Luca Giuliano

IL VALORE DELLE PAROLE

L'analisi automatica dei testi in Web 2.0

Data Science

Luca Giuliano

IL VALORE DELLE PAROLE

L'analisi automatica dei testi in Web 2.0

Data Science

Luca Giuliano
Il valore delle parole. L'analisi automatica dei testi in Web 2.0.
Roma : Dipartimento di Scienze statistiche [2013] 116 p.
ISBN 978-88-908757-0-0

Questo libro è stato realizzato con iBooks Author per la visualizzazione in iPad con iOS 5 o versione successiva. Per la segnalazione di errori e imprecisioni scrivere all'autore: Luca Giuliano, luca.giuliano@uniroma1.it. Una copia statica in PDF è disponibile sul sito del dipartimento di Scienze statistiche - Sapienza Università di Roma:

http://www.dss.uniroma1.it/it/node/5868

Istruzioni sull’attivazione dell’opzione “Gesti Multitasking” su iPad

Per una lettura ottimale di questo libro interattivo e per una piena fruizione dei link esterni ci dobbiamo prima di tutto assicurare che l’opzione Gesti multitasking per iPad sia abilitata.

Aprite **Impostazioni** e toccate **Generali**. Scorrete verso il basso per trovare “Gesti multitasking” e posizionatelo su ON.

Utilizzando quattro o cinque dita per strisciare in orizzontale (come nella figura al centro, qui sotto) si vedranno scorrere le applicazioni aperte. Provate scorrendo da destra a sinistra, dal momento che molto probabilmente in questo momento siete nell’ultima applicazione utilizzata.

Questo gesto vi permetterà di ritornare facilmente alla lettura del libro nel punto in cui vi siete interrotti dopo aver seguito un link esterno nel browser Safari.

Le altre funzioni di Gesti multitasking, spesso poco note agli utenti di iPad, sono:

- Chiudere le applicazioni e tornare alla schermata principale con un gesto di “pizzico” effettuato con quattro o cinque dita (figura a sinistra).
- Rivelare la barra delle applicazioni aperte con un colpo in su effettuato con quattro o cinque dita (figura a destra). Ripetendo il colpo verso il basso la barra delle applicazioni si chiude.

Capitolo 1

Mondi fatti di parole

Abituati come siamo a essere immersi nella comunicazione multimediale tendiamo a non dare la dovuta importanza alle parole e alla scrittura. Internet è prima di tutto un mondo di parole.

1

Informatica, statistica e linguistica

Ogni giorno una immensa quantità di dati e informazioni transitano sulle reti e l'avvento delle tecnologie Web 2.0 ha intensificato il volume di questo traffico con messaggi in Facebook e in Twitter, blog e commenti, contributi su piattaforme Wiki, partecipazione a forum e a gruppi di discussione, pagine web e news. Nel frattempo sta proseguendo a grandi passi la digitalizzazione degli archivi bibliografici, delle biblioteche e degli archivi storici a tutti i livelli, comprese le attività giudiziarie, di governo e le transazioni economiche.

Fig. 1.1 La rete trasporta ogni ora l'equivalente di venti miliardi di libri

Secondo il rapporto IDC – Global Go to Market Services 2011 (Digital Universe, italy.emc.com) in Internet si esprime ogni mese un flusso di informazioni pari a 150 exabytes. Tenendo presente che il 90% è costituito da materiale video, le "parole" (documenti txt, pdf e doc) che viaggiano sulla rete potrebbero essere stimate in 15 exabytes al mese (circa 15 milioni di terabytes). Si consideri che l'intera collezione a stampa della Library of Congress, dotata di 29 milioni di libri (senza tenere conto delle immagini e dei tracciati audio), è stata stimata di circa 30 terabytes. Questo vorrebbe dire che la rete trasporta ogni ora una quantità di testi equivalente a venti miliardi di libri.

Si tratta di una massa enorme di informazioni sui comportamenti linguistici che era impensabile fino a qualche tempo fa. È un'opportunità di analisi preziosa per la lettura e la comprensione dei fenomeni sociali. Un'opportunità

che il Web 2.0 rende ancora più significativa per il protagonismo che ogni utente esprime, inserendo autonomamente in web i contenuti che lo interessano e condividendo con gli altri utenti giudizi, opinioni, scelte.

Le nostre capacità di gestione e calcolo non sono ancora in grado di sfruttare pienamente tutte le informazioni che potremmo estrarre da questa crescita esponenziale di testi digitalizzati. Tuttavia la rapidità con cui negli ultimi anni è migliorato il rapporto tra memorizzazione e gestione di grandi insiemi di dati fa ben sperare per il prossimo futuro.

Gli strumenti di analisi e visualizzazione dei dati testuali che ci apprestiamo ad esplorare in questo libro non sono sostitutivi dei software commerciali o di ricerca utilizzati nella linguistica computazionale e nella statistica linguistica. Per una visione più completa dell'argomento è utile consultare una bibliografia specifica. Un buon punto partenza per una guida al software disponibile è il portale BAMBOO DiRT – Digital Research Tools, in particolare la sezione: Analyze Texts.

Tra i software di maggior rilievo sviluppati nel mondo della ricerca possiamo segnalare: Alceste, Alias-i, CATMA, DTM-VIC, Gate, IRaMuTeQ, Lexico3, TaLTaC2, TXM – Textométrie, T-LAB, WordStat.

L'incontro tra informatica e linguistica è fortemente intrecciato con la statistica e la matematica. La sintesi avviene sia nella ricerca di base sia nelle applicazioni tecnologiche, in particolare nei settori della traduzione automatica, del riconoscimento e sintesi digitale del parlato e della gestione dei grandi sistemi informativi. Lo sviluppo dell'informatica dagli anni '50 in poi ha esercitato un grande fascino sui linguisti e, in generale, sugli studiosi di discipline affini come la sociolinguistica e la psicolinguistica, dando vita alla linguistica computazionale (Jurafsky e Martin, 2000; Chiari, 2007). Oggi più che mai l'analisi quantitativa del linguaggio rappresenta una sfida straordinaria per la metodologia della ricerca nelle scienze sociali (Bolasco, 2013).

Tra gli anni ’60 e gli anni ’70 uno statistico francese, Jean-Paul Benzécri, mettendo a frutto le nascenti disponibilità di calcolo automatico, iniziò ad applicare ai dati linguistici i modelli statistici dell’*Analyse des Données*, mettendo in evidenza come la matematica del finito e della combinatoria che si rifà a strutture algebriche astratte permettesse di ricostruire tracce di senso più semplici soggiacenti alla complessità presente in un corpus di testi. Così facendo Benzécri stava realizzando il sogno degli statistici del passato, come Maurice Kendall e John Tukey, che avevano anticipato la possibilità di compiere analisi multidimensionali su matrici di centinaia di variabili. L’approccio di Benzécri possedeva i tratti di una sintesi efficace e affascinante, soprattutto se applicata alle proprietà squisitamente qualitative delle parole e del lessico.

Fig. 1.2 Jean-Paul Benzécri

Già prima che l’informatica fornisse un approccio quantitativo all’analisi linguistica e dei testi, un contributo importante in questo senso era stato dato dalle osservazioni induttive e dalle misure del matematico russo Viktor Bunyakowski che nel 1848 tracciava un abbozzo di aritmetica del linguaggio che tenesse conto della frequenza delle parole e della loro lunghezza. Altri contributi analoghi vennero poi successivamente dagli inventori della stenografia, come Jean-Baptiste Estoup (1916), e dalla psicolinguistica di Adolf Busemann con il suo studio sul linguaggio dei bambini (1925).

Negli anni Trenta, George K. Zipf (1929) con il suo “principio di frequenza relativa” delle forme grafiche aprì la strada alla lessicometria e alla vera e propria statistica linguistica.

Le scienze dell’uomo, oggi, hanno tutto da guadagnare dalle immense risorse offerte dall’informatica. L’analisi dei processi di comunicazione non può prescindere ormai dalla sempre più diffusa digitalizzazione dell’informazione. Bibliografia, classificazione dei testi e dei documenti, editoria elettronica, gestione della conoscenza basata su dati non strutturati, come ad esempio nella gestione della documentazione giuridica, pongono continuamente problemi nuovi le cui soluzioni non possono che scaturire dalla collaborazione intensa tra gli informatici, gli statistici e gli esperti delle discipline sostantive.

La gestione intelligente della memoria

La digitalizzazione dei documenti e la pervasività della comunicazione mediata dal computer hanno reso potenzialmente disponibili grandi masse di informazioni e dati. Tuttavia questo non ci deve far dimenticare che, per quanto dettagliate e precise siano le nostre informazioni, esse saranno sempre incomplete rispetto alla realtà e complessità dei fenomeni oggetto di studio. La nostra conoscenza è sempre intrinsecamente incerta. Per questo, informatica e statistica devono essere affiancate dalla scienza dell'incerto, la probabilità.

L'analisi dei testi e del linguaggio non si sottraggono a questo limite. Ogni osservazione sul lessico e sulle strutture linguistiche rimanda necessariamente alle regole che governano la lingua come fenomeno sociale e pertanto pone problemi di inferenza che si risolvono soltanto con il supporto dei modelli matematici dell'incertezza. D'altra parte l'informazione stessa – già nella classica definizione di Claude Shannon – è inversamente proporzionale alla probabilità: un evento improbabile è più informativo di un evento probabile.

Nella interpretazione di un testo e nell'analisi del suo contenuto semantico, l'informazione si esprime in una rete di significati di complessità crescente e non lineare che – per essere efficiente – deve uniformarsi al modelli della mente e del pensiero nei quali l'originalità, l'imprevisto e la sorpresa rappresentano inevitabilmente caratteristiche dell'intelligenza e della creatività.

Fig. 1.3 Il Memex di Vannevar Bush

Il primo a cogliere il senso complessivo dei vantaggi offerti dalla tecnologia per la "gestione intelligente della memoria" fu Vannevar Bush con il suo "memex" (*memory extension*) uno strumento di consultazione e indicizzazione dei documenti d'archivio inventato negli anni Trenta del secolo scorso. Il

memex avrebbe dovuto riprodurre i percorsi associativi e non lineari della mente (Bush, 1945).

Il passo successivo fu compiuto nel 1960 da Ted Nelson con il progetto Xanadu, il prototipo di una *Literary Machine* mai giunta a compimento (Nelson, 1990) ma che avrebbe trovato una realizzazione pratica e rivoluzionaria qualche anno dopo nel World Wide Web di Tim Berners Lee (1991) e poi, successivamente, nella piattaforma Wiki di Ward Cunnigham (1995).

La nozione di hyperlink introdotta da Ted Nelson in breve tempo è entrata a far parte delle nostre modalità naturali di produzione intellettuale a tal punto da averne dimenticato il contenuto innovativo. Oggi per noi il link è come un nuovo segno di interpunzione nel testo scritto: una convenzione normativa che aiuta a rappresentare nel testo scritto l'organizzazione reticolare del pensiero.

I documenti di ricerca, se non sono stati concepiti nel momento della loro rilevazione in una forma strutturata e con dati classificabili a priori (modelli di rilevazione, questionari, test), sono irriducibilmente ancorati alla loro forma fisica originaria: la lettera, l'articolo di giornale, il manifesto, il diario, la fotografia ecc. Fino all'avvento della digitalizzazione, i documenti "naturali" prodotti per qualsiasi motivo nella vita sociale non potevano essere sottoposti all'analisi empirica del sociologo se non con immensa fatica ed enorme impiego di tempo. W.I. Thomas e F. Znaniecki nel 1920 portarono a termine il loro lavoro *The Polish Peasant in Europe and America* analizzando in sei anni almeno 1.000 lettere e 8.000 documenti vari tratti da giornali dell'epoca, ma nessuno sa esattamente quanti fossero, quali furono scartati e perché. Questi problemi furono messi in luce molto tempo dopo la pubblicazione, in un convegno del 1938 durante il quale si seppe che gran parte della documentazione originale era andata distrutta (Madge, 1966).

Una base di dati fondata sull'analisi dei documenti naturali, tipicamente d'archivio, ha un carattere difficilmente ispezionabile se non riproducendo il percorso, spesso singolare e intuitivo, dello studioso che lo ha compiuto. Oggi molti documenti non strutturati sono prodotti "naturalmente" in forma digitale (forum in Internet, email, messaggi in Facebook e Twitter, blog, news online) e pertanto possono essere analizzati sia automaticamente con l'analisi lessicometrica sia in modo semi-automatico (CAQDAS - *Computer Assisted Qualitative Data Analysis Software*) utilizzando strumenti che consentono la trasparenza delle scelte metodologiche compiute dai ricercatori fino a seguirne passo per passo il percorso di concettualizzazione, operativizzazione e classificazione (Giuliano e La Rocca, 2008).

Dall'informazione al dato testuale

La rete, come si è detto, è un "archivio vivente" che contiene una massa ingente di informazione. Uno dei risultati più interessanti ed efficaci della evoluzione di Internet verso l'interattività del Web 2.0 è la possibilità di avere a disposizione oltre ai testi anche gli strumenti online con i quali analizzare questi di flussi di informazione per scavare in essi al fine di rintracciare schemi interpretativi e ricostruirne il senso. Si tratta, ovviamente, di strumenti di primo approccio, in alcuni casi abbastanza semplificati. L'analisi automatica dei dati testuali e il *text mining* richiedono approcci più sofisticati di quanto lascino intendere gli strumenti illustrati in questo libro. Tuttavia, come si potrà osservare, essi riservano più di qualche sorpresa e, come spesso accade con il software, in alcuni casi le loro potenzialità vanno al di là degli obiettivi di chi li ha inventati, rivelandosi strumenti aperti all'intervento creativo dell'utente.

L'informazione abbandonata a se stessa rimane offuscata da un apparente disordine. Applicando alcune metodologie di analisi automatica dei testi possiamo imparare a gestire meglio e in modo più proficuo il patrimonio di dati, per ora nascosti, che sono già a nostra disposizione selezionando ciò che è di nostro interesse e visualizzandone sinteticamente i risultati.

Prima di procedere con l'illustrazione di alcuni di questi strumenti selezionati per la loro semplicità, efficacia e affidabilità, dobbiamo però impadronirci delle nozioni fondamentali che riguardano la preparazione dei testi e il modo migliore per utilizzarli in modo consapevole e appropriato.

Che cosa intendiamo quando parliamo di “testi online”? Nella rete troviamo un po’ di tutto: testi poetici e letterari, articoli di giornali, documenti legislativi, articoli scientifici, pagine web, articoli di Wikipedia, messaggi dei social network, ricette di cucina, barzellette, testi di canzoni. Non vi sono limiti alla varietà di testi che possiamo trovare in Internet. Ciascuno di questi testi, anche preso singolarmente, può essere oggetto di analisi automatica. Tuttavia il risultato non sarà necessariamente interessante, non più di quanto possa esserlo entrare in una biblioteca e contare i libri che si trovano negli scaffali o entrare in una pinacoteca e misurare le cornici dei quadri esposti. Il nostro interesse verso l’analisi di uno o più testi deve essere guidato da qualche domanda che vogliamo rivolgere a essi e da una o più risposte provvisorie che rappresentano le nostre ipotesi di lavoro.

I testi, pertanto, sono analizzabili solo se costituiscono un corpus: una qualsiasi collezione di testi confrontabili tra loro sotto un qualche punto di vista.

In molti casi il corpus si costituisce facilmente: per esempio, la totalità degli interventi in un forum, in una chat o in un gruppo di discussione di un social network (corpus chiuso). Ciascuno degli interventi costituisce un testo o un frammento di testo. In alcuni casi

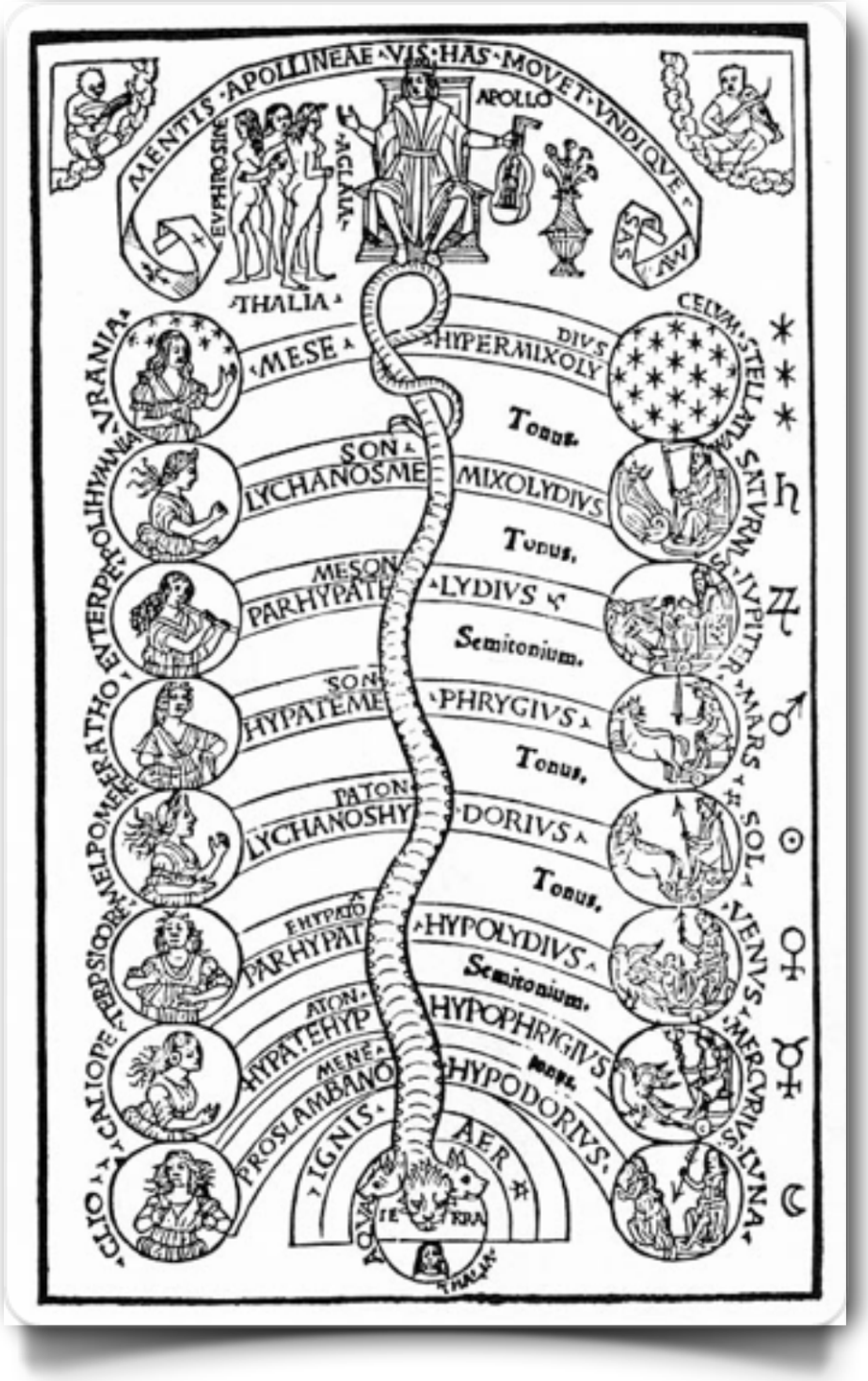


Fig. 1.4 Corpus Hermeticum

gli interventi potrebbero essere raggruppati secondo le caratteristiche degli scriventi (maschi e femmine, gruppo di età, interessi prevalenti ecc.) oppure semplicemente secondo la scansione temporale. In altri casi può essere necessario costituire il corpus secondo determinati criteri di selezione e rappresentatività (corpus campionato). Il corpus è sempre una conseguenza delle scelte e delle decisioni operative di chi compie l'analisi. È impossibile dire a priori se un corpus è costituito adeguatamente senza assumere come riferimento lo scopo per cui verrà analizzato.

Fonti e corpora online

Come si é detto le fonti online per il reperimento dei testi utili alla costruzione di un corpus possono essere le più diverse e, per la maggior parte, sono note (social network, Wikipedia, quotidiani online, blog, RSS ecc.). Meno noti potrebbero essere alcuni siti e portali utili per il reperimento di pubblicazioni digitali e pertanto ne elenchiamo alcuni dei più interessanti, soprattutto quelli che non possono mancare nella mappa dell'analista di testi online.

Fig. 1.5 Project Gutenberg

Project Gutenberg vanta oltre 30.000 volumi nel suo catalogo ed è stato sicuramente il primo ad occuparsi della digitalizzazione dei libri a stampa.
È nato nel 1971, ben prima dell'avvento di Internet, per iniziativa di Michael Hart. È considerato il primo archivio di ebook e raccoglie, su base volontaristica, testi che sono di pubblico dominio oppure per i quali sono scaduti i diritti d'autore (secondo la legge USA). I testi, in massima parte in lingua inglese, sono scaricabili in diversi formati dall'HTML al formato ASCII.

Liber Liber è nato nel 1994 su iniziativa di Marco Calvo e altri ed è una organizzazione di volontari che si propongono di contribuire alla diffusione della cultura e del sapere condiviso online. La biblioteca digitale porta il

Fig. 1.6 Bartleby.com

nome di Progetto Manuzio ed è in collegamento con il progetto Gutemberg per la lingua italiana.

The Online Books Page è un progetto di biblioteca digitale diretto da John Mark Ockerbloom, ricercatore della University of Pennsylvania. Ha superato il milione di volumi mettendo in rete le risorse di diverse università e diversi progetti, compresi i progetti Gutemberg e Manuzio, conseguendo una vasta disponibilità di volumi in tutte le lingue del mondo.

Internet Sacred Text Archive, fondato da John Bruno Hare è il più grande archivio di testi che appartengono alla tradizione sacra, alla mitologia e al folklore di ogni luogo e di ogni tempo.

Bartleby.com è un progetto iniziato nel 1993 alla Columbia University. Il nome deriva da un racconto di Herman Melville (*Lo scrivano*), uno dei più famosi della letteratura americana. Vi si trovano testi letterari, scientifici e politici, come i discorsi inaugurali dei Presidenti USA.

PoliTxt è un archivio di testi politici. Oltre ai discorsi inaugurali dei Presidenti USA vi si trovano le trascrizioni dei dibattiti televisivi.

WebCorp Live è uno strumento gestito dalla Birmingham City University per la ricerca linguistica ma può essere utilizzato per diverse applicazioni anche per affinare i risultati dei motori di ricerca e per la costruzione di corpora tematici. Fondamentalmente è un generatore di concordanze che funziona sulla base delle pagine web e dei

motori di ricerca più noti (Google, Bing e Yahoo!) con la possibilità di selezionare le news e la lingua di interesse. Il sito offre diverse opzioni, soprattutto se si utilizza l'interfaccia di ricerca avanzata. La finestra di selezione dei siti secondo gli argomenti (ad es. UK Tabloid Newspaper) o l'estensione di nazionalità (.it, .fr, .cn) permette di ottenere dei risultati molto focalizzati. Anche la finestra di inserimento delle parole o delle frasi per la ricerca, permette di generare risultati molto flessibili e sofisticati, facendo uso di semplici pattern di ricerca che si basano su wildcards (caratteri jolly) ed espressioni regolari.

Normalizzazione del corpus

In genere il corpus da elaborare richiede una fase preliminare di conversione del testo in formato ASCII (*plain text*) e di normalizzazione. In qualche caso è lo strumento di analisi stesso a compiere alcune semplici operazioni di pulitura e standardizzazione. Il semplice e diretto “copia e incolla” in una finestra di acquisizione nella maggior parte delle situazioni è sufficiente, per esempio, a eliminare la formattazione, i marcatori di HTML o le immagini. L’analisi automatica dei testi dedica una particolare attenzione alla “preparazione del corpus” e ai problemi di trascrizione e transcodifica di caratteri accentati e caratteri speciali. Un altro importante passaggio riguarda il trattamento dei caratteri in maiuscolo. Le maiuscole permettono di risolvere alcuni problemi di ambiguità lessicale (rosa / Rosa; bari, Bari) e tuttavia la loro presenza duplica le forme grafiche come conseguenza della punteggiatura: la lettera maiuscola che segnala l’inizio di un periodo. Per l’analisi dei testi online tutto questo deve necessariamente essere messo da parte perché sarebbe oggetto di una trattazione scientifica troppo specialistica. Una delle semplificazioni più drastiche cui è necessario ricorrere in questi casi è la riduzione in minuscolo di tutti i caratteri.

Nelle maggior parte delle applicazioni online il corpus è costituito da una raccolta di testi in sequenza, senza che sia possibile distinguere delle partizioni. Una partizione, per esempio, in un corpus costituito dai discorsi del giuramento dei presidenti USA sarebbe rappresentata dalla suddivisione tra i discorsi dei diversi presidenti; oppure, in un corpus costituito dai post in Facebook, una partizione potrebbe suddividere i post dei maschi da quelli delle femmine.

Parole, forme e occorrenze

Gli elementi costitutivi di un testo sono le parole. Nel trattamento automatico dei testi le parole si presentano sempre attraverso la loro forma grafica e cioè come una stringa di caratteri delimitata da due separatori. Nelle applicazioni disponibili online oggetto di questo libro l'utente non ha la possibilità di scegliere quali caratteri considerare come separatori perché è il software stesso a compiere questa scelta.

Le forme grafiche intese come unità di conto vengono definite occorrenze (*word tokens*). L'analisi automatica di un testo fornisce come primo risultato un conteggio delle forme grafiche con le rispettive occorrenze. L'elenco delle forme grafiche sarà rappresentato dalle forme grafiche distinte (*word types*). L'inventario delle forme grafiche si presta a vari tipi di ordinamento. I più consueti sono l'ordinamento alfabetico o secondo le occorrenze decrescenti.

Non tutte le parole di un testo possono essere considerate come equivalenti dal punto di vista semantico. Nell'analisi automatica del linguaggio si distinguono le parole piene dalle parole vuote (in alcuni casi indicate come *stop words*). Le parole vuote vengono definite di volta in volta come parole che non esprimono un contenuto interessante ai fini dell'analisi (e spesso sono parole grammaticali o di semplice legame nella frase), mentre le parole piene sono quelle che contribuiscono significativamente all'interpretazione del testo.

Nell'analisi automatica del linguaggio spesso le parole dei testi sono filtrate a priori attraverso una lista di *stop words*. Non vi sono criteri univoci per definire un elenco di parole da inserire nelle *stop words*. Sicuramente vi sono le parole grammaticali, come si è detto, ma in alcuni casi vi possono essere anche i verbi ausiliari o parole ritenute banali o non essenziali per determinati scopi. In una lista di *stop words* in inglese, francese e italiano utilizzata nei motori di ricerca possiamo trovare parole comuni come le seguenti ma che certo non possono essere definite come parole vuote dal punto di vista linguistico:

Quando è possibile è bene controllare l'elenco delle *stop words*, se è il caso modificarle, non tenerne conto e applicare criteri personali secondo gli scopi dell'analisi.

English	Français	Italiano
able, about, accordance, actually, adopted, affected, affecting, afterwards, against, (…) wish, world, zero	Avoir, devrait, doit, droite, début, elle, encore, est, fait, (…) voie, vous, vu	Buono, comprare, consecutivo, dentro, deve, fine, fino, gente, indietro, (…) voi, volte, vostro

Fig. 1.7 Esempi di *stop words* utilizzati dai motori di ricerca (Fonte: Ranks.nl)

Capitolo 2

Textalyser uno strumento per iniziare

Textalyser è un buon esempio di come si possa realizzare un programma di base per l'analisi automatica del testo con semplicità ma anche con rigore.

2

Textalyser

Textalyser offre diverse possibilità di analisi del testo: conteggio delle occorrenze, delle proposizioni, delle sillabe e dei segmenti ripetuti. Inoltre applica alcuni indicatori di leggibilità.

Il testo può essere inserito con copia e incolla in una finestra di interrogazione, specificando il link di una pagina web o con un upload del file direttamente dal proprio computer.

Per esaminare le opzioni di Textalyser prendiamo come riferimento il sito che riporta i discorsi inaugurali dei presidenti degli Stati Uniti (Bartleby.com) e scegliamo il discorso inaugurale di Barack Obama del 20 gennaio 2009. Potremmo acquisire il testo direttamente dall'indirizzo web ma è preferibile utilizzare il "copia e incolla" nella finestra apposita perché la pagina web contiene forme grafiche che non sono parte costituente del discorso (in altri casi il testo potrebbe essere suddiviso in diverse pagine e pertanto non verrebbe caricato completamente dal link alla pagina principale). Textalyser opera da solo una normalizzazione molto semplificata riducendo tutto il testo in caratteri minuscoli (per operazioni di pulitura più sofisticate è necessario utilizzare un elaboratore di "puro testo" come il Blocco note e Notepad++).

In figura 2.1 vediamo il risultato dell'analisi con le opzioni di default di Textalyser (*minimum characters* 3; *number of words* 10; *stoplist English*). L'output fornisce subito le principali misure lessicometriche : il *complexity factor* (o *type/token ratio*) è il rapporto tra le forme grafiche o parole distinte (*types*=V) e il totale delle occorrenze (*tokens*=N).

Questo rapporto è uno dei parametri fondamentali per valutare l'adeguatezza del corpus per l'analisi statistica. La soglia di accettabilità, secondo la quale si ritiene adeguata l'estensione lessicale del testo come rappresentativo del linguaggio, è inferiore o uguale al 20%. Valori superiori si riscontrano in testi piccoli come quello di questo

Total word count (tokens, occorrenze, N)	1.309
Number of different words (types, parole distinte, V)	839
Complexity factor (Lexical Density, N/V)	64,1%
Readability (Gunning-Fog Index: 6-easy 20 hard)	8,9
Total number of characters	16.637
Number of characters without spaces	7.837
Average Syllables per Word	1,58
Sentence count	137

Fig. 2.1 Misure lessicometriche del discorso del giuramento del presidente Barack Obama del 20 gennaio 2009 (elaborazione: Textalyser)

esempio. Il valore è ulteriormente “distorto” anche perché nell’analisi di default sono state considerate solo le parole con un minimo di tre caratteri e sono state escluse le *stop words*. Selezionando 1 carattere come lunghezza minima delle parole e nessuna *stoplist* le occorrenze sono N = 2.381 e le parole distinte V = 903 con una *type/token ratio* di 37,9%. Si tratta di un valore ancora lontano dall’essere adeguato. Tuttavia l’analisi è comunque interessante se ci limitiamo a utilizzare le evidenze quantitative come sussidio per compiere delle valutazioni prevalentemente qualitative.

L’indicatore di leggibilità (Gunning fog Index) è calcolato sulla base della lunghezza delle proposizioni di cui è composto il testo e della lunghezza delle parole in sillabe. Non ha rilevanza per un’analisi automatica del testo che ha come obiettivo di estrarre il contenuto, ma può essere ugualmente interessante per la comparazione dei testi

(per l'elaborazione di un indice di leggibilità dei testi in italiano vedi Lucisano e Piemontese, 1988; il servizio Èulogos e l'accesso libero per una dimostrazione su Translated.net).

