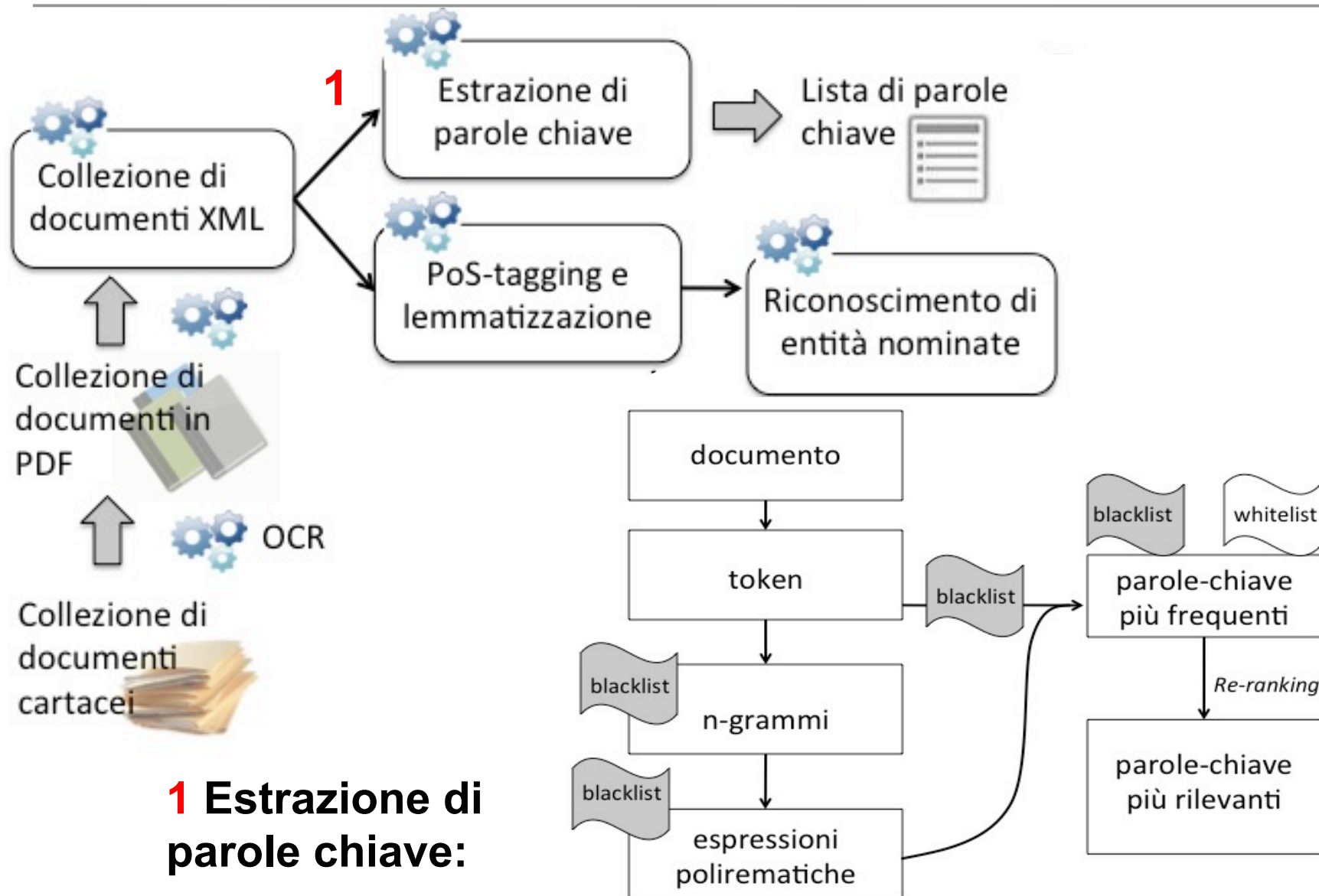


```
<xml archive="I.pdf">
  <file id="1">
    <head>
      <url>I.pdf_doc_number_1</url>
      <date>1901-01-01</date>
      <title>Title of the document</title>
      <description>Description of the Document</description>
      <location>Geographic location</location>
      <abstract>Content of the 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>
  <file id="2">
    <head>
      <url>I.pdf_doc_number_2</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</content>
    <original_text>Text of the document not in Italian - if present</original_text>
  </file>
</xml>
```

