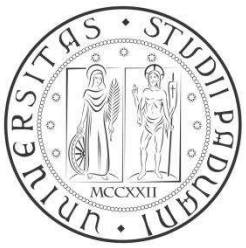
UNIVERSITÀ DEGLI STUDI DI PADOVA
Facoltà di Scienze Statistiche
Corso di Laurea Triennale in Statistica e Tecnologie Informatiche



Tesi di Laurea

Analisi Statistica di Dichiarazioni Politiche tramite "Correlated Topic Model"

Relatore
Livio Finos

Candidata
Sara Baldan

Correlatore
Dario Solari

Anno Accademico 2010/2011

SOMMARIO

INTRODUZIONE

Il presente lavoro si occupa di analizzare collezioni testuali con metodi statistici. Lo scopo primario è valutare uno dei modelli esistenti, il **correlated topic model**, applicandolo su una raccolta di testi in lingua italiana.

Iniziamo con una breve presentazione dell'analisi del testo e della sua evoluzione nel corso del tempo, citando alcuni degli strumenti maggiormente usati, per introdurre il lettore all'argomento.
Il capitolo seguente entra nel merito del *correlated topic model*, descrivendone gli aspetti teorici e i metodi di stima. È presente anche un breve paragrafo per familiarizzare con la distribuzione normale logistica, elemento caratteristico del modello in esame.
Il terzo capitolo guida il lettore attraverso un esempio pratico, in cui il modello viene applicato ad alcune dichiarazioni politiche. Viene descritta in dettaglio non solo la stima dei parametri del modello, ma anche la precedente fase di preparazione dei testi, effettuata con l'aiuto del software statistico **R**. I codici utilizzati sono consultabili sul sito *http://associazionerospo.org/about/opensource/automazioni/*.
Sempre in questo capitolo, si suggeriscono gli elementi da prendere in considerazione per comprendere al meglio il risultato, e si forniscono alcuni grafici molto utili per l'interpretazione.
L'ultimo passaggio è dedicato alle conclusioni, dove vengono discussi vantaggi e svantaggi del modello, assieme ad alcune proposte per studi futuri e a considerazioni sulle ultime tecniche sviluppate.

Infine, per chi fosse interessato e desiderasse saperne di più, nella bibliografia è presente una breve rassegna di articoli che ampliano e approfondiscono l'argomento.

1. L'ANALISI DEL TESTO

L'**analisi testuale**, o analisi del testo, è un insieme di tecniche che permette di analizzare ed esplorare un singolo testo o una raccolta, anche molto ampia, di testi. Non è un'invenzione recente[1], eppure al giorno d'oggi viene sempre più utilizzata. Questo accade in buona parte – ma non soltanto – per la crescente diffusione di Internet, che tramite pagine web, social network, newsgroup, chat e forum rende disponibile un'immensa quantità di informazioni; quantità peraltro in continua crescita. Per riuscire a orientarsi in questo oceano di risorse si deve ricorrere a strumenti quali i motori di ricerca, ed è per questo che si rendono necessarie procedure sempre più avanzate per analizzare, elaborare e catalogare semi-automaticamente le informazioni principali di un testo. Sebbene questa sia una delle applicazioni principali dell'analisi testuale, non è l'unica. Altri rilevanti ambiti di applicazione riguardano, per esempio, interviste, questionari, rassegne stampa, ma anche opere letterarie e teatrali.

Tradizionalmente, vi sono due differenti approcci per condurre un'analisi del testo, uno linguistico e uno statistico. Secondo l'approccio linguistico, un testo contiene un certo numero – grande ma finito – di **elementi significativi**, perciò è possibile stilarne un elenco. Per elementi significativi si intendono tutte le forme grafiche[2] e i segmenti del testo che sono portatori di senso. L'analisi del testo si basa sullo studio di questo elenco, che prevede sia parole singole ("credito", "politica") sia parole composte ("campagna elettorale", "mezzi di comunicazione"). Un limite di questo approccio è l'enorme mole di lavoro richiesta per la stesura dell'elenco (difficoltà oggi parzialmente schivata con l'uso di software *ad hoc*). Inoltre, è opportuno precisare che in questo approccio manca una procedura ben definita che garantisca "trasparenza", riproducibilità e affidabilità. Infatti i risultati dell'analisi sono strettamente connessi alla figura del ricercatore, che individua gli elementi essenziali e i punti chiave. Il

[1] Secondo Krippendorff (1983), autore di uno dei più noti e autorevoli manuali per le scienze sociali, il primo esempio documentato di analisi del contenuto si può rintracciare nella Svezia del XVII secolo, dove fu applicata a testi religiosi. (Tuzzi, 2003, pag. 20)

[2] Si definisce **forma grafica** una sequenza di caratteri appartenenti all'alfabeto della lingua delimitata da due separatori (spazi e punteggiature).

risultato è perciò fortemente **discrezionale**, poiché dipende dalla sensibilità, dalla conoscenza dell'argomento trattato e dal background socio-culturale di chi svolge l'analisi. Nonostante questa caratteristica sia stata inizialmente considerata un limite, è oggi opinione diffusa che non solo sia impossibile condurre un'analisi testuale senza una dose di discrezionalità, ma che quest'ultima possa essere addirittura il vero punto di forza dell'analisi.

Infatti il ruolo delle decisioni del ricercatore rimane determinante per la qualità dell'analisi, in quanto non è possibile (e probabilmente nemmeno auspicabile) rendere automatici tutti i procedimenti di analisi, anche a causa dell'incapacità del ricercatore "di spiegare la sua conoscenza nei termini di un programma per computer" (Krippendorf, 1983, p. 172). Pertanto, lo strumento informatico è uno strumento prezioso per guidare il ricercatore nell'analisi, ma non può sostituirlo, poiché, come fa notare Franco Rositi, l'automazione e i supporti informatici devono "aiutare il giudizio, e non eliminarlo". (della Ratta-Rinaldi, 2005)

L'approccio statistico nasceva in contrapposizione a quello linguistico, quando si cercava una tecnica che liberasse completamente i risultati dalla discrezionalità del ricercatore. L'idea di fondo è molto semplice: più un termine è presente nel testo, più lo rappresenta. Questo metodo ha un primo limite molto evidente: non riesce a tenere conto dei **poliformi**[3]. Si pensi a uno degli esempi suddetti, "campagna elettorale", e a come il significato sia completamente differente se si considerano le due parole divise: "campagna" ed "elettorale"! Un altro limite evidente è la presenza, in qualsiasi testo, di molte parole che in gergo vengono definite **vuote** (o **stopwords**), ovvero termini fondamentali per la corretta strutturazione del discorso e quindi per la sua comprensibilità, ma che non danno alcuna informazione sul significato latente del messaggio. Si pensi alle congiunzioni, agli avverbi, ai verbi, agli ausiliari, …; tutte parole grammaticali che, sebbene in alcuni casi indicative del carattere del discorso[4], non

[3] Un **poliforme** è una sequenza di parole che esprime un contenuto autonomo, differente da quello espresso singolarmente dalle parole che lo compongono.

[4] «Il sovrautilizzo di preposizioni come *in* o *di* sottolinea il carattere descrittivo del discorso; una prevalenza di *non*, *per* e *con* sottolinea particolari intenzionalità del parlante, mentre quella dei *ma* e *se* evidenzia elementi legati ad incertezza.» (Bolasco, 1999, p. 193).

sono certo comparabili alle parole cosiddette **piene**, ricche di significato ai fini dell'analisi. Se lo studio si basa esclusivamente sulla frequenza delle parole, come nell'approccio statistico più classico, è chiaro che queste espressioni grammaticali devono essere filtrate affinché i risultati siano ragionevoli e comprensibili.

Gli studiosi si sono spesso trovati in difficoltà sulla scelta del metodo da utilizzare: meglio quello **qualitativo** (che prevede un approccio linguistico) o quello **quantitativo** (e quindi un approccio statistico)?

Negli ultimi anni del Novecento si è andata sviluppando un'opinione che ha raccolto sempre più consensi, e in effetti sembra la soluzione più ragionevole. Quest'idea si fonda sull'affermazione che in un campo così vasto come l'analisi del testo non ci può essere una netta distinzione tra "qualitativo" e "quantitativo", poiché la stessa analisi testuale presenta sia componenti qualitative (il testo è di per sé qualitativo per eccellenza) sia quantitative (gli strumenti di analisi sono tipicamente statistico-matematici). Le due componenti sono perciò complementari, e un utilizzo saggio di entrambe conduce a risultati migliori.

> Nel caso dell'analisi del contenuto la contrapposizione "quantitativo" versus "qualitativo" è un problema, se non proprio falso, sicuramente mal posto, perché ogni approccio di tipo statistico deve operare mediate strumenti di tipo quantitativo. Non è però quantitativo l'oggetto di studio e per poter trattare statisticamente le informazioni si passa attraverso una forma di codifica. Nell'analisi testuale convivono contesti e significati di parole, di natura puramente qualitativa, con ranghi, frequenze e distribuzioni di probabilità, che sono invece quantitativi, nel rispetto della natura di entrambi. (Tuzzi, 2003)

Così, sono nate e si sono sviluppate tecniche miste – secondo un approccio che potremmo definire "quantiqualitativo" (Tuzzi, 2003) – che rientrano nel vasto ambito delle tecniche statistiche di analisi testuale. Il principale punto di forza di queste tecniche è che riescono a fornire dei buoni risultati anche con grandi moli di dati; quando invece la valutazione della stessa quantità di dati, sottoposta a un approccio linguistico puro, richiederebbe un impiego di tempo e di lavoro davvero notevole. Risulta evidente, dunque, come lo sviluppo di questo nuovo approccio sia in buona

parte una logica conseguenza della crescita esponenziale della quantità di informazioni disponibili al giorno d'oggi, crescita dovuta soprattutto alla diffusione di Internet.

1.1 Alcuni Strumenti di Analisi

Il procedimento classico di analisi del testo comincia con (I) una fase di lettura e di "pulizia" del corpus[5], che serve a prepararlo alla fase successiva: (II) l'analisi vera e propria. Infine vi è (III) l'elaborazione e l'esposizione dei risultati. Il processo di preparazione del corpus può essere più o meno laborioso – a discrezione del ricercatore – e può consistere nella *lemmatizzazione*[6], nello *stemming*[7], nell'individuazione delle forme complesse (ad esempio i *poliformi*) e altro ancora.

Una volta conclusa questa prima parte, si procede con gli strumenti di analisi che si è deciso di utilizzare. Alcuni tra i più affermati sono l'analisi delle corrispondenze, l'analisi delle co-occorrenze, l'individuazione delle parole caratteristiche, del linguaggio peculiare, delle corrispondenze lessicali e altri ancora; tuttavia riteniamo opportuno non entrare nel merito di questi metodi, dato che sono già consolidati e perciò esiste una documentazione consistente, anche con esempi di applicazione, molto più esauriente di quanto potrebbe mai essere questo lavoro. Nel caso in esame, infatti, la fase di analisi viene svolta con l'utilizzo di un preciso strumento statistico: un modello per l'analisi statistica di documenti e altri dati discreti. Questo modello viene presentato nel capitolo seguente.

---

[5] Un **corpus** è una collezione di testi selezionati e organizzati per facilitare le analisi linguistiche.

[6] La **lemmatizzazione** è il processo di riduzione di una forma flessa di una parola alla sua forma canonica, detta *lemma*. Per esempio, in italiano il verbo *camminare* può apparire come *cammina*, *camminò*, *camminando* e così via. La forma canonica, *camminare*, è il lemma della parola ed è la forma di riferimento per cercare la parola all'interno di un dizionario.

[7] Lo **stemming** è il processo di riduzione della forma flessa di una parola alla sua forma radice, detta *tema*. Il tema non corrisponde necessariamente alla radice morfologica (lemma) della parola: normalmente è sufficiente che le parole correlate siano mappate allo stesso tema (ad esempio, che *andare*, *andai*, *andò* mappino al tema *and*), anche se quest'ultimo non è una valida radice per la parola.

2. IL CORRELATED TOPIC MODEL

Il **correlated topic model**[8] (CTM) è un'evoluzione del **latent Dirichlet allocation**[9] (LDA), un modello con variabili latenti (cioè variabili delle quali non siamo in grado di rilevare le realizzazioni) secondo cui le parole di ogni documento sono realizzazioni di misture di *argomenti* (*topics*). In particolare, ogni *argomento* si concretizza in una distribuzione di probabilità di tipo **multinomiale** su un vocabolario prefissato (l'insieme delle parole utilizzabili nel corpus). Ogni *argomento* è compatibile con ogni documento del corpus (cioè ogni documento può trattare di qualsiasi *argomento*), ma la sua frequenza (che possiamo intendere come "rilevanza": se si parla molto di quell'*argomento*, esso sarà rilevante per il documento in esame) all'interno del documento varia stocasticamente tra i documenti, poiché tale frequenza – che viene vista come un vettore in cui l'*i*-esimo elemento rappresenta la rilevanza del *topic i* all'interno del documento – si assume essere realizzazione di una variabile casuale di Dirichlet.

Il limite maggiore del modello LDA è la sua impossibilità di prevedere **correlazione** tra gli *argomenti*. È naturale pensare che, se in un documento si parla di concerti, nello stesso documento sarà più facile leggere anche di gruppi musicali o di orchestre piuttosto che di cucina, perché l'argomento relativo ai concerti è più attinente a temi musicali piuttosto che a temi culinari. Tuttavia la struttura del modello LDA non è in grado di cogliere questa relazione, in quanto i vari elementi di un vettore casuale di Dirichlet si assumono essere indipendenti. L'unico modo per risolvere questa pesante restrizione è ricorrere a un'altra ipotesi di distribuzione delle frequenze. Ciò avviene nel CTM, dove le frequenze si ipotizzano essere realizzazioni di una variabile casuale **normale logistica**. In questa maniera siamo in grado di incorporare una struttura di covarianza tra le frequenze degli *argomenti* di un documento, e ne otteniamo un modello più realistico.

[8] Blei e Lafferty, 2007.

[9] Blei, Ng e Jordan, 2003.

2.1 ASSUNZIONI

Il *correlated topic model* può essere visto come un **modello grafico probabilistico** (o **rete bayesiana**), qui rappresentato in figura 1, il cui funzionamento si può descrivere con il seguente processo:

– dato K (numero di *topics*), D (numero di documenti), V (numero di parole del vocabolario), N_d (numero di parole del d-esimo documento), β_i (vettore di lunghezza V: distribuzione di probabilità sul vocabolario dell'*argomento* i-esimo);

1. si estrae η_d da una variabile casuale normale $(K-1)$-variata $\mathcal{N}_{K-1}(\mu, \Sigma)$; da η_d si ricava $\theta_d = \mathrm{f}(\eta_d)$ [10], che è il vettore di frequenza dei *topics* nel documento d;

2. per ogni parola n del documento d si ricava l'*assegnazione tematica* $z_{d,n}$ da una multinomiale $\mathcal{Z}_{d,n}$ di parametro θ_d. L'assegnazione tematica indica l'*argomento* dal cui vocabolario (o meglio, dalla cui distribuzione sul vocabolario) si estrarrà la parola n. In parole povere, $z_{d,n}$ indica il *topic* da cui proviene la parola n. Al termine di questo passo, dunque, avremo N_d realizzazioni di $\mathcal{Z}_{d,n}$;

3. grazie a $z_{d,n}$ si ottiene il corrispondente $\beta_{z_{d,n}}$. Esso diventa il parametro di una multinomiale $\mathcal{W}_{d,n}$ – che permette quindi di estrarre dalla distribuzione sul vocabolario per l'*argomento* $\beta_{z_{d,n}}$ – la cui realizzazione restituisce la vera e propria parola $w_{d,n}$, , che diventerà l'n-esima parola del d-esimo documento;

4. il d-esimo documento w_d è l'insieme ordinato delle N_d parole trovate in questo modo.