Le dieci parole più frequenti (fig. 2.2) ci offrono un'indicazione di prima approssimazione sulle modalità retoriche del discorso di Obama, in particolare se le confrontiamo con le prime dieci parole che caratterizzano il secondo discorso del giuramento di George W. Bush del 20 gennaio 2005.

G.W. Bush	Occorrenze	% Freq.	B. Obama	Occorrenze	% Freq.
our	50	4,4	our	67	5,1
freedom	24	2,1	you	14	1,1
liberty	14	1,2	nation	12	0,9
you	12	1,1	new	11	0,8
america	12	1,1	those	11	0,8
your	12	1,1	must	8	0,6
every	10	0,9	every	8	0,6
nation	9	0,9	what	8	0,6
own	9	0,8	these	8	0,6
country	8	0,7	less	7	0,5

Fig. 2.2 Dieci parole più frequenti del discorso del giuramento dei presidenti G. W. Bush e Barack Obama del 20 gennaio 2005 e 2009 (elaborazione: Textalyser; *stoplist* attiva)

N-grams e keyword prominence

Per la vera e propria estrazione del contenuto sono molto più significativi i 3-grams, segmenti ripetuti di testo composti di tre forme grafiche, tra i quali troviamo subito espressioni che offrono delle precise indicazioni sui temi più rilevanti dei due discorsi (fig. 2.3).

G.W. Bush	Occ.	Prominence	B. Obama	Occ.	Prominence
the united states	4	72,7	and we will	3	76,1
we have seen	3	**93,5**	a new era	2	43,0
of our time	2	56,5	who seek to	2	44,3
do not accept	2	60,7	of our nation	2	44,5
america will not	2	72,4	we will not	2	60,6
america's influence is	2	72,8	that our power	2	63,1
of the world	2	75,7	of our economy	2	77,6
of ending tyranny	2	76,8	**say to you**	2	**83,3**
in our world	2	83,1	**a new age**	2	**85,8**
we have proclaimed	2	**84,6**	**time has come**	2	**87,2**
of liberty in	2	**87,5**	on this day	2	76,1

Fig. 2.3 Segmenti ripetuti composti di tre forme (3-grams) del discorso del giuramento dei presidenti G. W. Bush e Barack Obama del 20 gennaio 2005 e 2009 (elaborazione: Textalyser)

La *keyword prominence* misura la posizione media del segmento nel testo; più il segmento è vicino all'inizio del testo e più alto è il valore. Ad esempio la forma *we have seen* nel discorso di Bush con 3 occorrenze e 99.5 di *prominence* è presente nella prima parte del discorso e pertanto ha un valore più alto.

Dal confronto dei due discorsi si osserva immediatamente come i segmenti ripetuti siano molto diversi nella loro "composizione". Il discorso di Obama si caratterizza per un tono partecipativo (*and we will*; *our power*; *our economy*) e in asserzioni che mettono in evidenza la novità della sua elezione e il cammino da compiere: *time is come*; *a new age*; *say to you* sono i tre segmenti con rilevanza maggiore. Nel discorso inaugurale di Bush vi è una forte accentuazione del confronto tra libertà e tirannia, del ruolo dell'America nel presente e della sua vulnerabilità: quel *we have seen* che si riferisce implicitamente al tragico 11 settembre 2001.

```
Textalyser Results
The complete results, including complexity factor, and other features

Total word count :                                          1209
Number of different words :                                  729
Complexity factor (Lexical Density) :                      60.30%
Readability (Gunning-Fog Index) : (6-easy 20-hard)           10.1
Total number of characters :                               12123
Number of characters without spaces :                       7184
Average Syllables per Word :                                1.66
Sentence count :                                             108
```

Fig. 2.4 Secondo discorso del giuramento del presidente Barack Obama del 21 gennaio 2013 (elaborazione: Textalyser; min. car. 1; *stoplist* attiva - Fonte: Bartleby.com - testo a scorrimento)

Capitolo 3

Wordle visualizzare il peso delle parole

Wordle - Beautiful Word with Clouds è stato concepito come uno strumento per produrre rappresentazioni grafiche esteticamente gradevoli delle parole più utilizzate in un testo ma, utilizzato con cura, permette di ottenere anche una sintesi efficace del contenuto.

3

Wordle

Wordle fu sviluppato da Jonathan Feinberg come applet Java nel 2008 quando era ancora un ricercatore presso la IBM. Ne ritroveremo una versione semplificata tra gli strumenti di visualizzazione dei testi di Many Eyes. Tuttavia il modo più diretto e divertente di utilizzare la "nuvola delle parole" è sicuramente da questo sito in cui l'applicazione è ad accesso libero e con una quantità di opzioni che la rende molto flessibile (Feinberg, 2010).

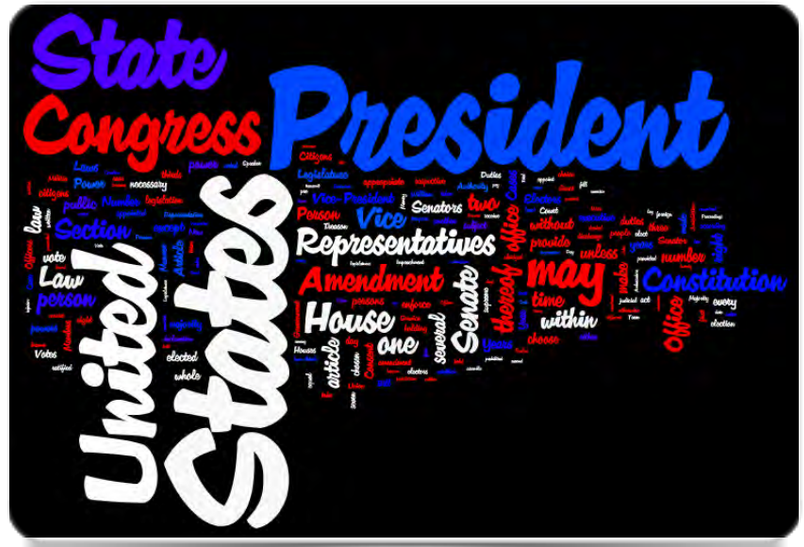


Fig. 3.1 US Constitution (Elab. Wordle)

In Wordle le parole più "pesanti" (con un numero maggiore di occorrenze) sono rappresentate proporzionalmente in una dimensione più grande. Il suo maggiore limite è che – dal punto di vista grafico – la dimensione può trarre in inganno, perché alcune lettere dell'alfabeto occupano uno spazio maggiore (per esempio la "o" e la "a") rispetto ad altre (tipicamente la "i" e la "l").

Wordle, cliccando su *Create* dalla home page, apre una finestra di acquisizione del testo con un semplice "copia e incolla", oppure con l'inserimento nell'apposito spazio di un indirizzo web o di una pagina dotata di RSS feed.

Cliccando su *Go*, il programma visualizza la nuvola delle parole con le opzioni predefinite. In questa fase di acquisizione è bene decidere subito, nella preparazione del corpus, se lasciare le parole così come sono nel testo (maiuscole e minuscole) oppure modificare tutto in maiuscolo o in minuscolo per evitare la duplicazione delle

parole per il solo effetto della diversa grafia in cui sono scritte (*Web*/*web*). Comunque questa opzione può essere attivata in seguito dal menu a discesa *Language*.

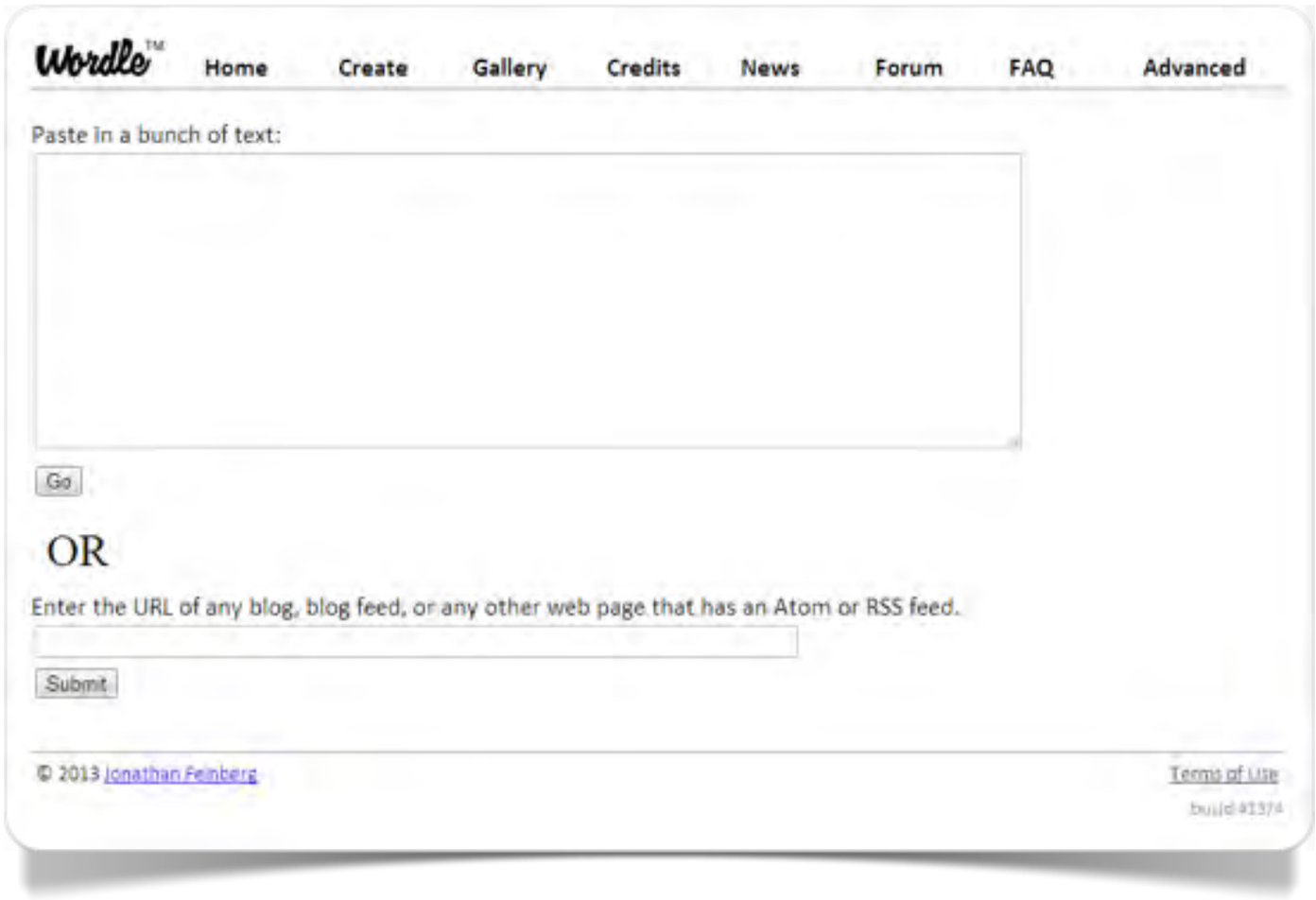


Fig. 3.2 Wordle: interfaccia di acquisizione del testo

In questo esempio (fig. 3.3) abbiamo scelto *The Universal Declaration of Human Rights* riducendo le forme grafiche in minuscolo. La prima visualizzazione è abbastanza efficace, anche per la posizione che ha assunto *right*, sovrastato da *equal* e appoggiato su alcune parole chiave fondamentali come *freedom*, *social*, *protection*, *law*,

everyone. Questo risultato, sebbene gradevole ed efficace, è del tutto casuale (spesso è necessario effettuare diverse prove per ottenere un'immagine soddisfacente in termini di comunicazione).

Fig. 3.3 *The Universal Declaration of Human Rights* (elab. Wordle con opzioni predefinite)

Opzioni di visualizzazione

La nuvola delle parole può essere modificata con le opzioni dei menu a discesa utilizzando diversi tipi di carattere, colori e ordinamento delle parole, ma queste modalità di rappresentazione non hanno alcun significato interpretativo, sono soltanto un espediente estetico.

Wordle di default riconosce le lingua in cui è scritto il testo ed elimina automaticamente le parole vuote (*stop words*). Questa opzione può essere modificata a cura dell'utente dal menu a discesa *Language*. Il carattere ~ (tilde; 007E in Unicode o 126 in caratteri ASCII) è interpretato dal programma come un carattere di unione (*will~be*): le parole sono rappresentate insieme nel grafico senza che la tilde stessa sia visualizzata (la stessa cosa vale per il carattere Unicode 00A0 o ASCII 160). Questa operazione di formazione delle "parole composte" o polirematiche deve essere effettuata in sede di preparazione del testo.

Il menu *Layout* permette di intervenire sul numero delle parole da rappresentare (di default: 150) e sulla loro disposizione nell'immagine: in ordine alfabetico o casuale; tutte in orizzontale, tutte in verticale o miste; con il bordo della nuvola arrotondato o in linea retta. Una parola può essere cancellata dal grafico posizionandosi su di essa e cliccando con il tasto destro del mouse. In questo caso è utile eliminare la parola *article*, con 30 occorrenze (una per ogni articolo della dichiarazione), ma priva di contenuto ai fini della rappresentazione grafica del documento. Nei menu a discesa *Font* e *Color* si possono selezionare i caratteri grafici e la palette dei colori. La visualizzazione in figura 3.4 è stata realizzata con eliminazione della parola *article*, con *font Teen*, massimo di parole 80 e combinazione dei colori *Firenze*.

Fig. 3.4 *The Universal Declaration of Human Rights* (elab. Wordle con opzioni Font/Teen, Maximum words 80, Color palette Firenze)

Integrazione tra Textalyser e Wordle

Wordle, dalla pagina di acquisizione *Advanced*, permette anche di elaborare i dati testuali che provengono da un formato “tabella” (o Excel) in due colonne separate dal segno di interpunzione “due punti”. Nel nostro esempio (fig. 3.5) è stata inserita una tabella dei 3-grams più significativi del discorso di Barack Obama già analizzato con Textalyser (fig. 2.3). In questo esempio le occorrenze sono state sostituite con l’indicatore di *keyword prominence*. Il risultato che si ottiene è molto simile a una sintesi riassuntiva dei temi più rilevanti.

Fig. 3.5 Discorso del giuramento di Barack Obama del 20 gennaio 2009; 3-grams secondo la *keyword prominence* (elab. Textlayser e Wordle)

Capitolo 4

Tagxedo visualizzazione creativa

Tagxedo - Word Cloud with Styles è stato sviluppato da Hardy Leung tenendo conto dell'esperienza di Wordle ma con intenti più artistici e ludici.

4

Tagxedo

Fig. 4.1 Cats (elab. Tagxedo)

La visualizzazione di Tagxedo è sempre basata sulla nuvola di parole ma le opzioni grafiche e le potenzialità di gestione del'output lo rendono molto più versatile. La prima novità introdotta da Tagxedo, già alla sua prima apparizione, fu la disponibilità per l'utente di scegliere la forma della nuvola di parole, con gradevoli risultati estetici e una notevole efficacia comunicativa, come dimostrano gli esempi disponibili in Gallery.

Le modalità di acquisizione del testo, oltre a quelle classiche del copia e incolla, dell'upload di un file o dell'inserimento di un URL, comprendono anche l'accesso tramite RSS feed di Google News o gli ultimi 100 post di un ID di Twitter.

Fig. 4.2 Tagxedo: interfaccia di acquisizione del testo

Tagxedo Creator (fig. 4.2) presenta una serie di opzioni che permettono di fare praticamente tutto con la visualizzazione delle parole: dalla scelta della palette dei colori (*Theme*), al font dei caratteri, all'orientamento delle parole (verticale, orizzontale o misto), con la possibilità di bloccare una delle opzioni quando il risultato è ritenuto soddisfacente per proseguire nella sperimentazione delle opzioni successive.

L’opzione *Shape* offre una vasta scelta di forme, mentre con l’opzione *History* l’applicazione tiene traccia di tutte le visualizzazioni effettuate con il testo in lavorazione.

Fig. 4.3 Tagxedo Creator

Opzioni di visualizzazione

Il menu *Word – Layout options* (fig. 4.4) apre una serie di schede con modalità di intervento molto sofisticate. Nella scheda *Skip* è possibile selezionare, tra le prime 200 forme grafiche del testo in ordine di frequenza, quali utilizzare per la visualizzazione finale. Diversamente da Wordle, Tagxedo utilizza solo una lista di *stop words* in inglese, ma con *Skip* è sempre possibile cancellare dal vocabolario le parole che l'utente ritiene non significative.

Nella scheda *Layout* l'opzione *Normalize Frequency* assegna una frequenza "teorica" a ciascuna parola secondo l'ordine di occorrenze. Il valore di *Spread* (di default fissato a 40) è il rapporto tra la frequenza più alta e la più bassa. Più alto è lo *spread* e maggiore è la dimensione che assumono le parole con frequenza più alta.

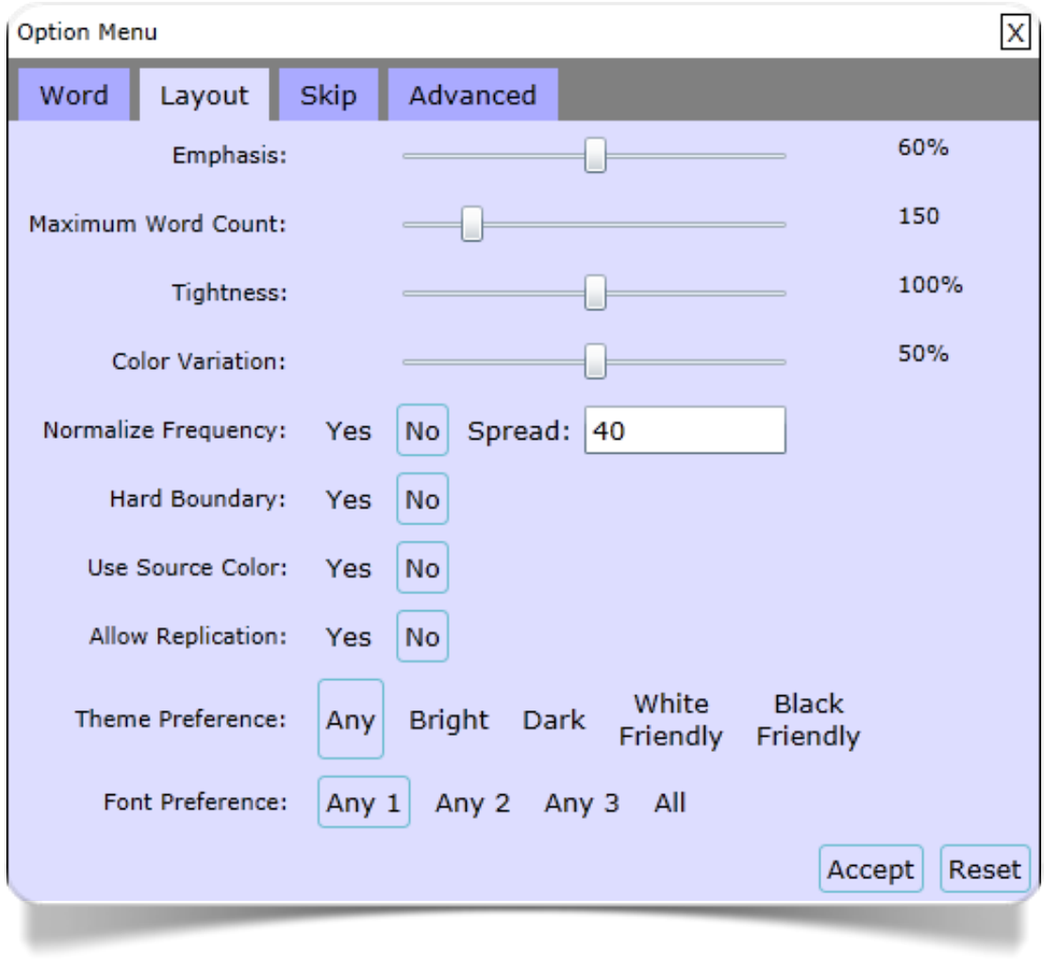


Fig. 4.4 Tagxedo creator: Option menu

Tagxedo non dispone di un vero e proprio tutorial, ma la pagina Facebook offre una quantità di suggerimenti utili e di spiegazioni. Altri esempi e applicazioni si trovano in *101 Ways to Use Tagxedo*.

La figura 4.5 si riferisce all'iniziativa del *New York Times online* che, in occasione del decennale dell'11 settembre, ha chiesto ai suoi lettori di rispondere alla domanda: *Where Were You on September 11, 2001?* Le risposte sono state 38.000.

Fig. 4.5 Visualizzazione della survey del *New York Times*: *Where Were You on September 11, 2001*? (elaborazione di Hardy Leung con Tagxedo)

Capitolo 5

Many Eyes strumenti per l'analisi dei testi

Many Eyes è un laboratorio online della IBM che mette a disposizione degli utenti alcuni potenti strumenti di analisi e visualizzazione sia di dati testuali che di dati numerici. Qui ci occupiamo della visualizzazione dei dati testuali.

5

Many Eyes

Fino a ottobre 2014 su Many Eyes si potevano liberamente utilizzare, elaborare e visualizzare i testi e i file già caricati sul database del sito oppure, dopo aver effettuato l'iscrizione gratuita procedere all'upload del proprio corpus preparato in precedenza. Ora con l'aggiornamento in corso questo e ancora possibile ma con molte limitazioni. Gli strumenti di visualizzazione dei testi - attualmente - si sono ridotti al solo Word Cloud Generator.

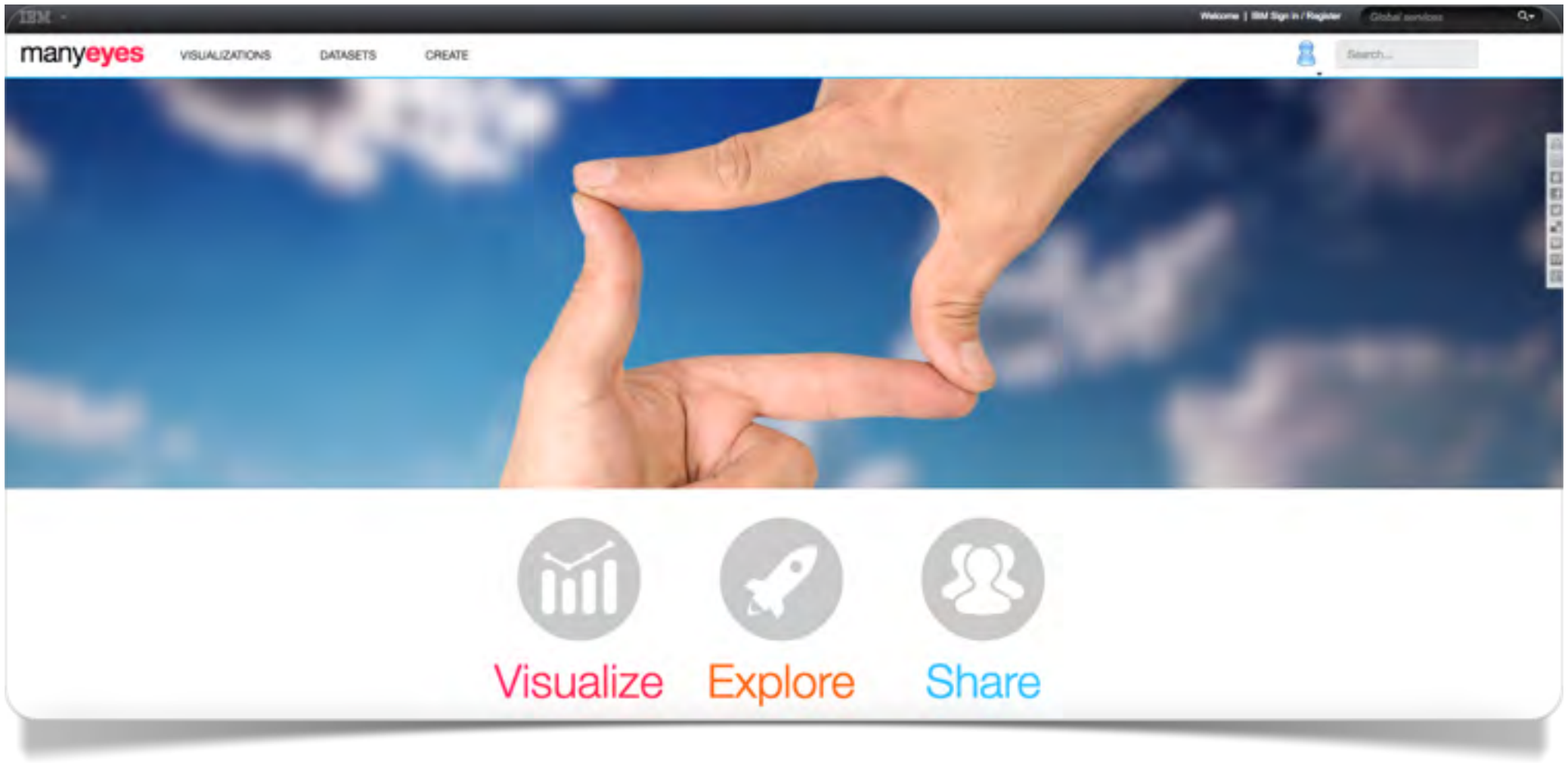


Fig. 5.1 La home page di Many Eyes 2015

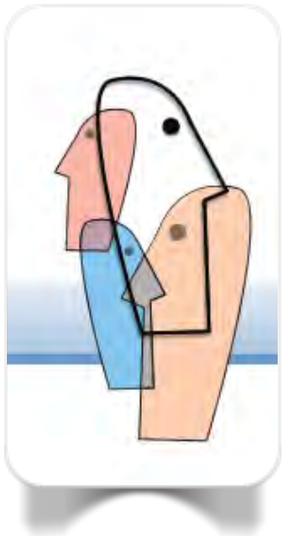


Fig. 5.2 Logo di Many Eyes fino al 2014

Nell'aggiornamento, almeno allo stato attuale, sono stati privilegiati gli strumenti di analisi e visualizzazione dei dati numerici o dati strutturati.

Negli esempi seguenti saranno illustrate tutte le modalità di visualizzazione che in passato erano liberamente accessibili. Dal punto di vista metodologico cerchiamo, in questo modo, di mantenere almeno la memoria di tools che, purtroppo, non sono replicabili in nessun altro sito di elaborazione dei testi online. Le immagini fanno pertanto riferimento alla versione di Many Eyes antecedente all'aggiornamento. La speranza è che tutte le operazioni vengano ripristinate al più presto sul sito.

Dal menu *Partecipate - Create a visualisation* le indicazioni del tutorial ci avrebbero guidato nella selezioni di un corpus (ora non più disponibile) di sicuro interesse: *Facebook Statuses Containing 'muslim', 'obama', and '9 11'.* Si trattava di una collezione di stati personali estratti da Facebook in prossimità del decimo anniversario dell'attentato alle Twin Towers con riferimenti al presidente USA, a "musulmano" e all'11 settembre.

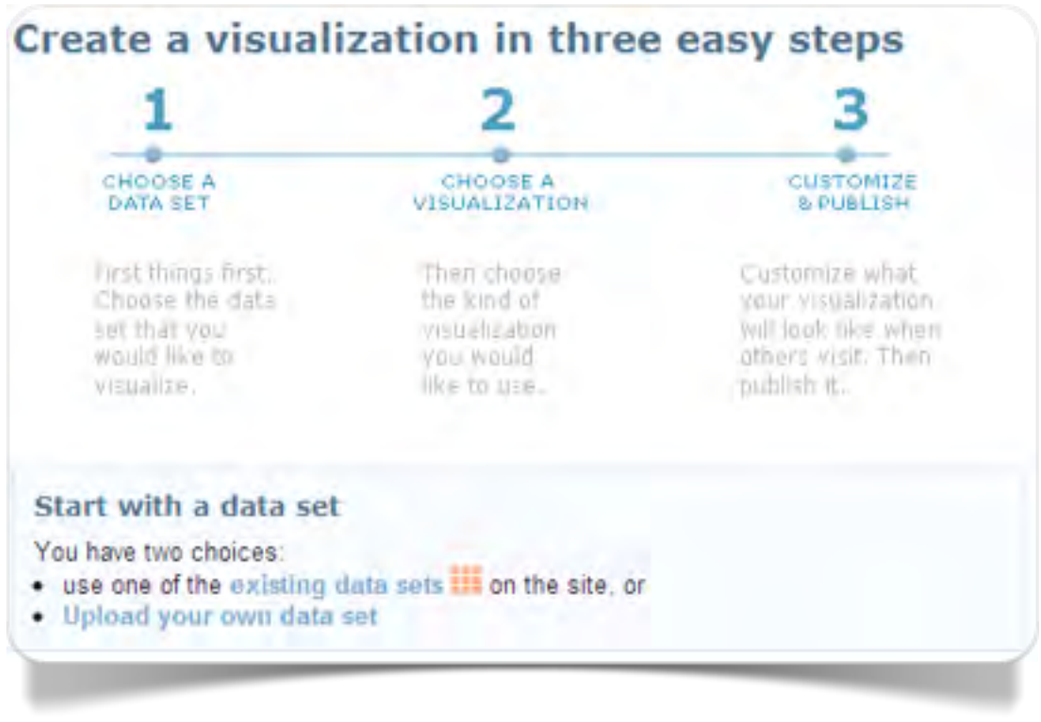


Fig. 5.3 Many Eyes: steps di creazione delle visualizzazioni

Dalla pagina di visualizzazione del file di testo cliccando su *View as text* era possibile esaminare le caratteristiche dell'intero corpus. Cliccando su *Visualize* all'utente erano offerti diversi strumenti a seconda della

modalità di organizzazione del data set. La modalità di visualizzazione erano quattro:

Word Tree permetteva di analizzare il testo sulla base di una classificazione ad albero delle concordanze, collocando la parola selezionata come pivot nel contesto delle parole che la seguono o la precedono.

Tag Cloud visualizzava le parole e i 2-grams (segmenti ripetuti di due parole consecutive) in una dimensione che dipende dalle occorrenze: una variante di Wordle per il confronto tra due testi (partizioni) dello stesso corpus.