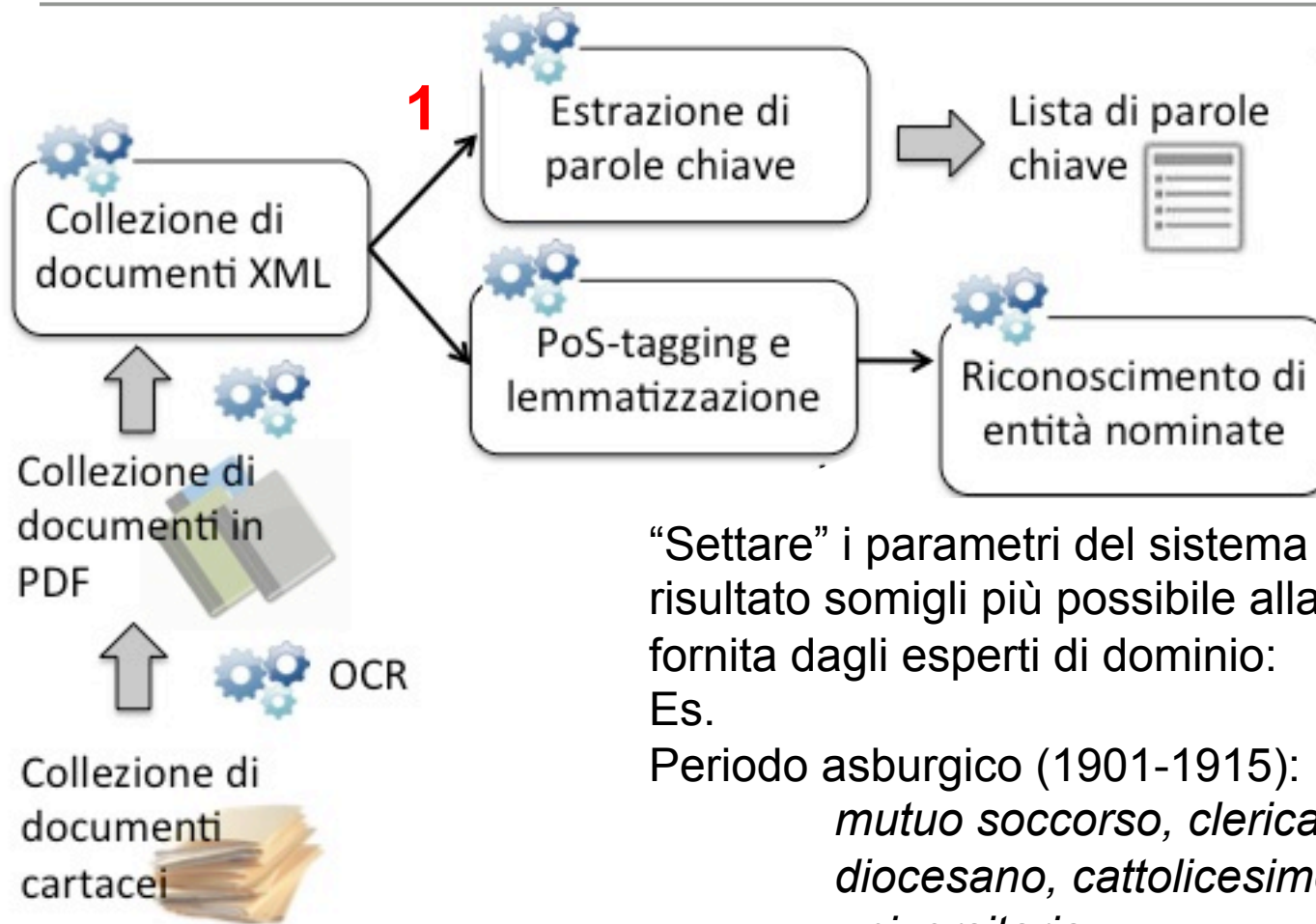
Dal documento cartaceo alla piattaforma web



Dal documento cartaceo alla piattaforma web



Dal documento cartaceo alla piattaforma web



1 Estrazione di parole chiave:

“Settare” i parametri del sistema in modo che il risultato somigli più possibile alla mappa concettuale fornita dagli esperti di dominio:

Es.