Il procedimento viene quindi ripetuto per ogni documento del corpus $\{w_1, \ldots, w_D\}$.

Ricordiamo che per tornare al modello LDA è sufficiente modificare il punto 1., dove in particolare θ_d viene estratto da una variabile casuale di Dirichlet piuttosto che da una normale logistica.

[10] Vedi paragrafo 2.2: *La Normale Logistica*.

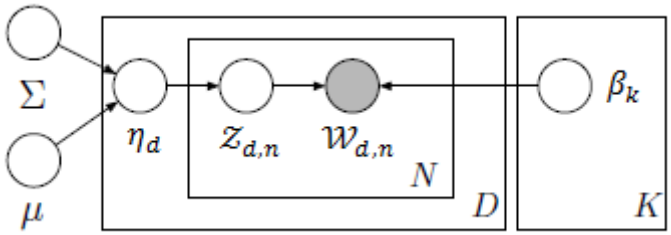


Figura 1: rappresentazione grafica probabilistica del modello CTM. I cerchi rappresentano variabili casuali latenti (se bianchi) od osservabili (se colorati). I rettangoli indicano più replicazioni.

Si noti che in questo modello le uniche variabili direttamente osservabili sono le parole $w_{d,n}$, tutte le altre sono invece **variabili latenti**. Inoltre, i parametri di nostro interesse – di cui vogliamo ottenere una stima – sono soltanto μ, Σ e β_i (per ogni i); dato che sono necessari e sufficienti per stimare le parole di un documento.

2.2 La Normale Logistica

La distribuzione normale logistica viene usata nel CTM per incorporare la correlazione tra *topics* all'interno di uno stesso documento.
Di fatto, è molto semplice ricavare questa distribuzione a partire da una normale multivariata. Infatti possiamo definire che

$$u \text{ è realizzazione di } \mathcal{U} \sim \mathcal{L}_H(\mu, \Sigma)$$

con $\mathcal{L}_H$ indicante distribuzione normale logistica H-variata, μ vettore H-variato, e Σ matrice $H \times H$ di varianza e covarianza; se

$$u = f(v) = \frac{e^v}{1+\sum_{j=1}^{H} e^{v_j}} \qquad \Leftrightarrow \qquad v = f^{-1}(u) = ln\frac{u}{1-\sum_{j=1}^{H} u_j}$$

dove quindi f indica l'inversa della funzione *logit*, v_j e u_j indicano il j-esimo elemento rispettivamente dei vettori v e u, e

$$v \text{ è realizzazione di } \mathcal{V} \sim \mathcal{N}_H(\mu, \Sigma),$$

con $\mathcal{N}_H$ che indica la distribuzione normale H-variata.

Nel CTM, H è il numero di *topics* meno 1 (ovvero $K-1$). Infatti, non dimentichiamo che il vettore u serve per trovare θ, cioè una distribuzione di probabilità – dove la somma degli elementi dev'essere 1. Per questo estraiamo dalla normale soltanto $K-1$ elementi, visto che la K-esima probabilità si può ottenere con

$$u_K = 1 - \sum_{j=1}^{K-1} u_j, \text{ e si ha che } \theta = u^* = [u^T \ u_K] = [u_1\, u_2 \ \dots\ u_{K-1}\ u_K].$$

La differenza fondamentale tra $\mathcal{U}$ e $\mathcal{V}$ è che, mentre quest'ultima può variare in tutto lo spazio reale $\mathbb{R}^{K-1}$, la prima invece varia soltanto in $\mathbb{S}^{K-1}$ definito come

$$\mathbb{S}^{K-1} = \{u \in \mathbb{R}_+^{K-1} : u_1 + \dots + u_{K-1} < 1\}.$$ [11]

Ciò equivale a dire che u^* varia in

$$\mathbb{S}^* = \{u^* \in \mathbb{R}_+^K : u_1 + \dots + u_{K-1} + u_K = 1\}$$

ed è dunque adatto a rappresentare uno spazio di probabilità. È chiaro che $\mathbb{S}^{K-1}$ rappresenta il supporto di $\mathcal{U}$.

La distribuzione di probabilità della normale logistica è facilmente ricavabile da quella della normale, semplicemente sostituendo v con la corrispondente espressione di u. Si ottiene allora:

$$p(u) = |2\pi\Sigma|^{-\frac{1}{2}} \left(\prod_{j=1}^{K-1} u_j\right)^{-1} \exp\{-\frac{1}{2}[ln\frac{u}{u_K} - \mu]^T \Sigma^{-1} [ln\frac{u}{u_K} - \mu]\} \qquad \text{per } u \in \mathbb{S}^{K-1}.$$

2.3 Stima dei Parametri

I parametri del CTM si possono stimare tentando di massimizzare la verosimiglianza del corpus in funzione di $\beta_{1,\dots,K}$ e della $\mathcal{N}_{K-1}(\mu, \Sigma)$. Tuttavia, per riuscirci dovremmo prima marginalizzare la struttura di verosimiglianza relativa alle variabili latenti, e questo non è possibile a causa della natura stessa di tali variabili. Per risolvere il problema si procede allora con il metodo **expectation–maximization** (EM), nella sua variante "**variational**". L'EM tradizionale prevede due passi: il primo (anche detto E-

[11] Si noti che con $\mathbb{R}_+^{K-1}$ indichiamo lo spazio $(K-1)$-dimensionale $(0, +\infty) \times \dots \times (0, +\infty)$.

step, da *expectation*) in cui si calcola la distribuzione a posteriori[12] delle variabili latenti condizionandosi ai dati e ai parametri correnti del modello; e il secondo (anche detto M-step, da *maximization*) in cui si effettua una stima di massima verosimiglianza dei parametri condizionandosi ai dati e alla distribuzione delle variabili latenti – che viene vista come valore atteso delle statistiche sufficienti – trovata al punto precedente. Il procedimento viene reiterato finché si giunge a convergenza.

La distribuzione a posteriori delle variabili latenti, espressa per il documento w, è

$$p(\boldsymbol{\eta}, z \mid w, \boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{p(\boldsymbol{\eta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} p(z_n \mid \boldsymbol{\eta}) p(w_n \mid z_n, \boldsymbol{\beta}_{1:K})}{\int p(\boldsymbol{\eta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n \mid \boldsymbol{\eta}) p(w_n \mid z_n, \boldsymbol{\beta}_{1:K}) \, d\boldsymbol{\eta}}.$$

Si intuisce facilmente che questa quantità non è direttamente calcolabile. Anche se sfruttassimo un modello con distribuzioni coniugate (come avviene nell'LDA, con la Dirichlet e la multinomiale), cosa che permetterebbe una buona semplificazione dell'espressione, probabilmente non riusciremmo comunque a calcolarla analiticamente a causa della presenza di un'operazione dell'ordine di K^N (data dalla somma dei K valori z_n all'interno del prodotto delle parole w_n). Nel CTM la situazione si complica ulteriormente, perché la distribuzione Gaussiana non è coniugata alla multinomiale e l'integrale a denominatore diventa intrattabile. Per questo motivo si ricorre al **variational EM**, che permette di aggirare l'ostacolo. Nel *variational EM* il primo step dell'EM classico non utilizza più la "vera" distribuzione a posteriori, ma una sua approssimazione ottenuta tramite metodi variazionali[13] (*variational methods*).

La funzione obiettivo del *variational EM* è

$$L\left(\mu, \Sigma, \beta_{1,\dots,K};\ w_{1,\dots,D}\right) \geq \sum_{d=1}^{D} E_{q_d}\left[\log p\left(\eta_d, z_d, w_d \middle|\ \mu, \Sigma, \beta_{1,\dots,K}\right)\right] + H(q_d)$$

[12] Si supponga di conoscere la distribuzione a priori $p(\theta)$ di un'osservazione X con verosimiglianza $p(X|\theta)$. Allora la **distribuzione a posteriori** si definisce come: *distribuzione a posteriori* $\propto$ *distribuzione a priori* $\times$ *verosimiglianza*, ovvero $p(\theta|X) \propto p(\theta) \times p(X|\theta)$.

[13] L'idea che sta dietro i **variational methods** è di ottimizzare i parametri liberi di una distribuzione sulle variabili latenti, così che la distribuzione ottenuta sia vicina a quella posteriore "reale". Questa vicinanza viene misurata con la divergenza di *Kullback-Leibler*.

dove $H(q_d)$ denota l'**entropia** della *distribuzione variazionale* q_d. La funzione obiettivo è stata ottenuta vincolando la verosimiglianza del corpus alla disuguaglianza di Jensen. Tale disuguaglianza, espressa per un singolo documento, è

$$\log p(w_{1,\dots,N}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta}) \geq E_q[\log p(\boldsymbol{\eta}|\boldsymbol{\mu},\boldsymbol{\Sigma})] + \sum_{n=1}^{N} E_q[\log p(z_n|\boldsymbol{\eta})] + \sum_{n=1}^{N} E_q[\log p(w_n|z_n,\boldsymbol{\beta})] + H(q).$$

Si noti che i valori attesi vengono calcolati in funzione di q (distribuzione variazionale delle variabili latenti, la cui entropia è $H(q)$).

Per trovare la soluzione, è necessario introdurre dei *parametri variazionali* λ_i, ν_i^2, tali che $E_q[e^{\eta_i}] = e^{\lambda_i + \nu_i^2/2}$ per $i \in \{1, \dots, K\}$.
L'algoritmo *variational EM*, allora, nell'E-step cercherà di massimizzare il vincolo in funzione dei *parametri variazionali*, effettuando dell'*inferenza variazionale* per ciascun documento. Durante l'M-step, invece, il vincolo verrà massimizzato in funzione dei parametri del modello. Questo equivale a trovare la stima di massima verosimiglianza dei parametri $\{\beta, \mu, \Sigma\}$ utilizzando il valore atteso delle statistiche sufficienti – valore atteso che si calcola rispetto alle *distribuzioni variazionali* trovate nell'E-step.

Per ogni iterazione, avremo

$$\widehat{\boldsymbol{\mu}} = \frac{1}{D}\sum_d \lambda_d \,,$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{D}\sum_d I\nu_d^2 + (\lambda_d - \hat{\mu})(\lambda_d - \hat{\mu})^T \,,$$

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} \propto \sum_d [\phi_d]_i \, m_d \,,$$

dove $[\phi_d]_i$ è l'elemento i-esimo di ϕ_d, vettore medio della multinomiale di assegnazione tematica $\mathcal{Z}_d$, e m_d è un vettore V-variato che conta quante volte ciascuna parola del vocabolario compare nel documento d.
I due passi vengono ripetuti finché il vincolo imposto sulla verosimiglianza converge, e i relativi $\{\hat{\beta}, \hat{\mu}, \hat{\Sigma}\}$ diventano le stime dei parametri.

3. IL CASO *OPENPOLIS*

Openpolis è un'associazione senza fini di lucro costituita nel 2008, che gestisce vari progetti: *openparlamento*, *openpolis*, *voisietequi*.

La filosofia su cui si basa l'associazione è di garantire a ogni cittadino un'informazione politica imparziale e consapevole, allo scopo di aumentare la partecipazione pubblica alla vita politica, tramite «l'accesso alle informazioni pubbliche, la trasparenza dei dati, la partecipazione democratica» senza che questa sia influenzata da interessi politici o economici di parte.

Come si può leggere sul sito, «il patrimonio dell'associazione è costituito dalla credibilità che i nostri progetti hanno saputo guadagnarsi nel tempo e dalla crescita continua della comunità di utenti e delle migliaia di persone che ogni giorno consultano i nostri siti.»

Lo statuto completo dell'associazione si può visionare alla pagina *www.openpolis.it/chisiamo*.

In questo lavoro ci interessiamo all'omonimo progetto gestito dall'associazione: **openpolis**, i cui obiettivi e modalità di funzionamento sono descritti dettagliatamente nel sito. Per chiarezza, di seguito riportiamo un estratto di tale descrizione.

«*www.openpolis.it* raccoglie le informazioni su tutti i politici italiani dal più piccolo Comune al Parlamento Europeo e le mette in rete a disposizione di tutti gratuitamente. Ad ognuno dei circa 150.000 rappresentanti in carica è dedicata una scheda dove viene ricostruito il profilo (dati anagrafici, carriera nelle istituzioni, nei partiti, in aziende, etc.) e dove vengono raccolte le sue dichiarazioni pubbliche.

Il database è enorme e i dati sono soggetti inevitabilmente a cambiamenti frequenti, quindi il metodo della redazione distribuita presso tutti gli utenti è l'unico in grado di assicurare un certo grado di affidabilità e aggiornamento delle informazioni. Sono i cittadini stessi che verificano, correggono, aggiungono e aggiornano i contenuti. Il singolo utente può "adottare" il proprio rappresentante, o la propria comunità, per monitorare da vicino le attività e aggiornare le informazioni che lo riguardano, diventa cioè responsabile – insieme ad altri – della correttezza e della qualità dei contenuti delle schede dei politici

"adottati". Inoltre gli utenti possono raccogliere le dichiarazioni pubbliche (documentate con link alla fonte) dei rappresentanti in modo da documentare nel tempo le posizioni politiche su vari temi. [...]»

3.1 LA RACCOLTA DELLE DICHIARAZIONI

Il nostro obiettivo è testare l'efficacia del CTM valutandolo su un caso concreto[14], e a questo scopo utilizziamo l'insieme delle dichiarazioni dei politici, consultabili su *www.openpolis.it*, fino al 1 dicembre 2010 e a partire da molti anni prima (alcune dichiarazioni risalgono addirittura agli anni '80 e '90). Tale raccolta, che costituirà il nostro corpus, ci è stata gentilmente concessa dall'associazione stessa. A questo punto disponiamo di 12.807 dichiarazioni nella forma di altrettanti file di testo. Precisamente, un generico file "*dichiarazione.txt*" presenta la seguente struttura:

```
titolo

data

autore

testo della dichiarazione
```

Come esempio, vediamo la dichiarazione contenuta nel file "*357126.txt*":

```
Standard and Poor's smentisce Berlusconi e Alemanno - NESSUN BUCO nei conti del Comune di Roma

21/06/2008

WALTER VELTRONI,125671

<hr />

<b>Standard and Poor's smentisce Alemanno e Berlusconi sul debito del Comune di Roma.</b><br />

<hr />

<b>Non 10 miliardi come ripetuto in un mantra dalle destre, ma 6,9 miliari di euro. E non declassa il Campidoglio. Lo spiega in un intervista a La Stampa Myriam fernandez de Heredia, responsabile per Standard and Poor's dei giudizi sul merito di credito del settore pubblico in Europa.</b><br />
```

[14] L'applicazione è ancor più interessante se si considera che – a quanto ci è dato sapere – il CTM non è ancora stato usato su un corpus in lingua italiana.

La litania sul megadebito ripetuta come un disco rotto dal neosindaco e rilanciato ieri da Silvio Berlusocni in un imbarazzante show da Bruxelles, dove ha accusato l'ex sindaco di Roma e segretario del PD : "Non c'è nessuna città d'Europa che ha lasciato un deficit di 16 mila miliardi di vecchie lire". E S&P riaccende i riflettori sul debito di Milano, governata dalla Lega e dagli uomini di Berlusconi fin dal 1993.