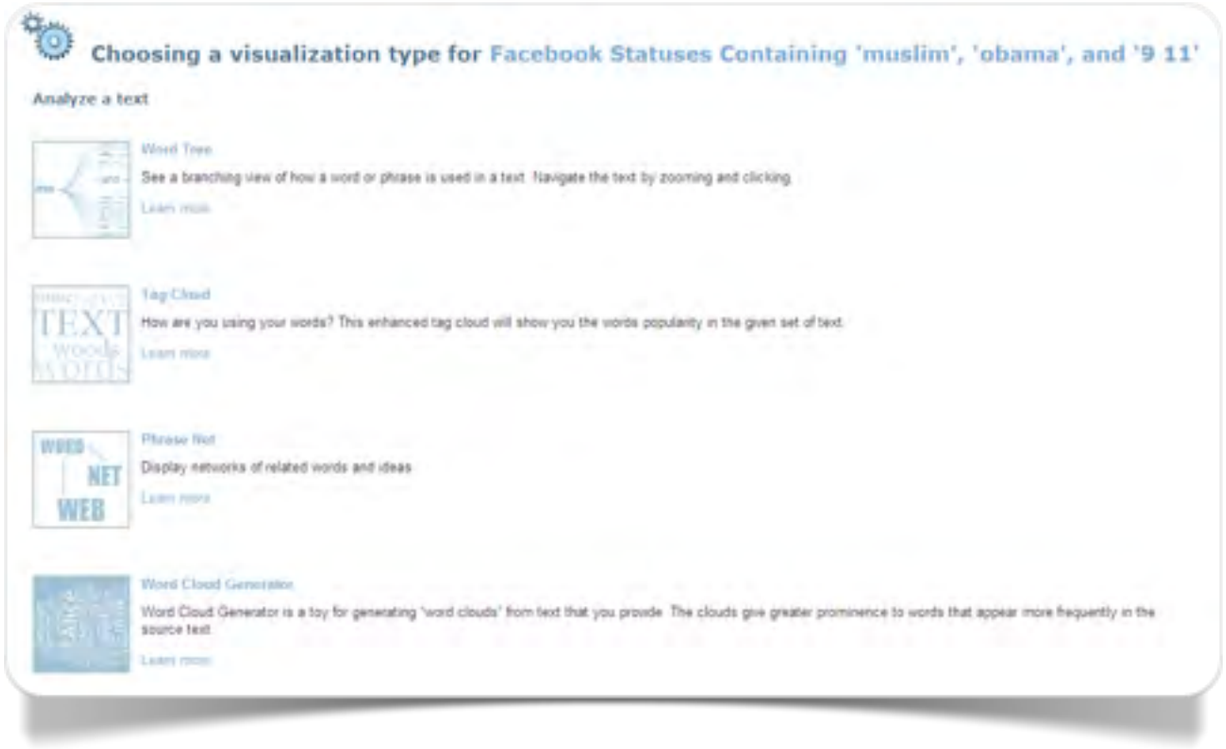


Fig. 5.4 Scegliere il tipo di visualizzazione

Phrase Net era un diagramma della rete di parole in relazione tra loro tramite parole “ponte” o “link”.

L’unico che rimane attivo attualmente è :

Word Cloud Generator, una versione semplificata di Wordle, sviluppata da Jonathan Feinberg quando era ricercatore della IBM. La versione attuale, tuttavia, è del tutto snaturata e inutilizzabile per una visualizzazione seria di un testo. Per esempio non permette di selezionare la soglia di occorrenza delle parole e non elimina in modo automatico le parole vuote. Sono rimste attive solo le opzioni sul font e sull’orientamento delle parole nella “nuvola”.

Word Tree

In *Word Tree* le concordanze erano classificate per formare un “albero di parole”: diramazioni delle sequenze da una parola base (parola pivot) che ne costituisce il tronco, come accade nelle fronde di un albero (inserendo nel campo *Search* la parola selezionata con l’opzione *Start*), oppure ramificazioni delle sequenze che precedono la parola, come accade nelle radici dell’albero (attivando l’opzione *End*).

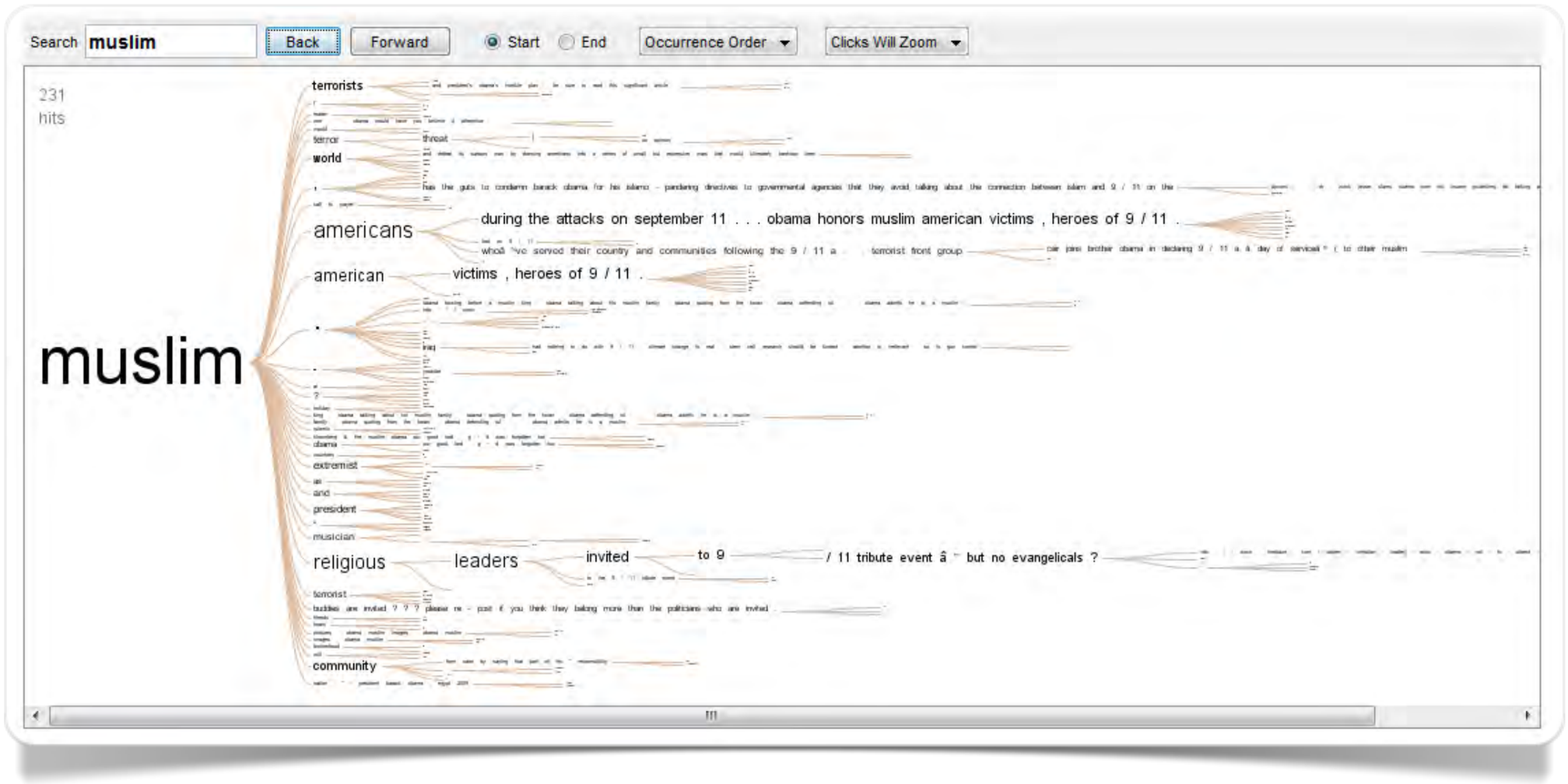


Fig. 5.5 Word Tree della forma *muslim* [Start] nel corpus 9-11 (elab. IBM’s Many Eyes)

Anche in questo caso, come in una *word cloud*, le parole e le sequenze erano rappresentate con grandezze proporzionali alle occorrenze (*muslim* = 231 occorrenze). La visualizzazione era dinamica e modificabile con ordinamento delle sequenze in ordine alfabetico o di frequenza. Inoltre tutte le sequenze si presentavano in modalità navigabili in modo da evidenziarne il contesto semantico. La figura 5.6, per esempio, è stata ottenuta dalla visualizzazione della figura 5.5 cliccando su *world* (*muslim world* = 6 occorrenze).

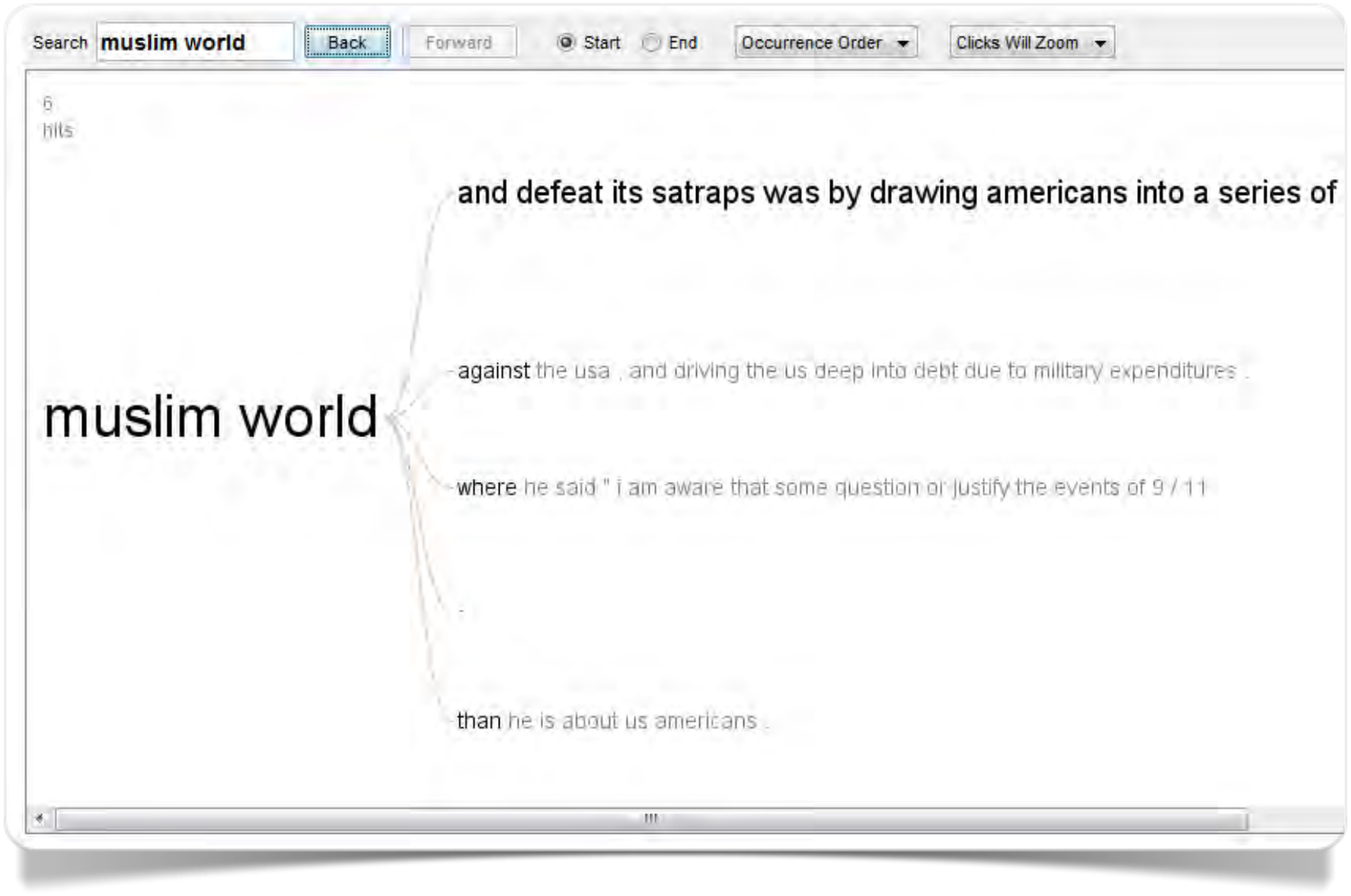


Fig. 5.6 Word Tree della forma *muslim world* [Start] nel corpus 9-11 (elab. IBM's Many Eyes)

Tag Cloud

Tag Cloud permetteva di visualizzare la frequenza delle parole e dei 2-grams (segmenti ripetuti di due parole consecutive). Posizionando il mouse su una parola si evidenziavano le occorrenze delle parole nel contesto. La novità più interessante di *Tag Cloud* rispetto a *Wordle* e *Word Cloud Generator* era la possibilità di confrontare due testi all'interno di un corpus, separandoli con un apposito marcatore. Nelle'sempio di figura 5.7 sono messi a confronto i discorsi inaugurali di George W. Bush del 20 gennaio 2005 e di Barak Obama del 20 gennaio 2009.

La visualizzazione delle parole e la loro posizione permette di osservare immediatamente le parole che sono presenti solo nel discorso di Bush (in rosso) o di Obama (in blu). Le parole comuni sono affiancate e la grandezza della parola ne indica la frequenza in ciascun testo. Posizionando il mouse sulla parola (ad es. *liberty*) si osservano le occorrenze (15 in Bush e 2 in Obama) e le relative concordanze per ciascun testo.

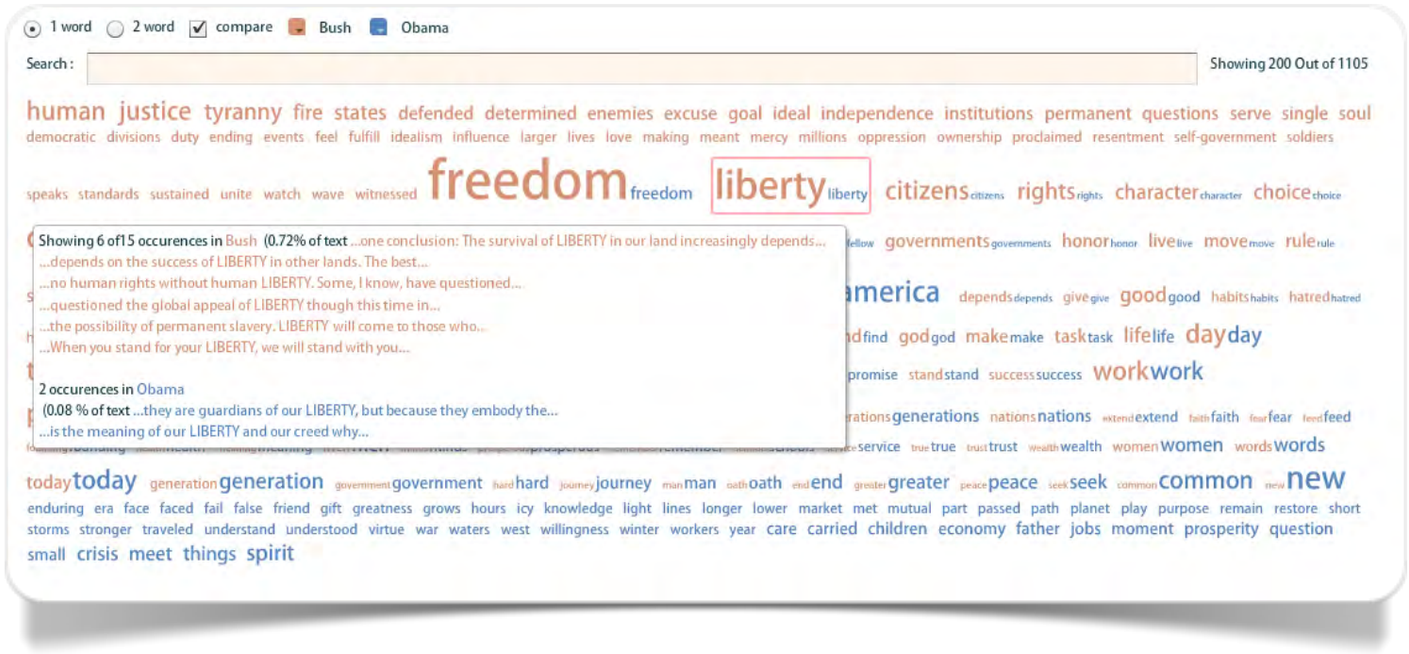


Fig. 5.7 – Confronto tra le 200 parole più frequenti dei discorsi inaugurali dei presidenti degli Stati Uniti G.W. Bush - 20 gennaio 2005 - e B. Obama - 20 gennaio 2009 (elaborazione Tag Cloud in IBM's Many Eyes)

Phrase Net

Phrase Net visualizzava il diagramma della rete di parole in relazione tra loro tramite la selezione di una parola o un carattere “ponte”. Lo strumento prevedeva una selezione di link predefiniti, ma con l’opzione della finestra *Enter your own* (fig. 5.8) era possibile adattare le opzioni per qualsiasi lingua. Il risultato della figura 5.9 è stato ottenuto selezionando come link lo spazio [*space*] tra una parola e l’altra con una visualizzazione delle 30 sequenze più frequenti. Il setting comprendeva una lista di *common words* (*stop words*) da escludere. La visualizzazione, navigabile e personalizzabile, permetteva di rendere più comprensibile e interpetabili le relazioni più intricate, come in questo caso.

Phrase Net era particolarmente efficace nella individuazione dei segmenti ripetuti che hanno una funzione centrale nella ricostruzione dei concetti che rappresentano il contenuto del testo.

In questo caso (fig. 5.9) emerge con chiarezza la centralità di *muslim* rispetto a *Obama* (*muslim american/americans*, *muslim religious leaders*, *muslim terrorist*, *muslim islamic attack*). La direzione della freccia indica la parola che precede e segue nel segmento (*9 11 memorial*).

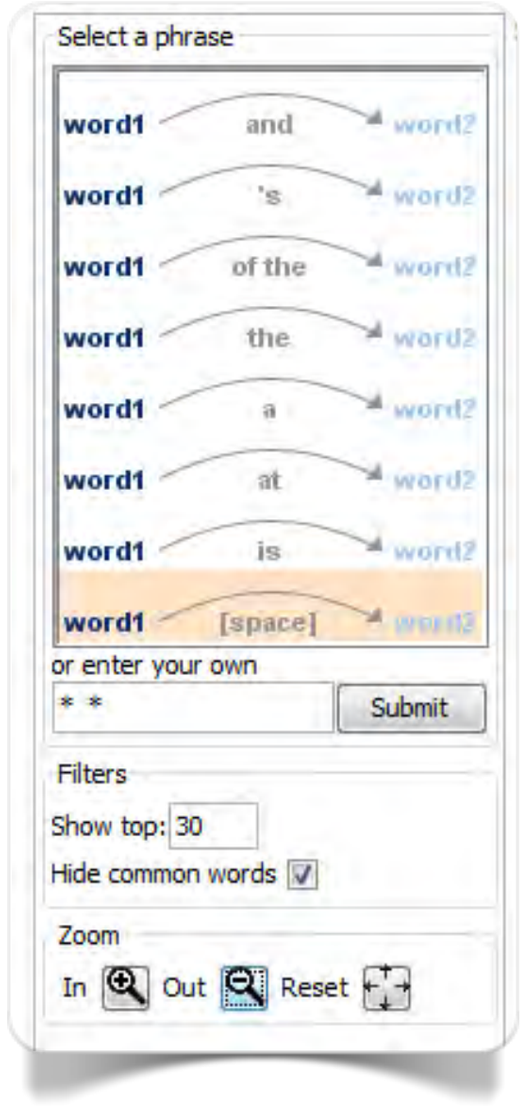


Fig. 5.8 Phrase Net: opzioni di selezione (IBM’s Many Eyes)

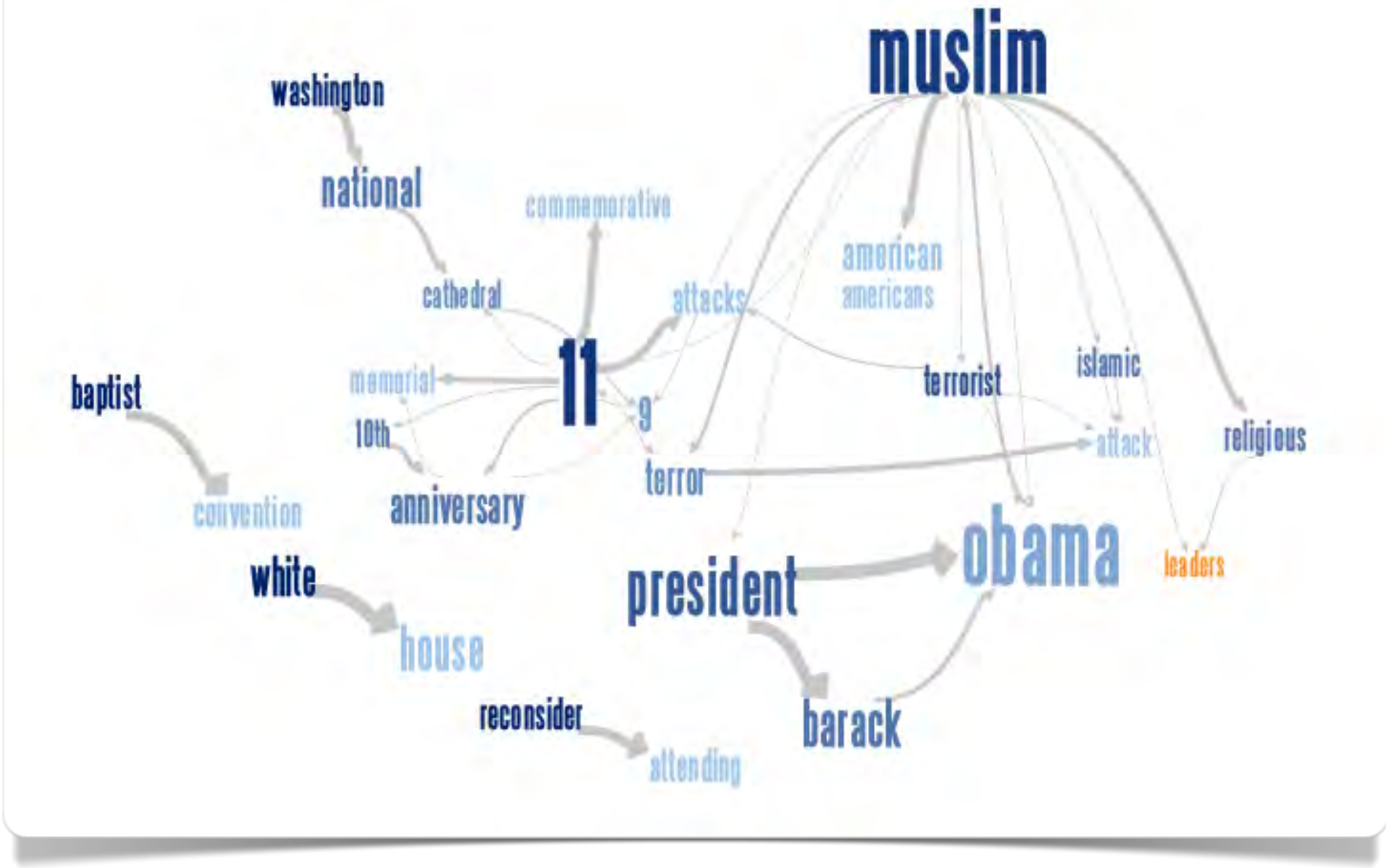


Fig. 5.9 Phrase Net del corpus 9-11 : opzione: [space] *30 top matches* (IBM's Many Eyes)

Word Cloud Generator

Word Cloud Generator era una versione semplificata di Wordle e pertanto non richiede ulteriori spiegazioni. Nella figura 5.10 possiamo osservare la visualizzazione delle 100 parole più frequenti del *corpus 9-11*. Sono state eliminate le parole “banali” che sono servite per selezionare gli status di Facebook: *Obama*, *muslim* e *9 11*.

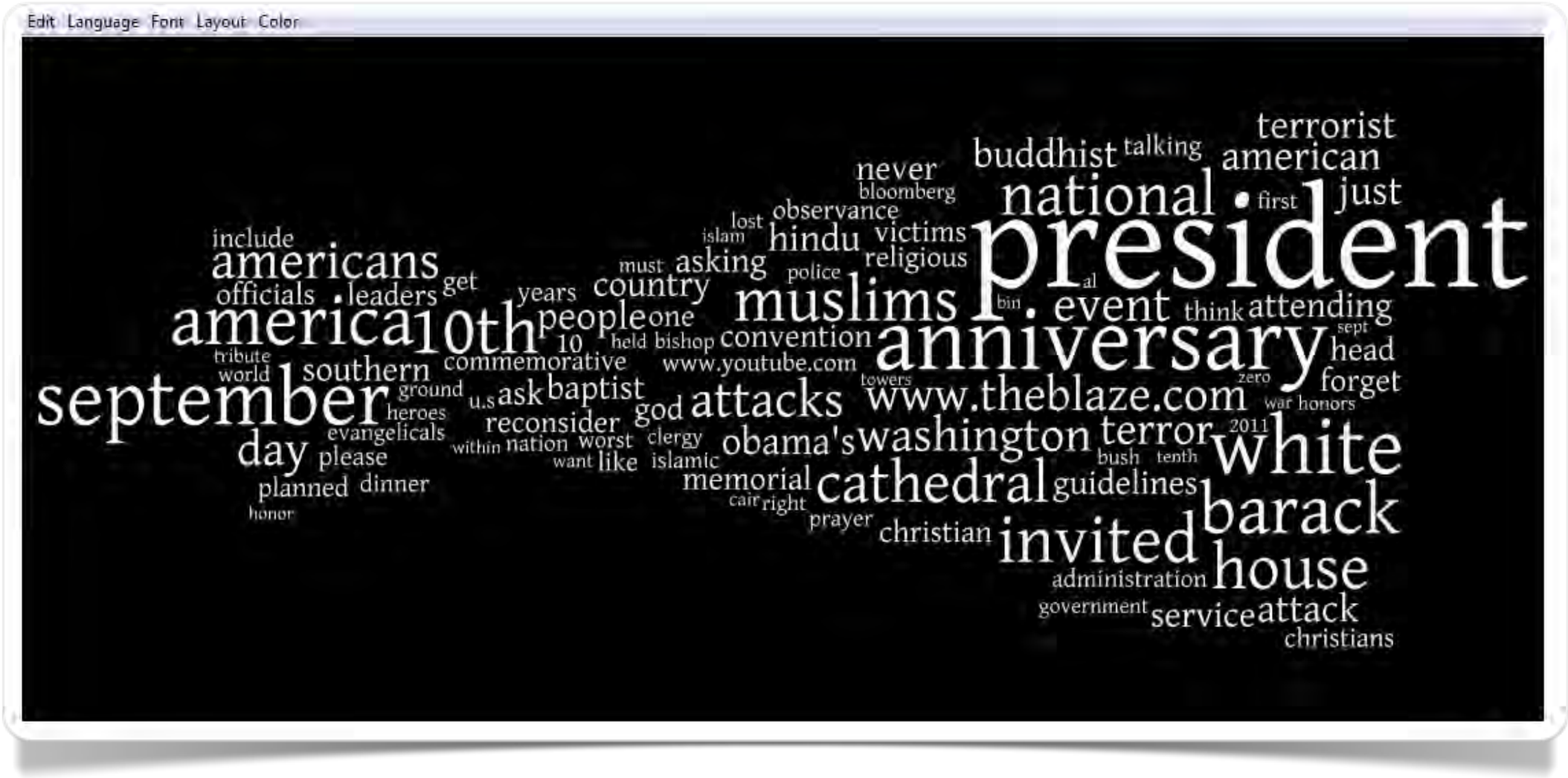


Fig. 5.10 Word Cloud Generator del corpus 9-11 (elab. IBM’s Many Eyes)

Capitolo 6

Netlytic Internet Community Text Analyzer

Netlytic è un software online creato con lo scopo di fornire uno strumento per l'analisi di email, newsgroup, forum, blog e, in generale, dei messaggi che fanno capo ai social network.

6

Netlytic

Netlytic di Anatoliy Gruzd (School of Information Management, Dalhousie University, Canada) è ad accesso gratuito ma, per l'elaborazione dei testi, richiede l'iscrizione al sito (Gruzd, 2010). Prevede anche una forma di accesso a pagamento che permette di gestire fino a 100 dataset di 300.000 record ciascuno. Per gli accessi gratuti ad uso limitato si possono utilizzare gli account di Gmail o di Yahoo!

I dataset possono essere importati da *New Dataset* nel seguente modo (System Overview):

1. Attraverso RSS feeds.

2. Con upload da sistemi di cloud storage come Google Drive o Dropbox.

3. Da account Twitter, utilizzando parole chiave, hashtag o @username.

4. Da YouTube e Instagram, importando i commenti ai video e alle immagini.

5. Da file di testo in formato csv utilizzando oppure con i formati standard della posta elettronica.

Per esempio Netlytic identifica ogni partizione di testo che inizia con *From:* come un messaggio email. Il modo più semplice per preparare un corpus per l'analisi consiste nel controllare che questo formato sia rispettato (anche con un *From:* fittizio come intestazione di ciascun messaggio/testo). Nell'esempio che segue, il corpus è costituito da 529 messaggi (13 maggio – 7 luglio 2009) di un gruppo di discussione in Facebook sul tema: *Fire fighters are heroes serving the community, soldiers are trained assassins serving the big corporations!* Il copia e incolla da Facebook non ha il formato richiesto, pertanto il file deve essere preparato inserendo *From:* tra un messaggio e l'altro:

From:

Claudia Green FIRE FIGHTERS ARE HEROES SERVING THE COMMUNITY, SOLDIERS ARE TRAINED ASSASSINS SERVING THE BIG CORPORATIONS! WAKE UP NAIVE AND NARROW MINDED SUCKERS! PATRIOTISM TRANSLATES STUPIDITY!

From:

Kyle Eberhardt Lady we fight for eachother and your family not the government that employs us or thier cronnies. All the BS is coming from within Washington DC your own government. The Us millitary is the last half way honest thing this country has. Revolution is coming and most of us side with the PEOPLE that mean you. So help us take back our country, because Washington is not going to help you put out the fire when the time comes . Shure they'll take your money and promiss to do a good job but look how that's turn out

From:

Jerry Weed What an ignorant person you are. If you live in the US and think what we do for you is wrong then get out of our country! Go live over in Iraq and Afghanistan and then tell us we're naive and narrow minded.

From:

Annabel Ward To be fair that's probably what the Iraqis think "get out of our country"-not really a great line of defence for a nation that goes wherever it chooses.

[...]

Dopo aver effettuato l'upload più adeguato allo scopo, selezionando il dataset da analizzare troviamo due strumenti di analisi principali:

Keyword Extractor: individua le forme grafiche più frequenti dopo aver eliminato le parole banali e le *stop words* il cui elenco, di oltre 500 termini, è consultabile da un link attivo nella finestre di elaborazione con l'icona .

Categories: classifica alcune forme grafiche o segmenti di testo secondo determinati criteri di contenuto in modo da permetterne una rapida visualizzazione, anche quantitativa.