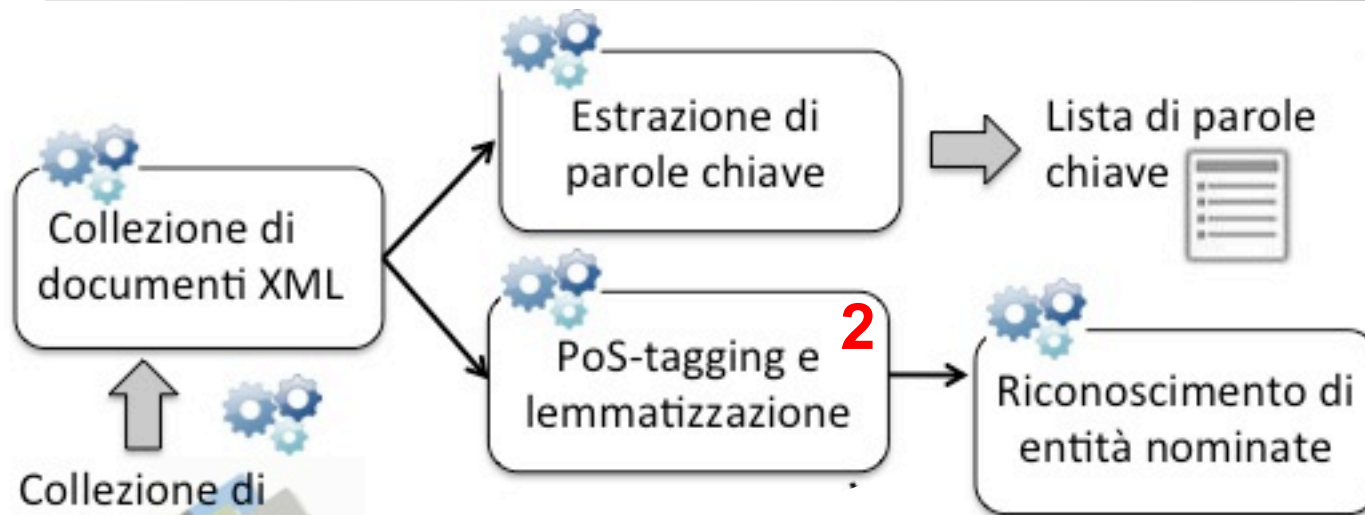
Periodo asburgico (1901-1915):

mutuo soccorso, clericalismo, comitato diocesano, cattolicesimo, questione universitaria, ...

Dal partito popolare al fascismo (1919-1927):

profughi, socialismo, ricostruzione, guerra, pace, Versailles, regionalismo, Trentino, ...

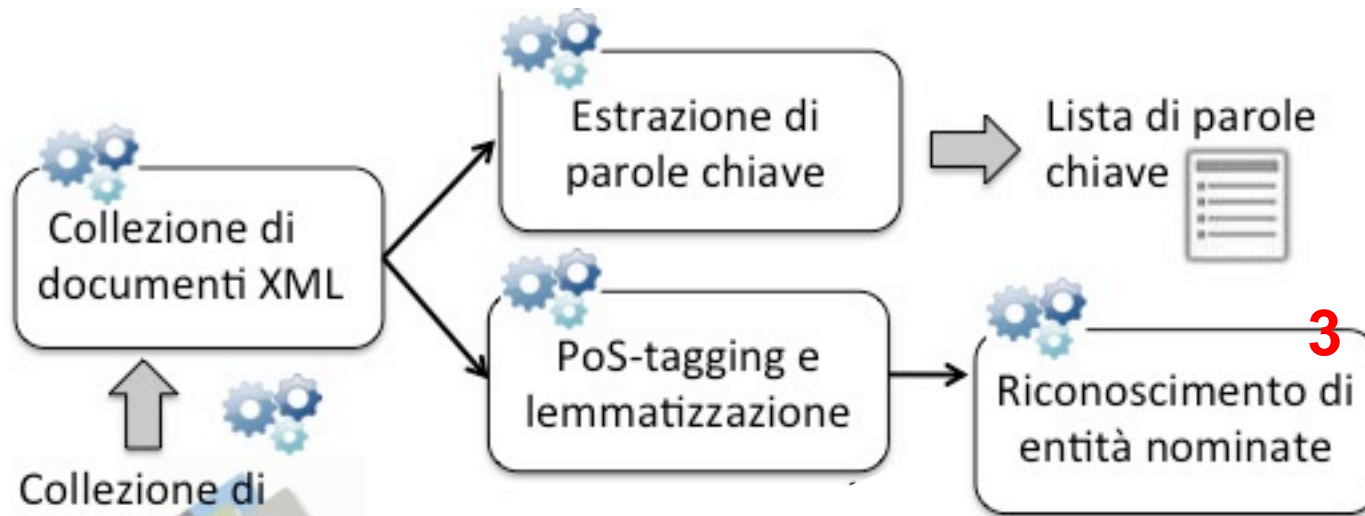
Dal documento cartaceo alla piattaforma web