Alle accuse del premier ha replicato anche l'Unità smontando in un pezzo molto dettagliato il bluff della destra.

E dopo le dichiarazioni di S&P il primo a parlare è stato Marco Causi, deputato Pd ed ex assessore al bilncio capitolino fin dal 2001: "L'intervista della dottoressa Fernandez, responsabile di Standard and Poors per il settore pubblico in Europa, pubblicata oggi sulla Stampa fa giustizia della grande mistificazione costruita negli ultimi giorni intorno ai conti del Comune di Roma. Standard and Poors conferma infatti in 6,9 miliardi il livello del debito comunale e afferma che la relazione della Ragioneria generale dello Stato chiesta da Alemanno, "non dà nessuna informazione nuova, sono dati che già avevamo".

E' esattamente questo che non mi stanco di ripetere da settimane. Non c'è nessun dato che la giunta Veltroni abbia occultato e il livello del debito del Comune è cresciuto durante la gestione Veltroni meno di quello nazionale e meno del tasso d'inflazione.

Non ha alcun senso utilizzare termini come buco o bancarotta. Chi lo fa è in malafede e dimostra così di voler soltanto montare una palese strumentalizzazione politica".

E anche Milano per S&P finisce sotto accusa.

Se n'è accorto Causi: "Standard and Poors (cita il pezzo ndr) sottolinea che il comune di Milano ha un debito elevato. Lo stiamo monitorando, li stiamo monitorando entrambi Roma e Milano" e ricorda così alla politica italiana, e soprattutto al Presidente del Consiglio, che invece di montare inaccettabili montature politiche, sarebbe meglio riflettere su come aiutare non solo Roma ma tutte le grandi città italiane che, dopo anni di restrizione e di tagli, rischiano di non avere più fiato sufficiente a realizzare le importanti infrastrutture, soprattutto di trasporto, di cui hanno bisogno".

Anche per Paolo Gentiloni, responsabile della comunicazione del Partito democratico, il giudizio di S&P interviene a "cancellare le accuse rumorose e gonfiate contro Veltroni e gli anni della sua amministrazione". L'agenzia di rating, per Gentiloni, dice che "i conti romani sono esattamente quelli conosciuti, non c'è nessun buco e tantomeno dei debiti occultati".

I buchi veri - prosegue Gentiloni - ce li hanno lasciati gli amici di Alemanno, con i guasti nella sanità della regione. Ora questa vicenda, fatta solo di accuse propagandistiche, dovrà finire. Alemanno, dopo aver cercato scuse per non farlo, sarà costretto a governare e dovrà dimostrare di saper mantenere le sue promesse. E Berlusconi deve smettere di cercare di delegittimare il leader del maggior partito dell'opposizione".

```
Mentre <b>Silvio Di Francia</b> si chiede: <b>“Ora tappezzeranno Roma di manifesti
per dire scusate, non sappiamo leggere i bilanci ?.</b> Dispiace che tra chi ha
effettuato le ripetute ricognizioni sul debito annunciate dal sindaco Alemanno
nessuno abbia avuto in mente di interpellare Standard & Poor’s l’agenzia che
monitora il debito di diverse città in tutto il mondo. Senza voler ascoltare le
nostre spiegazioni potevano interpellare chi per mestiere dà giudizi sul credito
del settore pubblico per sentirsi dire che non c’è nessun buco nascosto. <br />

<br />
```

Per avere un’idea delle caratteristiche del corpus sono utili alcune rappresentazioni grafiche. Per esempio, il numero di dichiarazioni presenti per ogni politico si può vedere tramite un diagramma, qui rappresentato in figura 2 con gli autori già ordinati per frequenze decrescenti. È naturale pensare che i politici con più dichiarazioni siano anche i più “discussi” nel sito, tuttavia nel grafico non è facile individuarli tra i più di 1800 politici rappresentati. La figura 3 è un dettaglio della precedente, che mostra solo i politici per cui sono presenti almeno 100 dichiarazioni: i personaggi che rispettano questa condizione sono esattamente venti.

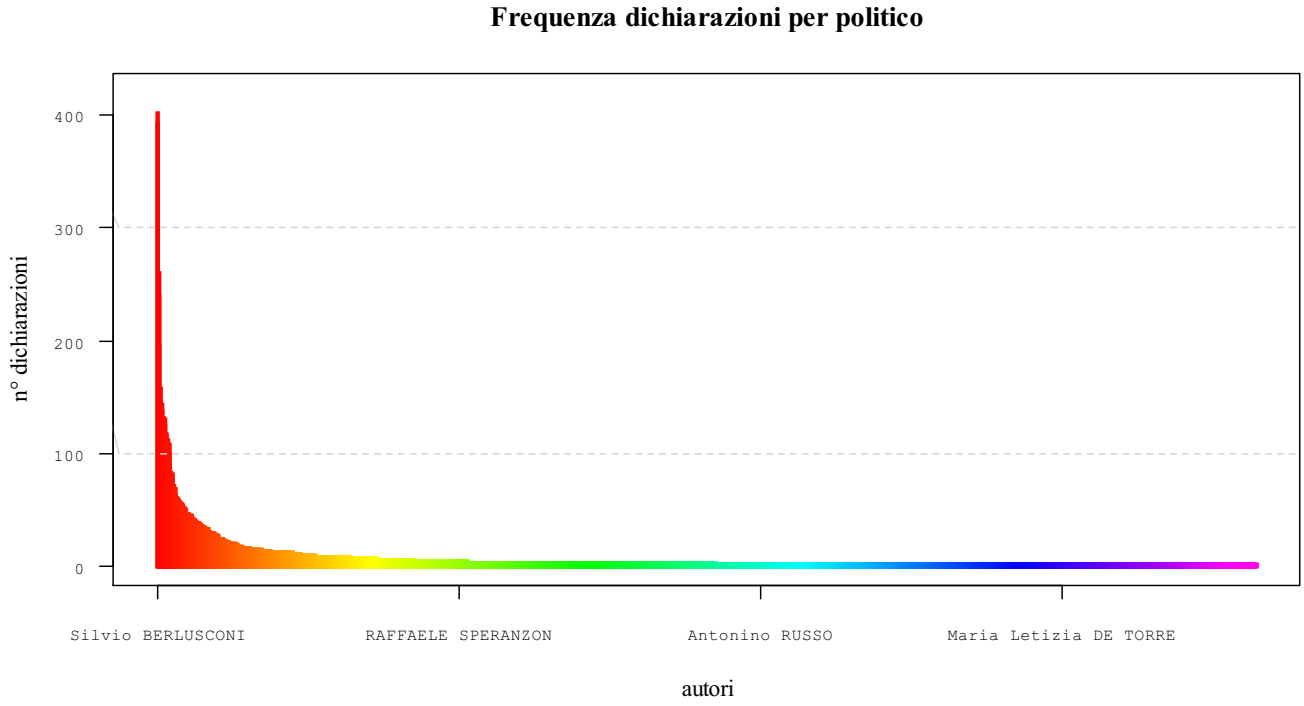


Figura 2: diagramma di frequenza delle dichiarazioni per ogni politico. Si vede che il politico con più documenti inseriti è il nostro premier, Silvio Berlusconi.

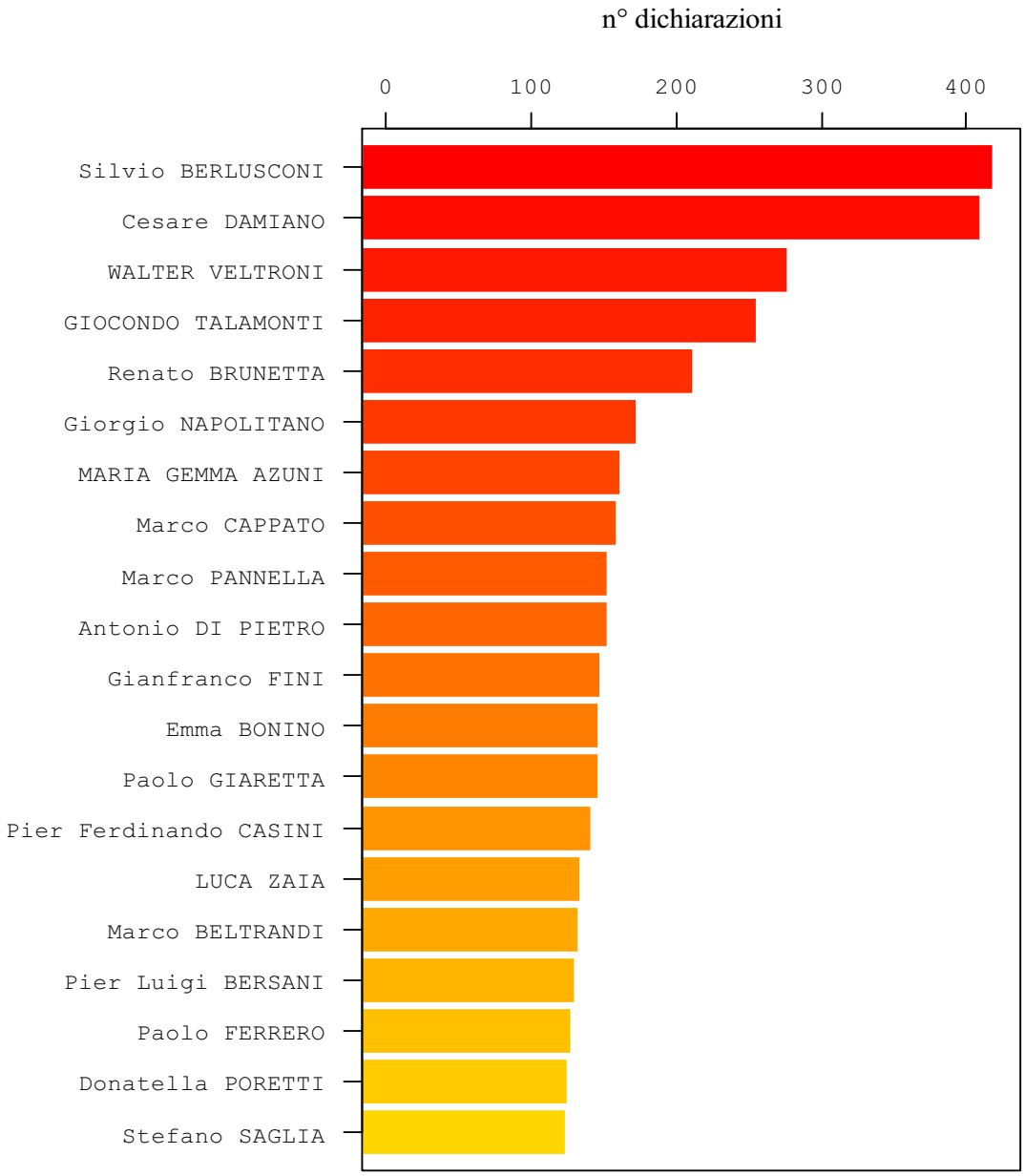


Figura 3: diagramma di frequenza delle dichiarazioni per i politici più presenti in *openpolis*.

Può essere interessante anche vedere la frequenza di dichiarazioni nel tempo, rappresentata in un diagramma in figura 4, e più in dettaglio per gli ultimi mesi in figura 5.

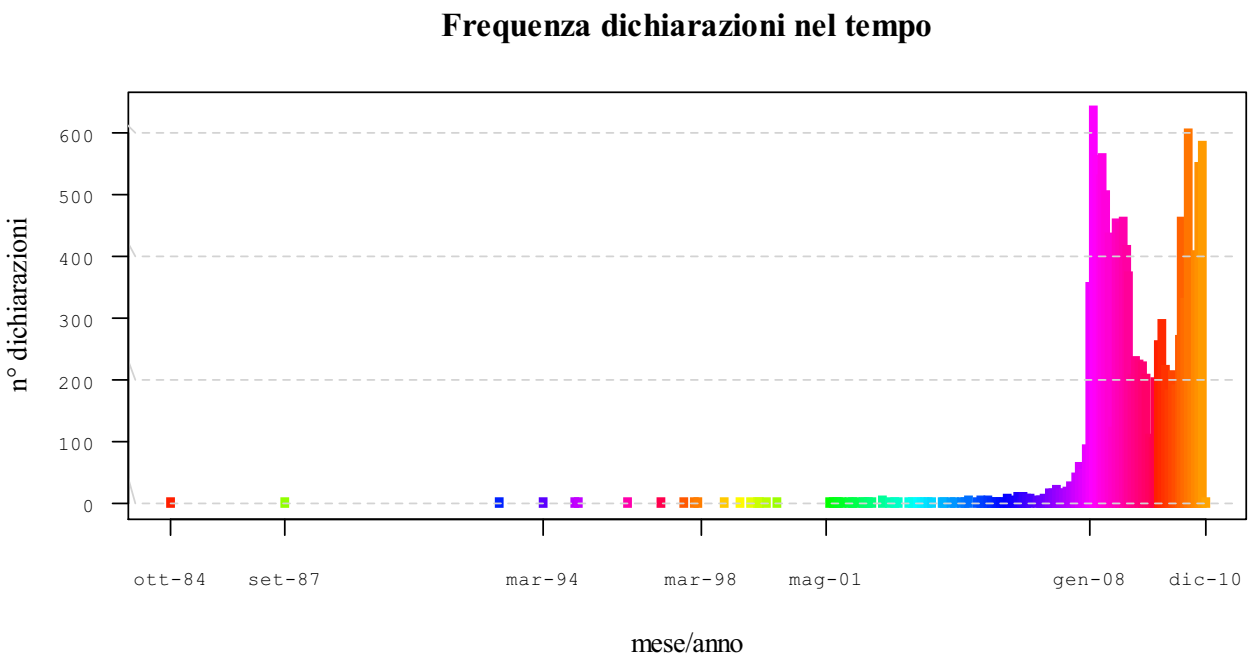


Figura 4: diagramma di frequenza delle dichiarazioni nel tempo, da ottobre 1984 a dicembre 2010.

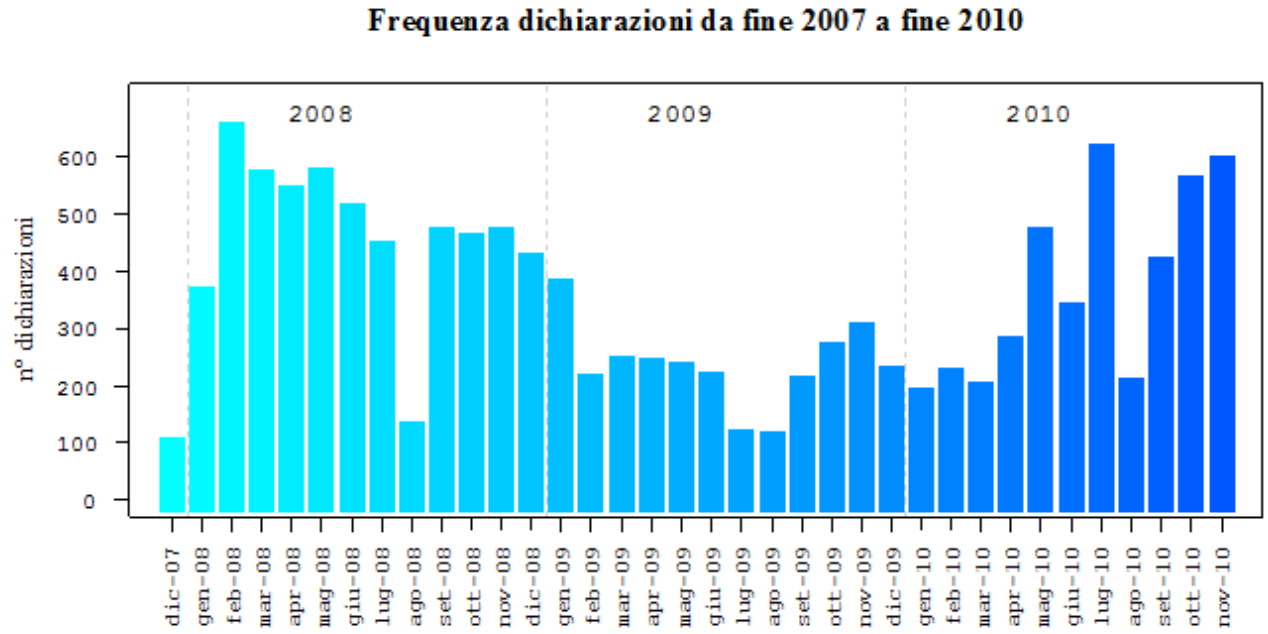


Figura 5: diagramma di frequenza delle dichiarazioni nel tempo, da dicembre 2007 a novembre 2010. La figura non mostra il mese di dicembre 2010 in quanto conta soltanto 2 dichiarazioni.