Netlytic dispone di diverse opzioni per la normalizzazione del testo al fine di permettere all'utente di eliminare le fonti di rumore molto frequenti nella comunicazione mediata dal computer. Questa operazione può essere condotta preliminarmente nella fase di acquisizione e upload del corpus oppure al termine di una prima analisi esplorativa all'interno degli strumenti illustrati in precedenza. L'applicazione esemplificativa al corpus di Facebook permetterà di comprendere più in dettaglio questi passaggi.

Fig. 6.1 La schermata di Netlytic con le schede dei diversi passaggi di analisi e visualizzazione

Keyword Extractor

Dal menu con lo strumento *Keyword Estractor* otteniamo il risultato di figura 6.2 (molto simile a *Tag Cloud* di Many Eyes) in cui le parole più frequenti sono rappresentate con una dimensione “proporzionalmente” più grande. I nomi propri sono stati eliminati manualmente operando direttamente sulla crocetta rossa a destra di ciascuna forma grafica. La visualizzazione grafica di questo risultato (fig. 6.3), che prende in esame le 25 forme grafiche più

Fig. 6.2 Forme grafiche estratte (Top 100) con Keyword Extractor nel gruppo di discussione Facebook “Fire Fighters are heroes”; 13 maggio- 7 luglio 2009 (elaborazione Netlytic)

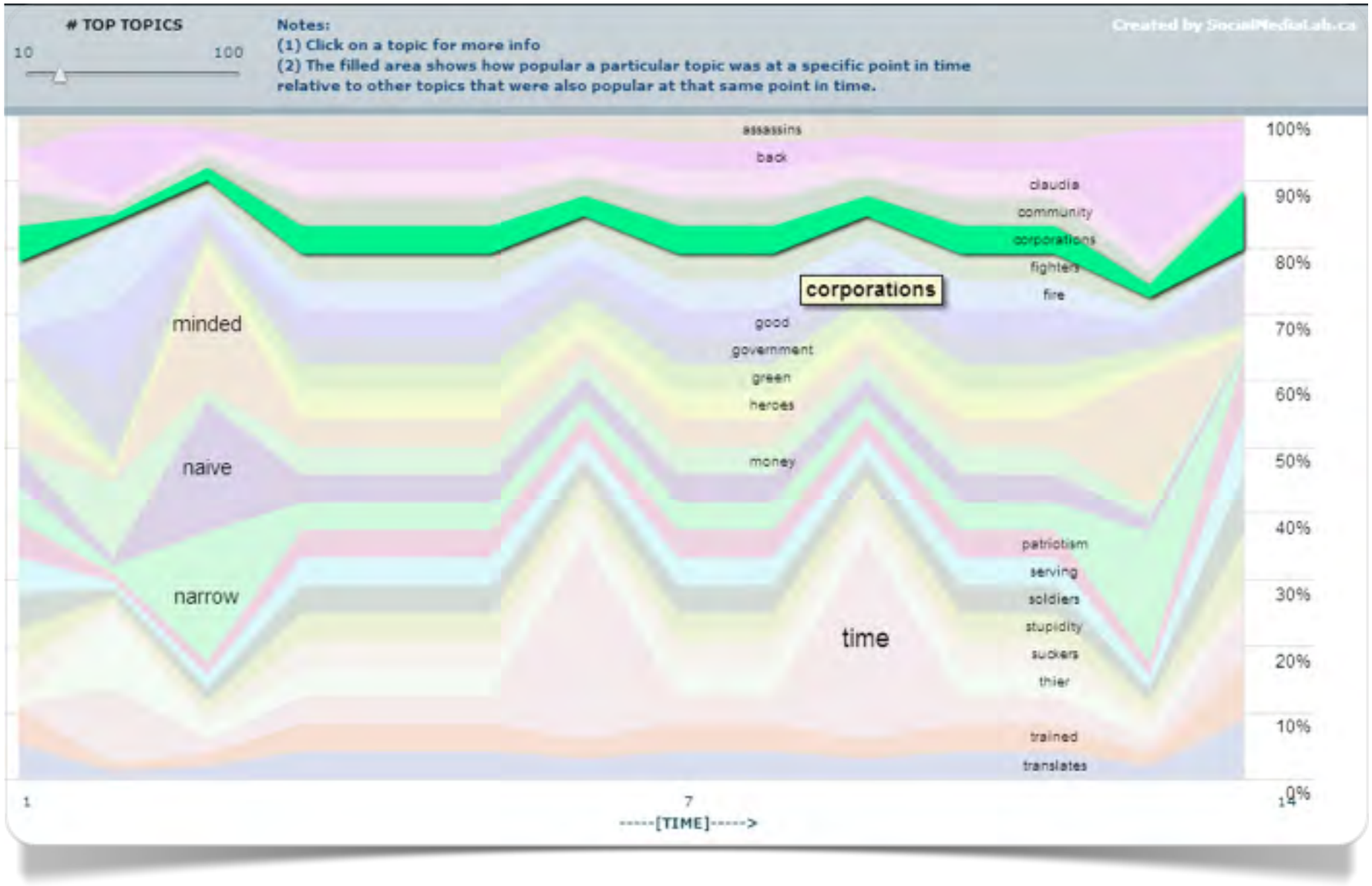


Fig. 6.3 Visualizzazione secondo il flusso temporale delle forme grafiche estratte (Top 25) con Keyword Extractor nel gruppo di Facebook “Fire Fighters are heroes”; 13 maggio- 7 luglio 2009 (elaborazione Netlytic)

rilevanti, permette di apprezzare la presenza dei “concetti” estratti per tutta la durata delle discussione. Il grafico mostra che all’inizio di maggio le parole chiave più rilevanti della discussione erano *minded*, *naïve*, *narrow* (con riferimento critico al modo in cui era affrontato il tema), mentre alla fine di giugno il focus si fa più articolato con

parole come *corporation*, *fighter/s* e *money*. La percentuale indica quanto è frequente una particolare parola in uno specifico momento rispetto alle 25 parole utilizzate più frequentemente nello stesso momento.

Dalle forme presenti nella figura 6.3 possiamo osservare come alcune argomentazioni facciano riferimento ai presidenti USA (Bush e Obama) e alla multinazionale americana Hulliburton specializzata nello sfruttamento dei giacimenti petroliferi e fortemente indiziata di azioni illegali nella guerra in Iraq. Significativa è la presenza di parole come fighter/s, money, narrow, stupidity al fine di esprimere l'indignazione di molti partecipanti alla discussione per l'associazione che alcuni fanno tra le parole soldiers e assassins.

Fig. 6.4 Visualizzazione delle forme grafiche estratte (top 50) in Twitter: 13-18 luglio 2014; keyword: #Israel (elab. Netlytic)

La percentuale sull'asse dell'ordinata rappresenta la distribuzione dell'argomentazione in un momento specifico secondo la frequenza della parola rappresentata. Nel secondo esempio (fig. 6.4 e fig. 6.5) possiamo osservare i risultati dell'analisi di 12.000 tweets selezionati con l'hashtag #Istrael nella settimana dell'intervento armato nella striscia di Gaza tra il 13 e il 18 luglio 2014.

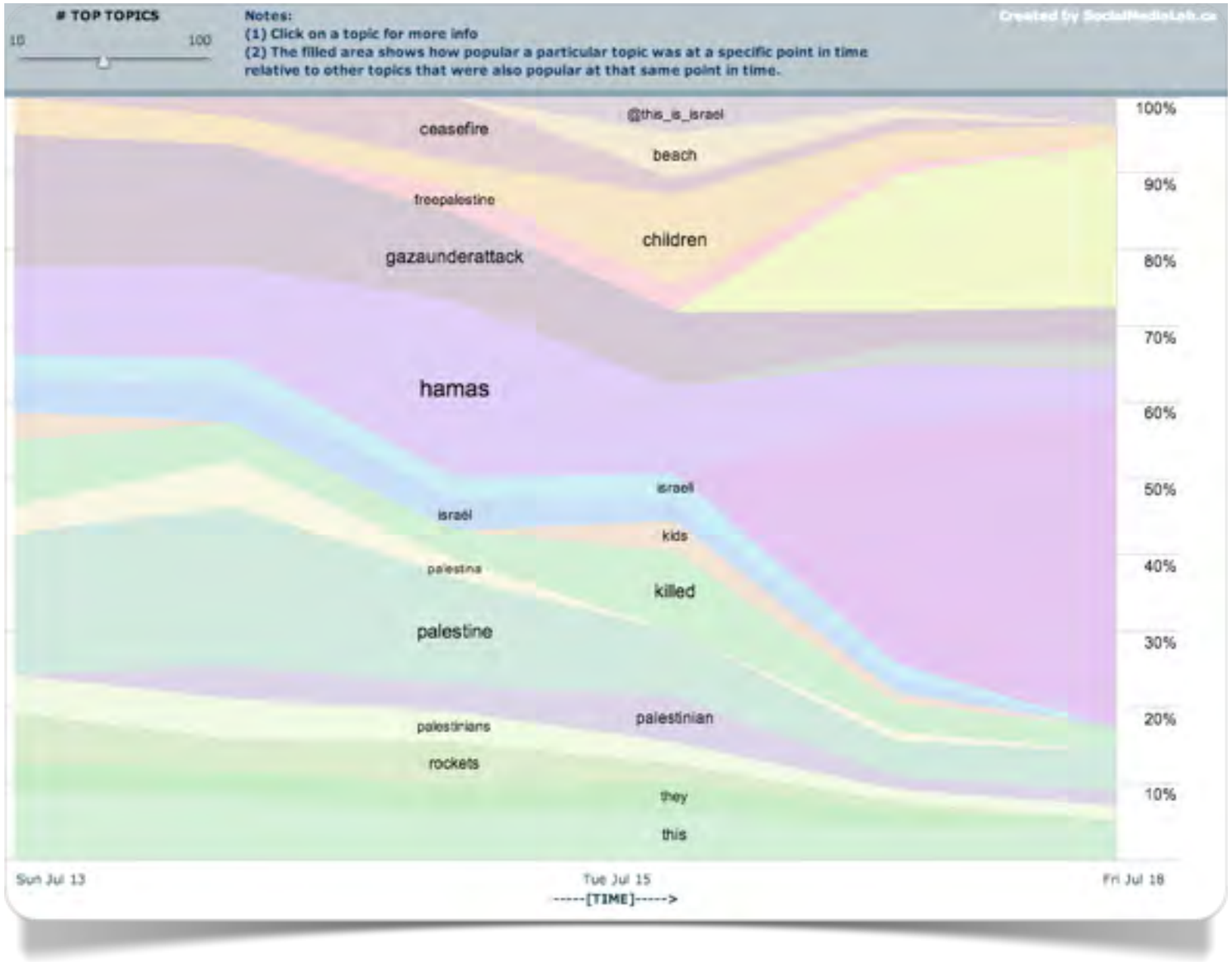


Fig. 6.5 Visualizzazione secondo il flusso temporale delle forme grafiche estratte (top 50) in Twitter: 13-18 luglio 2014; keyword: #Israel (elab. Netlytic)

Categories

Come si è detto lo strumento *Categories*, cui si accede dal menu *Text Analysis*, permette di classificare parole, polirematiche o frasi secondo criteri linguistici o semantici fino a ricostruire, anche secondo una dimensione quantitativa, i tratti di sintesi dei contenuti trattati o del tono generale della comunicazione (Pennebaker e Graybeal, 2001).

Le categorie principali predefinite sono: *agreement*, *certainly*, *disagreement*, *evalutation*, *opinion*, *positive*, *reference*, *self*, *uncertainty*, *us*. L'utente può aggiungere o modificare le categorie preesistenti creandosi una lista personale di parole e di criteri di classificazione. Questo strumento permette di applicare al corpus una vera e propria procedura semi-automatica di Analisi del contenuto (Losito, 2002; Krippendorf, 2004).

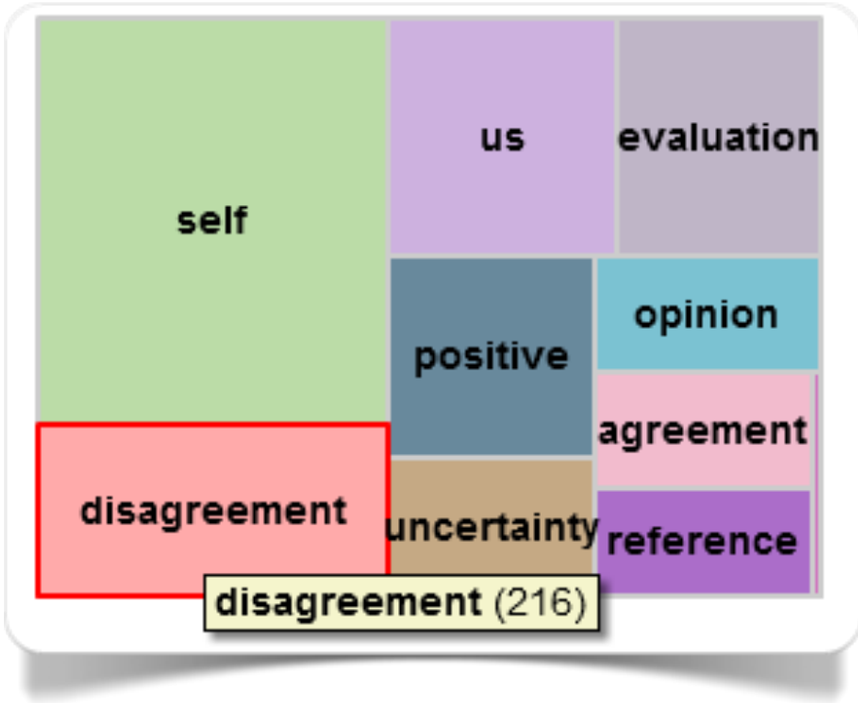


Fig. 6.6 Visualizzazione delle Cognitive & Social Categories del corpus di Facebook "Fire Fighters are heroes"; 13 maggio- 7 luglio 2009 (elaborazione con Netlytic)

In questo caso l'elaborazione con lo strumento *Categories* evidenzia bene che ci troviamo all'interno di un confronto di opinioni in cui gli elementi di disaccordo sono maggiori rispetto a quelli di accordo. Per la visualizzazione del grafico (fig. 6.3) è necessario cliccare sul menu *Home*, selezionare *Visualize* in corrispondenza del corpus in elaborazione e poi selezionare la scheda *Categories*.

Nella Gallery 6.1 possiamo osservare l'analisi per categorie dei tweets con riferimento al conflitto tra Hamas e Israele e alcuni screenshot in cui sono approfondite le categorie *Condition* e *Feelings (bad)*.

Gallery 6.1 Visualizzazione dell'opzione Categories nel corpus Twitter del 13-18 luglio 2014; keyword: #Israel (elab. Netlytic)

Fig. 6.1.1 Categorie

Visualizzazione delle concordanze

In tutti gli strumenti presentati, puntando con il cursore sulle categorie o sulle parole, è sempre possibile attivare la visualizzazione delle concordanze (fig. 6.7) e del contesto complessivo (messaggio o testo) in cui appare la parola per facilitare l'interpretazione del risultato ottenuto (Haythornthwaite e Gruzd, 2007).

Fig. 6.7 Visualizzazione delle concordanze e del contesto della parola heroes nel corpus di Facebook "Fire Fighters are heroes"; 13 maggio- 7 luglio 2009 (elaborazione con Netlytic)

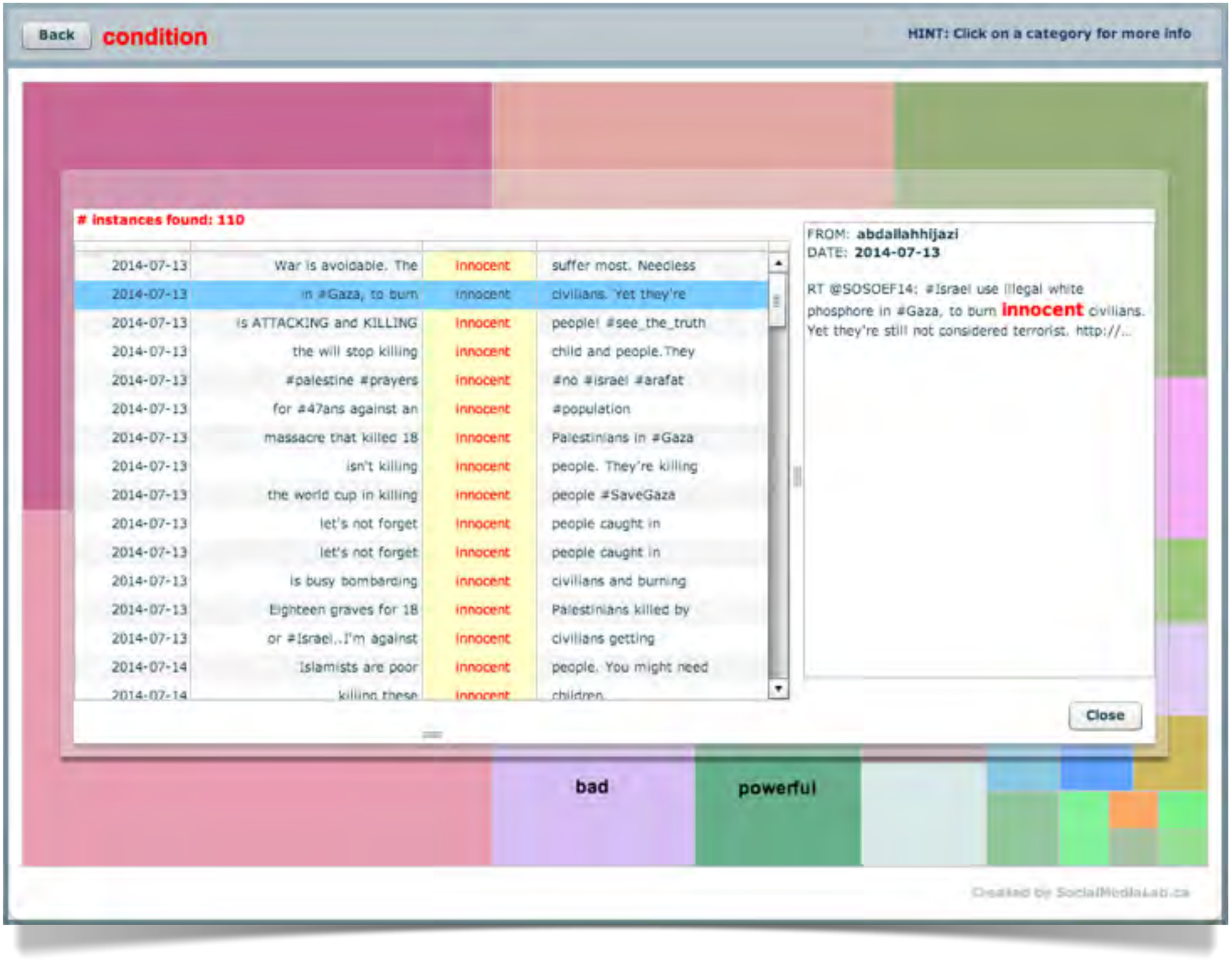


Fig. 6.8 Visualizzazione delle concordanze e del contesto della categoria *Innocent* nel corpus Twitter 13-18 luglio 2015; keyword: #Israel (elab. Netlytic)

Network Analysis

Dal menu possiamo accedere anche allo strumento *Network Analysis* che, pur non essendo strettamente inerente all'analisi automatica dei dati testuali, permette di visualizzare gli scambi di comunicazione tra gli utenti oppure, più semplicemente, le co-occorrenze dei nomi propri all'interno dei messaggi o dei frammenti in cui è suddiviso il corpus.

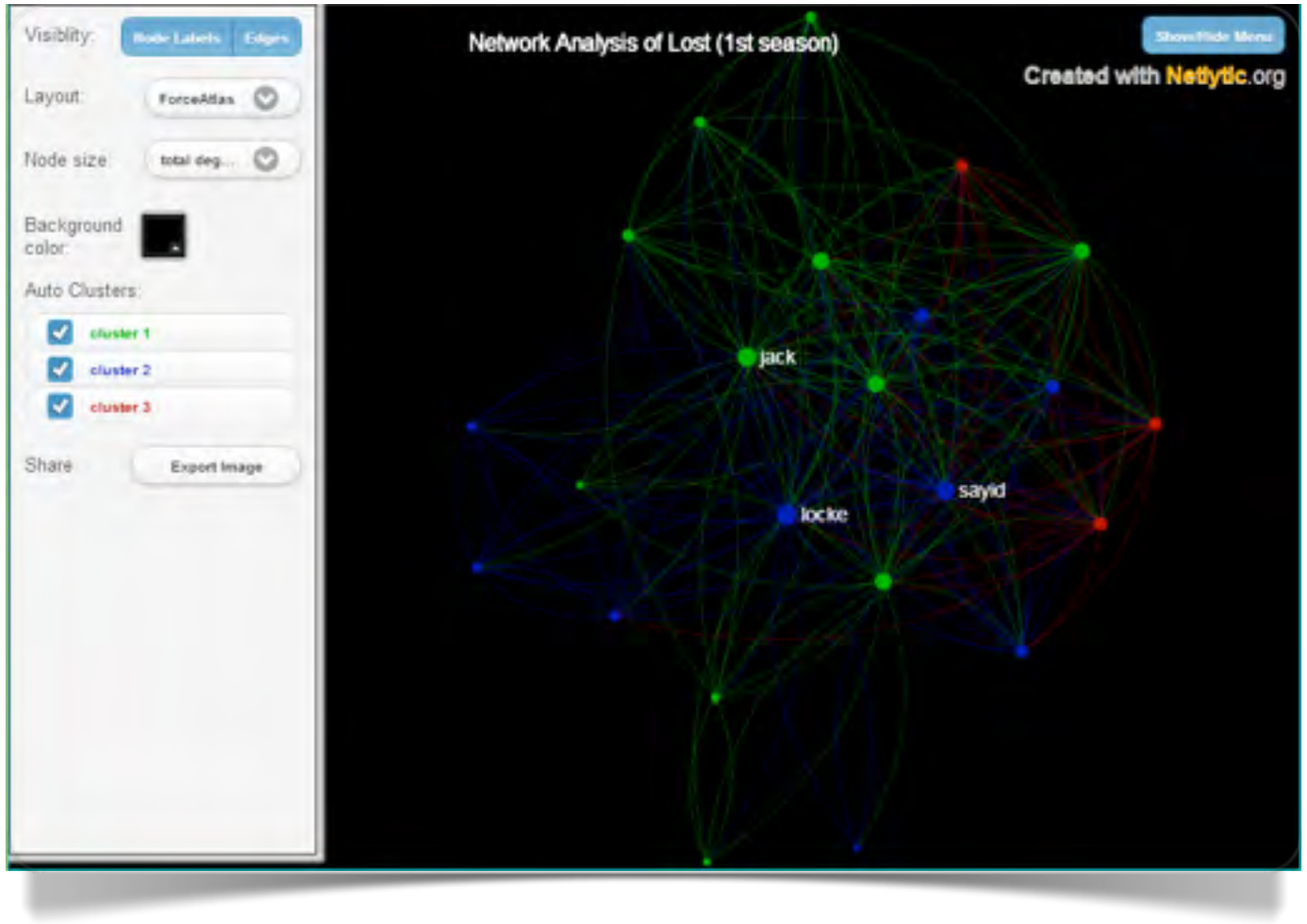


Fig. 6.9 Visualizzazione del network di *Lost* (ABC Studios - elaborazione con Netlytic)

Nella figura 6.9 possiamo osservare il network della prima stagione della serie televisiva *Lost* (ABC Studios) in cui è evidente la centralità delle interazioni intorno a tre personaggi: Jack, Sayid e Locke. Nella figura 6.10 vediamo invece il network dei tweet sul conflitto tra Hamas e Israele.

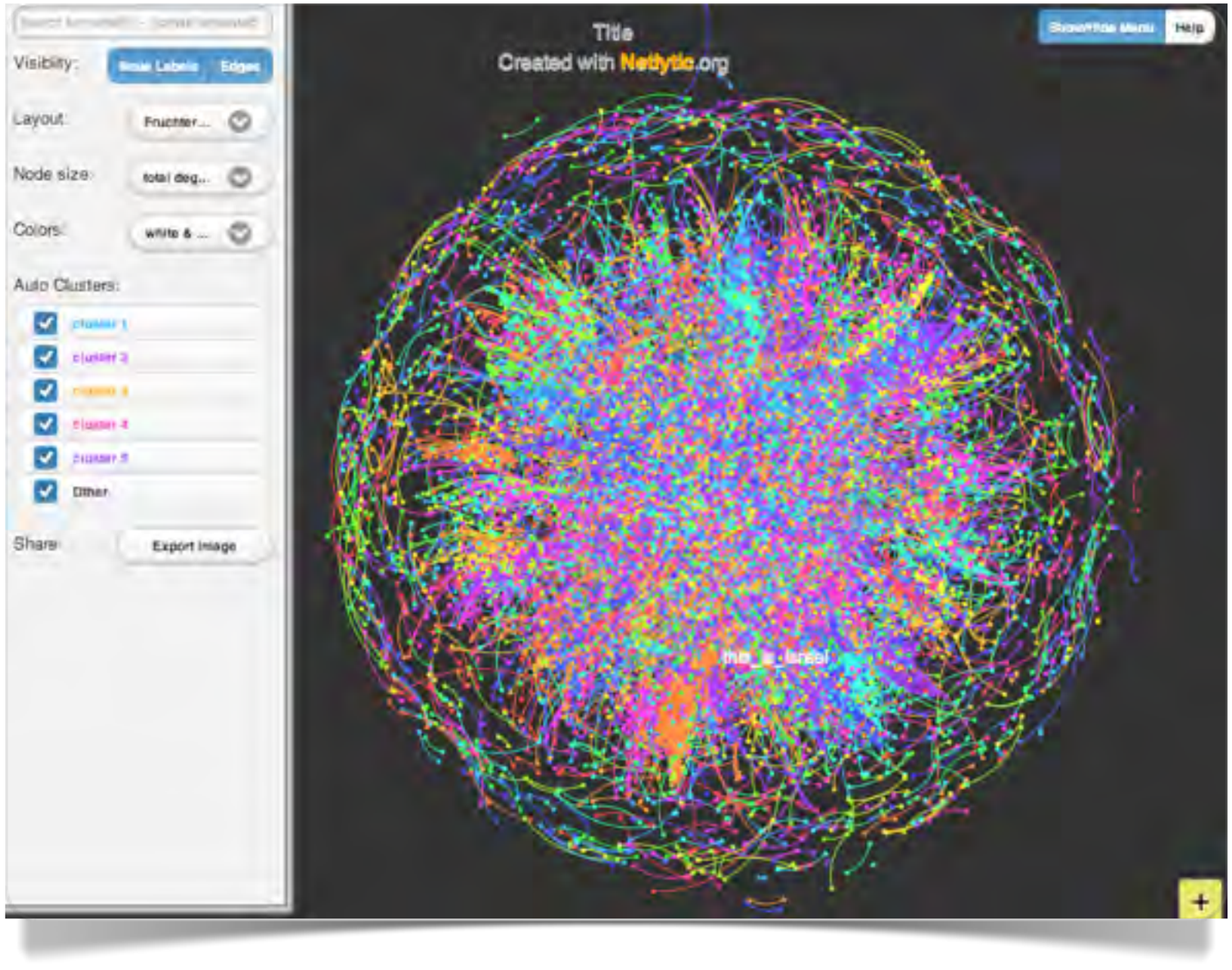


Fig. 6.10 Network Analysis del corpus Twitter 13-18 luglio 2015; keyword: #Israel; *Who mentions whom* (elab. Netlytic)

Capitolo 7

TAPoR Text Analysis Portal of Research

7

TAPoR è un progetto sponsorizzato da sei università canadesi (Montréal, Alberta, Toronto, New Brunswick, Hamilton).

TAPoR & TAPoRwere

TAPoR – Text Analysis Portal of Research, iniziato nel 2002 e coordinato da Geoffrey Rockwell, è un progetto in pieno sviluppo e si propone di diventare un portale di sperimentazione ad alto livello per l'analisi dei testi digitali (Rockwell, 2003). Dal 1 giugno 2012 è attivo un nuovo portale sperimentale: Tapor 2.0.

Il portale è ricchissimo di documentazione e aggiornamenti sullo stato della ricerca nella linguistica computazionale e nell'analisi automatica dei testi. Nella sezione TAPoR Texts sono raccolti testi di esempio e una selezione della documentazione didattica. L'utente può iscriversi gratuitamente e creare, con il proprio account, una sezione di testi personali; se lo ritiene opportuno può renderli pubblicamente utilizzabili da tutti. Con il login si accede a due ambienti principali di lavoro - TAPoRware e Voyant – che interagiscono tra loro all'interno del portale, però alcune opzioni più complesse sono tuttora instabili e pertanto – in questa presentazione introduttiva – è preferibile seguire l'accesso esterno agli strumenti di analisi e visualizzazione più affidabili. Una guida alle funzioni principali è disponibile in inglese, francese e italiano.

Fig. 7.1 TAPoR project - logo

TAPoRwere – Prototype of Text Analysis Tools è stato sviluppato da Geoffrey Rockwell, Lian Yan, Andrew Macdonald and Matt Patey. Gli strumenti sono compatibili con i documenti in HTML, XML e Plain Text e possono essere elaborati sia direttamente da fonte URL in web che con upload di file dal computer personale. In generale in questo primo ambiente vi si trovano strumenti introduttivi ma anche applicazioni sperimentali ancora non del tutto messe a punto (e che attualmente sembra siano stati

abbandonati) come Raw Grep (un generatore di concordanze che utilizza come pivot stringhe di testo anziché parole), Keyords Finder (identificatore di parole chiave che si basa sul principio di massima frequenza delle “parole contenuto” e degli n-grams), Word Brush (un visualizzatore di parole “a scomparsa” con applicazioni solo estetiche). Ultimamente è stato aggiunto un visualizzatore di frequenze in cui le parole sono rappresentate da gocce d’acqua: Voyant Term Fountain.