² Part-of-Speech tagging e lemmatizzazione:

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

Dal documento cartaceo alla piattaforma web



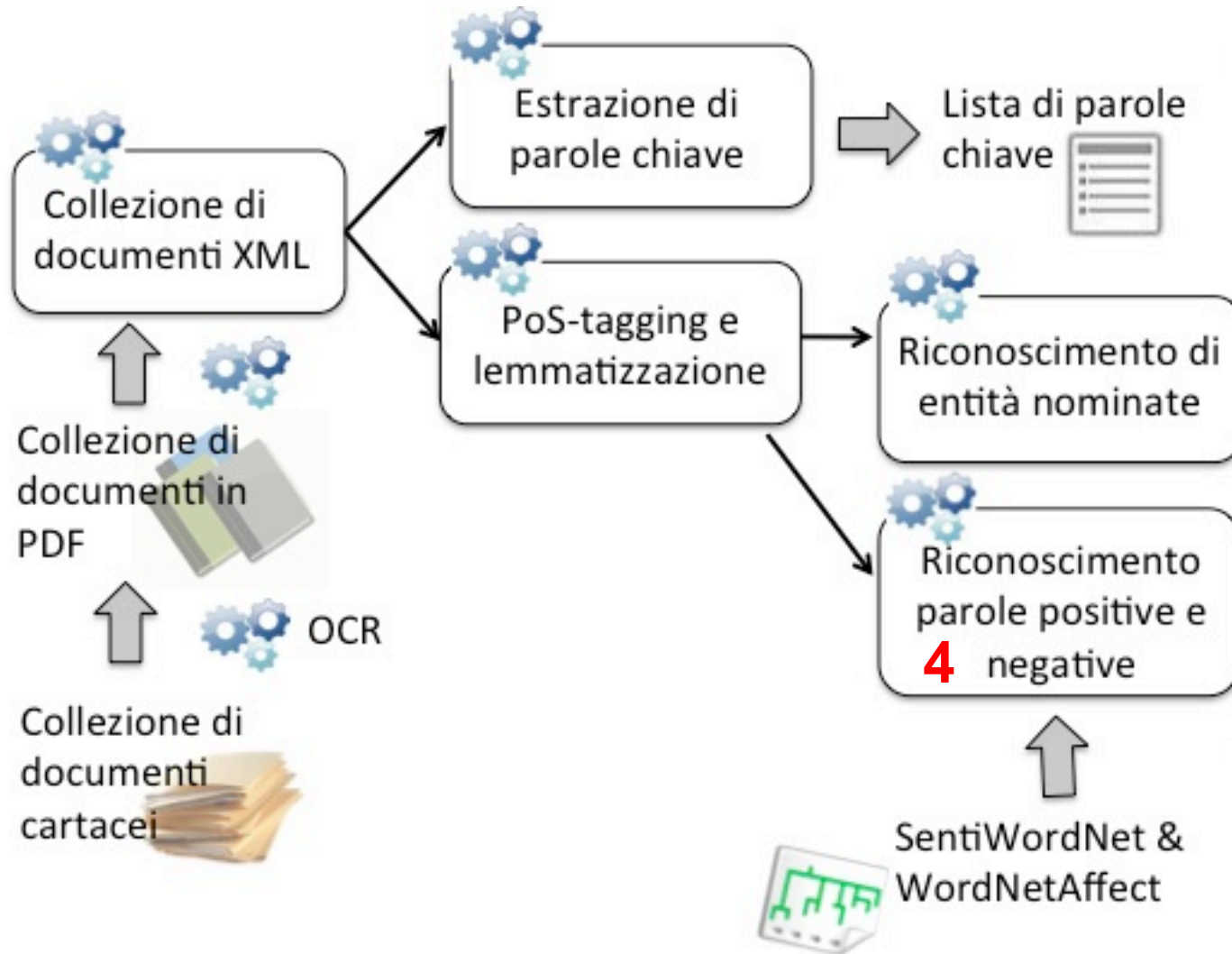
3 Riconoscimento di entità nominate:

☒ Person ☐ Organization ☒ Location

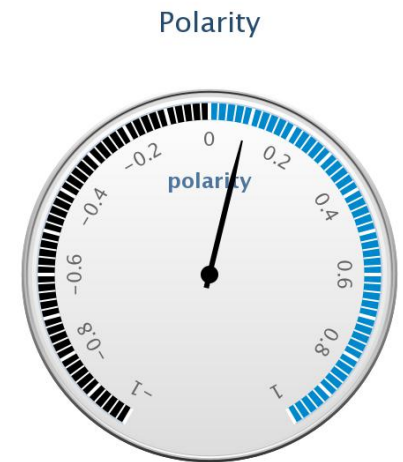
☐ Show TextPro format

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 Sardinia .
Dopo di lui nacquero Mario , che fu sacerdote , Marcella e Augusto .

Dal documento cartaceo alla piattaforma web

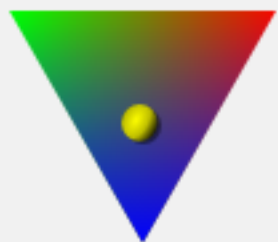


4
Riconoscimento
parole positive e
negative:



Riconoscimento parole positive e negative

- Utilizzare due lessici per la polarità, *SentiWordNet* e *WordNetAffect*, “mapparli” sul lessico italiano e applicarli a testi italiani: il valore di polarità è dato dal rapporto tra parole positive e parole negative presenti nel testo



P: 0.25 O: 0.5 N: 0.25

singular#1 **remarkable#1**

unusual or striking; "a remarkable sight"; "such poise is singular in one so



P: 0.375 O: 0.625 N: 0

remarkable#2 noteworthy#2

worthy of notice; "a noteworthy fact is that her students rarely complain";

Riconoscimento parole positive e negative

- Ma quanto è difficile annotare la polarità su testi storici?
- Esperimento per l'annotazione della polarità di “*sindacato*” e “*sindacalismo*” nel corpus di De Gasperi, 525 frasi
- Annotazione effettuata tramite *crowdsourcing* con la piattaforma CrowdFlower
- 5 annotazioni raccolte per ogni occorrenza, 4 possibili etichette: positivo, negativo, neutro, non so.
- Gold standard contenente 60 frasi annotate da due esperti, per valutare la qualità dell'annotazione data in *crowdsourcing* e quella calcolata automaticamente in base al lessico derivato da SentiWordNet e WordNetAffect.