La distribuzione del numero di dichiarazioni nel tempo è visibilmente non uniforme, con un gran numero di dichiarazioni concentrate negli ultimi 3 anni. Questo non ci sorprende, dal momento che il sito è nato nel 2008. Viceversa, la presenza di articoli antecedenti la data di creazione del progetto si può spiegare ipotizzando che alcuni utenti abbiano fornito dichiarazioni raccolte in passato ma ritenute comunque rilevanti per spiegare e confrontare avvenimenti di attualità.

3.2 LETTURA ED ELABORAZIONE DEI DATI

Per l'elaborazione dei nostri dati usiamo il software statistico **R** (*www.r-project.org*). Di seguito descriviamo i passaggi effettuati omettendo i comandi usati, che sono consultabili liberamente su *http://associazionerospo.org/about/opensource/automazioni/*.
Il primo passo è naturalmente la lettura dei file, in questa fase i testi vengono ordinati per data crescente, e viene effettuato anche un controllo volto all'eliminazione di eventuali dichiarazioni identiche. Al termine di questa fase avremo una lista contenente cinque vettori: uno per le dichiarazioni, il secondo con il numero di riferimento di ogni dichiarazione, e gli ultimi tre contenenti il titolo, l'autore e la data.

Per esempio, la lettura del file *"357126.txt"* porterà ad avere:

```
doc$titolo
 "Standard and Poor's smentisce Berlusconi e Alemanno – NESSUN BUCO nei
conti del Comune di Roma"

doc$numero
 357126

doc$data
 "21/06/2008"

doc$autore
 "WALTER VELTRONI,125671"

doc$testo
 "<hr /><b>Standard and Poor's smentisce Alemanno e Berlusconi sul debito
del Comune di Roma.</b><br /><hr /><b>Non 10 miliardi come ripetuto in un
mantra dalle destre, ma 6,9 miliari di euro. [...]"
```

Si noti che nella fase di lettura del testo sono stati anche eliminati i caratteri di "a capo", in modo da avere tutto il testo in una sola riga.

Per semplificare i passaggi seguenti, trasformiamo tutti i testi in minuscolo con la funzione `tolower()`.

PRIMA

```
doc$testo
 "<hr /><b>Standard and Poor’s smentisce Alemanno e Berlusconi sul debito del Comune di Roma.</b><br /><hr /><b>Non 10 miliardi come ripetuto in un mantra dalle destre, ma 6,9 miliari di euro. [...] Senza voler ascoltare le nostre spiegazioni potevano interpellare chi per mestiere dà giudizi sul credito del settore pubblico per sentirsi dire che non c’è nessun buco nascosto. <br /><br />"
```

DOPO

```
doc$testo
 "<hr /><b>standard and poor’s smentisce alemanno e berlusconi sul debito del comune di roma.</b><br /><hr /><b>non 10 miliardi come ripetuto in un mantra dalle destre, ma 6,9 miliari di euro. [...] senza voler ascoltare le nostre spiegazioni potevano interpellare chi per mestiere dà giudizi sul credito del settore pubblico per sentirsi dire che non c’è nessun buco nascosto. <br /><br />"
```

Il passo successivo prevede di “fare pulizia” nel corpus, ovvero risolvere imprecisioni ed effettuare aggiustamenti per renderlo più omogeneo e, di conseguenza, elaborabile. Il primo problema da risolvere è quello della codifica dei testi. Infatti le dichiarazioni risultano avere codifiche diverse, e pertanto una buona parte viene visualizzata in modo errato. Di seguito riportiamo degli stralci di alcune dichiarazioni, così come ricevute da *openpolis*, per chiarire la questione.

```
[1] "Un pacchetto di misure per combattere il terrorismo in tutte le sue forme, comprese quelle che si annidano sul web: è quanto promette entro l'autunno Franco Frattini, commissario europeo con delega per la Giustizia. [...] \"Tutto questo non ha niente a che vedere con la libertà di espressione\". Per questo, a suo dire, è necessario definire reato penale la creazione di risorse internet contenenti informazioni potenzialmente utilizzabili dai terroristi, poiché \"troppo spesso scopriamo siti che contengono tutte le istruzioni per la creazione di una bomba\"."

[2] "<p>&quot;Io non mi sieder&ograve; mai pi&ugrave; ad un tavolo in cui ci sia il signor Bossi. Non sosterr&ograve; mai pi&ugrave; un governo che conti su Bossi come sostegno. [...]

[3] "ROMA - \"Sono pronto a chiedere la chiusura del policlinico Umberto I di fronte all'impossibilitÃ  di ristrutturarlo\". Dopo l'inchiesta
```

```
dell'Espresso, il presidente della Regione Lazio, Piero Marrazzo, entra in
campo cosÃ¬ nella partita sul futuro dell'ospedale universitario piÃ'
grande d'Europa. \"L'Umberto I - sostiene il governatore - non Ã¨ proprio
un modello di architettura e organizzazione sanitarie moderne. Ãˆ stato
progettato alla fine dell'Ottocento e inaugurato nel 1904: pensato come
una cittadella della salute con tanti padiglioni, risponde a criteri
vecchi e va ristrutturato. [...]
```

La soluzione usata – laboriosa ma efficace – è una funzione che sostituisce ad ogni gruppo di caratteri mal codificato il carattere che invece dovrebbe essere visualizzato[15]. Come è prevedibile, è stato possibile individuare la "corrispondenza" tra caratteri mal codificati e carattere effettivo solo grazie a un controllo umano. Nella solita dichiarazione avremo:

PRIMA

```
doc$testo
 "<hr /><b>standard and poor's smentisce alemanno e berlusconi sul debito del comune di roma. [...] e' esattamente questo che non mi stanco di ripetere da settimane. [...]"
```

DOPO

```
doc$testo
 "<hr /><b>standard and poor's smentisce alemanno e berlusconi sul debito del comune di roma. [...] è esattamente questo che non mi stanco di ripetere da settimane. [...]"
```

Il passaggio seguente riguarda la correzione degli errori ortografici (perlomeno, di quelli più frequenti), l'eliminazione dei tag HTML[16] e di eventuali spazi bianchi "multipli". Per gli errori ortografici, la funzione utilizzata è del tutto analoga a quella usata per correggere gli errori di codifica. Per l'eliminazione di codice HTML e spazi bianchi in eccesso, invece, abbiamo usato due semplici funzioni disponibili nelle librerie di **R**.

[15] Nel caso di lettere accentate abbiamo usato soltanto accenti gravi, anche quando questo è ortograficamente errato (ad esempio, *perché* diventa *perchè*). Seppur errata, infatti, è la soluzione che più si avvicina alla scrittura comune.

[16] Per **tag HTML** intendiamo frammenti di codice HTML rimasti nella dichiarazione per errore. Sono elementi "di disturbo" nella nostra analisi perché non danno alcuna informazione sul contenuto del documento, essendo utili soltanto a definirne la formattazione.

PRIMA	DOPO
doc$testo "<hr /> <hr />standard and poor's smentisce alemanno e berlusconi sul debito del comune di roma. <hr />non 10 miliardi come ripetuto in un mantra dalle destre, ma 6,9 **miliari** di euro. [...] la litania sul megadebito ripetuta come un disco rotto dal neosindaco e rilanciato ieri da silvio **berlusocni** in un imbarazzante show da bruxelles [...] e dopo le dichiarazioni di s&p il primo a parlare è stato marco causi, deputato pd ed ex assessore al **bilncio** capitolino fin dal 2001 [...]"	doc$testo "standard and poor's smentisce alemanno e berlusconi sul debito del comune di roma. non 10 miliardi come ripetuto in un mantra dalle destre, ma 6,9 **miliardi** di euro. [...] la litania sul megadebito ripetuta come un disco rotto dal neosindaco e rilanciato ieri da silvio **berlusconi** in un imbarazzante show da bruxelles [...] e dopo le dichiarazioni di s&p il primo a parlare è stato marco causi, deputato pd ed ex assessore al **bilancio** capitolino fin dal 2001 [...]"

Infine, un altro aggiustamento importante ha previsto l'individuazione dei **poliformi**[17] più frequenti nel corpus, tramite la funzione `textcnt()` del package **tau** (che sta per *text analysis utilities*), che opera in modo da trovare le sequenze di parole utilizzate più volte in un testo (la soglia minima di frequenza può essere decisa dal ricercatore). L'obiettivo è la sostituzione degli spazi all'interno dei poliformi con il simbolo di sottolineatura "_" in modo che, durante il calcolo per la stima del modello CTM, essi vengano considerati come un'unica parola (per esempio, da *partito democratico* a *partito_democratico*). I segmenti trovati vengono salvati all'interno di un vettore, e quindi è facile sostituirli usando la solita funzione `gsub()`. Inoltre, abbiamo aggiunto uno spazio prima e dopo ogni simbolo di punteggiatura, per facilitare una fase successiva (l'individuazione delle parole del vocabolario). Nel nostro caso abbiamo deciso di non effettuare lo *stemming*[18], se però avessimo preso la decisione opposta avremmo probabilmente stemmato i termini più o meno a questo punto.

Il risultato finale di questa procedura sulla dichiarazione numero 357126 è il seguente.

[17] Vedi capitolo 1.

[18] Vedi paragrafo 1.1.

standard and poor ' s smentisce alemanno e berlusconi sul debito del comune di roma . non 10 miliardi come ripetuto in un mantra dalle destre , ma 6 , 9 miliardi di euro . e non declassa il campidoglio . lo spiega in un intervista a la stampa myriam fernandez de heredia , responsabile per standard and poor ' s dei giudizi sul merito di credito del settore pubblico in europa . la litania sul megadebito ripetuta come un disco rotto dal neosindaco e rilanciato ieri da silvio berlusconi in un imbarazzante show da bruxelles , dove ha accusato l ' ex sindaco di roma e segretario del partito_democratico : non c ' è nessuna città d ' europa che ha lasciato un deficit di 16 mila miliardi di vecchie lire . e s & p riaccende i riflettori sul debito di milano , governata dalla lega e dagli uomini di berlusconi fin dal 1993 . alle accuse del premier ha replicato anche l ' unità smontando in un pezzo molto dettagliato il bluff della destra . e dopo le dichiarazioni di s & p il primo a parlare è stato marco causi , deputato partito_democratico ed ex assessore al bilancio capitolino fin dal 2001 : " l ' intervista della dottoressa fernandez , responsabile di standard and poors per il settore pubblico in europa , pubblicata oggi sulla stampa fa giustizia della grande mistificazione costruita negli ultimi giorni intorno ai conti del comune di roma . standard and poors conferma infatti in 6 , 9 miliardi il livello del debito comunale e afferma che la relazione della ragioneria generale dello stato chiesta da alemanno , non dà nessuna informazione nuova , sono dati che già avevamo . è esattamente questo che non mi stanco di ripetere da settimane . non c ' è nessun dato che la giunta veltroni abbia occultato e il livello del debito del comune è cresciuto durante la gestione veltroni meno di quello nazionale e meno del tasso d ' inflazione . non ha alcun senso utilizzare termini come buco o bancarotta . chi lo fa è in malafede e dimostra così di voler soltanto montare una palese strumentalizzazione politica " . e anche milano per s & p finisce sotto accusa . se n ' è accorto causi : " standard and poors (cita il pezzo ndr) sottolinea che il comune di milano ha un debito elevato . lo stiamo monitorando , li stiamo monitorando entrambi roma e milano e ricorda così alla politica italiana , e soprattutto al presidente del consiglio , che invece di montare inaccettabili montature politiche , sarebbe meglio riflettere su come aiutare non solo roma ma tutte le grandi città italiane che , dopo anni di restrizione e di tagli , rischiano di non avere più fiato sufficiente a realizzare le importanti infrastrutture , soprattutto di trasporto , di cui hanno bisogno " . anche per paolo gentiloni , responsabile della comunicazione del partito_democratico , il giudizio di s & p interviene a " cancellare le accuse rumorose e gonfiate contro veltroni e gli anni della sua amministrazione " . l ' agenzia di rating , per gentiloni , dice che " i conti romani sono esattamente quelli conosciuti , non c ' è nessun buco e tantomeno dei debiti occultati " . i buchi veri - prosegue gentiloni - ce li hanno lasciati gli amici di alemanno , con i guasti nella sanità della regione . ora questa vicenda , fatta solo di accuse propagandistiche , dovrà finire . alemanno , dopo aver cercato scuse per non farlo , sarà costretto a governare e dovrà dimostrare di saper mantenere le sue promesse . e berlusconi deve smettere di cercare di delegittimare il leader del maggior partito dell ' opposizione " . mentre silvio di francia si chiede : ora tappezzeranno roma di manifesti per dire scusate , non sappiamo leggere i bilanci ? . dispiace che tra chi ha effettuato le ripetute ricognizioni sul debito annunciate dal sindaco alemanno nessuno abbia avuto in mente di interpellare standard & poor ' s l ' agenzia che monitora il debito di diverse città in tutto il mondo . senza voler ascoltare le nostre spiegazioni potevano interpellare chi per mestiere dà giudizi sul credito del settore pubblico per sentirsi dire che non c ' è nessun buco nascosto.

Concluse tutte queste fasi, effettuiamo ancora una volta un controllo sui testi per eliminare eventuali valori nulli o mancanti ("NA") e doppioni di dichiarazioni. Il nostro corpus, adesso, contiene 12.575 testi.

3.3 APPLICAZIONE DEL MODELLO

Una volta effettuati gli aggiustamenti preliminari e aver ottenuto il corpus definitivo, è opportuno effettuare alcune valutazioni sulla sua adeguatezza dal punto di vista numerico.