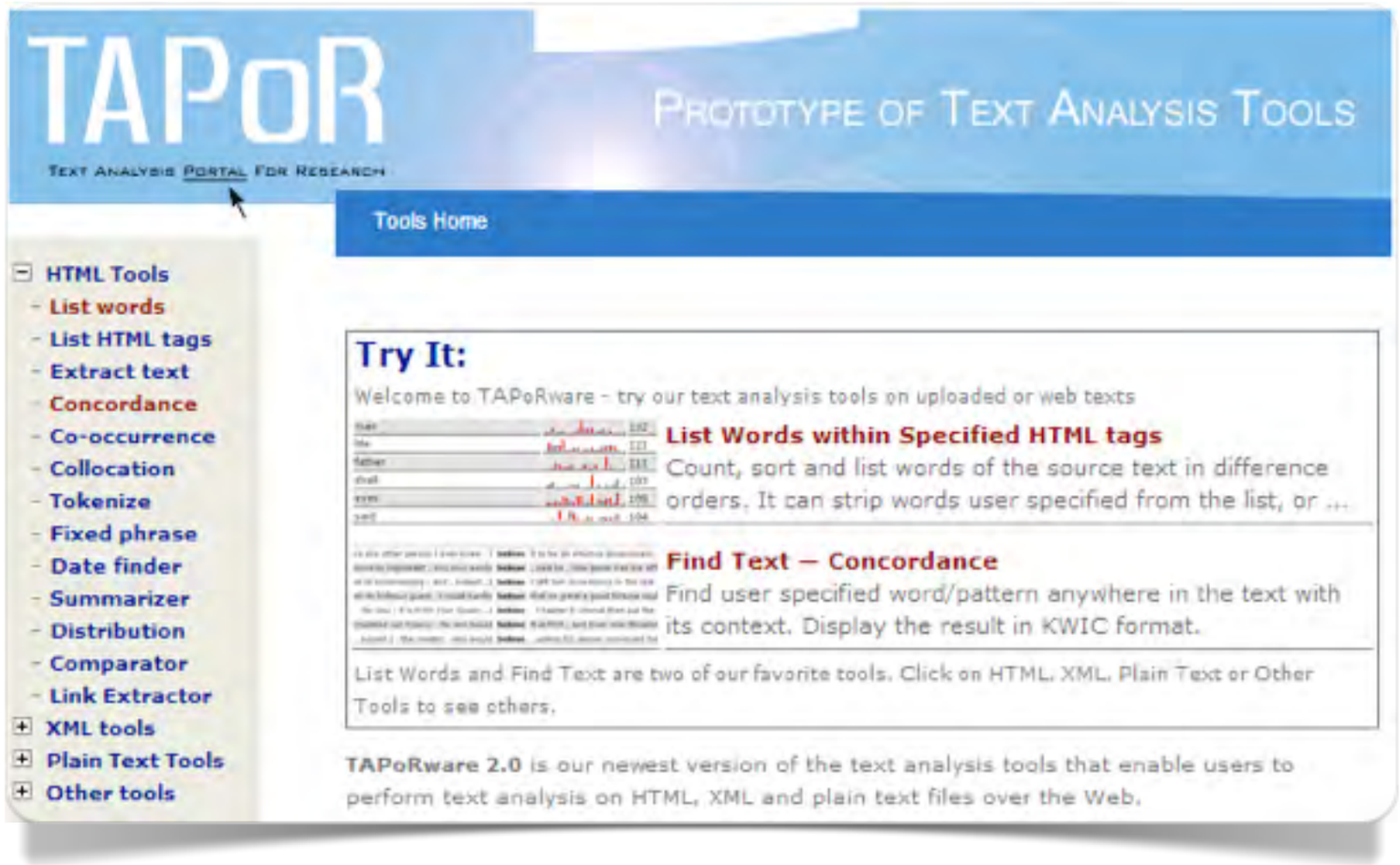


Fig. 7.2 TAPoRwere project - Main Page

Negli esempi che seguono utilizzeremo sempre la fonte URL, la più accessibile e generalizzabile in termini di utilizzo. L'upload di file personali è comunque intuitiva e permette di elaborare corpora personali anche di medio-grandi dimensioni (2 Mb in plain text in pochi minuti).

Il progetto prosegue con nuovi tools per l'analisi dei testi su Tapor 2.0 e sulla piattaforma Voyant.

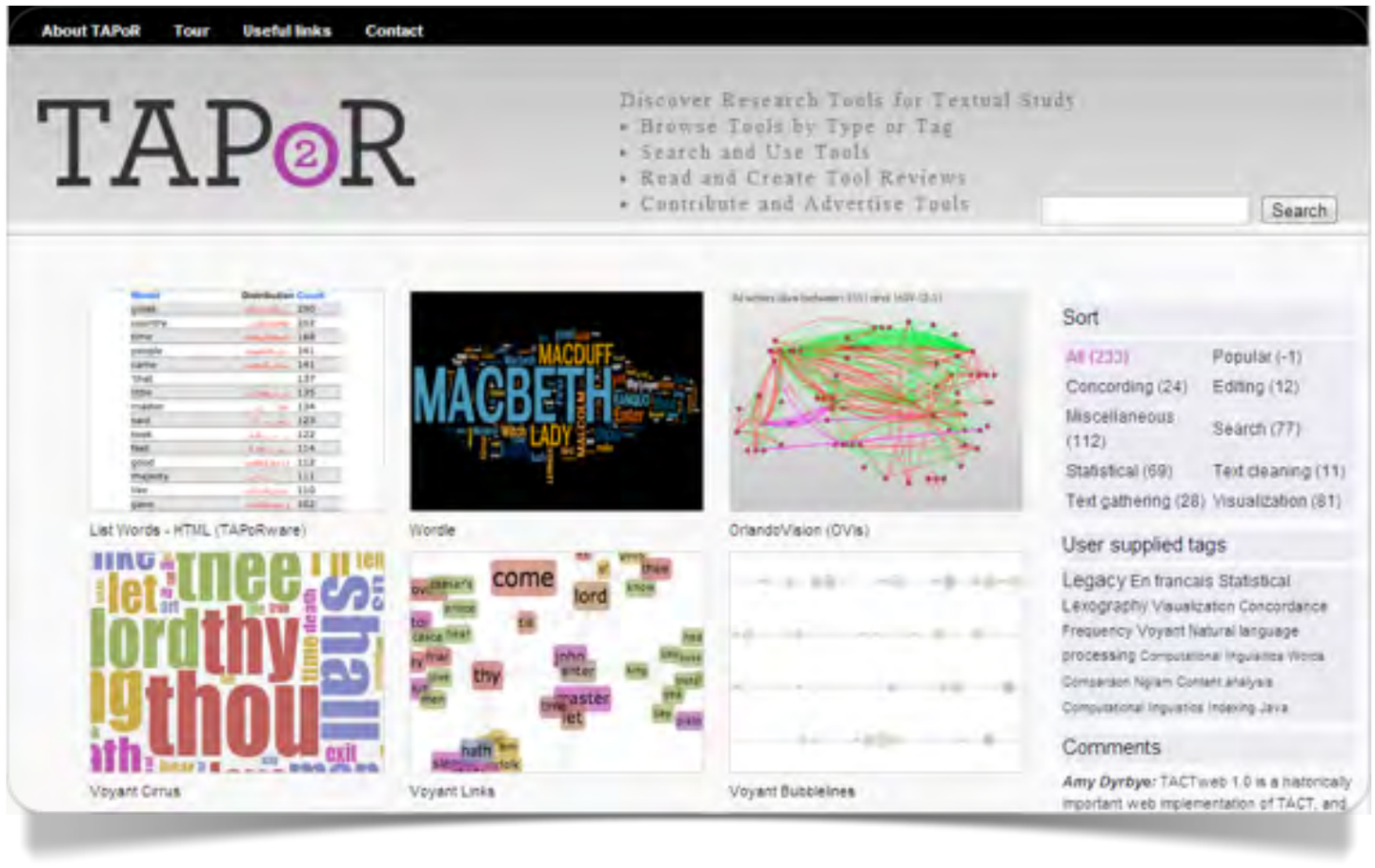


Fig. 7.3 Tapor 2.0 - Main Page

List Words

List Words effettua un conteggio delle occorrenze e delle frequenze del vocabolario (fig. 7.4). È possibile applicare la *Glasgow Stop Words list* generata con metodi automatici che traggono origine dagli studi di *Text Mining* e *Information Retrieval* (Lo Tsz-Wai *et alii*, 2005) oppure una lista di parole da escludere dall'analisi preparata autonomamente dall'utente. Vi sono diverse opzioni per l'output, tra le quali il vocabolario in ordine di frequenza decrescente o crescente, in ordine alfabetico o in ordine di apparizione nel testo.

Nell'esempio vediamo l'output della voce *Social Science* in Wikipedia. Nella seconda colonna della figura 7.5 possiamo osservare la distribuzione di un certo numero di parole a scelta (5, 20 o 50 parole più frequenti) in tutto il corpus suddiviso in 20 frammenti (5% del corpus per frammento).

Fig. 7.4 List Words - Interfaccia di acquisizione del testo

Summary: There are 2444 unique words other than those in the stop list, there are 6231 words other than those in the stop list. There are 9614 words in total including the stop words.

Words	Distribution	Count
social		208
science		172
sciences		79
studies		54
human		46
sociology		42
research		40
history		38
study		38
economics		37
political		36
geography		35
psychology		34
theory		33
environmental		32
edit		31
law		30
public		30
anthropology		30
university		30
philosophy		26

Fig. 7.5 Le forme più frequenti nella voce Social Science di Wikipedia (elab. Taporware)

Concordance

Lo strumento Concordance produce come risultato le concordanze rispetto a una parola pivot o a una espressione regolare Unix (word/pattern). Le opzioni prevedono la scelta del contesto nel quale collocare le concordanze, dalle parole vicine fino ai paragrafi ,e l'indicazione del numero di parole dell'intorno da considerare. Nell'esempio (fig. 7.6) vediamo l'output delle concordanze di *methods* nella voce Social Science in Wikipedia.

Summary: 17 entries found.

] Positivist social scientists use	**methods**	resembling those of the natural
share in its aims and	**methods**	. Contents 1 Social science
, quantitative research and qualitative	**methods**	are being integrated in the
of applied mathematics . Statistical	**methods**	were used confidently . In
generally attempted to develop scientific	**methods**	to understand social phenomena in
way , though usually with	**methods**	distinct from those of the
geography use many technologies and	**methods**	to collect data such as
the social sciences , uses	**methods**	and techniques that relate to
and testing theories . Empirical	**methods**	include survey research , statistical
inception , sociological epistemologies ,	**methods**	, and frames of enquiry
use a diversity of research	**methods**	, drawing upon either empirical
critical theory . Common modern	**methods**	include case studies , historical
share in its aims and	**methods**	. Social scientists employ a
scientists employ a range of	**methods**	in order to analyse a
ancient historical documents . The	**methods**	originally rooted in classical sociology
market research . Social research	**methods**	may be divided into two
Â· Humanities human science	**Methods**	Historical method Â· Empiricism

Fig. 7.6 Concordanze di *methods* nella voce "Social Science" di *Wikipedia* (elab. Taporware)

Co-occurrence

Co-occurrence produce come output le co-occorrenze di due parole entro un contesto definito dalla distanza tra di esse. Nell'esempio (fig. 7.7) vediamo la co-occorrenze di *philosophy* e *science* nel contesto di 5 parole rispetto alla voce Social Science in *Wikipedia*.

8 co-occurrences found

Scientific method History of ***science*** **Philosophy** of ***science*** ***Science*** policy Humanities

Law Linguistics Media studies Methodology **Philosophy** Political ***science*** Psychology Public administration

disciplines such as psychology , **philosophy** , computer ***science*** , linguistics

however , with the positivist **philosophy** of ***science*** in the 19th

Scholarly method , Teleology , **Philosophy** of ***science*** , and Philosophy

Philosophy of ***science*** , and **Philosophy** of social ***science*** [edit

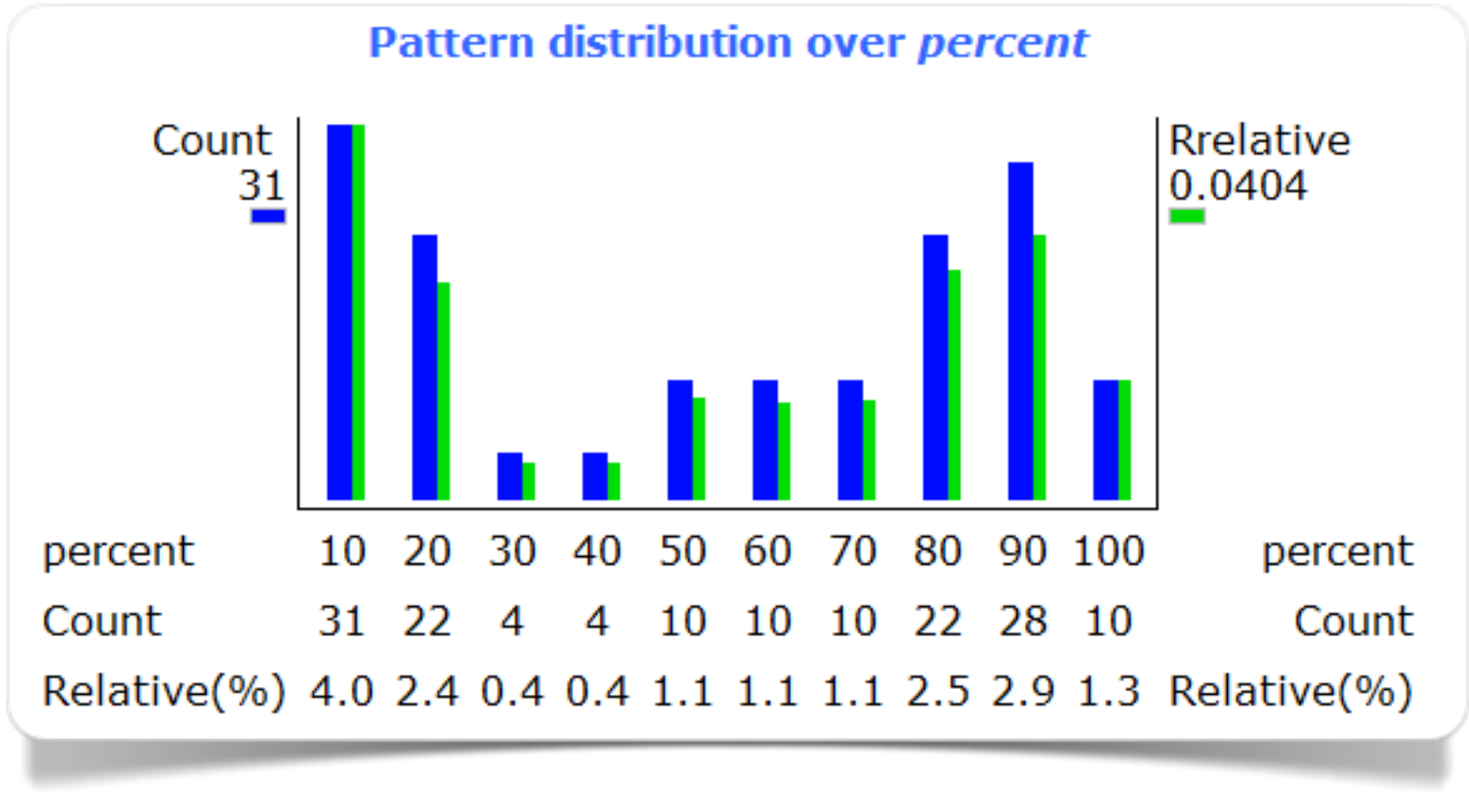
. Boston studies in the **philosophy** of ***science*** , 232 .

social ***science*** : a post-empiricist **philosophy** of social ***science*** . CUP

Fig. 7.7 Co-occorrenze di *philosophy* e *science* nella voce "Social Science" di *Wikipedia* (elab. Taporware)

Distribution

Lo strumento Distribution permette di osservare la distribuzione delle frequenze assolute e relative della parola selezionata all'interno del corpus suddiviso in 10 frammenti (10% del testo per ciascun frammento, o secondo altre percentuali di frammentazione). Nell'esempio (fig. 7.8) vediamo la distribuzione di frequenza della parola *science* nella voce Social Science in Wikipedia in blocchi di testo del 10%.



percent	10	20	30	40	50	60	70	80	90	100	percent
Count	31	22	4	4	10	10	10	22	28	10	Count
Relative(%)	4.0	2.4	0.4	0.4	1.1	1.1	1.1	2.5	2.9	1.3	Relative(%)

Fig. 7.8 Distribuzione delle frequenze assolute e relative della parola *science* nella voce "Social Science" di Wikipedia (elab. Taporware)

Comparator

Comparator mette a confronto due testi che appartengono a un unico corpus al fine di evidenziare le parole in comune e quelle esclusive di ciascun testo. Nell'esempio il corpus è costituito dai due discorsi inaugurali del presidente USA George W. Bush (acquisiti da Bartleby.com): il primo, del 20 gennaio 2001; il secondo, del 20 gennaio 2005. I due discorsi di svolgono in contesti molto diversi: il primo ha un tono molto tradizionale e si

Common words

Words	Text 1 counts	Text 1 relative	Text 2 relative	Text 2 counts	Relative ratio (text1/text2)	Word distribution in text 1	Word distribution in text 1
new	5	0.0028	0.0004	1	6.2638		
common	5	0.0028	0.0004	1	6.2638		
times	3	0.0017	0.0004	1	3.7583		
service	3	0.0017	0.0004	1	3.7583		
schools	3	0.0017	0.0004	1	3.7583		
nation's	3	0.0017	0.0004	1	3.7583		
lives	3	0.0017	0.0004	1	3.7583		
faith	3	0.0017	0.0004	1	3.7583		
defend	3	0.0017	0.0004	1	3.7583		
deep	3	0.0017	0.0004	1	3.7583		
birth	3	0.0017	0.0004	1	3.7583		
promise	5	0.0028	0.0009	2	3.1319		
courage	5	0.0028	0.0009	2	3.1319		
public	4	0.0022	0.0009	2	2.5055		
ideals	4	0.0022	0.0009	2	2.5055		

Fig. 7.9 Comparazione tra le frequenze delle parole nei discorsi del giuramento di G. W. Bush del 20 gennaio 2001 (text 1) e del 20 gennaio 2005 (text 2) (elab. Taporware)

appella agli elementi di continuità della nazione americana; il secondo avviene in uno scenario internazionale fortemente segnato dalle guerre nel Medio Oriente, dalle ferite dell'11 settembre 2001 e della lotta al terrorismo.

Nella figura 7.9 possiamo osservare le prime 15 parole comuni ordinate secondo i valori decrescenti del rapporto tra occorrenze del testo 1 (2001) e occorrenze del testo 2 (2005). Sono escluse, come di consueto, le *stop words*. La colonna dei valori relativi mette in evidenza le parole che – sebbene comuni, sono più rappresentative del testo 1 rispetto al testo 2: *new*, *common*, *times*, *service*, *schools*, ecc.

Words in text 1 only

Words	Text 1 count	Text 1 relative				Word distribution in text 1	Word distribution in text 2
story	7	0.0039					
purpose	4	0.0022					
civility	4	0.0022					
children	4	0.0022					
beyond	4	0.0022					
spirit	3	0.0017					
small	3	0.0017					
responsibility	3	0.0017					
principles	3	0.0017					
place	3	0.0017					

Fig. 7.10 Frequenze delle parole presenti esclusivamente nel discorso del giuramento di G. W. Bush (text 1) del 20 gennaio 2001 (elab. Taporware)

Nelle figure 7.10 e 7.11 vediamo le prime 10 parole presenti esclusivamente nel discorso del 2001 (Text1) o, rispettivamente, del 2005 (text 2). È evidente in quest'ultimo discorso la presenza di riferimenti all'attentato delle *Twin Towers* e alla guerra: *the history we have seen together*; *we have seen our vulnerability*; *there can be no human rights without human liberty*; *the ultimate goal of ending tyranny in our world*.

Words in text 2 only

Words			Text 2 relative	Text 2 count		Word distribution in text 1	Word distribution in text 2
seen			0.0026	6			
human			0.0026	6			
came			0.0022	5			
tyranny			0.0018	4			
task			0.0013	3			
soul			0.0013	3			
self			0.0013	3			
rule			0.0013	3			
questions			0.0013	3			
permanent			0.0013	3			

Fig. 7.11 Frequenze delle parole presenti esclusivamente nel discorso del giuramento di G. W. Bush (text 2) del 20 gennaio 2005 (elab. Taporware)

Capitolo 8

HyperPo Digital Text Reading Environment

8

HyperPo (ora non più disponibile, ma interessante da esaminare) era un ambiente completo di analisi dei testi sviluppato da Stéfan Sinclair in grado di eseguire analisi linguistiche su testi in inglese, francese, spagnolo, tedesco e italiano.

HyperPo

In *HyperPo* l'acquisizione del testo (fig. 8.2) avveniva in tre modalità: tramite URL, "copia e incolla" oppure con l'upload di un file in plain text (Sinclair, 2003).

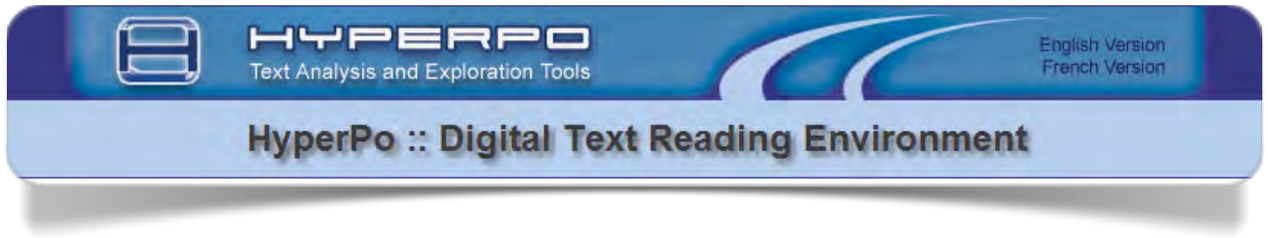


Fig. 8.1 HyperPo Home Page (fino al 2012)

Tra i menu erano disponibili gli strumenti:

- *Words Tools*
- *Document Tools*
- *Visualisation Tools* (di questo strumento è attualmente utilizzabile solo Mandala Browser).

Negli esempi utilizziamo il testo di riferimento: *The Universal Declaration of Human Rights*.

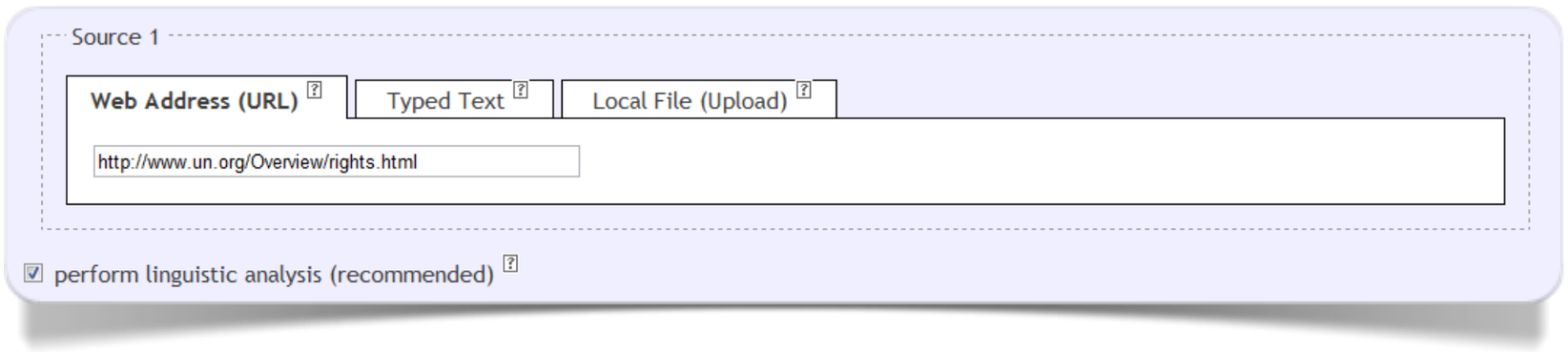


Fig. 8.2 Hyperpo – Modalità di acquisizione dei testi

Word Tools: Frequencies

Uno strumento interessante del menu Word Tools era *Study a word* nel quale erano riassunte tutte le informazioni di approfondimento su una determinata parola. Per esempio, nell'output della parola *right* (con l'opzione *match morphological variant*) vediamo le forme flesse *right* e *rights* con le frequenze assolute, le frequenze relative per mille, e lo z-score, che è una misura standardizzata delle frequenze (fig. 8.3). Nella stessa pagina erano rappresentate altre modalità di analisi e visualizzazione di base come *Word Series* (segmenti ripetuti), *Word Collocates* (co-occorrenze), *Distribution of Word* (grafico di distribuzione della parola nel testo suddiviso in parti predefinite) e *Word in Context* (concordanze rispetto a 5 parole che precedono e seguono la parola pivot).

① The Universal Declaration of Human Rights

Frequencies

Type	Raw	Relative	Z-Score
rights	25	13	24.652
right	33	17	32.652

Fig. 8.3 Frequenze della forma right* (Elab. HyperPo)

Document Tools

Lo strumento *Frequencies* nel sub-menu *Document Context* elencava le parole più frequenti in ordine di occorrenze, di frequenze relative, di z-score oppure in ordine alfabetico. Tra le opzioni era possibile ottenere il vocabolario dei lemmi; selezionare le parole con un senso compiuto (*content words*) o le *function words*, come gli articoli e le proposizioni. Nell'esempio (fig. 8.4) sono elencate le 10 parole di contenuto più frequenti.

① The Universal Declaration of Human Rights

Type	Raw	Relative	Z-Score
article	60	31	59.652
right	33	17	32.652
everyone	30	16	29.652
top	30	16	29.652
rights	25	13	24.652
human	16	8	15.652
equal	11	6	10.652
freedom	11	6	10.652
freedoms	10	5	9.652
law	10	5	9.652

Fig. 8.4 Vocabolario delle 10 parole di contenuto più frequenti (Elab. HyperPo

Con l'opzione *Statistical Summary* si ottenevano le misure lessicometriche di base del testo (fig. 8.5), tra le quali il conteggio delle occorrenze (*tokens*), il conteggio delle parole distinte (*types*) e il rapporto di densità o estensione lessicale (*type/token ratio*).

① The Universal Declaration of Human Rights

description	value
Characters	
total alphabetic characters	9391
shortest word length	1
longest word length	15
average word length	4.9
standard deviation of word length	2.9
Words	
count of all words	1932
count of unique words	564
lexical density (unique/all)	29.2

Fig. 8.5 Misure lessicometriche di base (Elab. HyperPo)

Visualization Tools: Mandala Tokens

Mandala è un'interfaccia che permette di navigare tra le parole di un testo cercando di individuarne le relazioni sulla base delle loro proprietà morfologiche o statistiche. Stefan Sinclair e i suoi collaboratori hanno scelto la metafora del "mandala" (il ciclo/circonferenza buddista che rappresenta il processo di formazione del cosmo) perché ritenevano che fosse adeguata ad esprimere il senso di una tendenza centrale che orienta il baricentro delle parole secondo i poli di attrazione che si aggiungono di volta in volta (Gainor et alii, 2009; Brown et alii, 2010).

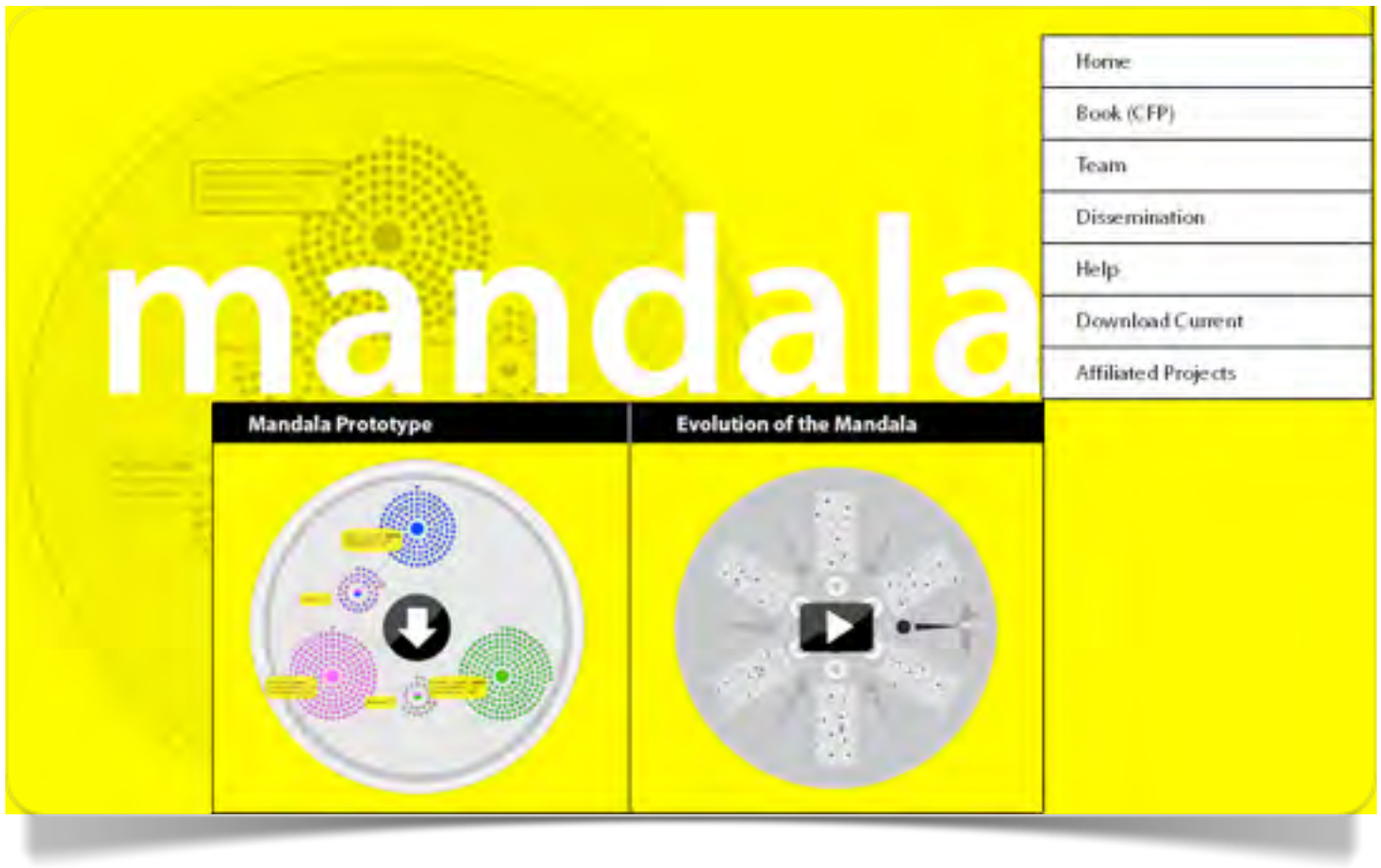


Fig. 8.6 Mandala Browser Home page

L'obiettivo è l'analisi esplorativa, il riconoscimento di strutture e relazioni, la formulazione di ipotesi.

Mandala non è più disponibile in HyperPo, ma il progetto prosegue su sistema operativo OS X in Mandala Browser, sempre con una equipe di sperimentazione diretta da Sinclair (fig. 8.6).