Riconoscimento parole positive e negative

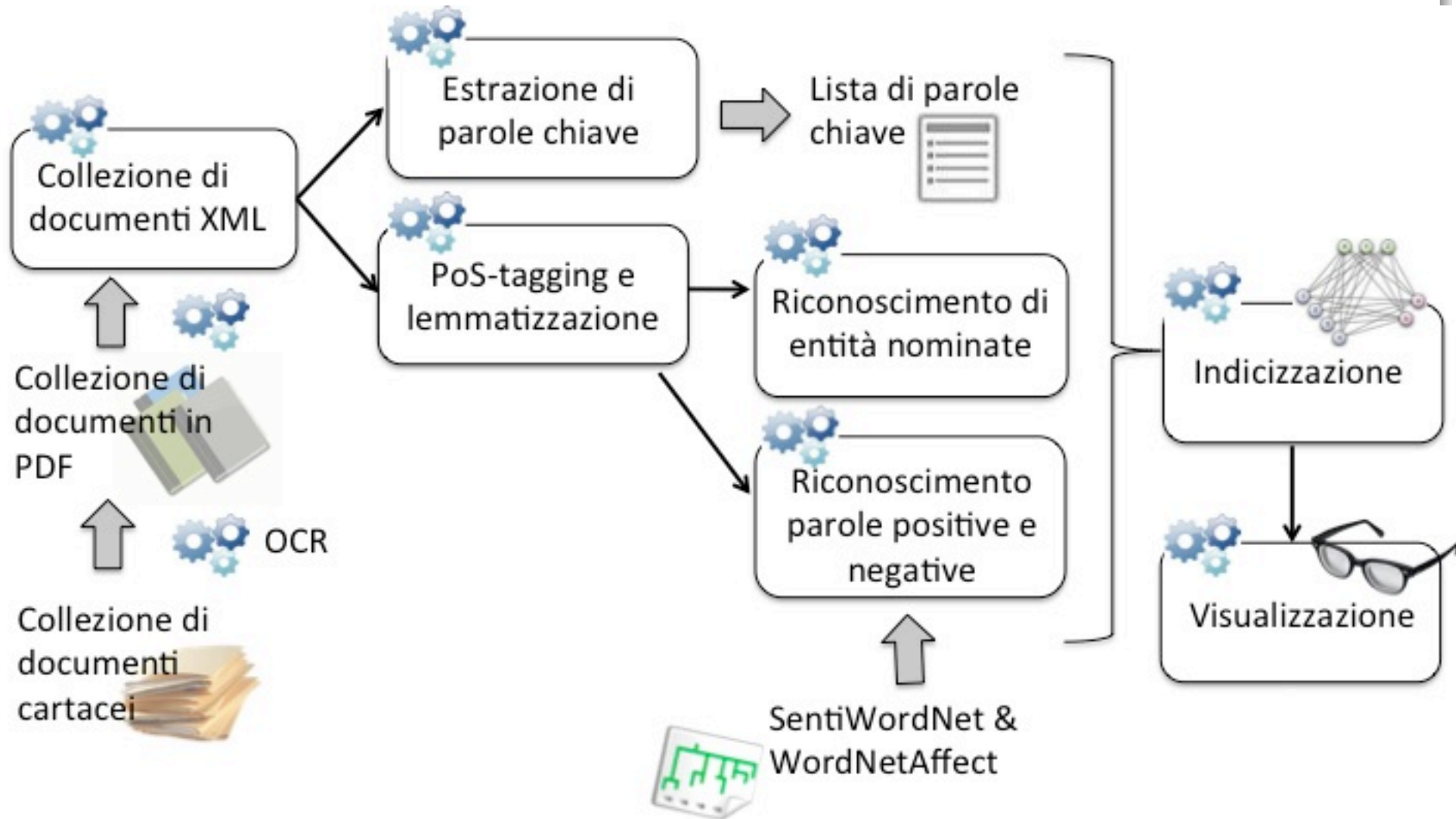
- Ma quanto è difficile annotare la polarità su testi storici?

Accuratezza	Crowdsourcing	SentiWNW/WNAffect