In linea generale,

ai fini dell'analisi statistica, un corpus è considerato di piccole dimensioni se ha una lunghezza inferiore alle 15mila forme grafiche (100 Kbytes), di medie dimensioni se varia tra le 15mila e le 50mila (300 Kbytes), medio-grande se varia tra le 50mila e le 100mila (700 Kbytes) e grande oltre le 100mila. (Tuzzi, 2003)

Il nostro corpus ha un totale di quasi 4 milioni di forme grafiche, perciò lo possiamo definire molto grande. Inoltre,

dal momento che nei casi applicativi non è possibile estendere a piacere il campione a disposizione, è necessario poter valutare se il corpus ha le caratteristiche per permettere uno studio su base statistica dei contenuti. Al fine di valutare la ricchezza lessicale si confronta il rapporto tra ampiezza del vocabolario ed estensione del corpus con un valore empirico (Bolasco, 1999). Se il rapporto tra il numero di parole diverse e il numero di parole totali supera il 20% il corpus è da considerarsi non sufficientemente esteso per un approccio su base statistica, in quanto il vocabolario è troppo vasto. (Tuzzi, 2003)

Nel nostro caso tale rapporto vale $80475/3970979 \simeq 0.02$, valore molto inferiore a quello suggerito come limite. Ricordiamo anche che

è stato valutato empiricamente che, se il rapporto tra numero di *hapax*[19] e numero di parole diverse[20] supera il 50% e, quindi, il vocabolario è costituito per oltre la metà da *hapax*, il corpus non è trattabile statisticamente in quanto costituito da troppe parole originali. (Tuzzi, 2003)

È facile calcolare il numero di *hapax*, che risulta essere 31.719, e il rapporto tra questo numero e il numero di parole distinte è $31719/80475 \simeq 0.4$.

A questo punto, per applicare il CTM possiamo usare alcune funzioni contenute all'interno dei pacchetti **tm** (a significare *text mining*) e **topicmodels** (che contiene funzioni sia per il CTM, sia per il modello LDA).

Innanzitutto si definisce un oggetto di tipo *Corpus* (necessario per usare le funzioni che ci interessano), quindi si procede alla creazione di una matrice chiamata **Document-Term Matrix** (matrice "*Documenti* × *Parole*"). Essa ha sulle colonne tutte le parole del vocabolario del corpus, e sulle righe il numero di riferimento dei documenti. La generica cella (j, k) contiene la frequenza con cui la parola k compare nel documento j; perciò se il valore della cella è 0 significa che la parola non compare mai nel documento, se è 1 vuol dire che compare una sola volta, e così via.

Infine, per ottenere il modello si usa il comando `CTM()`, che richiede in ingresso una matrice di tipo *Document-Term* e il numero di *topics* del modello (nella funzione questo parametro è chiamato k). Altri parametri sono già preimpostati e possono essere modificati dall'utente. Il risultato è un oggetto di tipo *CTM* che contiene, tra l'altro:

- $\hat{\mu}$ e $\hat{\Sigma}$ stime dei parametri della normale logistica la cui realizzazione dà θ_d (vettore di frequenza dei *topics* nel d-esimo documento);
- il vocabolario;
- $\log(\hat{\beta}_i)\ \forall\, i$, ricordiamo che β_i è il vettore contenente la distribuzione dell'*argomento* i-esimo sul vocabolario;

[19] Si definisce **hapax** una parola che compare una sola volta nel corpus.

[20] Con **parole diverse** si intendono tutti i termini del vocabolario. Le parole del vocabolario, naturalmente, sono le stesse che compaiono nel corpus, ma mentre nel vocabolario ogni parola compare una sola volta, nel corpus una parola può comparire molte volte.

- $\hat{z}_{d,n}\ \forall\, d, n$, che rappresenta l'*argomento* da cui più probabilmente proviene la parola n del documento d.

Un altro comando utile è `posterior()`, che ha bisogno di un oggetto di tipo *TopicModel* (*CTM* o *LDA*) in ingresso. La sua funzione è di fornire la distribuzione a posteriori sia degli *argomenti* per dichiarazione, che del vocabolario per *argomento*.

Inoltre, abbiamo creato la funzione `mostraparolefreq()` che stampa a video le parole più frequenti di un *topic*, in particolare quelle la cui frequenza cumulata è maggiore o uguale a un valore stabilito dall'utente[21]. Pertanto, `mostraparolefreq()` necessita del numero indicatore del *topic* che si vuole esaminare, e della soglia minima che la frequenza cumulata deve raggiungere (il valore di default è 0.30).

3.4 IL RISULTATO

Per il corpus di *openpolis* abbiamo usato un numero di *topics* k = 20 e una matrice *Document-Term* "ridotta", ovvero ottenuta eliminando dalla matrice originale le parole con frequenza molto alta o molto bassa[22]. Così facendo le dimensioni della matrice si sono notevolmente ridotte: da 12.575 × 80.641 si passa a 12.529 × 2.577 [23].
Inoltre, per assicurare la riproducibilità dei risultati abbiamo utilizzato anche i parametri `seed` e `initialize = "seeded"`.

[21] Non si confonda `mostraparolefreq()` con `get_terms()`, già presente nella libreria: la prima estrae tante parole quante ne servono per raggiungere la soglia stabilita di frequenza cumulata; la seconda estrae le prime *h* parole più frequenti per il *topic*, con *h* fissato e stabilito dall'utente.

[22] Uno strumento di ausilio per la scelta dei termini da eliminare è la **matrice Term Frequency – Inverse Document Frequency**. La sua struttura è identica a quella di una matrice *Document-Term*, ma le celle contengono un valore – nel nostro caso – tra 0 e 14 circa. Più il valore è piccolo, meno la corrispondente parola è "rilevante" per il nostro studio. Infatti la matrice TF–IDF viene creata appositamente per dare poco peso sia alle parole eccessivamente frequenti, che a quelle eccessivamente rare. Nel nostro caso abbiamo scartato tutte le parole il cui peso nella TF–IDF era minore a 0.05, cioè circa la mediana della distribuzione della matrice stessa. In seguito, per ridurre ulteriormente il costo computazionale, sono stati eliminati anche i termini con frequenza totale nel corpus minore di 50.

[23] Si noti che, oltre al numero di colonne (parole), risulta diminuito anche il numero di righe (dichiarazioni). Il motivo è che sono state eliminate le righe la cui somma era diventata uguale a 0 in seguito all'esclusione di alcuni vocaboli – concettualmente, una riga con somma zero indica un documento che non ha nessuna parola.

Di seguito riportiamo l'output dei termini più frequenti di alcuni dei *topics* stimati[24], ottenuto facilmente tramite la funzione `mostraparolefreq()`.

TOPIC 5 – *Frequenza cumulata: 0.2885*

donne	rifiuti	violenza	donna	polizia	provinciale	femminile	forzedell ordine	quote	raccoltadif ferenziata
0.099	0.039	0.026	0.025	0.022	0.018	0.018	0.016	0.013	0.012

TOPIC 6 – *Frequenza cumulata: 0.3028*

napoli	rifiuti	rete	campania	internet	bassolino	governatore
0.076	0.064	0.058	0.050	0.026	0.016	0.013

TOPIC 8 – *Frequenza cumulata: 0.3036*

scuola	scuole	università	studenti	terni	alitalia	istruzione	scolastico	gelmini
0.086	0.037	0.032	0.031	0.030	0.029	0.023	0.019	0.018

TOPIC 10 – *Frequenza cumulata: 0.2635*

pace	alemanno	obama	esteri	israele	militari	pace	cina	militare	milano	frattini
0.029	0.022	0.021	0.018	0.016	0.016	0.029	0.014	0.013	0.012	0.012
dirittiumani	**onu**	**afghanistan**	**expo**	**hamas**	**visita**	**larussa**	**fascismo**	**bush**	**gaza**	
0.012	0.011	0.011	0.010	0.009	0.008	0.007	0.007	0.007	0.007	

TOPIC 18 – *Frequenza cumulata: 0.2745*

sicilia	immigrati	mafia	criminalità	mezzo giorno	lombardo	palermo	magistrati	procura	indagini
0.029	0.026	0.025	0.023	0.023	0.020	0.013	0.013	0.012	0.012
isola	**anti mafia**	**prostitu zione**	**inchiesta**	**marrazzo**	**castel nuovo**	**camorra**	**clande stini**	**procuratore**	**cuffaro**
0.010	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.007	0.007

TOPIC 20 – *Frequenza cumulata: 0.1108*

presentazione	provinciale	scrive	ribadito	festa	dichiara	insediamento	notizie	dover
0.009	0.008	0.007	0.007	0.007	0.007	0.006	0.006	0.006
promesse	**andrea**	**polizia**	**stefano**	**commenta**	**condivisione**	**gesto**	**manifestazione**	**giovedì**
0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005

[24] L'elenco completo di tutti i *topics* è visibile nell'Appendice, sezione A1.

Si noti che alcuni *topics* trattano di una tematica ben precisa (per esempio, i numeri 5, 6 e 8) mentre altri sono meno definiti (ad esempio, il 20esimo).

Per capire al meglio i risultati ottenuti, bisogna considerare anche le parole scartate inizialmente. Ricordiamo infatti che abbiamo ignorato sia le parole più frequenti del corpus, sia quelle meno frequenti; e mentre sembra ragionevole considerare queste ultime non significative ai fini del nostro studio proprio perché compaiono raramente, lo stesso discorso non vale per i termini più frequenti. Se delle parole vengono utilizzate molto di frequente, immaginiamo due possibilità: possiamo pensare che siano parole grammaticali (indispensabili per la comprensibilità del testo ma **vuote**[25] agli scopi della nostra ricerca) oppure che ci sia un "abuso" di parole **piene** provenienti da *argomenti* ricorrenti nella stragrande maggioranza delle dichiarazioni.
Il package **tm** di **R** implementa già una funzione `stopwords()` che, per alcune lingue, fornisce un elenco di parole vuote. Tuttavia, questo elenco sicuramente non è (e difficilmente potrà mai essere[26]) completo. Infatti, se andiamo a vedere quali sono le parole scartate dalla matrice TF–IDF[27], notiamo che sono sia parole piene che parole vuote, nonostante nel comando `DocumentTermMatrix()` avessimo specificato di escludere le *stopwords*.

Come accennato, individuando le parole piene tra quelle scartate si può avere un'idea delle tematiche globali del corpus. Nel nostro caso, i termini più rilevanti sembrano essere quelli riportati di seguito.

[25] Per la definizione di parole **vuote** e **piene** si rimanda al capitolo 1.

[26] È molto difficile pensare di poter realizzare un elenco comprendente tutte le parole vuote di una lingua. Si consideri che la stessa definizione di "parola vuota" varia a seconda del testo da esaminare.

[27] L'elenco completo è visibile nell'Appendice, sezione A2.

ambiente	amministrazione	berlusconi	bilancio	camera	centrodestra
centrosinistra	costituzione	crisi	democrazia	destra	diritti
diritto	economia	elezioni	europa	famiglie	futuro
giovani	giustizia	governo	impegno	imprese	interventi
intervento	istituzioni	italia	lega	legge	libertà
maggioranza	ministro	opposizione	paese	parlamento	partiti
partito	partitodemocratico	politica	politiche	popolodellalibertà	
premier	presidente	problema	problemi	prodi	pubblica
pubblici	pubblico	questione	regioni	repubblica	responsabilità
riforma	rischio	risorse	senato	sicurezza	sinistra
sociale	sociali	società	sviluppo	vita	voto

L'elenco ci suggerisce che tutti (o quasi) i testi del corpus parlino del governo, delle più importanti istituzioni e figure politiche del paese, dei principali partiti; di interventi e riforme su sicurezza, giustizia, ambiente; sviluppo della società; questioni pubbliche; crisi, economia, uso delle risorse; e ancora di futuro, giovani e famiglie; dell'Italia e dell'Europa, e altro ancora.

Se avessimo deciso di non eliminare queste parole dalla matrice *Document-Term* avremmo ottenuto un risultato poco significativo, caratterizzato da *topics* molto simili tra loro, con all'incirca le stesse parole.

Altri due elementi interessanti sono $\hat{\mu}$ e $\hat{\Sigma}$. Effettuando la trasformazione logistica su $\hat{\mu}$ si ottiene $\hat{\theta}$, stima del vettore θ (probabilità che un generico documento tratti dei vari *topics*).

$$\hat{\mu} = [-0.596 \ -0.356 \ -0.320 \ -0.038 \ -0.158 \ -0.088 \ -0.378 \ -0.969 \ -0.125$$
$$-0.587 \ -0.116 \ 0.274 \ -0.421 \ -0.414 \ -0.561 \ -0.244 \ -0.124 \ -0.008 \ -0.091]$$

$$\hat{\theta}^* = \frac{e^{\hat{\mu}}}{1+\sum_{i=1}^{19} e^{\hat{\mu}_i}} \ ; \ \hat{\theta} = [\hat{\theta}^*, \ 1 - \sum_{i=1}^{19} \hat{\theta}_i^*]$$

$$\hat{\theta} = [\underset{1}{0.035}\ \underset{2}{0.044}\ \underset{3}{0.046}\ \underset{4}{0.061}\ \underset{5}{0.054}\ \underset{6}{0.058}\ \underset{7}{0.043}\ \underset{8}{0.024}\ \underset{9}{0.056}\ \underset{10}{0.035}\ \underset{11}{0.056}\ \underset{12}{0.083}\ \underset{13}{0.041}\ \underset{14}{0.042}\ \underset{15}{0.036}\ \underset{16}{0.049}\ \underset{17}{0.056}\ \underset{18}{0.062}\ \underset{19}{0.057}\ \underset{20}{0.063}]$$

Da $\hat{\theta}$ si conclude che gli *argomenti* che hanno maggiore probabilità di essere presenti in un testo sono il numero 12 e il numero 20, mentre i meno probabili sono il primo, l'ottavo e il decimo.

Allo stesso modo, si può effettuare una trasformazione su $\hat{\Sigma}$ affinché sia analoga a una matrice di correlazione; così diventano visibili le correlazioni tra *argomenti* (figura 6).

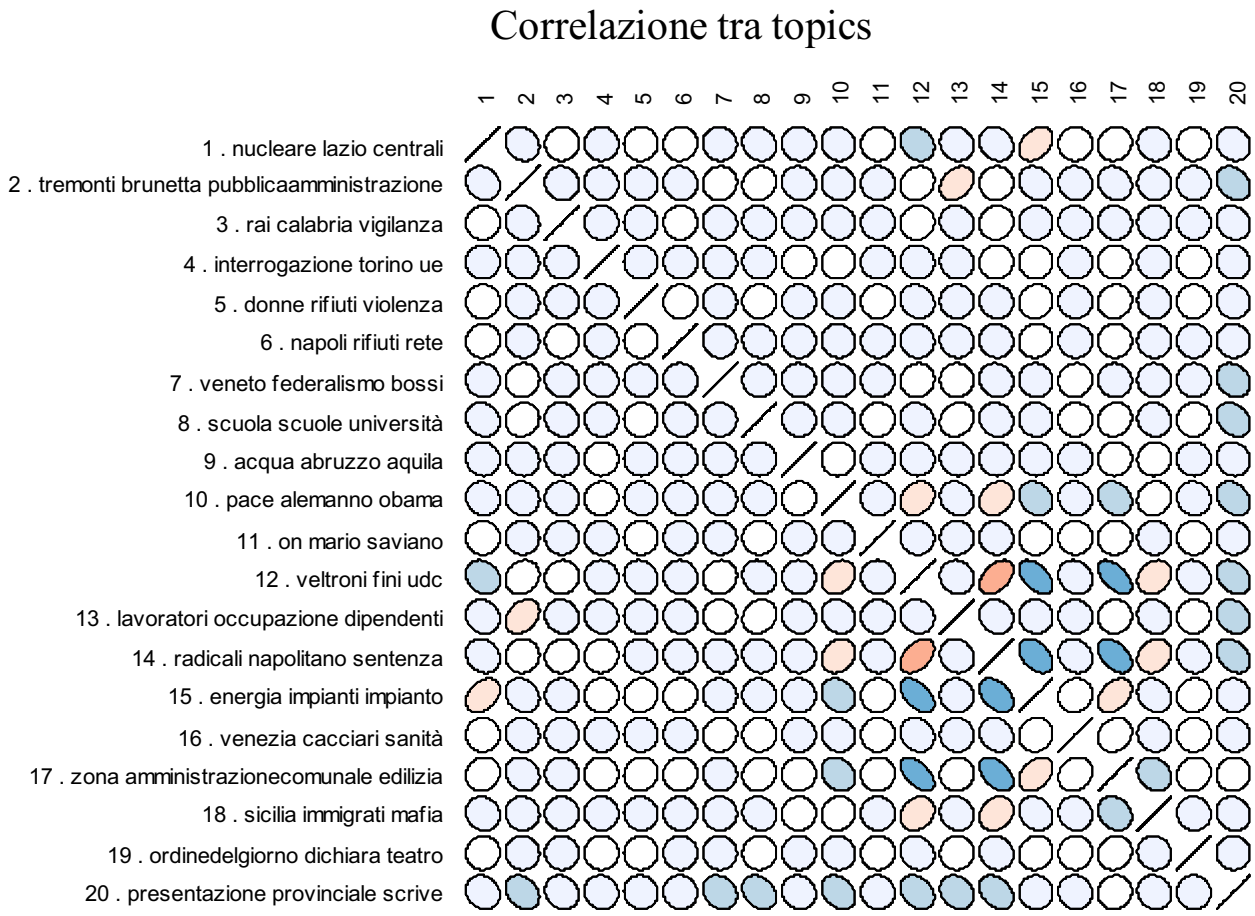


Figura 6: correlazione tra *argomenti*.