Il modello è applicabile a diverse collezioni di oggetti: in questo caso si tratta di parole (fig. 8.7). I punti neri che costituiscono il cerchio rappresentano le occorrenze (*word tokens*). Il cerchio grigio al centro rappresenta la tendenza centrale di tutte le

proprietà che definiscono le parole (frequenze, *word types*, lemmi, lunghezza delle parole ecc.). L'utente interagisce con il modello definendo degli assi (o magneti) sulla base di queste proprietà, delle loro misure, modalità o tipologie. Un magnete agisce da punto di attrazione delle parole che corrispondono alle proprietà che lo definiscono. Il risultato è una sorta di **diagramma di Venn** dinamico che porta le parole a raccogliersi in sottoinsiemi definiti dalle proprietà convergenti secondo i criteri scelti dall'utente in un display a destra dell'interfaccia.

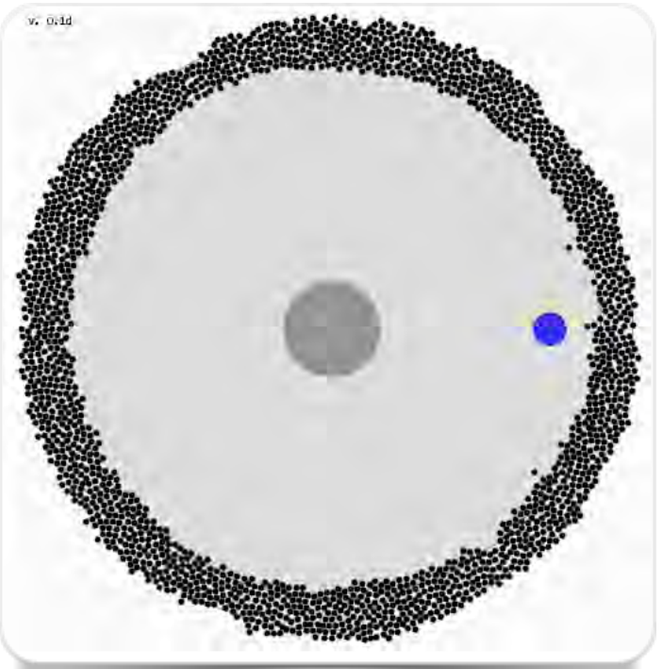


Fig. 8.7 Mandala Tokens della *Declaration of Human Rights* (Elab. HyperPo)

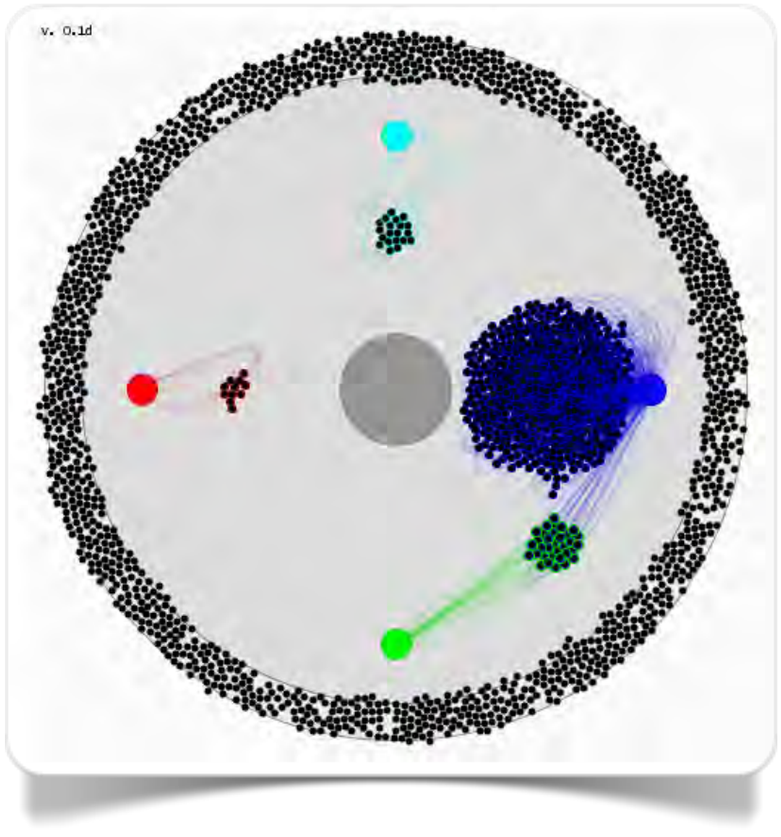


Fig. 8.8 Mandala Tokens della *Declaration of Human Rights* : parole e occorrenze di *human, right e rights* (Elab. HyperPo)

Nell'esempio della figura 8.8 il magnete blu è definito dalle parole con occorrenze maggiori di 30; il magnete verde è definito dalle occorrenze (n=33) della parola *right*; il magnete azzurro, dalle occorrenze (n=25) della parola *rights*; il magnete rosso, dalle occorrenze (n=16) della parola *human*. Possiamo osservare che solo il gruppo di pallini verdi, che rappresentano le occorrenze della parola *right*, tendono ad avvicinarsi e a evidenziare legami con il gruppo dei *tokens* blu che rappresentano le parole con occorrenze maggiori di 30.

Frequencies Centroid

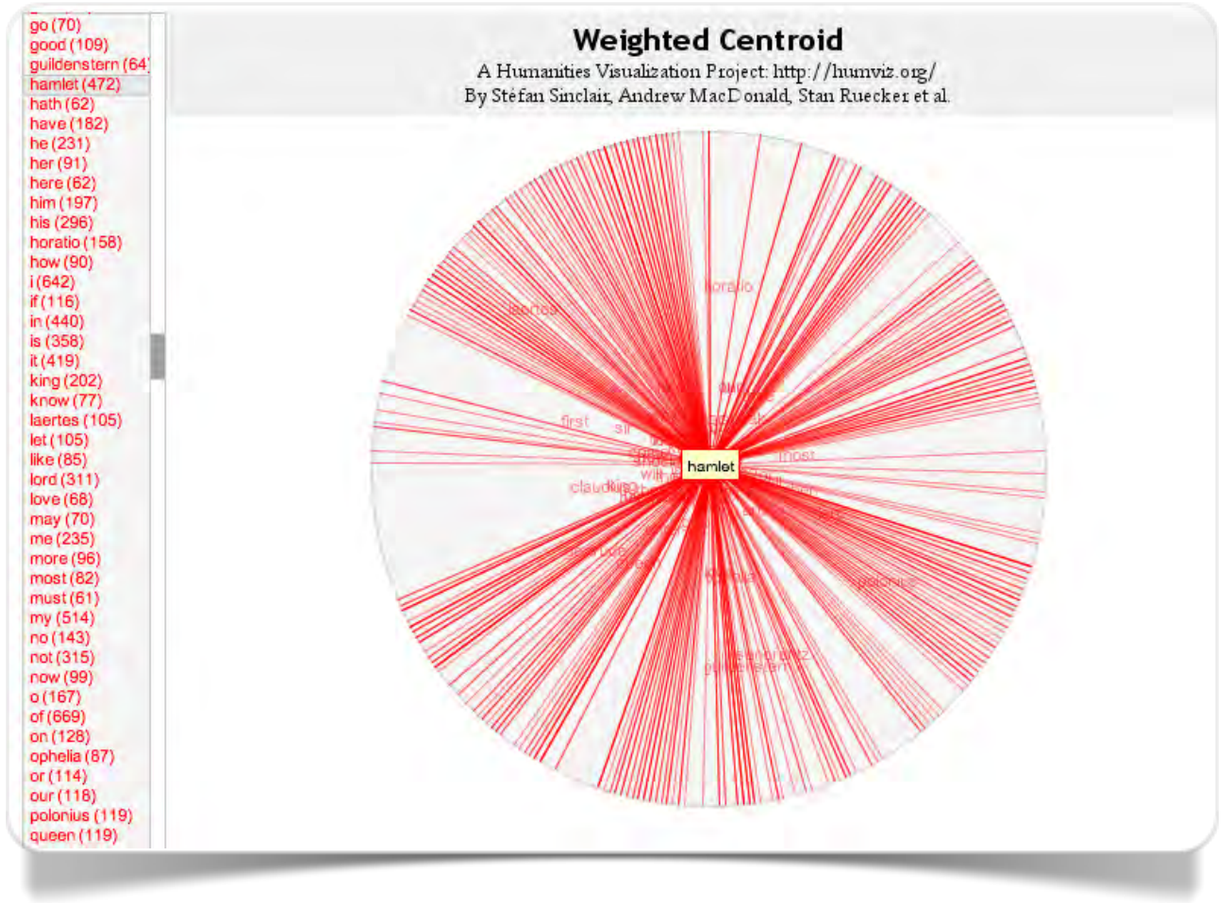


Fig. 8.9 Frequencies Centroid di *Hamlet* (Elab. HyperPo

Frequencies Centroid era uno strumento esplorativo dei testi che operava con modalità analoghe a Mandala Tokens. Il principio è piuttosto semplice: l'intero testo è disposto in senso orario lungo la circonferenza di un cerchio con inizio e termine alle ore 12:00. All'interno dell'ovale le parole sono disposte in una posizione che è determinata dal peso medio che la parola stessa ha rispetto alla distribuzione delle sue occorrenze nei capitoli o nella diverse parti di cui il testo e composto. Una parola si distanzia dalla circonferenza quanto più è utilizzata in parti diverse. Una parola utilizzata in modo frequente ed equilibrato in tutto il testo si andrà collocare al centro del cerchio.

Nell'esempio di figura 8.9, in cui è rappresentato *Amleto, principe di Danimarca* di William Shakespeare, *Hamlet*, con 472 occorrenze, è ovviamente presente in tutta la tragedia, ma con maggiore frequenza a ore quattro, sei, otto e dieci in cui è più attivo. A ore undici si svolge il duello mortale con *Laertes*.

Il personaggio di *Polonius* (119 occorrenze) si distribuisce tra il II e III atto fino alla sua uccisione per mano di *Hamlet* al termine del III atto (fig. 8.10). *Laertes* (105 occorrenze) è presente sporadicamente fino alla scena quinta del IV atto, in cui dialoga con il re Claudius e con Ophelia, e poi soprattutto nel V atto con il quale si conclude la tragedia (fig. 8.11).

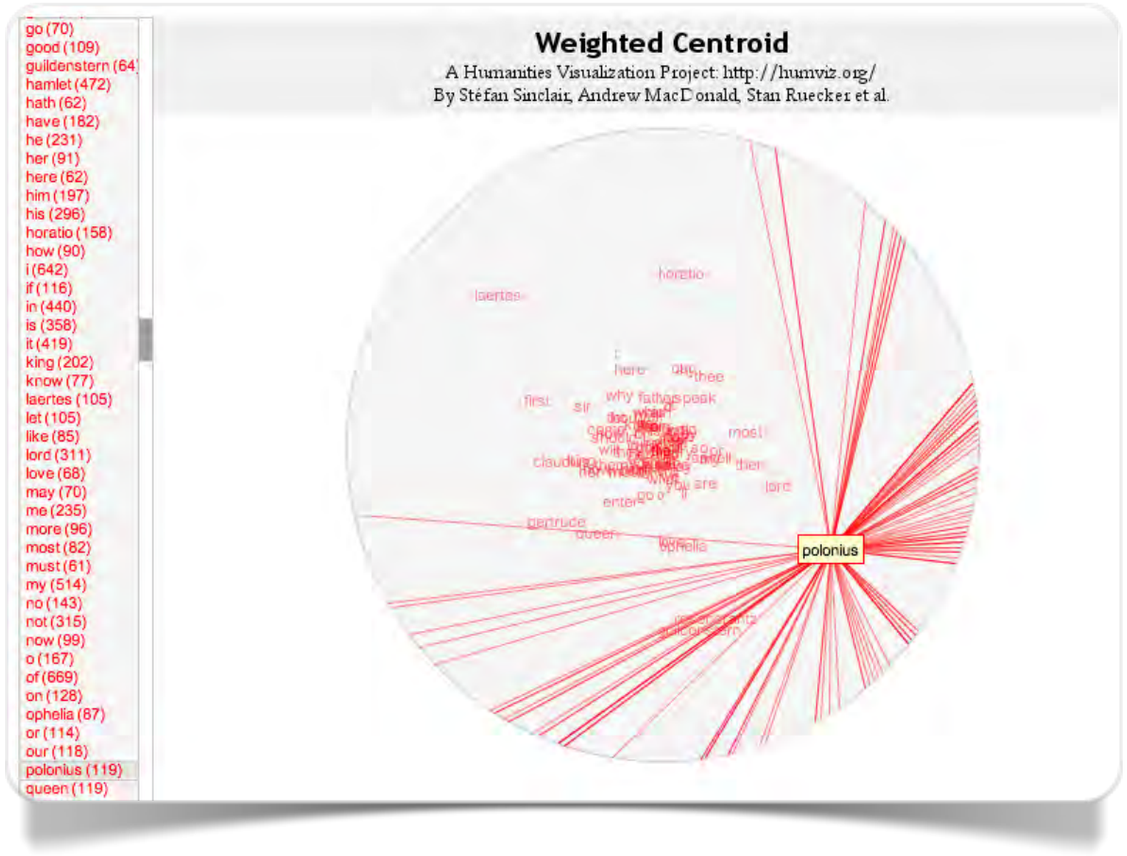


Fig. 8.10 Frequencies Centroid di *Polonius* (Elab. HyperPo)

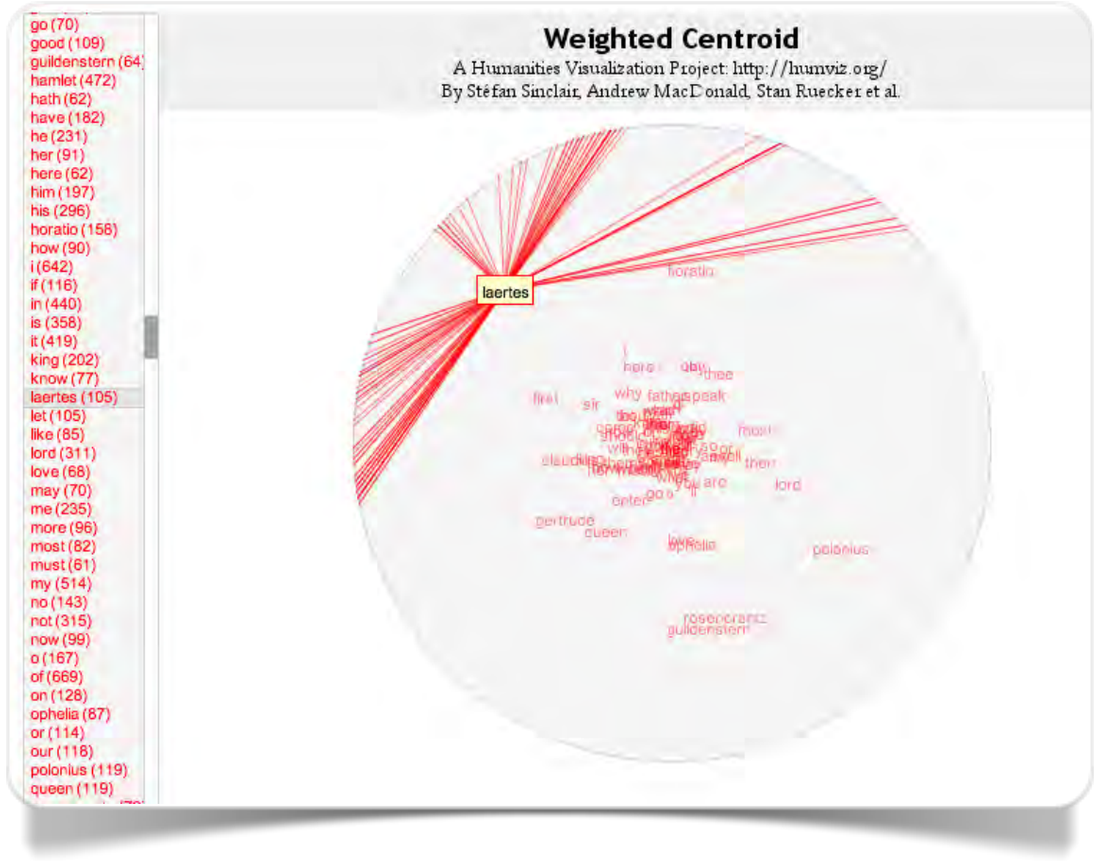


Fig. 8.11 Frequencies Centroid di *Laertes* (Elab. HyperPo)

Capitolo 9

Text Arc

Un esperimento molto simile a Frequencies Centroid (e che ha ispirato Stéfan Sinclair) è TextArc, sviluppato da W. Bradford Paley, un originalissimo artista di computer art che dall'inizio degli anni '70 si è dedicato alla rappresentazione dell'informazione nascosta, quella che non si presenta immediatamente alla nostra percezione.

9

Text Arc

Paley si definisce un *interaction designer* e la sua formazione deriva dalla fusione di diverse discipline: psicologia cognitiva, letteratura, informatica, cinema sperimentale (Paley, 2002). L'applicazione presenta un esempio sul testo di *Alice nel Paese delle Meraviglie*, di Lewis Carroll, ed è interfacciata con il progetto Gutemberg per la digitalizzazione dei testi. Pertanto ci sono centinaia di opere a disposizione per l'analisi. Come per Frequencies Centroid le parole si dispongono lungo la circonferenza (in questo caso di un ellisse. Le parole più frequenti sono evidenziate in un colore più intenso. Facendo passare il mouse su una parola si può osservare in linea retta il collegamento tra la parola e il punto in cui essa appare nella distribuzione del testo.

Nel caso di *Alice*, *mouse* appare 41 volte nel II e III capitolo, una volta nel primo e una volta nell'ultimo, pertanto la sua collocazione è situata accanto al bordo dell'ellisse a ore 2:00; *King* appare 64 volte solo dal capitolo VIII in poi e la parola è collocata a ore 10:00 in una posizione speculare; *Alice*, che ovviamente è presente dovunque, è collocata al centro dell'ellisse. Dando avvio alla lettura del testo,

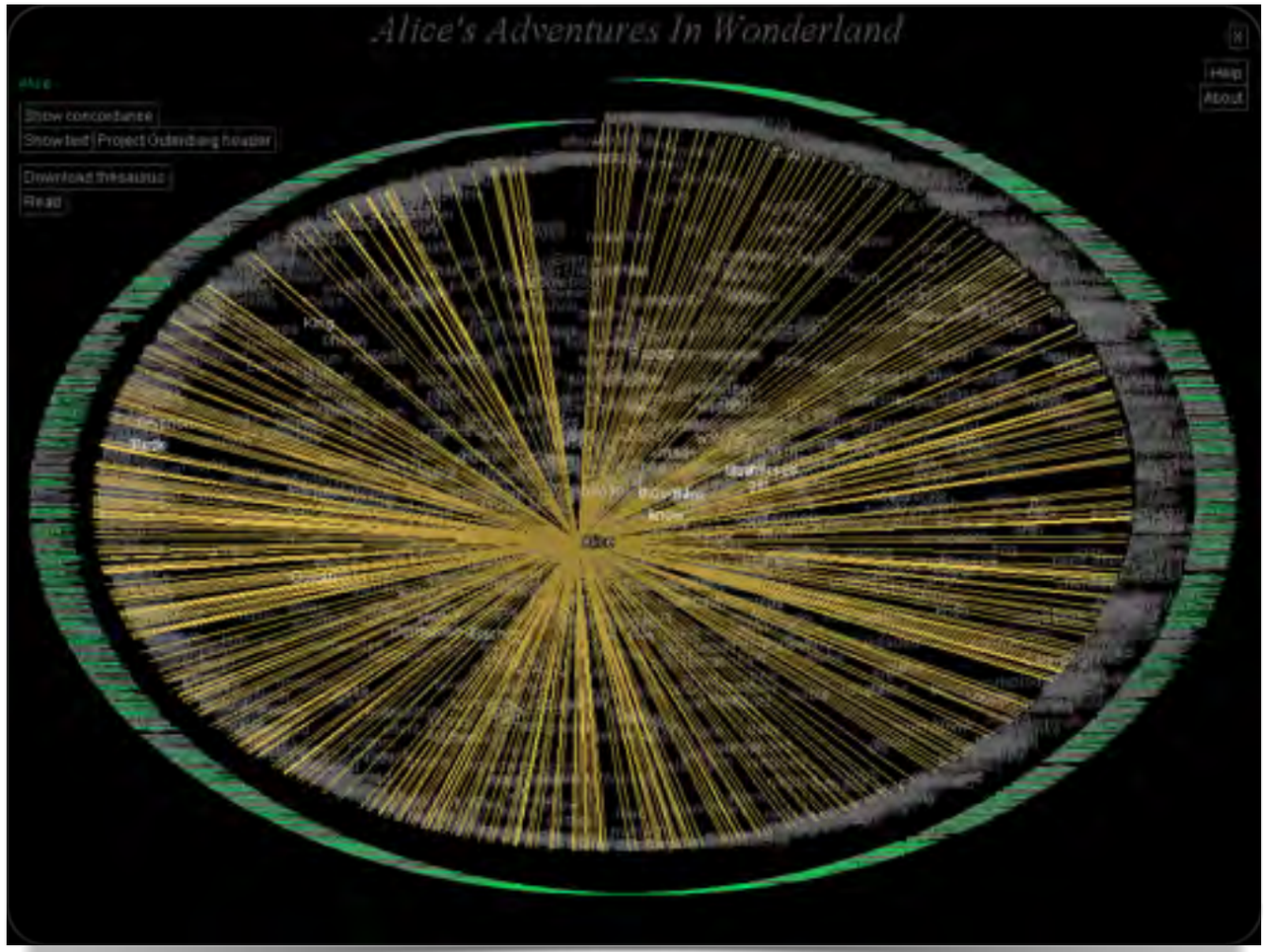


Fig. 9.1 Visualizzazzione della forma "Alice" in Alice nel Paese delle Meraviglie, di Lewis Carroll (Elab. TextArc)

che avviene cliccando sul bottone *Read* a sinistra del monitor, le parole scorrono lungo il racconto; selezionando *Show story line* si visualizzano i legami tra una parola e l'altra in successione con linee curve che (cliccando su *Sound*) sono accompagnate da una "musica" (che sottolinea il ritmo della narrazione secondo la rilevanza di un dato termine nell'insieme. Dallo stesso menu (*Show text*) è possibile visualizzare il testo e osservarne lo scorrimento, oppure visualizzare le occorrenze (*Show concordance*) delle parole rispetto alla loro posizione nel testo.

Movie 9.1 Visualizzazione animata di Alice nel Paese delle Meraviglie realizzata con TextArc

Capitolo 10

Voyant See through your text

Voyant è un progetto collaborativo Open Source e un ambiente completo di analisi dei testi, sviluppato da Stéfan Sinclair e Geoffrey Rockwell, con diversi tools che interagiscono tra di loro.

10

Voyant Workbench Screen

Voyant è una piattaforma di accesso alla gestione integrata dei testi (fig. 10.1). L'acquisizione può avvenire con diverse modalità: inserimento dell'URL; copia e incolla; upload di file dal proprio computer. È possibile anche analizzare un corpus composto da più testi (o frammenti). Voyeur accetta documenti in formato *plain text*, WinWord e Pdf.

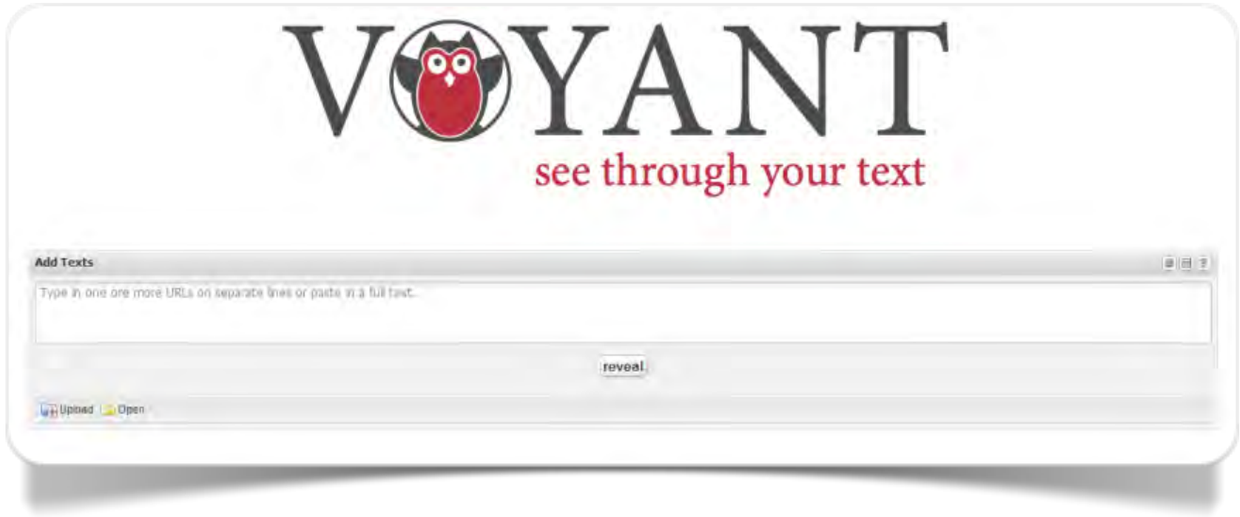


Fig. 10.1 Voyant: interfaccia di acquisizione del testo

Nell'esempio sottoponiamo ad analisi l'intero corpus delle opere di Shakespeare già presente nell'opzione *Open a pre-defined corpus*. La piattaforma di lavoro di Voyant visualizza il corpus in diverse modalità (fig. 10.2).

L'esplorazione avviene in sei finestre che possono essere nascoste e richiamate secondo necessità. In una prima fase si attivano tre finestre principali. La prima in alto a sinistra visualizza una classica nuvola di parole (che deriva dall'applicazione *Cirrus*). A destra si dispone il testo (*Corpus Reader*) e in basso a sinistra una descrizione sintetica (*Summary*) delle misure più rilevanti effettuate sul corpus e delle parole più significative di ciascuna opera.

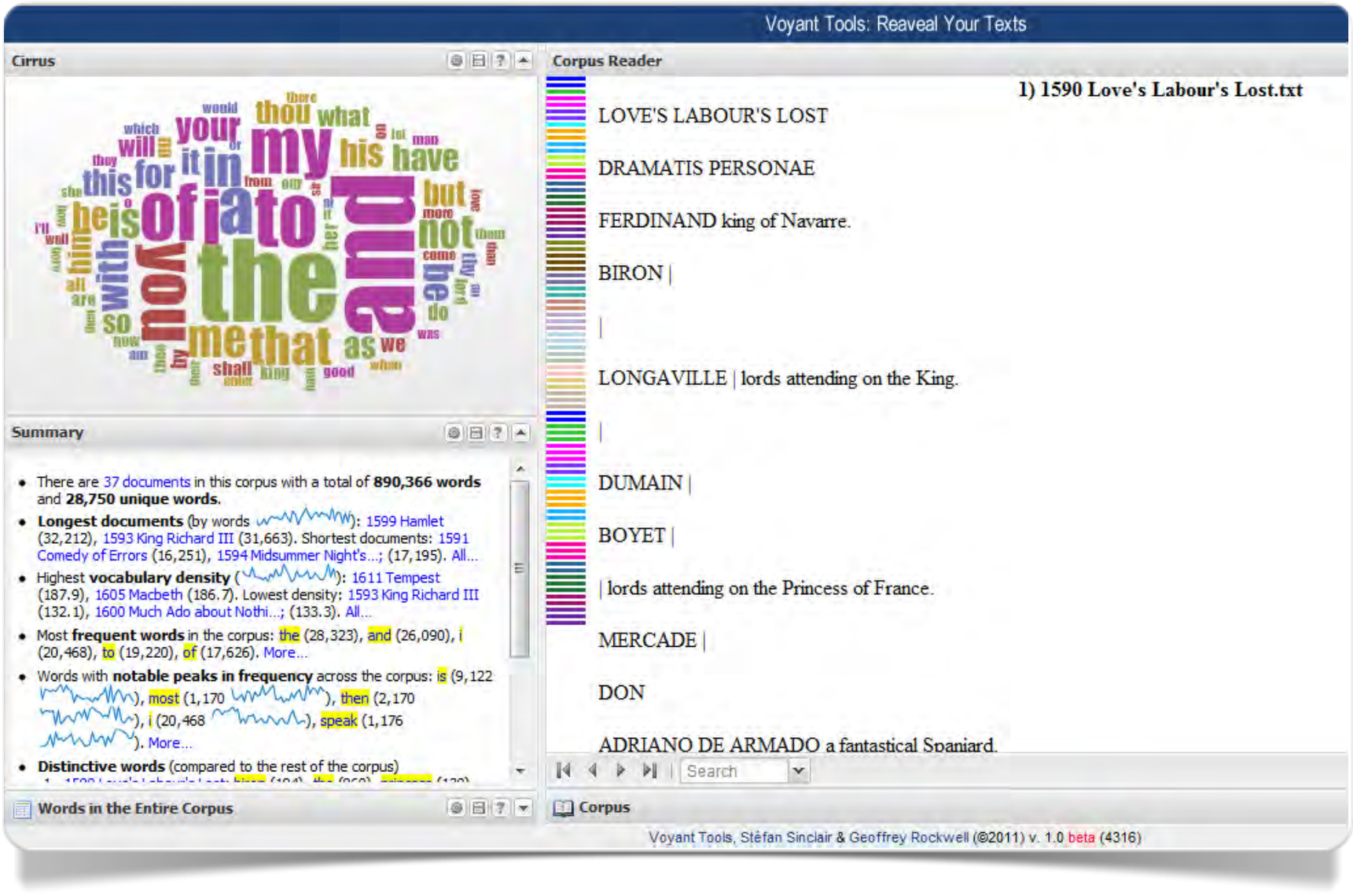


Fig. 10.2 Voyant: piattaforma di lavoro

Cirrus

In fase di acquisizione Cirrus visualizza la nuvola delle parole più frequenti in tutto il vocabolario del corpus. Questa opzione di default non è molto utile perché le parole più frequenti sono sempre parole forma (o parole vuote) e parole banali. Tuttavia, cliccando sulle opzioni in alto a destra della finestra di Cirrus, il menu per mette di selezionare la lista di *stop words* in diverse lingue (inglese, francese, tedesco, spagnolo, ungherese, italiano, norvegese).

Selezionando la lista di *stop words* in inglese, la nuvola di Cirrus cambia considerevolmente (fig. 10.3), anche se rimangono attive alcune parole vuote che non sono presenti nella lista come i pronomi arcaici: *thou*, *thee*, *thy*.

Fig. 10.3 Cirrus: nuvola delle parole più frequenti nel corpus Shakespeare (Elab. Voyant)

Summary

Nella finestra Summary (fig. 10.4) troviamo le classiche misure lessicometriche del corpus: 37 documenti (tutte le opere teatrali di Shakespeare) per un totale di 890.366 occorrenze (*tokens*) e 28.750 parole distinte (*types*). Il documento con il maggior numero di occorrenze è *Hamlet* (32.212).