Totale	68.30%	43%
Negativo	55.50%	22%
Positivo	80%	31%
Neutro	46.60%	86%
Non so	0	n.a.

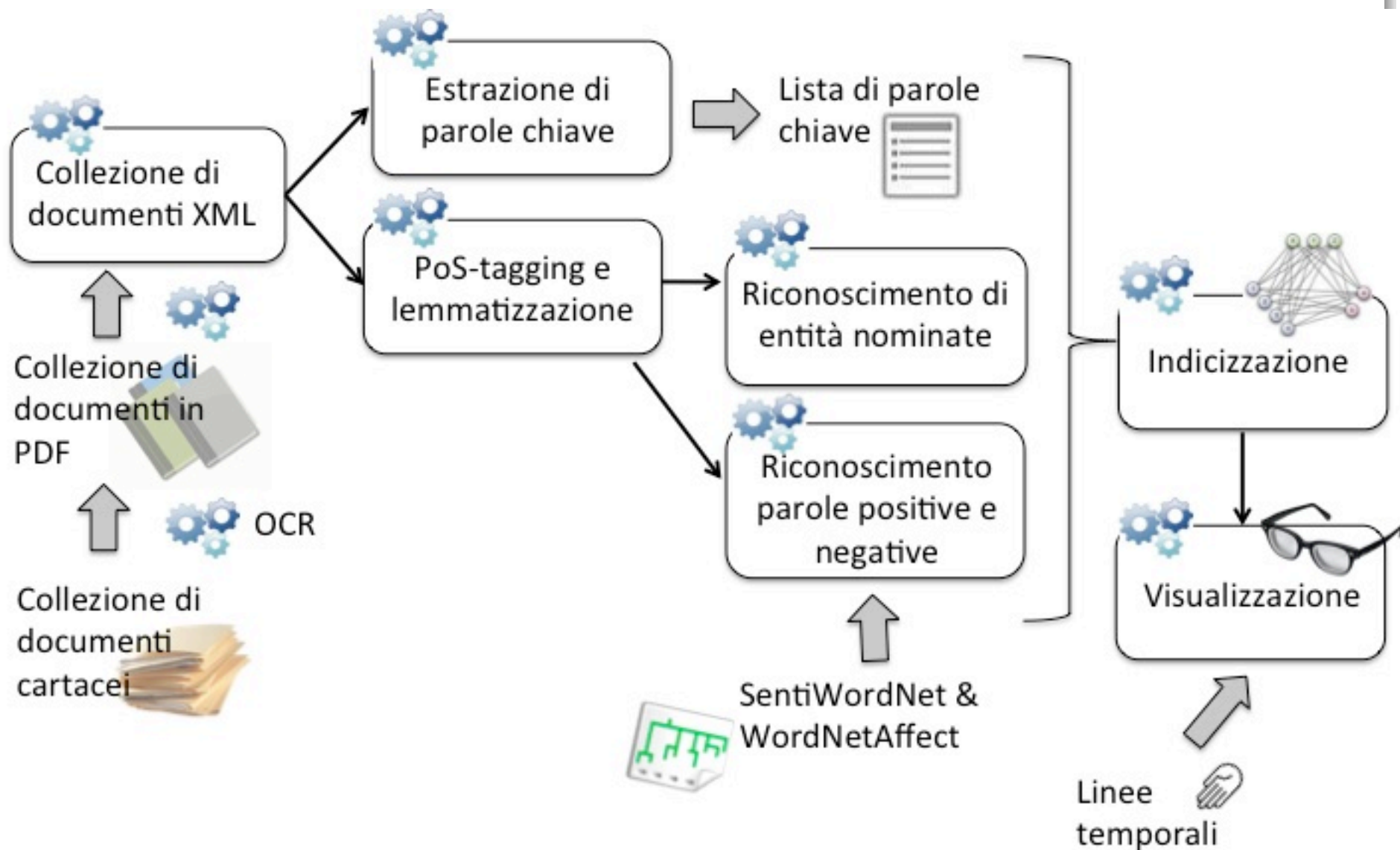
Accordo tra gli annotatori, sia esperti che via crowdsourcing, è basso:

*“Quando si costituì il sindacato agricolo-industriale se ne preannunziò il fallimento entro sei mesi ; e il **sindacato** prospera e fiorisce . Quando la cooperazione prese uno sviluppo imponente , Il Popolo eccitò i compagni ad entrarvi per mandare tutto alla malora.”*

Dal documento cartaceo alla piattaforma web



Dal documento cartaceo alla piattaforma web



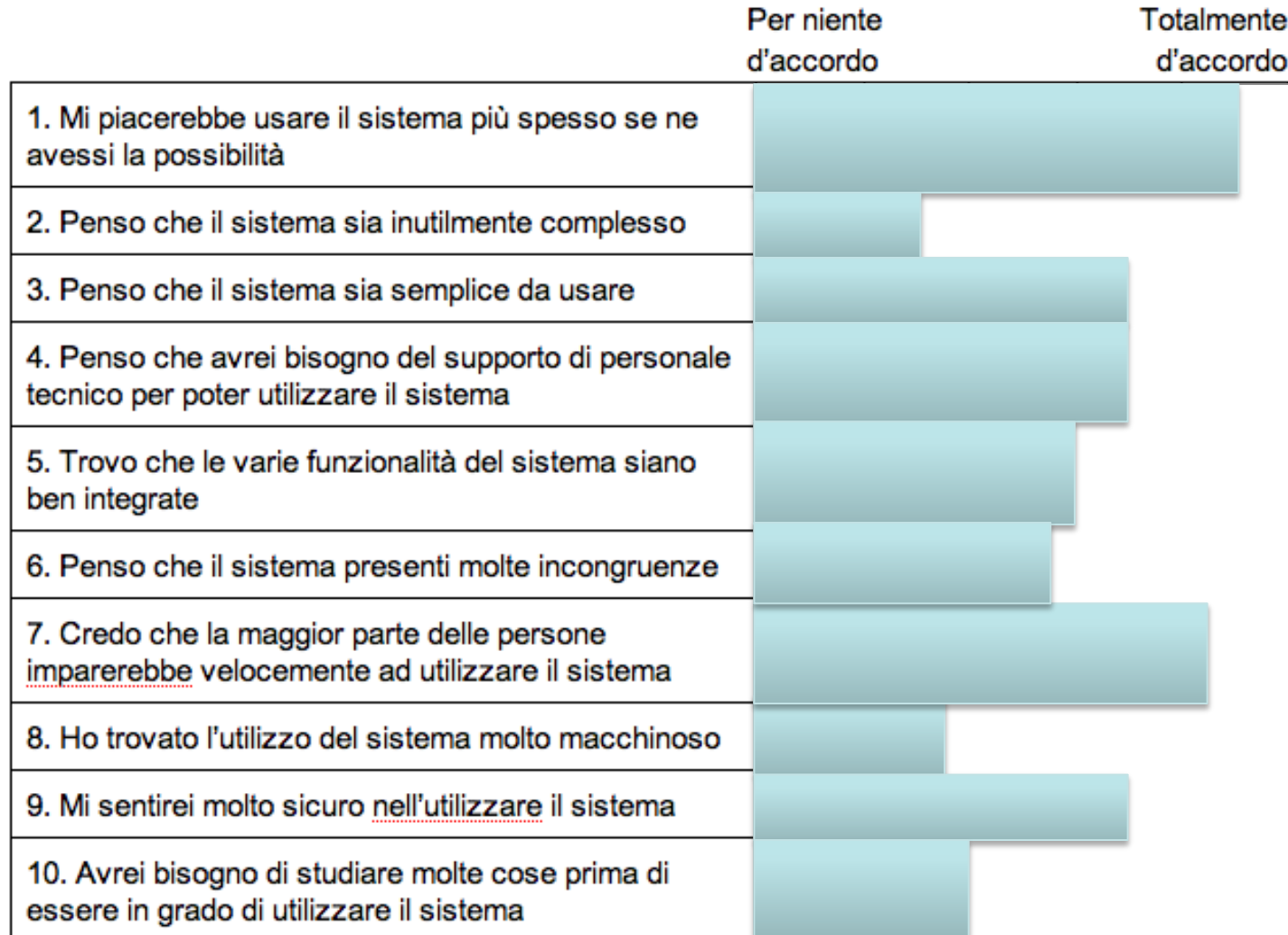
Linee temporali

- Libreria **TimelineJs**:
 - Inserire i campi desiderati in un Google spreadsheet
 - Pre-caricare nel sistema linee temporali diverse per uno stesso periodo
 - In A.L.C.I.D.E., 4 linee temporali: storia americana, storia asiatica e storia europea del '900 e biografia di De Gasperi.

Valutazione del sistema

- Valutazione basata sulla System Usability Scale (SUS)
 - Misurare *effectiveness*, *efficiency* e *satisfaction*
 - SUS copre vari aspetti dell'usabilità di un sistema, valuta 10 elementi su una scala da 1 a 5
 - Test da fornire agli utenti dopo che hanno avuto la possibilità di usare il sistema ma prima che si apra una fase di discussione

Valutazione del sistema: Risultati del questionario



Valutazione del sistema: Risultati del focus group

- Alcune funzionalità (es. *Timeline* o link delle entità nominate a *Wikipedia*) sono più adatte per scopi didattici, meglio dividerle dalle altre funzioni di ricerca
- L'estrazione dei concetti-chiave è un processo dai criteri “soggettivi”, utile ma di difficile interpretazione
- La polarità a livello di documento è poco utile, sarebbe meglio averla a livello di concetto
- Sarebbe utile poter avere un'analisi di come cambia il significato di una parola nel tempo (per es. guardando le co-occorrenze)

Progetto A.L.C.I.D.E.: Prossimi passi

- Implementare – se possibile – i suggerimenti degli esperti emersi nel focus group
- Utilizzare il corpus diviso per tipologia di discorso e per ruolo del parlante per fare esperimenti di classificazione automatica: come cambia il vocabolario, lo stile, la struttura del discorso per le diverse tipologie?
- Aggiungere altre collezioni di documenti, anche in inglese
- *Rendere la piattaforma accessibile a tutti (almeno una versione di base)*

Concludendo...



Intervento alla seduta plenaria del C.N.R. Roma, 15.02.1946

[...] Il popolo italiano si trascina, in questo momento, carico di affanni e attraverso stenti e privazioni verso l'erta della sua rinascita. Non possiamo dargli da sfamarsi e da coprirsi. Parrebbe ironia parlargli di cultura e di ricerca scientifica; ma l'uomo non vive di solo pane e sopravvive all'indigenza di oggi per la speranza del domani. Esso ha bisogno di credere nelle forze dello spirito che domineranno la materia. Una di queste forme è la scienza, la ricerca scientifica.

- **Grazie dell'attenzione! Domande?**