La maggior parte dei *topics* non sembrano particolarmente correlati, alcuni invece mostrano correlazione positiva (per esempio, i numeri 12 e 14, 1 e 15, 2 e 13) o negativa (come 14 e 15, 14 e 17).

Un altro elemento particolarmente interessante è la matrice 12.529 × 20 contenuta in `gamma`, che contiene la stima della distribuzione a posteriori sui *topics* per ogni dichiarazione. Essa diventa il punto di partenza per numerose valutazioni, ad esempio la si può usare per confrontare gli argomenti trattati dai diversi politici, eventualmente anche raggruppandoli per posizione politica; oppure per confrontare le tematiche affrontate dallo stesso politico nel tempo; o ancora per avere un'idea della variabilità delle tematiche discusse nel tempo.

Tuttavia, bisogna fare attenzione quando si effettuano queste analisi sul corpus di *openpolis*, per più ragioni. In primis, si tenga sempre presente che le dichiarazioni sono inserite nel sito dagli utenti, pertanto il corpus non è esaustivo, né tantomeno un campione casuale. Poi, se vogliamo studiare la variabilità dei *topics* nel tempo (sia complessivamente, sia per un dato politico) non dimentichiamo che – pur contando numerosissime dichiarazioni – *openpolis* è un corpus "giovane": i testi precedenti al 2008 sono solo il 4,5% del totale. Inoltre, quando ci concentriamo su un certo personaggio politico dovremmo accertarci che il numero di sue dichiarazioni presenti nel corpus sia sufficientemente grande per un approccio statistico, e che le dichiarazioni siano sufficientemente distribuite nel tempo[28]. Infine, nel caso si volesse fare un confronto sui politici divisi per posizione politica, si presti attenzione ai cambi di schieramento, anche passati.

Tenuto conto di queste considerazioni, riteniamo interessante riportare l'esito del confronto tra gli *argomenti* più probabili per i politici con almeno 100 dichiarazioni (figura 7), e del confronto tra gli *argomenti* più probabili per i mesi del 2010 (figura 8).

[28] Se pure un personaggio contasse 100 dichiarazioni, ma fossero tutte concentrate nell'arco di soli due mesi, sarebbe probabile che gli *argomenti* trattati nella maggior parte dei testi fossero gli stessi.

Probabilità dei *topics* per politico

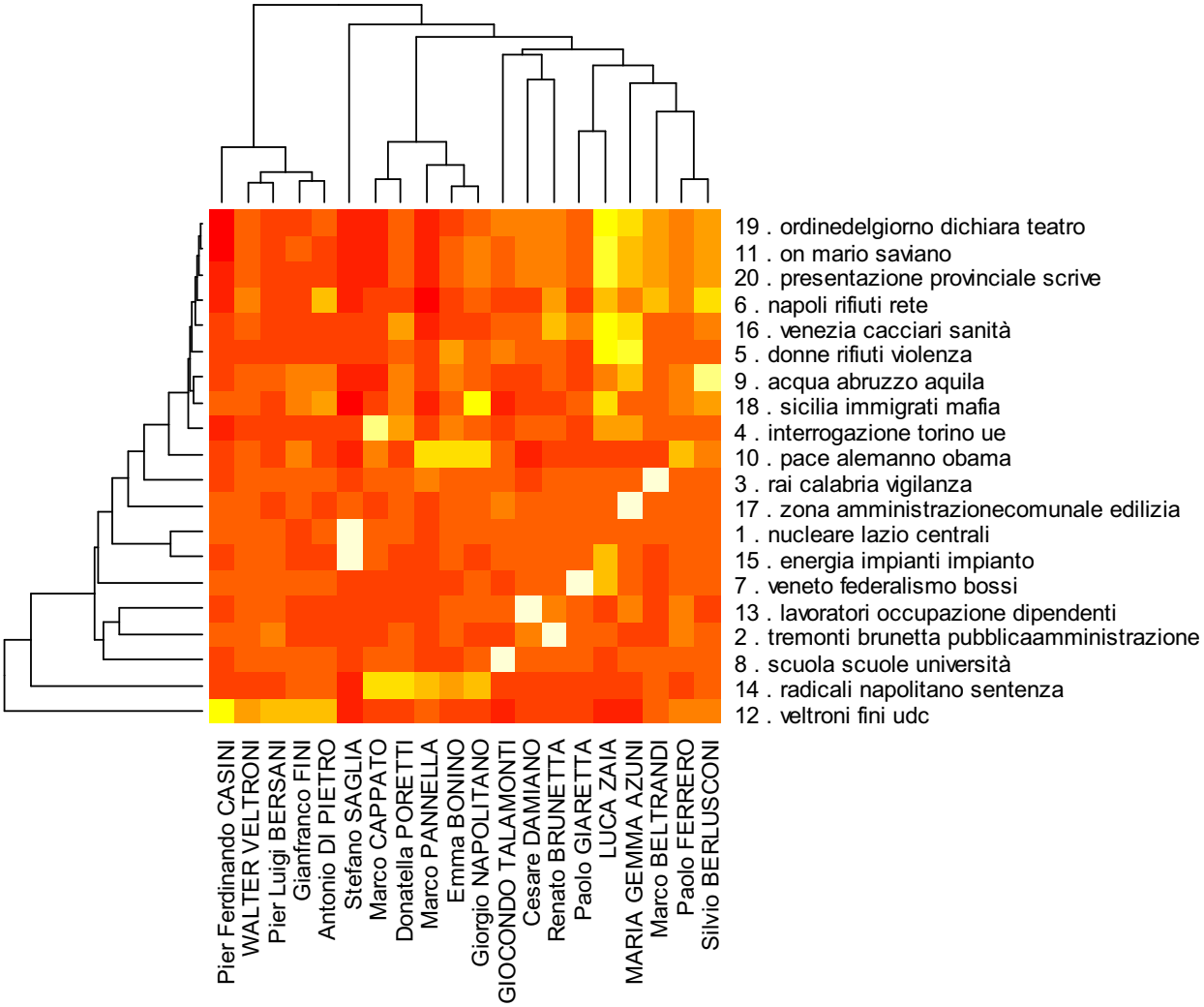


Figura 7: probabilità di trattare ogni *argomento* per i venti politici più presenti nel corpus. La casella più chiara indica una probabilità maggiore.

La figura 7 ci lascia tutto sommato soddisfatti. Per esempio, risulta che Giocondo Talamonti (consigliere comunale nel Comune di Terni e dirigente scolastico di un istituto superiore) sia molto interessato al *topic* su scuola e università. Così, Paolo Giaretta (senatore e membro del Partito Democratico Veneto) appare molto coinvolto quando si parla del Veneto e delle questioni attinenti, mentre sembra interessarsi meno all'*argomento* su Napoli e sul problema dei rifiuti. E ancora, il ministro Brunetta si mostra molto preoccupato per le questioni relative alla… pubblica amministrazione!, mentre il premier Silvio Berlusconi appare mediamente interessato a tutti gli *argomenti*, con una preoccupazione maggiore per il problema dei rifiuti a Napoli e della ricostruzione in Abruzzo dopo il terremoto.

Lo studio del dendogramma superiore ci suggerisce anche alcune "affinità" tra i politici. Per esempio, sembra che i *topics* trattati da Casini, Veltroni, Bersani, Fini e Di Pietro siano molto simili; così come quelli sviluppati da Cappato, Poretti, Pannella, Bonino e Napolitano. Se confrontiamo i raggruppamenti con l'appartenenza politica, ci accorgiamo che il secondo gruppo è formato da personaggi affini alla sinistra radicale (ricordiamo l'appartenenza del presidente Napolitano al Partito Comunista); mentre il primo ci sorprende un po': infatti apparentemente comprende persone sia di centrodestra, sia di centro, che di centrosinistra. La spiegazione di questo raggruppamento, però, è semplice, ed è che le posizioni "moderate" hanno a cuore le stesse tematiche, seppur con opinioni probabilmente diverse.
Gli altri politici, infine, non mostrano particolari raggruppamenti – evidentemente nel periodo considerato si sono dedicati a temi differenti.

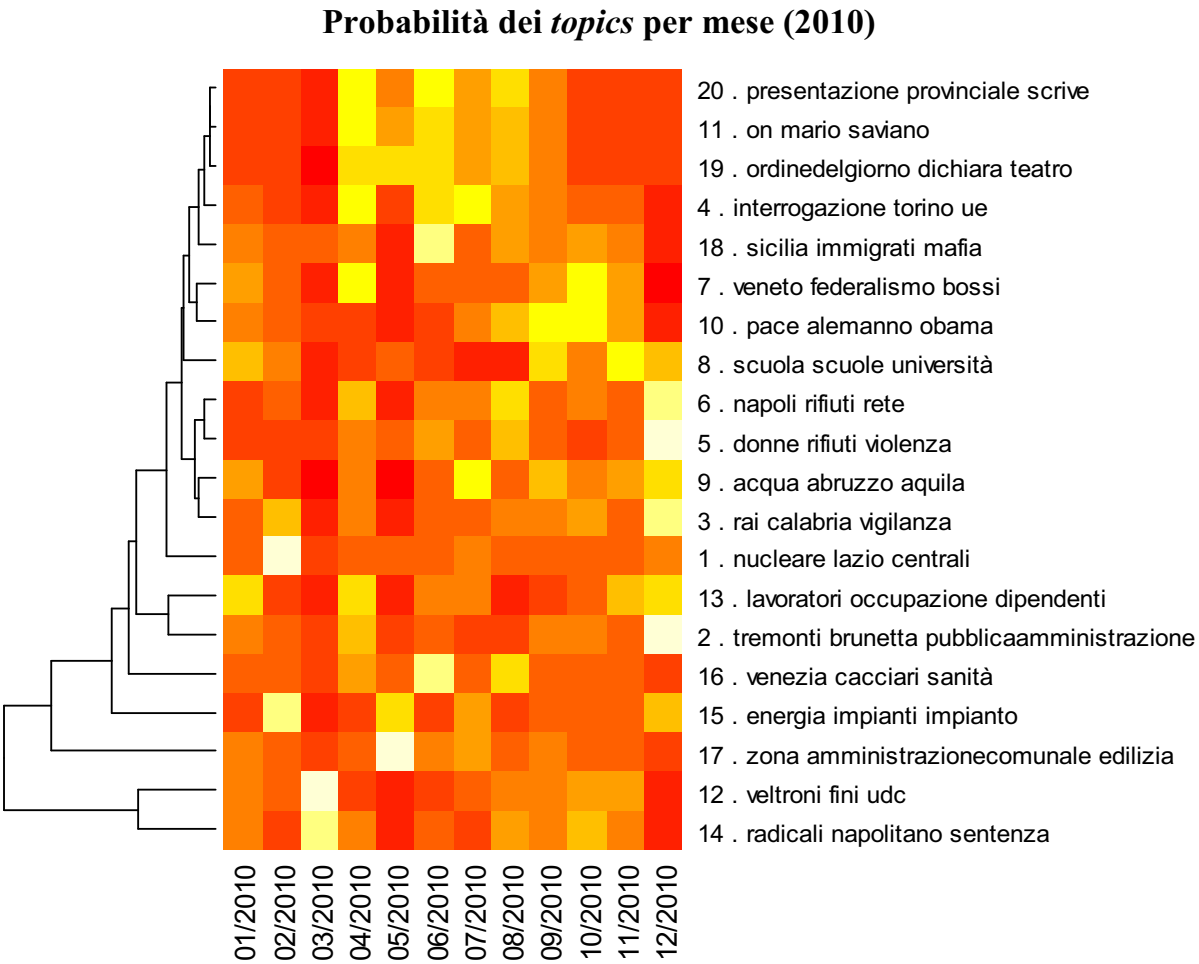


Figura 8: probabilità di trattare un certo *argomento* nel corso del 2010. La casella più chiara indica una probabilità maggiore.

Per valutare la figura 8 possiamo partire dal mese di dicembre 2010. Ricordiamo, infatti, che nel corpus sono presenti soltanto due dichiarazioni per tale mese. Se andiamo a leggerle ci accorgiamo che la prima parla di tasse e problemi finanziari, mentre la seconda parla di rifiuti e di produzione di energia.

> come nella maggior parte dei comuni della provincia di reggio , anche a montecchio l ' amministrazione ha deciso di aumentare l ' aliquota dell ' addizionale irpef fino allo 0 , 6 % a partire dal 2012 . la scelta portata avanti dalle amministrazioni reggiane di centrosinistra , tra cui montecchio , è iniqua e scellerata , ed antisociale : la settimana scorsa , i tre sindacati confederati cgil , cisl e uil di reggio hanno condannato senza se e senza ma la scelta portata avanti dalle amministrazioni reggiane , evidenziando come ciò sia accaduto solo in provincia di reggio e da nessun ' altra parte in tutta la regione . in consiglio comunale a montecchio , il sindaco e la giunta hanno giustificato la propria scelta , colpevolizzando , com ' era prevedibile , i tagli del governo , la cd . tassa tremonti . [...] l ' amministrazione ha fatto pertanto una chiara scelta , ovvero quella di non tutelare le persone , le famiglie in difficoltà economica , i pensionati , i giovani , le ragazze madri , di non tutelare le fasce di popolazione di cui la sinistra in italia dice di essere portatore d ' interessi . forse ciò non avviene a montecchio ! !
>
> i rifiuti ? adesso è venuta l ' ora di fare delle scelte di campo serie e definitive . [...] non tollero " afferma il consigliere di fiumicino - che i rifiuti diventino un business per pochi e un problema per i cittadini . per questo chiedo al sindaco di fiumicino e a quelli dei comuni limitrofi , di prendere finalmente in esame la tecnologia arrow bio , utilizzata con ottimi risultati in israele e in australia . una tecnologia che permette di abbattere i costi di raccolta e smaltimento rifiuti , con standard elevatissimi e livelli di differenziata del 75 per_cento , impensabili in italia . [...] e in più produce biogas , una fonte di energia alternativa pulita che viene poi utilizzata per il trasporto o per la produzione_di_energia_elettrica o termica , fertilizzante e acqua . la tecnologia di arrow bio è l ' acqua . acqua " prosegue satta - utilizzata per il trattamento dei rifiuti . l ' acqua per la separazione più efficace dei rifiuti . [...]