La *vocabulary density* o estensione del vocabolario (*type/token ratio*) è pari al 3,2% ed è sempre maggiore in ciascun singolo documento (18,6% nel Macbeth) rispetto al corpus. Dal punto di vista statistico il vocabolario è ritenuto "rappresentativo" del linguaggio di riferimento (in questo caso del testo shakespeariano) se è minore del 20%.

Le *distinctive words* sono le parole relativamente presenti in misura maggiore in un documento rispetto al complesso del corpus. In questo caso sono evidenziate le parole che identificano i personaggi di ciascuna opera teatrale.

Fig. 10.4 Summary : misure lessicometriche nel corpus Shakespeare (Elab. Voyant)

Word Trends

Cliccando sulle parole della nuvola di *Cirrus* oppure sulle parole evidenziate nel *Summary* si attiva la finestra *Word Trends*, in alto a destra del piano di lavoro, in cui si visualizza l'andamento delle frequenze relative (per 10.000 occorrenze) della parola selezionata nel corpus suddiviso in 37 documenti. Nella figura 10.5 si può osservare l'andamento della parola *king*. Sul sito del MIT, cui accede Voyant per acquisire il corpus, le opere complete di Shakespeare sono organizzate in sequenza secondo il seguente ordine: *Comedy* (17); *History* (10); *Tragedy* (10); *Poetry* (5, ma non considerate nel presente corpus). Per ciascuna categoria le frequenze relative maggiori si riscontrano nel documento n. 7 (*All's Well That Ends Well*, del 1590), nel documento n. 21 (*King Henry VI*, *part III*, del 1595) e nel documento n. 33 (*King Lear*, del 1605).

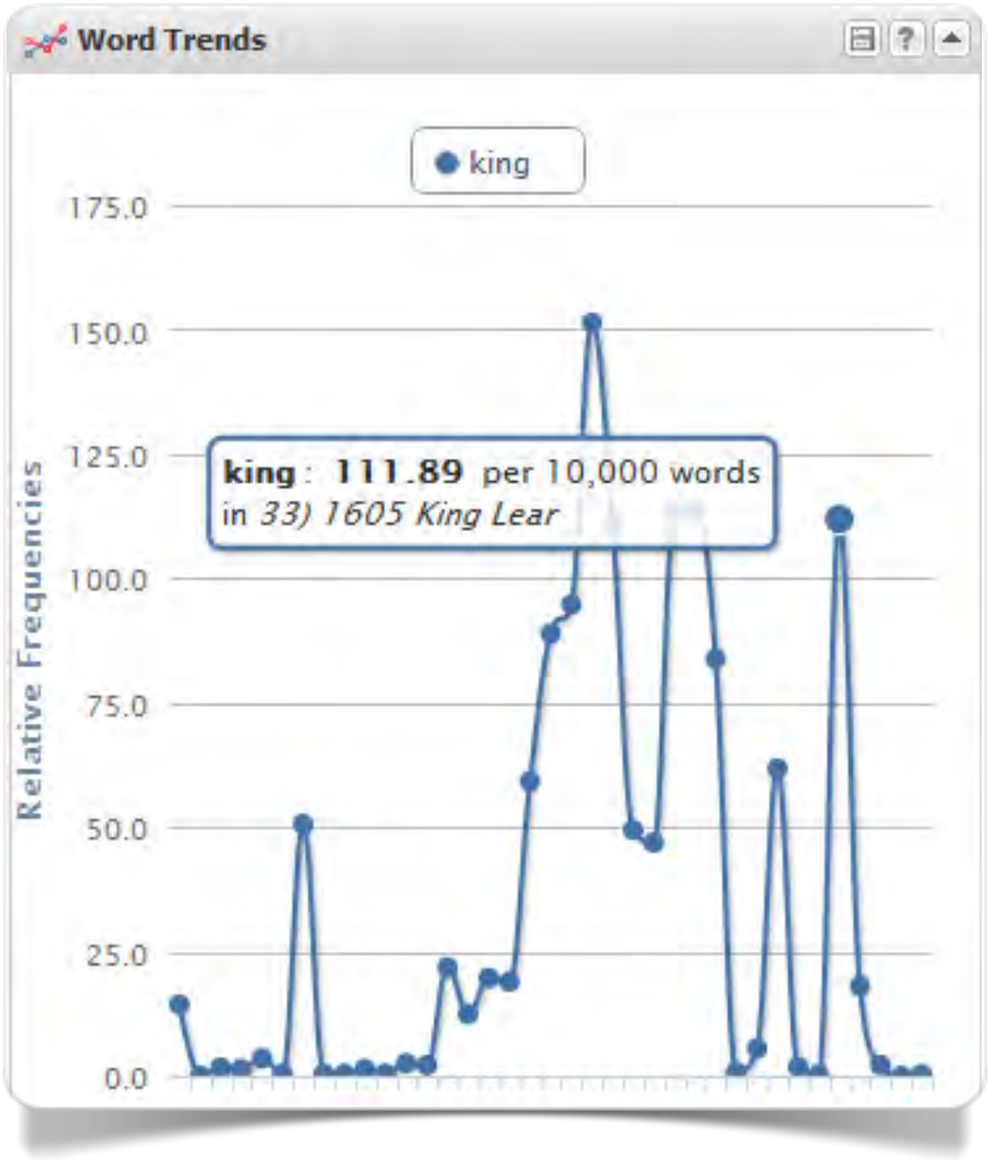


Fig. 10.5 Word Trends: frequenze realtive della parola king nel corpus Shakespeare (Elab. Voyant)

Nella finestra in basso a sinistra del piano di lavoro si visualizza la distribuzione delle occorrenze di tutte le parole nel corpus con i grafici delle frequenze relative (fig. 10.6)

La finestra in basso al centro visualizza una tabella riassuntiva delle misure lessicometriche dei documenti che costituiscono il corpus sottoposto ad analisi (fig. 10.7).

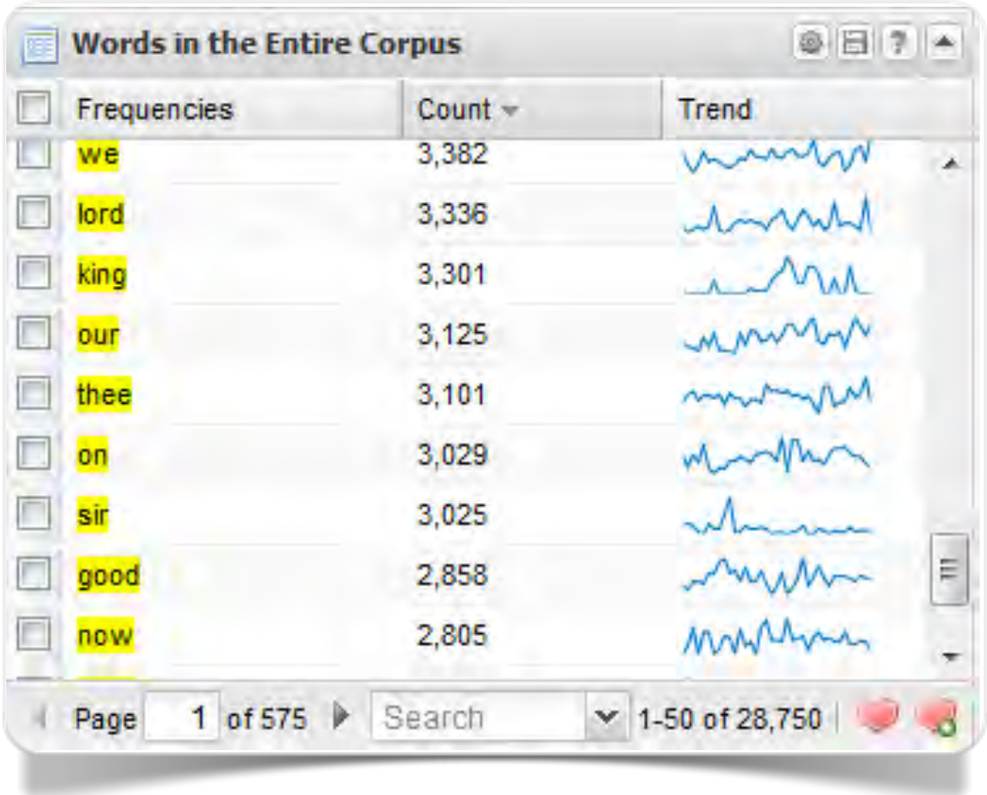


Fig. 10.6 Distribuzione delle occorrenze nel corpus Shakespeare (Elab. Voyant)

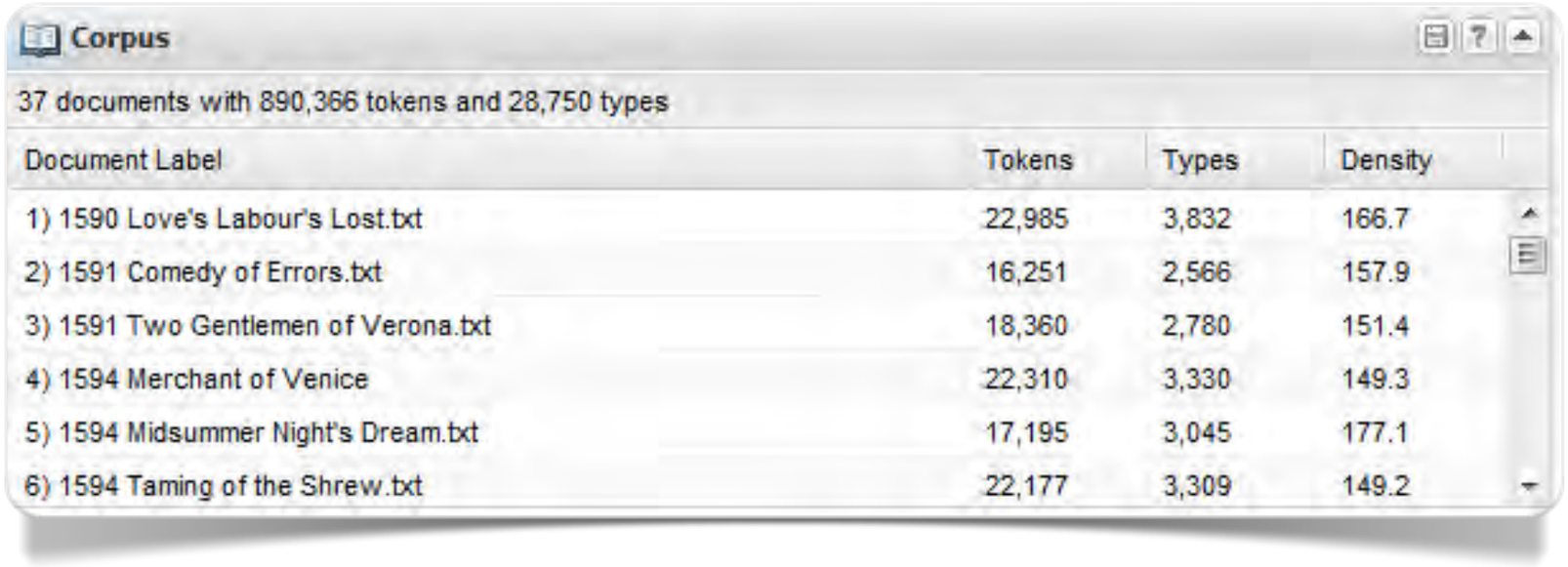


Fig. 10.7 Misure lessicometriche dei documenti del corpus Shakespeare (Elab. Voyant)

Bubblelines

Voyant mette a disposizione molti altri tools per l'analisi dei testi. Alcuni sono integrati nella piattaforma di lavoro (Cirrus, Corpus Grid, Corpus Summary, Reader, ecc.); altri sono accessibili singolarmente. Tra i più interessanti sono da segnalare *Bubblelines*, *Document Term Frequencies* e *Lava*.

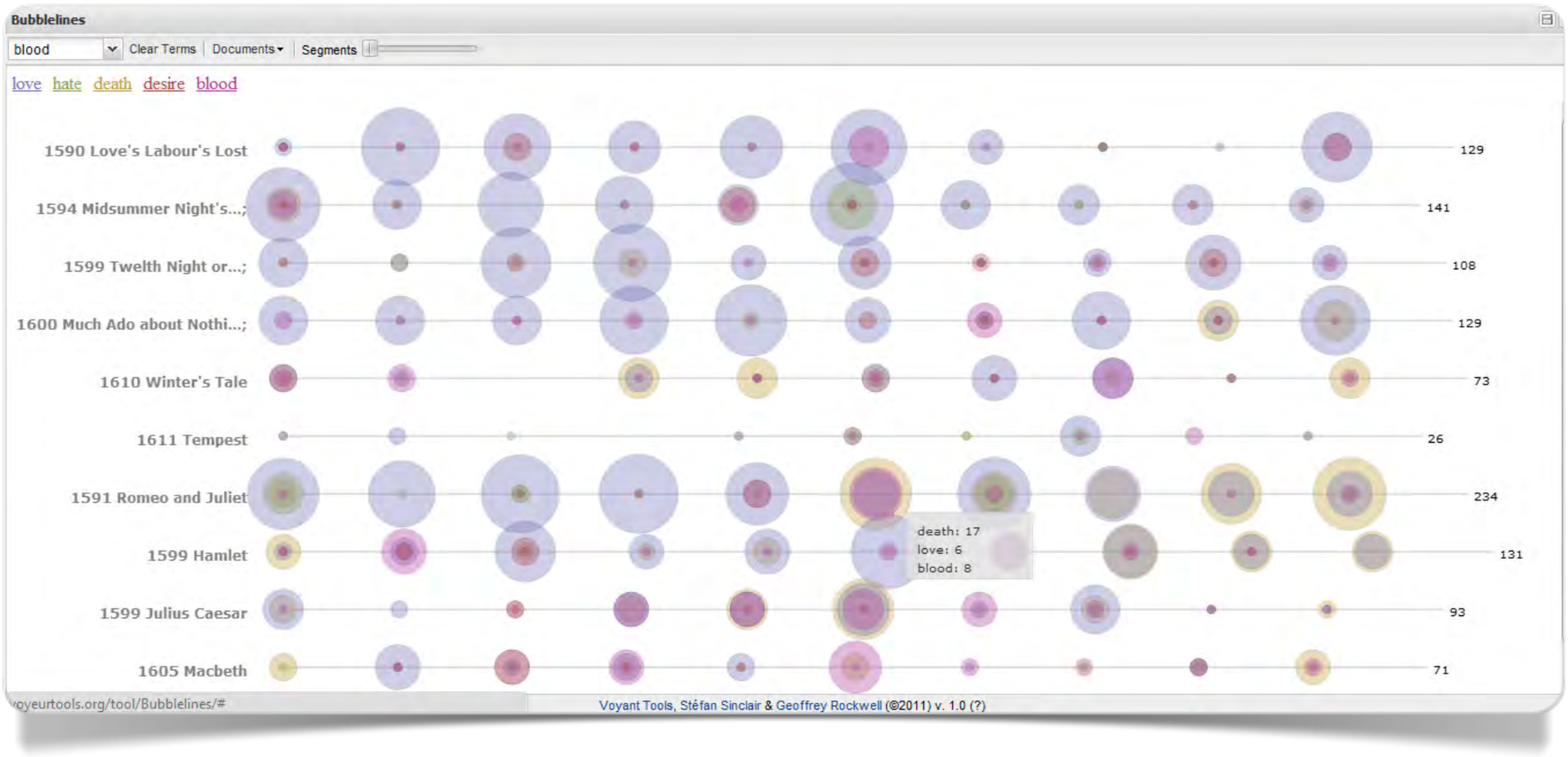


Fig. 10.8 Bubblelines: corpus Shakespeare (Elab. Voyant)

Bubblelines visualizza i documenti in analisi come linee rette disposte orizzontalmente. Ciascuna linea-documento è suddivisa in segmenti di lunghezza equivalente. Le parole sono identificate da un colore e rappresentate da bolle

di grandezza proporzionale alla frequenza disposte lungo le linee. Le parole co-occorrenti nella stessa frazione di testo si sovrappongono. L'obiettivo è di individuare analogie e differenze negli schemi comuni di ripetizione.

Nella figura 10.8 sono rappresentate le parole: *love*, *hate*, *death*, *desire*, *blood*. Le linee orizzontali, suddivise in dieci segmenti, rappresentano dieci opere: *A Midsummer Nigt's Dream*, *Twelfth Night*, *Much Ado about Nothing*, *Winter's Tale*, *The Tempest*, *Romeo and Juliet*, *Hamlet*, *Julius Caesar*, *Macbeth*. Come si può osservare, rispetto alle parole selezionate, *The Tempest* ha uno schema completamente diverso alle altre opere. Le prime quattro presentano delle somiglianze, in particolare rispetto alla parola *love* (in azzurro). *Romeo and Juliet* si distingue per la persistenza della associazione tra le parole *love* e *death*.

Document Term Frequencies

La visualizzazione delle frequenze delle parole nella interfaccia di Voyant è riferita a tutte le occorrenze del corpus (fig. 10.6). *Document Term Frequencies* (finestra in basso a destra) invece produce una tabella delle frequenze assolute e relative di una parola selezionata per ciascuno dei documenti che costituiscono il corpus. Nella figura 10.9 vediamo le occorrenze della parola *love* nelle dieci opere con valori più elevati in ordine decrescente.

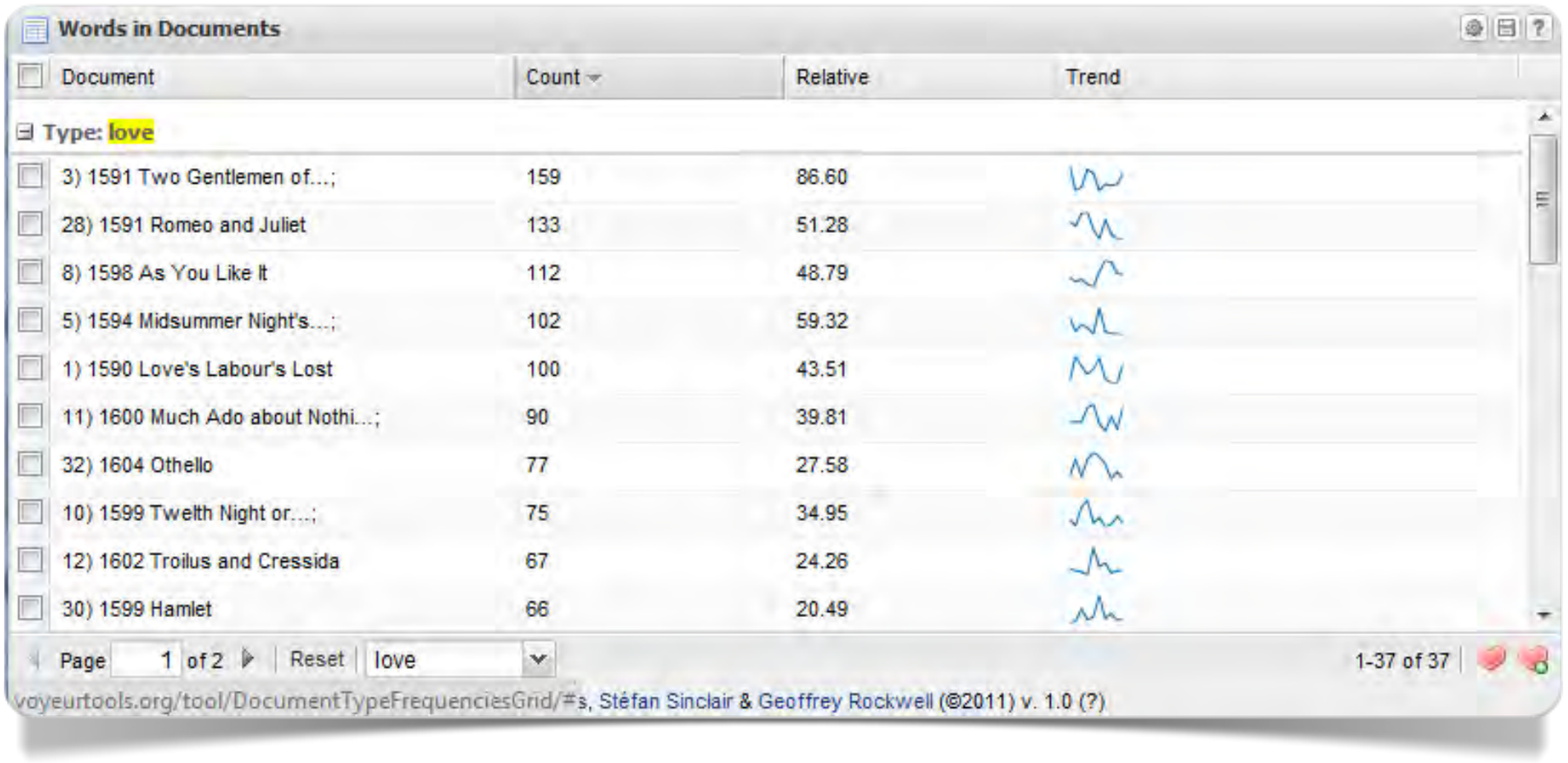


Fig. 10.9 Frequenze della parola love nel corpus Shakespeare (Elab. Voyant)

Lava

La visualizzazione del corpus in uno spazio tridimensionale è una sfida interessante e inusuale. Lava, nella prima schermata, presenta il corpus come un cilindro in verticale in cui ciascun segmento rappresenta un documento. Cliccando su uno dei cilindri si seleziona un documento e l'applicazione crea un anello, orientabile in diverse direzioni. Sulla circonferenza dell'anello si dispongono le parole in ordine decrescente di occorrenze. Nella figura 10.10 possiamo osservare l'anello del *Macbeth*. Si possono visualizzare più anelli e confrontarne le proprietà.

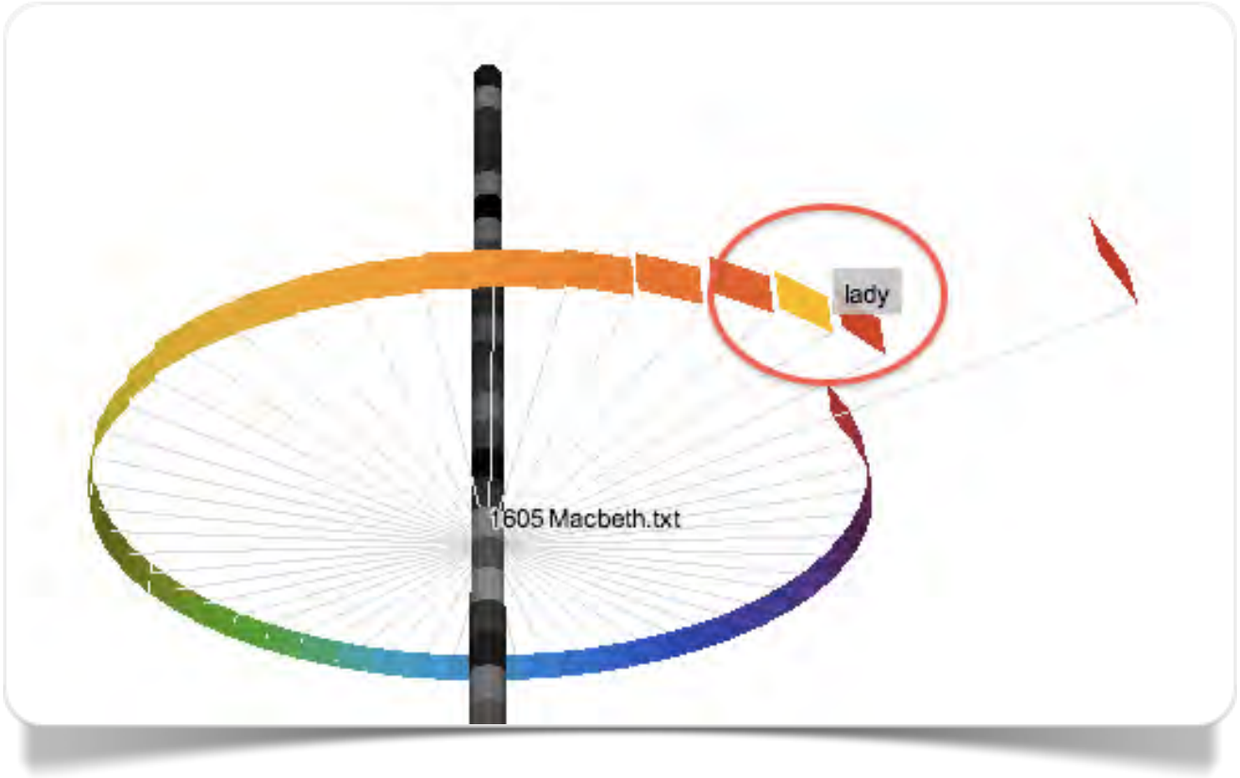


Fig. 10.10 Visualizzazione nello spazio del Macbeth rispetto al corpus Shakespeare (Elab. Voyant)

Capitolo 11

TerMine

Il National Centre for Text Mining (NaCTeM) è operativo presso l'Università di Manchester in collaborazione con l'Università di Tokyo e fornisce una serie di tools e servizi per la comunità accademica.

11

TerMine: C-Value

TerMine è stato sviluppato per l'identificazione delle terminologia specialistica, ma in generale è uno strumento per l'estrazione delle parole chiave e dei segmenti significativi nei documenti. Per la gran parte gli interessi dei ricercatori del NaCTeM sono rivolti alla ricerca in campo medico e alle potenzialità di sviluppo del Text Mining nel contesto della scoperta e della formulazione di ipotesi.

Fig. 11.1 NaCTeM – The National Centre of Text Mining – Logo

Il testo da analizzare può essere inserito in una finestra di acquisizione, caricato da una fonte esterna come file in plain text oppure indicando l'indirizzo web del documento da analizzare (fino a un massimo di 2 Mb). TerMine utilizza un metodo denominato C-value/NC-value che, per l'estrazione delle parole e dei segmenti rilevanti, tiene conto sia del contesto che di parametri linguistici e statistici (Frantzi et alii, 2000).

I termini estratti sono ordinati per valori di C-value decrescenti: i valori più alti corrispondono a termini di maggiore rilevanza (per esempio 2-grams e 3-grams hanno una rilevanza maggiore a parità di occorrenze rispetto alle parole singole). Nella figura 11.2 possiamo osservare i primi venti termini più rilevanti in un corpus costituito da 113 discorsi tenuti da Barack Obama durante la sua campagna elettorale dal 3 gennaio al 4 novembre 2008 in diverse città degli Stati Uniti. Nelle promesse elettorali del nuovo presidente sono evidenti i riferimenti alla riforma sanitaria: *health care* (1); *insurance company* (9); *health insurance* (13); *health care system* (18); *health care plan* (20).

Rank	Term	Score
1	health care	462.918915
2	senator mccain	371.409088
3	john mccain	296
4	wall street	273.105255
5	american people	270.333344
6	tax cut	149.899994
7	george bush	148.777771
8	21st century	146.806458
9	insurance company	134.5
10	main street	120
11	small business	119.099998
12	tax break	115
13	health insurance	84.799995
14	tax credit	84.428574
15	oil company	84.285713
16	capital gain tax	79.248123
17	special interest	79
18	health care system	76.581459
19	rescue plan	64.222221
20	health care plan	61.898499

Fig. 11.2 C-values dei 20 termini più rilevanti nel corpus dei discorsi di Barak Obama (Elab. TerMine)

Capitolo 12

LIWC Linguistic Inquiry and Word Count

Gli strumenti che abbiamo esaminato in questo capitolo sono in massima parte orientati all'analisi automatica del testo dal punto di vista della statistica linguistica. LIWC è un vero e proprio software di analisi del contenuto.

12

LIWC

LIWC (Linguistic Inquiry and Word Count) - sviluppato da James W. Pennebaker, Roger J. Booth, e Martha E. Francis – classifica le parole secondo determinate categorie (identificate da dizionari o lessici specifici) e ne confronta il risultato con dei valori di riferimento tratti dai testi campione. L'ipotesi di fondo degli autori è basata su un approccio "psicometrico" secondo il quale dall'uso delle parole si possono trarre indicazioni su alcuni aspetti cognitivi ed emozionali della personalità dei parlanti.

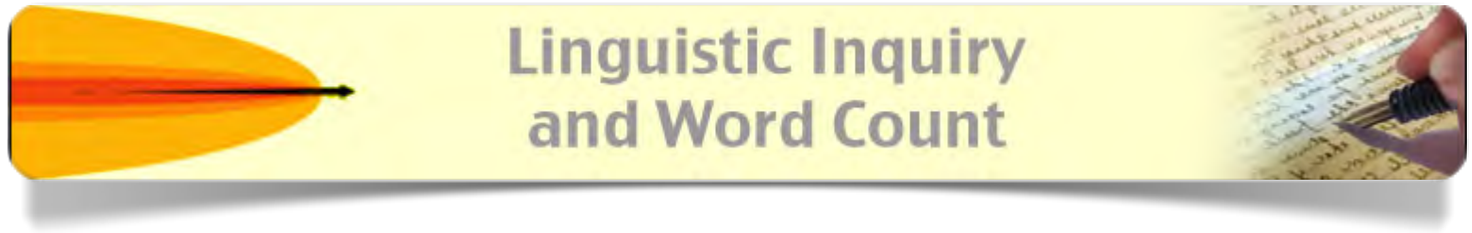


Fig. 12.1 LIWIC – Linguistic Inquiry and Word Count - Logo

I lessici di riferimento per queste misure sono stati costruiti utilizzando diverse fonti per un totale di 168 milioni di parole e 24.000 scriventi/parlanti. L'attribuzione delle parole alle categorie è avvenuta in diverse fasi (How it Works), anche con l'intervento di "giudici indipendenti". Oltre alle consuete categorie linguistico-grammaticali, nella classificazione troviamo processi sociali, affettivi, cognitivi e comportamentali.

Il software, per un utilizzo completo, richiede l'installazione sul computer e pertanto non sarebbe del tutto in linea con i presupposti degli strumenti raccolti in questo volume. Lo abbiamo incluso nella rassegna per la disponibilità di un punto di accesso online (Try Online) che permette di osservare alcuni dei risultati più rilevanti su un testo di prova inserito nella finestra di acquisizione. Nella figura 12.2 possiamo osservarne i risultato sul discorso di insediamento del presidente degli Stati Uniti Barack Obama del 21 gennaio 2013.

Nel discorso possiamo osservare come la dimensione delle *social words* sia decisamente più ampia (12.14) rispetto alla stessa dimensione nei testi personali e formali messi a confronto.

Rispetto al discorso del giuramento di John F. Kennedy del 20 gennaio 1961 (fig. 12.3) possiamo osservare come Obama non nasconda le difficoltà del suo secondo mandato, con un tono meno ottimistico e con maggiori riferimenti al sé.