In effetti, dal grafico pare che gli *argomenti* più rilevanti nel mese di dicembre 2010 abbiano riguardato tematiche finanziarie (*topic* 2, correlato positivamente al 13) e relative ai rifiuti (*topic* 5, correlato al 15 e al 6). Nel secondo documento si parla molto anche di acqua, ecco perché il *topic* 9 sembra rilevante.

Concentrandoci sull'*argomento* che parla di immigrazione (*topic* 18), vediamo che se ne è parlato molto nel mese di giugno; questo è piuttosto sensato perché proprio in quel mese è stato pubblicato il decreto che introduceva un test di lingua italiana a quanti richiedevano il permesso di soggiorno.

Allo stesso modo, il *topic* 1, che riguarda le centrali nucleari e la regione Lazio, appare un argomento "scottante" nel mese di febbraio 2010: proprio quando si è acceso il dibattito per la volontà di alcuni politici di costruire delle centrali nucleari nella regione[29]. Così, anche il *topic* 15 che tratta di produzione di energia (positivamente correlato al numero 1), risulta molto discusso in quel periodo.

Nelle figure 9 e 10 sono visibili i grafici analoghi per gli anni 2008 e 2009.

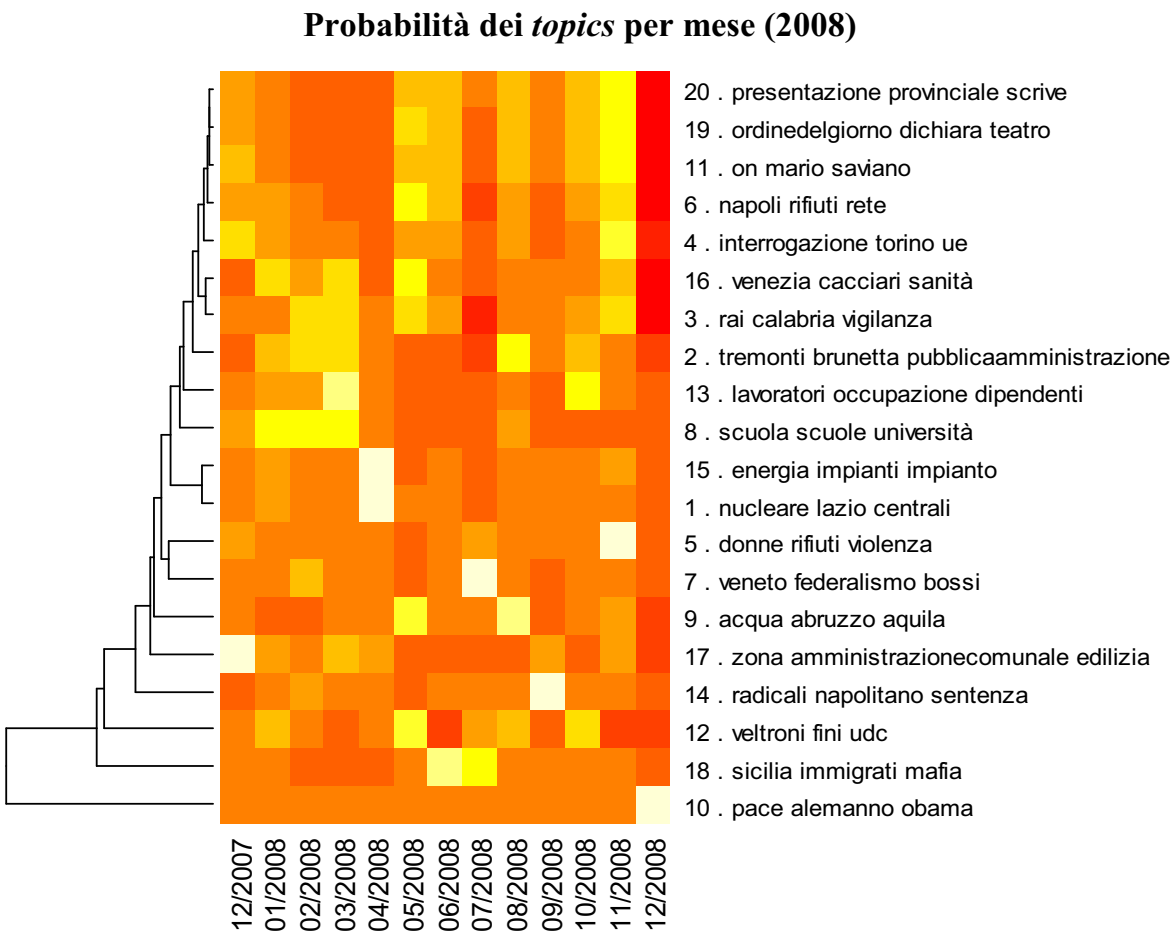


Figura 9: probabilità di trattare un certo *argomento* nel corso del 2008.

[29] Risale proprio a quel periodo il decreto legislativo n. 31 del 15/02/2010: «*Disciplina della localizzazione, della realizzazione e dell'esercizio nel territorio nazionale di impianti di produzione di energia elettrica nucleare, di impianti di fabbricazione del combustibile nucleare, dei sistemi di stoccaggio del combustibile irraggiato e dei rifiuti radioattivi, nonché misure compensative e campagne informative al pubblico, a norma dell'articolo 25 della legge 23 luglio 2009, n. 99*».

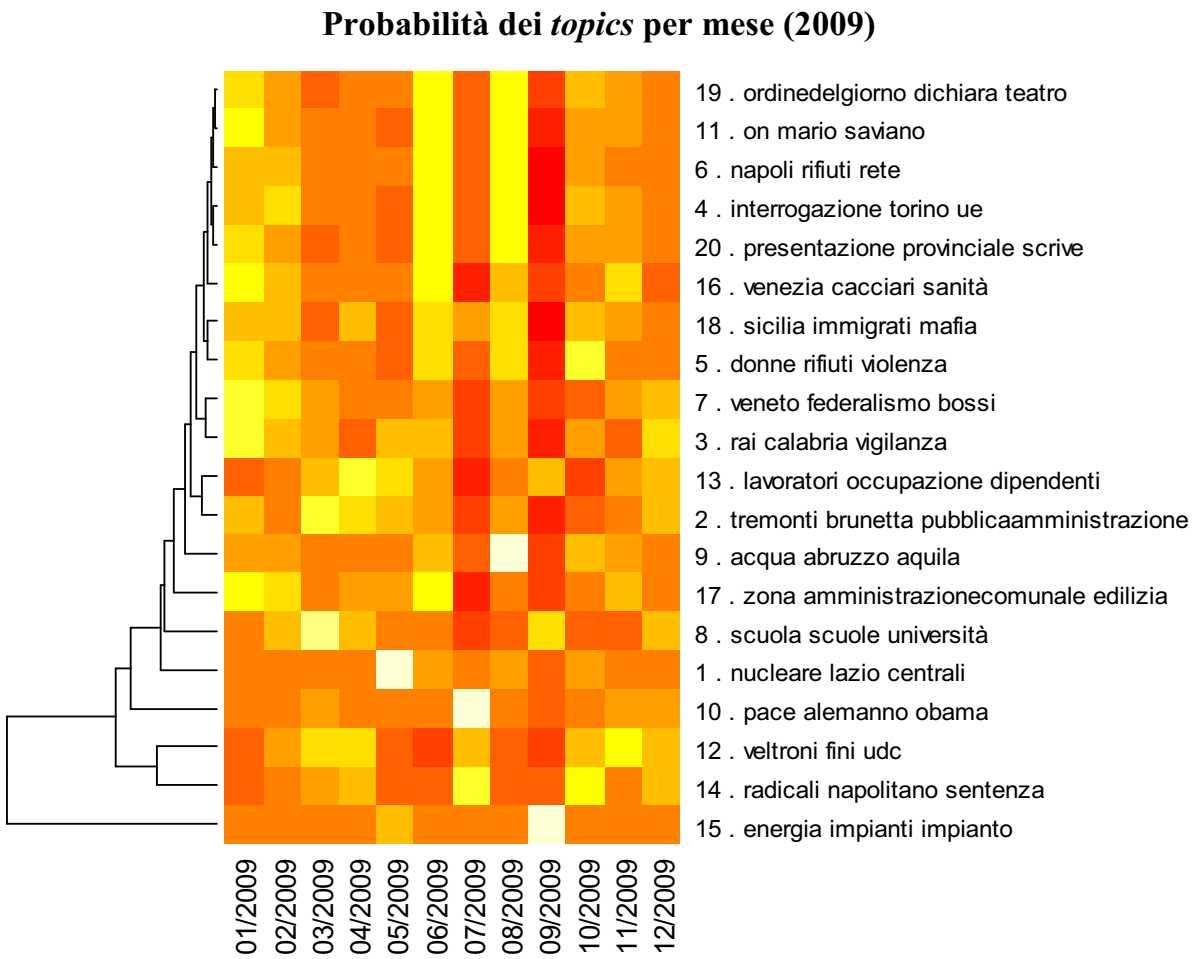


Figura 10: probabilità di trattare un certo *argomento* nel corso del 2009.

4. CONCLUSIONI

L'obiettivo del nostro studio era valutare le prestazioni del **correlated topic model**. Sulla base dei risultati ottenuti, esso risulta un valido strumento per l'analisi statistica di dati testuali. Nonostante alcuni aspetti richiedano ulteriori ricerche e approfondimenti, riteniamo che il modello sia utile perché informativo: per il corpus di *openpolis* ha reso possibile una classificazione automatica delle dichiarazioni dei politici in macro-argomenti. Un altro punto a favore è la capacità di ottenere risultati apprezzabili anche senza un'onerosa, e spesso poco automatizzata, pulitura del testo.
Inoltre – sebbene sia fondamentale trattare i risultati con le dovute cautele – nel nostro caso ha consentito di studiare come variano le tematiche trattate nel tempo, o a seconda del personaggio politico. Senza usare il *correlated topic model*, risultati di questo tipo si sarebbero potuti ottenere soltanto dopo molte ore di lettura e studio del corpus.

Di grande interesse pratico è la possibilità di disporre della correlazione tra *argomenti*. Prendiamo il caso di *openpolis*: ad ogni dichiarazione inserita, gli utenti assegnano dei *tags*[30], cosicché quando un utente consulta una dichiarazione gli vengono suggeriti altri articoli sullo stesso argomento.
Usando il *correlated topic model* non solo l'assegnazione dei *tags* può essere fatta in modo più automatico[31], ma il suggerimento agli utenti di dichiarazioni "simili" può riguardare sia gli *argomenti* trattati nell'articolo che l'utente sta leggendo, sia gli *argomenti* a essi correlati. Insomma, è un sistema per cogliere meglio il legame tra le informazioni.

Tra gli svantaggi del modello, invece, uno dei principali è la gravosità dal punto di vista computazionale – problema che si potrebbe forse aggirare individuando nuovi metodi di stima. L'elevato costo computazionale rende difficile applicare il modello a corpus molto grandi, cioè proprio nei casi in cui sarebbe più informativo. Altro problema è la

[30] I **tags** sono "etichette" che vengono assegnate manualmente (ovvero non in modo automatico ma in base a una valutazione umana) ad un articolo per indicare i temi trattati, e facilitare l'archiviazione, la catalogazione e il recupero delle informazioni.

[31] La possibilità di effettuare un controllo umano non dovrebbe comunque essere tolta: in questo modo si potrebbe, per esempio, "correggere" i *topics* nel caso presentassero parole vuote o altri elementi non significativi.

difficoltà nella scelta del numero di *topics*, *k*. La questione è, in realtà, piuttosto rilevante, perché la bontà dei risultati dipende molto dall'individuazione di un numero adeguato di *argomenti*. Sono già stati effettuati alcuni studi su questa problematica[32], ma a quanto ci risulta richiedono tutti di confrontare più modelli (dello stesso corpus) con *k* diversi, e se si considerano le difficoltà di calcolo già accennate è facile rendersi conto che non sempre è possibile un'elaborazione del genere.

Ulteriori svantaggi riguardano i limiti imposti dalla struttura stessa del *correlated topic model.* I più pesanti sono la sua incapacità di prevedere *argomenti* "annidati" (per esempio nel corpus di *openpolis* abbiamo delle tematiche comuni a tutti i testi[33], e in questo senso i *topics* stimati si possono considerare "*sotto-argomenti"*) nonché la "rigidità" nella struttura dei *topics* dal punto di vista temporale (data dal fatto che il CTM considera i documenti del corpus interscambiabili, non ordinati). Quest'ultima caratteristica è particolarmente insidiosa per i testi raccolti nel corso del tempo, in quanto un *argomento* che tratta di un avvenimento accaduto nel momento x è, generalmente, improponibile come *topic* per i testi precedenti l'istante x. In un caso come questo, la correlazione tra gli *argomenti* del CTM ci può fuorviare. A questo proposito lo stesso autore del CTM, Blei, ha suggerito nel 2009 [34] un nuovo modello, il **dynamic topic model**, che riesce a tenere conto dell'ordinamento temporale dei testi.

In realtà, oltre a quella appena menzionata sono state sviluppate molte altre varianti del CTM [35], sia dai suoi ideatori originali che da altri ricercatori, allo scopo di aggirare alcuni dei suoi limiti. È la dimostrazione dell'interesse crescente sugli strumenti di questo tipo.

In conclusione, i risultati del *correlated topic model* sul corpus di *openpolis* possono essere valutati positivamente, ma è d'obbligo ricordare l'esistenza di altri strumenti

---

[32] Si veda ad esempio *«A density-based method for adaptive LDA model selection»*, 2008, di J. Cao, T. Xia, J. Li, Y. Zhang e S. Tang.

[33] Rappresentate dalle parole piene escluse dalla matrice *Term Frequency – Inverse Document Frequency*, di cui parlavamo alle pagg. 29 e 30.

[34] *«Topic Models»*, 2009, di D. Blei e J. Lafferty (si veda anche la sezione *Approfondimenti* della Bibliografia).

[35] Si veda ad esempio *«Multilingual topic models for unaligned text»*, 2009, di J. Boyd-Graber e D. Blei, per applicare *topic models* su corpus con testi in più lingue.

disponibili già oggi per analisi di questo genere, dei quali è presente una piccola rassegna nella Bibliografia. D'altro canto l'espansione di Internet e la conseguente crescita delle informazioni disponibili in rete renderà sempre più importante lo studio di tecniche per l'elaborazione e la catalogazione di grandi moli di dati testuali. Di conseguenza, nell'immediato futuro ci aspettiamo sia un approfondimento degli strumenti già esistenti, sia lo sviluppo di nuovi.

Ampliando il tema dell'analisi testuale, la ricerca si sta interessando anche all'analisi di altri "elementi multimediali" (immagini, musica, filmati) sempre allo scopo di catalogarli in modo semi-automatico. È una sfida molto interessante, perché si prevede che, proprio come avviene per i testi, la quantità disponibile di questi elementi aumenterà a mano a mano che la tecnologia di Internet verrà implementata nelle altre tecnologie (telefoni cellulari, televisori, automobili, ma anche macchinari industriali). Modelli statistici per analisi di immagini e musica esistono già, ma di certo la strada da percorrere è ancora lunga e strettamente collegata ai progressi fatti nel campo dell'analisi testuale. Insomma, è una storia tutta da scrivere.