LIWC Dimension	Your Data	Personal Texts	Formal Texts
Self-references (I, me, my)	8.08	11.4	4.2
Social words	12.14	9.5	8.0
Positive emotions	3.74	2.7	2.6
Negative emotions	1.03	2.6	1.6
Overall cognitive words	6.31	7.8	5.4
Articles (a, an, the)	6.91	5.0	7.2
Big words (> 6 letters)	21.35	13.1	19.6

The text you submitted was 2141 words in length.

Fig. 12.2 Discorso del giuramento di Barack Obama del 21 gennaio 2013 (elab. LIWC)

LIWC Dimension	Your Data	Personal Texts	Formal Texts
Self-references (I, me, my)	5.10	11.4	4.2
Social words	9.99	9.5	8.0
Positive emotions	4.09	2.7	2.6
Negative emotions	2.16	2.6	1.6
Overall cognitive words	4.45	7.8	5.4
Articles (a, an, the)	8.41	5.0	7.2
Big words (> 6 letters)	19.11	13.1	19.6

The text you submitted was 1392 words in length.

Fig. 12.3 Discorso del giuramento di John F. Kennedy del 20 gennaio 1961 (elab. LIWC)

Capitolo 13

Tree Cloud

Le co-occorrenze tra le parole sono il tema di Tree Cloud, un'applicazione che riprende il modello di base della "nuvola di parole" per proporre una visualizzazione delle co-occorrenze tra le parole in base alla vicinanza tra di loro in un testo.

13

TreeCloud

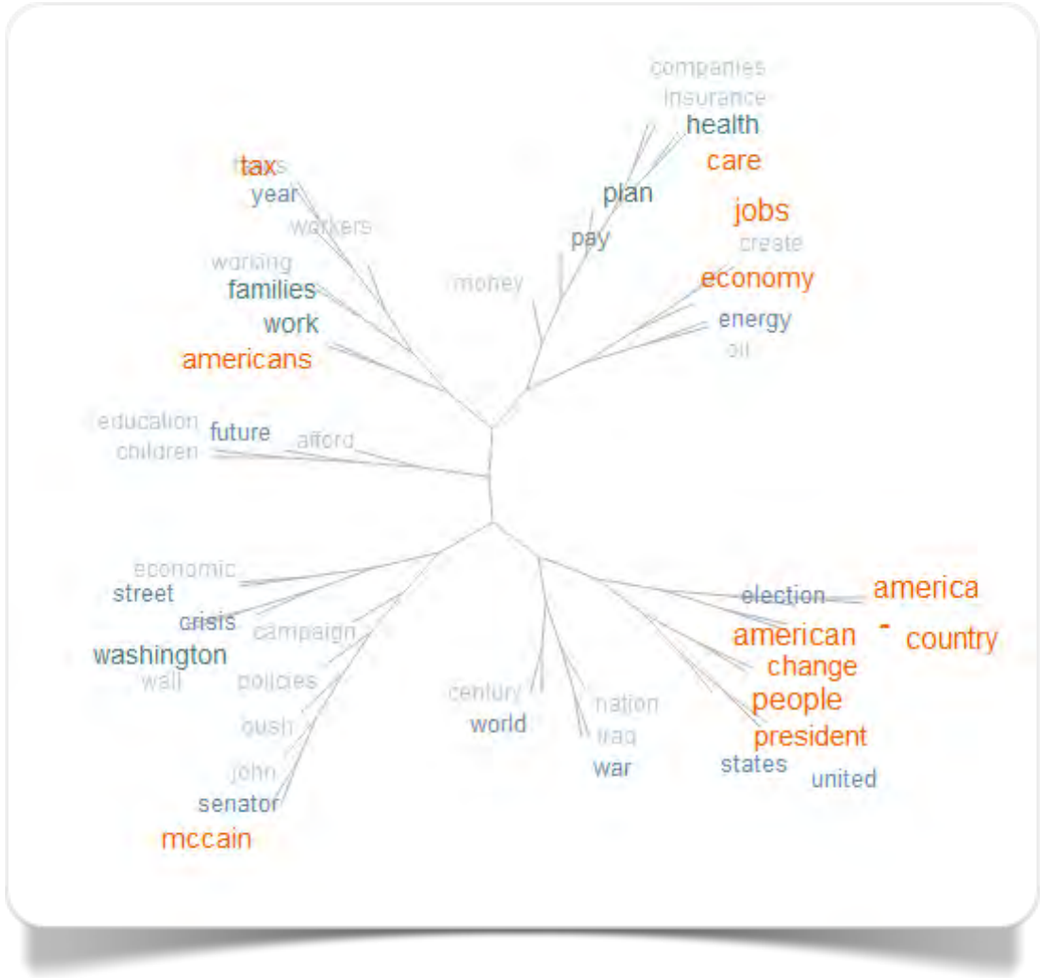


Fig. 13.1 Gli argomenti principali dei discorsi della campagna elettorale di Barak Obama (Elab. Tree Cloud)

L'algoritmo di Tree Cloud è stato sviluppato da Philippe Gambette e Jean Véronis per un software installabile su computer, ma permette l'elaborazione dei testi anche da una pagina di accesso con copia e incolla in *plain text*. Vi si accede da Create! Le parole si raggruppano ad albero secondo misure di prossimità e producono classificazioni "spontanee" in base alla loro vicinanza. Il risultato di una visualizzazione si presta pertanto ad essere interpretato come una sintesi degli argomenti trattati nel testo (Gambette e Véronis, 2010; Amstuz e Gambette, 2010).

Nella fig. 13.1 vediamo la visualizzazione che sintetizza il contenuto del corpus dei 113 discorsi della prima campagna elettorale di Barack Obama. Possiamo osservare come la struttura delle arborescenze si sia sviluppata in otto rami principali. A partire dalla posizione di mezzogiorno abbiamo: la politica sanitaria; la politica energetica; le elezioni presidenziali; la politica internazionale; i riferimenti all'avversario repubblicano; la politica conservatrice identificata con Washington; il futuro delle giovani generazioni; le famiglie americane e la politica fiscale.

Capitolo 14

Google Books Ngram Viewer

Google Books Ngram Viewer è stato realizzato da un team di ricercatori dell'Università di Harvard diretto da Jean-Baptiste Michel in collaborazione con Google.

14

Google Books Ngram Viewer

All'interno del progetto di digitalizzazione di Google Books - giunto attualmente a 15 milioni di volumi - per realizzare Ngram Viewer i ricercatori di Harvard e Jean-Baptiste Michel hanno costituito un sistema di corpora in diverse lingue (inglese, francese, spagnolo, tedesco, italiano, russo, cinese ed ebraico). Il corpus complessivamente utilizza il 4% dei libri pubblicati dal 1500 in poi per un totale di 500 miliardi di occorrenze (Michel et alii, 2010; Dekahaye e Gauvrit, 2013).

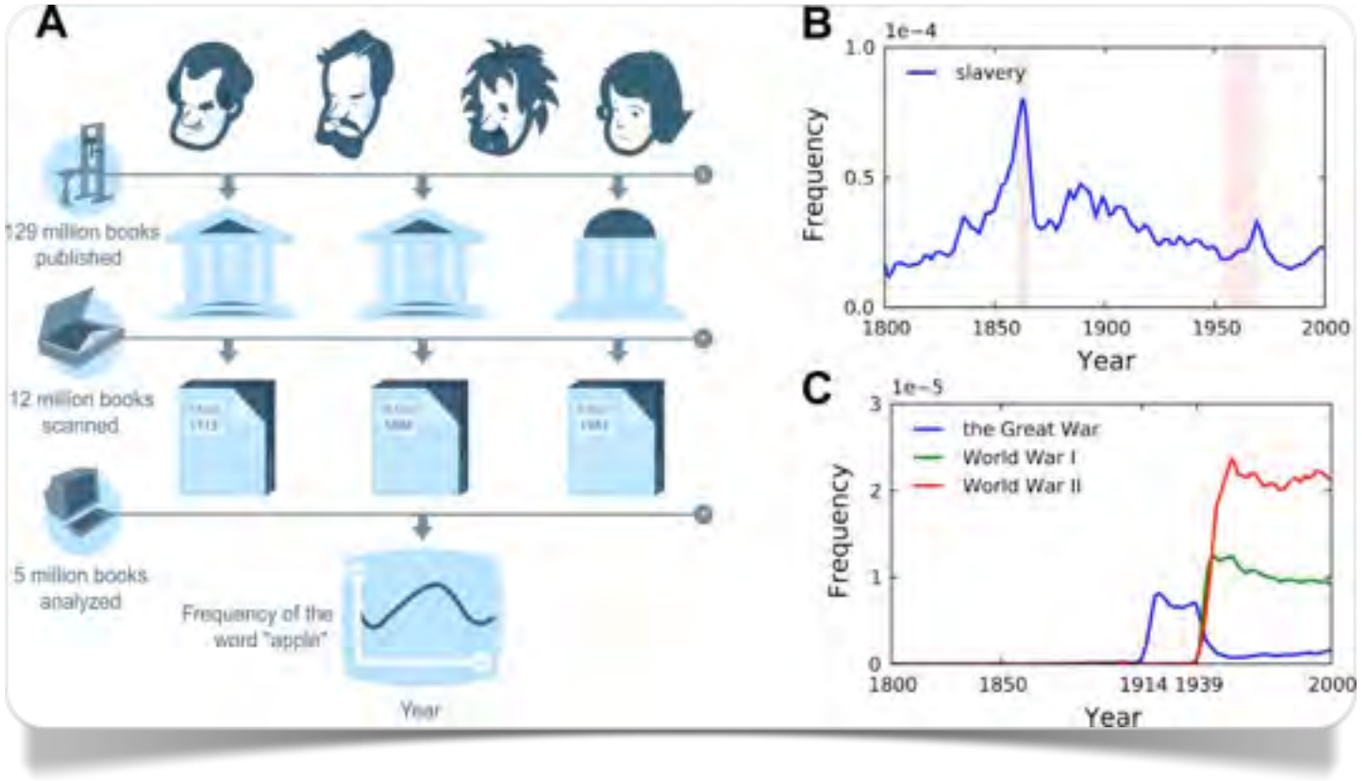


Fig. 14.1 Illustrazione dall'articolo originale di Jean-Baptiste Michel e coll. pubblicato su *Science* il 16 dicembre del 2010

Si tratta di 5.000.000 libri digitalizzati sui quali si possono effettuare ricerche longitudinali selezionando il tipo di corpus, le forme grafiche da inserire e il periodo storico di riferimento. La versione 2012 è stata migliorata rispetto a quella del 2009 ed è in corso di costante aggiornamento.

Google Ngrams Viewer lavora solo su un corpus precostituito. Tuttavia lo strumento è di assoluta rilevanza dal punto di vista linguistico e socio-culturale per effettuare comparazioni e valutazioni della presenza di un termine nell'arco temporale di 500 anni, sebbene gli stessi autori suggeriscano di utilizzare con cautela le consultazioni antecedenti al 1800. In figura 14.2 possiamo osservare un esempio in cui si mettono a confronto "socialismo", "comunismo" e "fascismo". Per questi termini (presenti senza ambiguità sia con l'iniziale maiuscola che minuscola) e preferibile utilizzare l'opzione *case insensitive*.

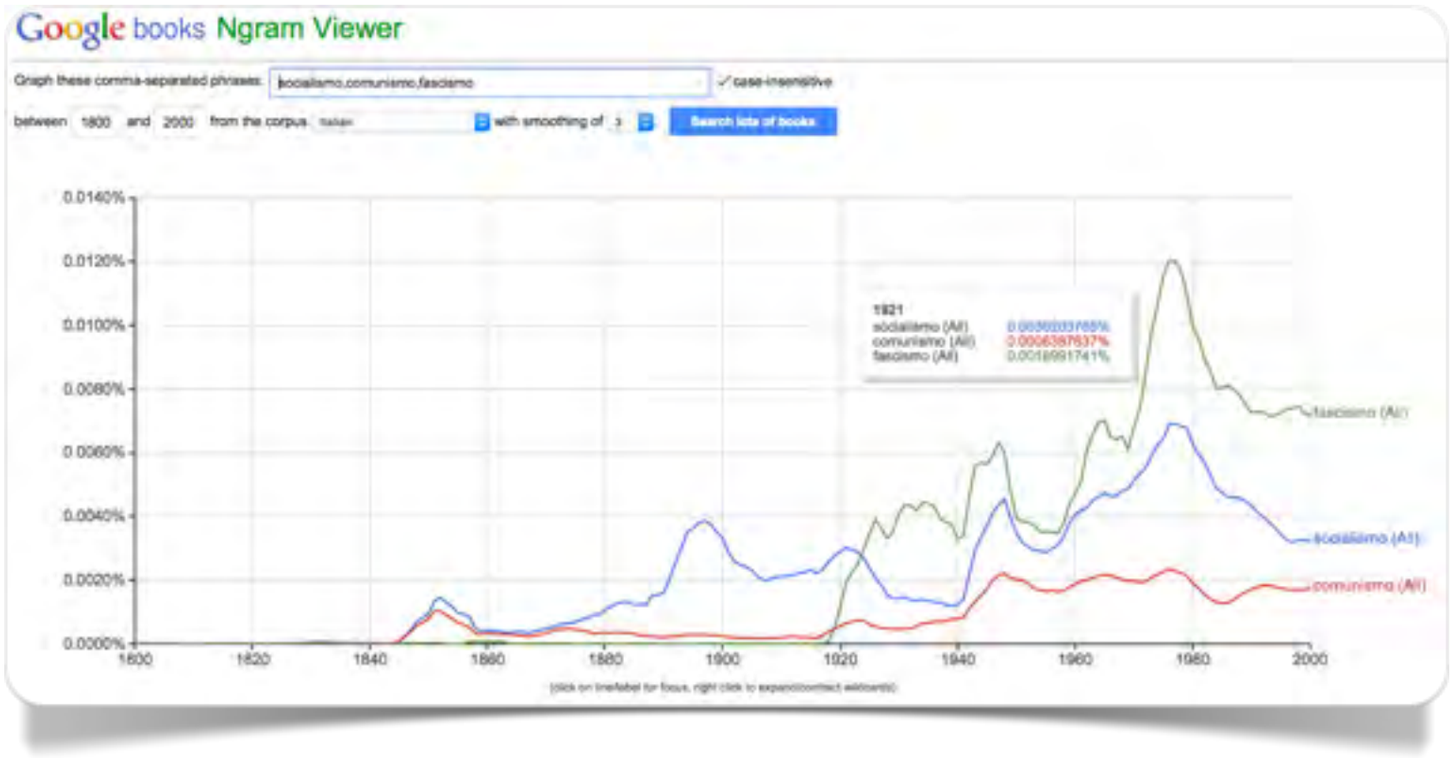


Fig. 14.2 Frequenze percentuali delle parole: socialismo, comunismo, fascismo (elab. Google Books Ngram Viewer)

Nelle figure 14.3 e 14.4 possiamo osservare le “fortune letterarie” di tre grandi poeti dell’Ottocento italiano. La visualizzazione cambia considerevolmente se si utilizzano nome e cognome (bigram) o solo il cognome (unigram).

Fig. 14.3 Frequenze relative percentuali dei bigrams: Giacomo Leopardi, Giosueè Carducci, Giovanni Pascoli (elab. Google Books NGram Viewer)

Il corpus è costituito da libri in senso stretto, con esclusione di periodici, giornali, documenti amministrativi ecc. La cautela di cui si è detto in precedenza è necessaria anche in considerazione dei numerosi errori di riconoscimento delle parole da parte dell’OCR durante la digitalizzazione dei testi. Su questo punto gli autori mettono in evidenza come gli aggiornamenti successivi abbiano apportato miglioramenti considerevoli alla qualità del corpus.

Per ciascuna parola, o per gli n-grams (fino a cinque parole consecutive) con almeno 40 occorrenze, il corpus è consultabile con diverse modalità, compresa la selezione delle categorie grammaticali. La visualizzazione di base

avviene con uno *smoothing* di 3, questo significa che il grafico rappresenta le frequenze relative di ciascuna forma (parola o n-gram) in media mobile: cioè il dato del 1950 è il valore medio di sette valori: quello riferito al 1950, dei tre anni che precedono e dei tre anni che seguono. La media mobile "ammorbidisce" la curva che rappresenta i dati. Nell'analisi delle serie storiche la media mobile ha l'effetto di ridurre le fluttuazioni casuali e permette di apprezzare meglio le tendenze a medio e lungo termine.

Lorem Ipsum dolor amet, consectetur

Capitolo 15

Il “doppio movimento” tra qualità e quantità

In generale tutti questi strumenti hanno potenzialità che vanno al di là degli scopi per i quali sono stati costruiti.

15

Total recall

Come spesso accade nell'informatica le applicazioni devono essere esplorate e utilizzate con una buona dose di saggezza ma anche di inventiva e creatività.

Nel trattamento dei testi digitalizzati si realizza un incontro sempre più fecondo tra scienze che coltivano l'unicità e la singolarità del loro oggetto di studio e scienze che tentano di generalizzare le osservazioni selezionando le proprietà fino a costituire classi di oggetti. Questa distinzione ha portato in passato alla separazione tra scienze umane e scienze della natura, scienze dell'interpretazione e scienze della spiegazione Ora la sintesi è possibile e spetta alle scienze umane e sociali accettare la sfida e muoversi nella direzione di una riconciliazione tra la presunta opposizione di qualità e quantità.

Con basi testuali costituite da centinaia di migliaia o da milioni di parole, il ricercatore, nella impossibilità di leggere direttamente il testo, è costretto a mettere in atto strategie lessicometriche e quantitative per individuare parole-chiave e unità semantiche che presentano qualche interesse rilevante e qualche peculiarità di presenza all'interno dei testi che sta analizzando. Poi,

Fig. 15.1 – *Total Recall* di Paul Verhoven (TriStar Pictures © 1990)

successivamente, potrà assume un punto di vista ermeneutico selezionando frammenti di testo che presentano un interesse particolare in base alla presenza delle parole chiave individuate. Questo “doppio movimento” dalla qualità alla quantità e dalla quantità alla qualità diventa essenziale e irrinunciabile nella gestione della della e-memory, della “memoria totale” di cui sembrerebbe non potremo fare a meno in un futuro non troppo lontano (Bell e Gemmel, 2009).

I pionieri (o i profeti) di questa rivoluzione si sono richiamati idealmente al racconto di Philip Dick in *We Can Remember It For You Wholesale* (1966), popolarizzato dal film *Total Recall* di Paul Verhoven (1990).

Nessuno di noi sarebbe pronto a scommettere su queste anticipazioni del futuro perché troppo spesso il futuro ha imboccato strade che nessuno aveva previsto intenzionalmente. Tuttavia i segnali ci indicano che questa è la direzione verso la quale stiamo andando.

Tracce digitali

Il Web 2.0 con la sua apertura a concetti come interattività, performance, opera aperta, interpretazione soggettiva e spazio cognitivo dell'utente, costruzione di senso, pragmatica esperienziale, scambio tra sistemi informativi, ha portato con sé tutti gli elementi essenziali che fanno capo all'interazione sociale, alle forme di comunicazione e alla "connettività totale" che caratterizzano la sfera della conoscenza, la valorizzazione economica del sapere e il perfezionamento delle strategie decisionali.

Sempre più numerose sono le applicazioni di condivisione delle informazioni personali (fotografie, comportamenti di consumo, letture, posizionamenti satellitari con GPS, hotel e ristoranti frequentati, liste di vini, film che vorremmo consigliare agli amici) e istituzionali (e-governement e open data). In questa rete di inter-connessioni di massa si sviluppano le tracce digitali della e-life e della memoria collettiva del prossimo futuro, in cui non vi sono più confini tra Real Life e Second Life (una distinzione che appartiene all'era "pre-Total Recall").

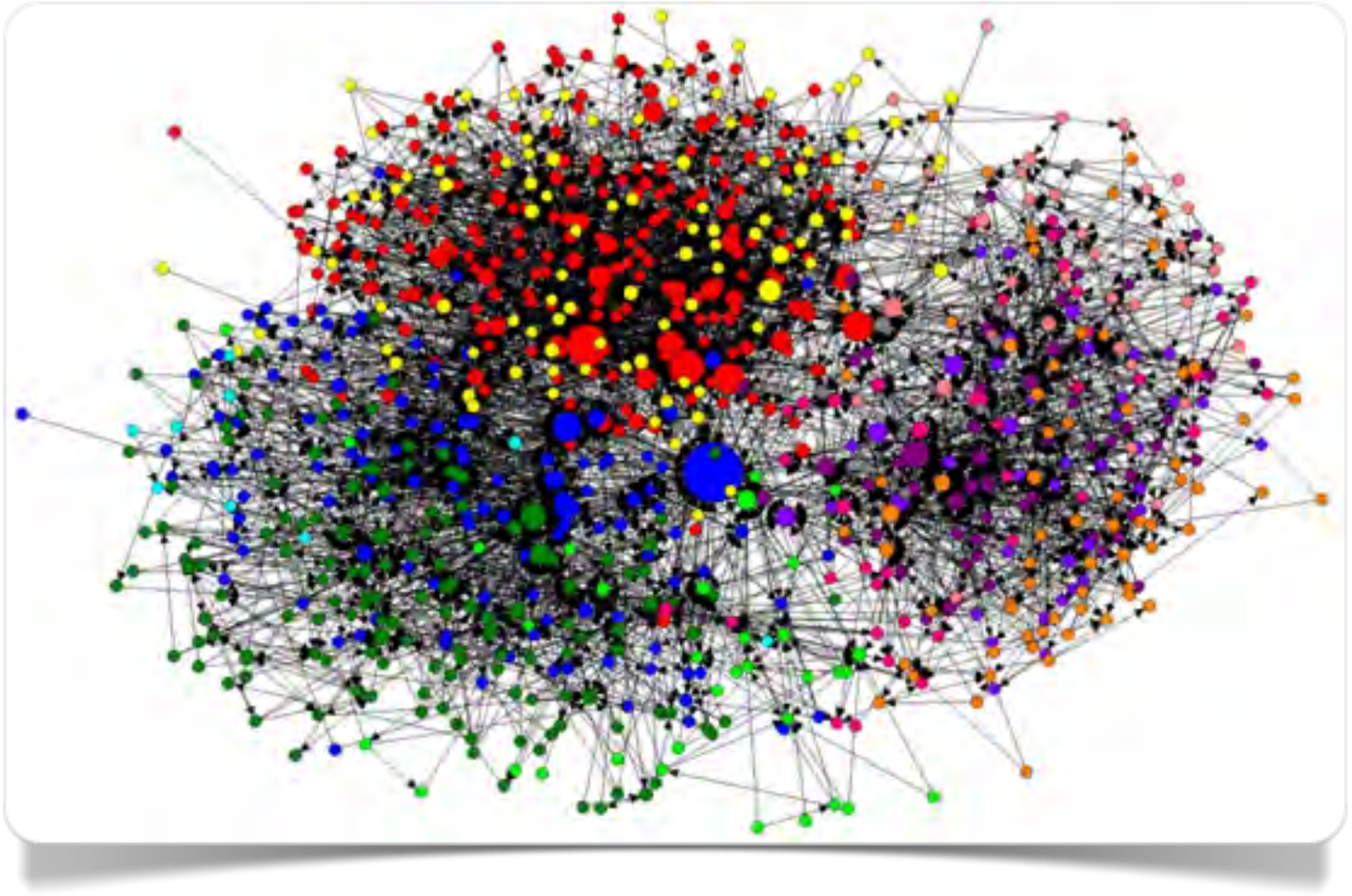


Fig. 15.2 *Social Network Analysis of the Orma* (Jean Ensminger © 2011)

L'analisi automatica dei testi svolge un

ruolo decisivo nella gestione dell'informazione e della comunicazione nella *Network Society*. Statistica e probabilità sono indispensabili per muoversi consapevolmente e con intelligenza in questo oceano di parole, in cui il disordine sembra regnare incontrastato e nel quale, invece, seguendo David Weinberger, sta venendo alla luce un nuovo tipo di ordine in cui i protagonisti e principali organizzatori di contenuti e significati sono gli utenti stessi (Weinberger, 2009).

Capitolo 16

Bibliografia e risorse in web

16

Bibliografia

Amstuz D., Gambette P. (2010) Utilisation de la visualisation en nuage arboré pour l'analyse littéraire. In S. Bolasco, I. Chiari, L. Giuliano (Eds). *Statistical Analysis of Textual Data* (JADT2010: 10th International Conference on statistical analysis of textual data, Rome: 9-11 June). Milano: LED, pp. pp. 227-238.

Brown S., Ruecker S., Antoniuk J., Farnel S., Gooding M., Sinclair S., Patey M., Gabriele S. (2010) Reading Orlando with the Mandala Browser: A Case Study in Algorithmic Criticism via Experimental Visualization. *Digital Studies / Le champ numérique*, 2 (1) - http://www.digitalstudies.org/ojs/index.php/digital_studies/article/viewArticle/191

Bell G., Gemmel J. (2010) *Total Recall. How the E-memory Revolution Will Change Everything*. Penguin Group, 2009.

Bolasco S. (2013) *L'analisi automatica dei testi. Fare ricerca con il text mining*. Milano: Carocci.

Bush V. (1945) As We May Think. *Atlantic Montly*, 176 (July), pp. 101-108.

Chiari I. (2007) *Introduzione alla linguistica computazionale*. Bari: Laterza.

Delahaye J-P., Gauvrit N. (2013) *Culturomics. Le numérique et la culture*. Paris: Odile Jacob.

Feinberg J. (2010) Wordle, in J. Steele e L. Iliinski, *Beautiful Visualization*. Sabastopol, CA: O'Reilly Media, pp. 37-58.

Frantzi K., Ananiadou S., Mima H. (2000) Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3 (2) pp. 115-130.

Gainor R., Sinclair S., Ruecker S., Patey M., Gabriele S. (2009) A Mandala Browser User Study: Visualizing XML Versions of Shakespeare's Plays. *Visible Language*, 43 (1), pp. 60-85.

Gambette P., Véronis J. (2010) Visualising a text with a tree cloud. In Locarek-Junge H. and Weihs C. (Eds). *Classification as a Tool for Research. Studies in Classification, Data Analysis, and Knowledge Organization*. Part 3, SpringerLink, pp. 561-569.

Giuliano L., La Rocca G. (2008) *L'analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso*. Milano: LED.

Gruzd A. (2010) Exploring Virtual Communities with the Internet Community Text Analyzer (ICTA). In Danile Ben Kei (Ed.). *Handbook of Research on Methods and Techniques for Studying Virtual Communities. Paradigms and Phenomena*. Hershey, PA: IGI Global.

Haythornthwaite C., Gruzd A. (2007) A Noun Phrase Analysis Tool for Mining Online Community, in *Proceedings of the 3rd International Conference on Communities and Technologies*, Michigan State University.

Jurafsky D., Martin J.H. (2000) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Sadle River, NJ: Pearson Prentice Hall.

Krippendorf K. (2004) *Content Analysis: An introduction to its Methodology*. Thousand Oaks, CA: Sage Pub. (2nd ed.).

Lo Tzai-Wai R., He B., Ounis I. (2005) Automatically building a stopword list for an information retrieval. *Journal of Digital Information Management*, 3(1), 17-97.

Losito G. (2002) *L'analisi del contenuto nella ricerca sociale*. Milano: Franco Angeli (IV ed.).

Lucisano P. e Piemontese M.E. (1988). Gulpease: una formula per la predizione delle difficoltà dei testi in lingua italiana. *Scuola e Città*, 3, 110-124.

Madge J. (1966) *Lo sviluppo dei metodi di ricerca empirica in sociologia*. Bologna: il Mulino (ed. orig. 1952).

Michel J-B., Yuan Kui Shen Y., Presser Aiden A., Veres A., Gray M.K., Pickett, J.P., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M.A., Lieberman Aiden E. (2010). Quantitative Analysis of Culture Using Millions of Digitized Books, www.sciencexpress.org /16 December 2010/Page 1/10.1126/science.1199644 (downloaded www.sciencemag.org - December 29, 2010).

Nelson T.H. (1981) *Literary Machine*. Swarthmore.

Paley W.B. (2002) TextArc: An alternate way to view a text (retrieved http://www.textarc.org/posters/TextArc_PosterNotes.pdf).

Pang L., Lee L. (2009) Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, Vol. 2, Nos. 1–2, pp. 1–135.

Pennebaker J.W. , Graybeal A. (2001) Patterns of Natural Language Use: Disclosure, Personality, and Social Integration. *Current Directions in Psychological Science*, 10(3), pp. 90-93.

Porter M.F. (1980) An algorithm for suffix stripping. *Program*, 14(3) pp 130–137.

Robertson S.E., van Rijsbergen C. J., Porter M.F. (1981). Probabilistic Models of Indexing and Searching, in R.N. Oddy R.N., S.E. Robertson, C.J. van Rijsbergen, P.W. Williams (Eds.). *Information Retrieval Research*, Proc. Joint ACM/BCS Symposium in Information Storage and Retrieval, Cambridge, June 1980, Butterworths, pp. 35-56.

Rockwell G. (2003) What is Text Analysis, Really? *Literary and Linguistic Computing*, 18(2), pp. 209-219.

Sinclair S. (2003) Computer Assisted Reading: Reconceiving Text Analysis. *Literary and Linguistic Computing*, 18(2), pp. 175-184.

Weinberger D. (2007) *Everything is miscellaneous: the power of the new digital disorder*. New York: Times Books.

Risorse web

Google Books Ngram Viewer (http://ngrams.googlelabs.com/)

LIWC – Linguistic Inquiry and Word Count (http://www.liwc.net/)

Many Eyes (http://www-958.ibm.com/software/data/cognos/manyeyes/)

NETLYTIC - Internet Community Text Analyzer (http://netlytic.org/)

Tagxedo (http://www.tagxedo.com/)

TAPoR – Text Analysis Portal of Research (http://portal.tapor.ca/portal/portal)

TAPoRwere – Prototype of Text Analysis Tools (http://taporware.ualberta.ca/~taporware/htmlTools/comparator.shtml?)

TerMine (http://www.nactem.ac.uk/software/termine/)

Textalyser (http://textalyser.net/)

TextArc (http://www.textarc.org/)

TreeCloud (http://www2.lirmm.fr/~gambette/treecloud/index.htm)

Voyant – See through your text (http://voyeurtools.org/)

Wordle (http://www.wordle.net/)

SAPIENZA
UNIVERSITÀ DI ROMA

Dipartimento di Scienze statistiche

Data Science Series

1. L.Giuliano, *Il valore delle parole. L'analisi automatica dei testi in Web 2.0*.
2. L. Giuliano. *The Value of Words. Automatic Text Analysis Tools in Web 2.0* (in preparazione).
3. D. Schiavon, *Wizard grafico. Una guida alla visualizzazione dei dati numerici* (in preparazione).

Luca Giuliano
Il valore delle parole. L'analisi automatica dei testi in Web 2.0.
Roma : Dipartimento di Scienze statistiche, [2013] 116 p.

ISBN 978-88-908757-0-0