APPENDICE

A1. TERMINI PIÙ FREQUENTI PER TOPIC

TOPIC 1					
nucleare	**lazio**	**centrali**	**siti**	**sviluppoeconomico**	**agenzia**
0.101	0.029	0.023	0.022	0.022	0.022
centrale	**lombardia**	**energia**	**sottosegretario**	**verdi**	
0.020	0.018	0.017	0.016	0.016	

Frequenza cumulata: 0.3046

TOPIC 2								
tremonti	**brunetta**	**pubblicaamministrazione**	**cgil**	**banche**	**renato**	**sindacato**	**confindustria**
0.047	0.042	0.026	0.023	0.022	0.018	0.018	0.014
fiat	**produttività**	**giulio**	**manovra**	**sciopero**	**cisl**	**letta**	**salari**	**pil**
0.014	0.011	0.011	0.011	0.010	0.010	0.009	0.008	0.008

Frequenza cumulata: 0.3034

TOPIC 3								
rai	**calabria**	**vigilanza**	**villari**	**centrale**	**informazione**	**regolamento**	**carbone**	**the**
0.057	0.023	0.023	0.023	0.022	0.019	0.014	0.014	0.012
vicenza	**serviziopubblico**	**on**	**sezione**	**to**	**cda**	**cortedeiconti**	**commissionedivigilanza**
0.012	0.010	0.010	0.010	0.010	0.010	0.009	0.009

Frequenza cumulata: 0.2892

TOPIC 4							
interrogazione	**torino**	**ue**	**milano**	**parlamentoeuropeo**	**bruxelles**	**firenze**
0.040	0.035	0.033	0.032	0.021	0.020	0.018
commissioneeuropea	**piemonte**	**velocità**	**direttiva**	**delturco**	**osservatorio**	**funzionari**	**stazione**
0.013	0.013	0.010	0.009	0.009	0.008	0.008	0.008

Frequenza cumulata: 0.3002

TOPIC 5									
donne	**rifiuti**	**violenza**	**donna**	**polizia**	**provinciale**	**femminile**	**forzedell ordine**	**quote**	**raccoltadif ferenziata**
0.099	0.039	0.026	0.025	0.022	0.018	0.018	0.016	0.013	0.012

Frequenza cumulata: 0.2885

TOPIC 6						
napoli	**rifiuti**	**rete**	**campania**	**internet**	**bassolino**	**governatore**
0.076	0.064	0.058	0.050	0.026	0.016	0.013

Frequenza cumulata: 0.3028

TOPIC 7									
veneto	**federalismo**	**bossi**	**galan**	**federalismofiscale**	**sindaci**	**leganord**	**zaia**	**province**	**giaretta**
0.088	0.042	0.036	0.027	0.025	0.022	0.021	0.018	0.018	0.017

Frequenza cumulata: 0.3142

TOPIC 8								
scuola	**scuole**	**università**	**studenti**	**terni**	**alitalia**	**istruzione**	**scolastico**	**gelmini**
0.086	0.037	0.032	0.031	0.030	0.029	0.023	0.019	0.018

Frequenza cumulata: 0.3036

TOPIC 9						
acqua	**abruzzo**	**aquila**	**ricostruzione**	**maroni**	**protezionecivile**	**bertolaso**
0.040	0.035	0.034	0.026	0.026	0.023	0.021
terremoto	**case**	**campi**	**guido**	**rom**	**sottosegretario**	**chiodi**
0.019	0.019	0.013	0.013	0.012	0.012	0.012

Frequenza cumulata: 0.3041

TOPIC 10										
pace	**alemanno**	**obama**	**esteri**	**israele**	**militari**	**pace**	**cina**	**militare**	**milano**	**frattini**
0.029	0.022	0.021	0.018	0.016	0.016	0.029	0.014	0.013	0.012	0.012
dirittiumani	**onu**	**afghanistan**	**expo**	**hamas**	**visita**	**larussa**	**fascismo**	**bush**	**gaza**	
0.012	0.011	0.011	0.010	0.009	0.008	0.007	0.007	0.007	0.007	

Frequenza cumulata: 0.2635

TOPIC 11								
on	**mario**	**saviano**	**alessandro**	**vicinanza**	**rampelli**	**provinciale**	**scritte**	**vorremmo**
0.016	0.010	0.010	0.009	0.009	0.007	0.006	0.006	0.006
luca	**tariffe**	**arte**	**assessori**	**ripristinare**	**luca**	**dover**	**sensibilità**	**tantissimi**
0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005

Frequenza cumulata: 0.1221

TOPIC 12							
veltroni	**fini**	**udc**	**dipietro**	**candidato**	**casini**	**bersani**	**dalema**
0.048	0.031	0.024	0.020	0.019	0.014	0.013	0.012
franceschini	**alleanzanazionale**	**lista**	**idv**	**leggeelettorale**	**ripristinare**		
0.012	0.011	0.010	0.010	0.010	0.005		
rutelli	**candidati**	**candidatura**	**referendum**	**gianfranco**	**italiadeivalori**	**tv**	
0.010	0.009	0.009	0.009	0.008	0.007	0.006	

Frequenza cumulata: 0.2884

TOPIC 13							
lavoratori	**occupazione**	**dipendenti**	**reddito**	**contratto**	**cassaintegrazione**	**disoccupazione**	**precari**
0.085	0.022	0.021	0.020	0.018	0.014	0.014	0.013
ammortizzatorisociali	**pensioni**	**damiano**	**lavoratore**	**contratti**	**sindacali**	**sacconi**	**cesare**	**pensione**
0.013	0.011	0.011	0.011	0.011	0.010	0.009	0.008	0.008

Frequenza cumulata: 0.3003

TOPIC 14								
radicali	**napolitano**	**sentenza**	**intercettazioni**	**pannella**	**testo**	**bonino**	**capodello stato**	**testamento biologico**
0.027	0.019	0.016	0.016	0.014	0.013	0.012	0.011	0.011
ddl	**eluana**	**magistrati**	**medico**	**senatrice**	**giudici**	**disegno dilegge**	**reati**	**englaro**	**cattolici**
0.011	0.011	0.010	0.008	0.008	0.008	0.007	0.007	0.007	0.007

Frequenza cumulata: 0.2222

TOPIC 15							
energia	**impianti**	**impianto**	**fontirinnovabili**	**sport**	**puglia**	**energetico**	**fonti**
0.041	0.033	0.019	0.015	0.013	0.013	0.011	0.011
emissioni	**luca**	**rinnovabili**	**sostenibile**	**agricoltura**	**energetica**	**ambientali**	**eolico**
0.011	0.005	0.010	0.010	0.010	0.009	0.009	0.009
fotovoltaico	**energierinnovabili**	**gas**	**installazione**	**tecnologie**			
0.009	0.009	0.008	0.008	0.008			

Frequenza cumulata: 0.2656

TOPIC 16								
venezia	**cacciari**	**sanità**	**ospedale**	**porto**	**mestre**	**ponte**	**san**	**cosenza**
0.065	0.028	0.026	0.025	0.024	0.019	0.018	0.013	0.012

asl	ospedali	marino	sanitaria	aeroporto	sanitario	comma	ospedali
0.012	0.011	0.010	0.010	0.010	0.010	0.010	0.011

Frequenza cumulata: 0.3027

TOPIC 17

zona	amministrazionecomunale	edilizia	beni	parco	urbanistica	siracusa	edifici	delibera	sorbello
0.021	0.015	0.013	0.012	0.012	0.011	0.010	0.010	0.010	0.010
residenti	**centro storico**	**bando**	**traffico**	**quartiere**	**centri**	**provinciale**	**san**	**valorizzazione**	**manutenzione**
0.010	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007

Frequenza cumulata: 0.2055

TOPIC 18

sicilia	immigrati	mafia	criminalità	mezzo giorno	lombardo	palermo	magistrati	procura	indagini
0.029	0.026	0.025	0.023	0.023	0.020	0.013	0.013	0.012	0.012
isola	**anti mafia**	**prostituzione**	**inchiesta**	**marrazzo**	**castel nuovo**	**camorra**	**clandestini**	**procuratore**	**cuffaro**
0.010	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.007	0.007

Frequenza cumulata: 0.2745

TOPIC 19

ordinedel giorno	dichiara	teatro	piemonte	stefano	ennesimo	industria	carceri	detenuti
0.019	0.015	0.014	0.013	0.012	0.011	0.011	0.011	0.010
carcere	**enti**	**onda**	**affermato**	**cota**	**polizia**	**festival**	**regolarmente**	**cameradeideputati**
0.010	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.006

Frequenza cumulata: 0.1843

TOPIC 20

presentazione	provinciale	scrive	ribadito	festa	dichiara	insediamento	notizie	dover
0.009	0.008	0.007	0.007	0.007	0.007	0.006	0.006	0.006
promesse	**andrea**	**polizia**	**stefano**	**commenta**	**condivisione**	**gesto**	**manifestazione**	**giovedì**
0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005

Frequenza cumulata: 0.1108

A2. Parole Scartate dalla Matrice TF–IDF

In grassetto i termini che consideriamo parole piene.

accordo	alcuni	almeno	altra	**ambiente**	**amministrazione**
andare	anni	anno	**assessore**	attenzione	attività
atto	attraverso	avanti	aver	base	bene
berlusconi	**bilancio**	bisogna	bisogno	**camera**	campo
casa	caso	**centro**	**centrodestra**	**centrosinistra**	certo
circa	città	**cittadini**	commissione	comunale	comune
comuni	comunque	conclude	condizioni	confronti	confronto
consigliere	**consiglio**	continua	conto	corso	cose
costituzione	credo	**crisi**	dare	dati	dato
davvero	decisione	**decreto**	**democrazia**	**destra**	detto
devono	dice	dire	**diritti**	**diritto**	discussione
dobbiamo	dovrebbe	**economia**	**elezioni**	esempio	euro
europa	ex	fa	**famiglie**	far	fatti
fatto	forse	forte	forza	fronte	fuori
futuro	generale	giorni	giorno	**giovani**	giunta
giustizia	**governo**	grado	grande	grandi	gruppo
idea	ieri	**impegno**	importante	**imprese**	incontro
infatti	**iniziativa**	inoltre	insieme	interno	**interventi**
intervento	**istituzioni**	**italia**	italiana	italiani	italiano
lavori	**leader**	**lega**	**legge**	**libertà**	livello
maggioranza	massimo	mentre	merito	mesi	mettere
mila	milioni	**ministro**	modo	momento	mondo
nazionale	nè	necessario	necessità	nessuno	nulla
nuova	nuove	**obiettivo**	occasione	oggi	**opposizione**
ore	ormai	**paese**	paesi	parlamentare	**parlamento**
parlare	parole	parte	particolare	**partiti**	**partito**
partitodemocratico	persona	personale	piano	**politica**	**politiche**
politico	**popolodellalibertà**	posizione	possa	possibile	possibilità
possono	posto	potrebbe	**premier**	**presidente**	**problema**
problemi	**prodi**	progetto	programma	**proposta**	propria
proprio	provincia	**pubblica**	**pubblici**	**pubblico**	punto
qualche	quali	qualità	quel	questione	rapporto
realtà	regionale	**regione**	**regioni**	**repubblica**	**responsabilità**
ricerca	**riforma**	riguarda	**rischio**	**risorse**	risposta
roma	ruolo	scelta	scelte	segretario	sempre
senato	senso	servizi	servizio	**settore**	sì
sicurezza	sindaco	**sinistra**	sistema	situazione	**sociale**
sociali	**società**	spiega	stata	state	stessa
storia	strada	**sviluppo**	tale	tema	tempi
territorio	tratta	troppo	tutta	tutte	unico
vero	verso	via	viene	vista	visto
vita	volta	**voto**	vuole		

BIBLIOGRAFIA E SITOGRAFIA

A. Tuzzi, 2003, *L'analisi del contenuto: introduzione ai metodi e alle tecniche di ricerca*, I ed., Roma: Carocci.

F. della Ratta – Rinaldi, *L'analisi testuale nello studio di caso*, http://www.ilpalo.com/parodie-scaricare-libri-narrativa/analisi-testuale.htm, ultima consultazione il 25/03/2011.

A. Vardanega, 2010, *Statistica testuale e interpretazione,* http://blog.agnesevardanega.eu/2010/08/23/statistica-testuale-e-interpretazione/, ultima consultazione il 25/03/2011.

F. Tomasi, 2002, *Manuale di informatica umanistica per l'applicazione delle pratiche computazionali ai testi letterari – Parte 9: Analisi del Testo*, http://www.griseldaonline.it/informatica/manuale.htm, ultima consultazione il 25/03/2011.

D. Blei & J. Lafferty, 2007, *A Correlated Topic Model of Science*, The Annals of Applied Statistics, Vol. 1, No. 1, p. 17–35.

D. Blei, A. Ng & M. Jordan, 2003, *Latent Dirichlet Allocation*, The Journal of Machine Learning Research, 3:993–1022.

J. Aitchison & S. M. Shen, 1980, *Logistic normal distributions: some properties and uses*, Biometrika 67, p. 261–272.

Wikipedia, www.wikipedia.org.

APPROFONDIMENTI

J. Huang & T. Malisiewicz, 2007, *Hierarchical Logistic Normal parameter estimation (Correlated Topic Model details*), http://www.cs.cmu.edu/~jch1/research/old/ctm.pdf.

D. Blei & J. Lafferty, 2009, *Topic Models*, Text Mining: Theory and Applications, London: Taylor and Francis.

K. Salomatin, Y. Yang & A. Lad, 2009, *Multi-field correlated topic modeling*, Proceedings of the Ninth SIAM International Conference on Data Mining, p. 628–637.

D. Newman, S. Karimi & L. Cavedon, 2009, *External Evaluation of Topic Models*, winner of "Best Paper Award" in the Fourteenth Australasian Document Computing Symposium.

L. Sangno, S. Jaeki & K. Yongjin, 2010, *An empirical comparison of four text mining methods*, The Journal of Computer Information Systems.

A. Chanen & J. Patrick, 2007, *Measuring Correlation Between Linguists' Judgments and Latent Dirichlet Allocation Topics*, Fifth Australasian Language Technology Workshop.

S. Williamson, C. Wang, K. Heller & D. Blei, 2010, *The IBP compound Dirichlet process and its application to focused topic modeling*, International Conference on Machine Learning.

W. Li & A. McCallum, 2006, *Pachinko allocation: DAG-structured mixture models of topic correlations*, Proceedings of the 23rd International Conference on Machine Learning.

COLLEGAMENTI UTILI

Strumenti e software utili per l'analisi testuale (ultima consultazione il 25/03/2011):

http://www.stabellini.org/
"Una serie di link su alcuni studi di Statistica, con particolare rilevo per la Statistica Testuale".

http://www.textanalysis.info/ncatsys.htm
Informazioni ed elenco di programmi sulla Text Analysis.

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicowww
Sito ufficiale del software "Lexico3".

http://www-958.ibm.com/software/data/cognos/manyeyes/page/Phrase_Net.html
"Many Eyes": esperimento della IBM per la ricerca di "patterns" in un testo.

http://tagcloud.oclc.org/tagcloud/TagCloudDemo
Sito per creare in pochi click la "tag cloud" di un testo.

http://www.wordle.net/
Sito per creare in pochi click la "word cloud" di un testo.

Documentazione della libreria *topicmodels* di **R**:

http://cran.r-project.org/web/packages/topicmodels/