

Tesi di Laurea

**ANALISI STATISTICA  
DI DICHIARAZIONI  
POLITICHE  
TRAMITE  
CORRELATED TOPIC  
MODEL**

Relatori: Livio Finos, Dario Solari

Candidata: Sara Baldan

# L'ANALISI DEL TESTO

The slide features a solid blue background. At the bottom, there are several overlapping, wavy, light blue lines that create a sense of movement and depth, resembling a stylized horizon or a series of waves.

# Cos'è l'analisi del testo

Un insieme di tecniche che permette di analizzare ed esplorare un singolo testo o una raccolta, anche molto ampia, di testi.

Il primo esempio documentato di analisi del contenuto (applicata a testi religiosi) si può rintracciare nella Svezia del XVII secolo.

Oggi è sempre più usata per la crescente diffusione di Internet, che tramite pagine web, social network, newsgroup, chat e forum rende disponibile un'immensa quantità di informazioni.

# Come condurre un'analisi del testo

## Tre fasi:

- 1) Preparazione del testo (lettura, individuazione poliformi, ...)
- 2) Analisi vera e propria (analisi delle corrispondenze, analisi delle co-occorrenze, individuazione delle parole caratteristiche, del linguaggio peculiare, delle parole omogenee, delle corrispondenze lessicali, ...)
- 3) Esposizione dei risultati

# Come condurre un'analisi del testo

Due approcci tradizionali:

- **linguistico**

il testo è visto come un insieme finito di **elementi portatori di senso** (parole o gruppi di parole) che possono essere elencati.

L'analisi si basa sullo studio di questi elementi.

- **statistico**

idea di base: più un termine è presente nel testo, più lo rappresenta.

L'analisi si basa sullo studio delle parole più frequenti.

# Come condurre un'analisi del testo

Limiti di ciascun approccio:

- **linguistico**

- › più grande è il corpus, più stilare l'elenco diventa laborioso!
- › mancanza di una procedura concordata e riproducibile
- › discrezionalità  
(punto debole o punto di forza?)

- **statistico**

- › non tiene conto delle **parole composte** (sequenze di parole con significato diverso)
- › non riesce a distinguere tra parole **vuote** e **piene**

# Quale metodo usare?

**QUALITATIVO**

(approccio linguistico)

oppure

**QUANTITATIVO**  
(approccio statistico)



Nell'analisi del testo non ci può essere una netta distinzione tra “qualitativo” e “quantitativo”: essa ha sia componenti **qualitative** (il testo è di per sé qualitativo per eccellenza) sia **quantitative** (gli strumenti di analisi sono tipicamente statistico-matematici).

Nuovo approccio misto:

QUALITATIVO

QUANTITATIVO

**“QUANTIQUALITATIVO”**  
(tecniche miste qualitative e quantitative)

# Approccio “quantiqualitativo”

Questo approccio presenta i vantaggi di entrambi i metodi.

- › possibilità di individuare **poliformi** e **sequenze di parole**
- › procedura ripetibile
- › economicità (di tempo, di calcolo): la fase di elencazione delle parole è svolta dal computer

# Analisi del testo con CTM

La fase di analisi viene svolta con l'uso del **CORRELATED TOPIC MODEL**

- 1) Preparazione del testo
- 2) Analisi vera e propria
- 3) Esposizione dei risultati



**CTM**

# Correlated Topic Model

Modello ad  
Argomenti  
Correlati

# Correlated Topic Model

- Evoluzione del Latent Dirichlet Allocation (LDA)
- Variabili latenti
- Ogni documento del corpus può trattare di uno o più argomenti
- Permette correlazione tra argomenti

# Rappresentazione Grafica

## Vocabolario:

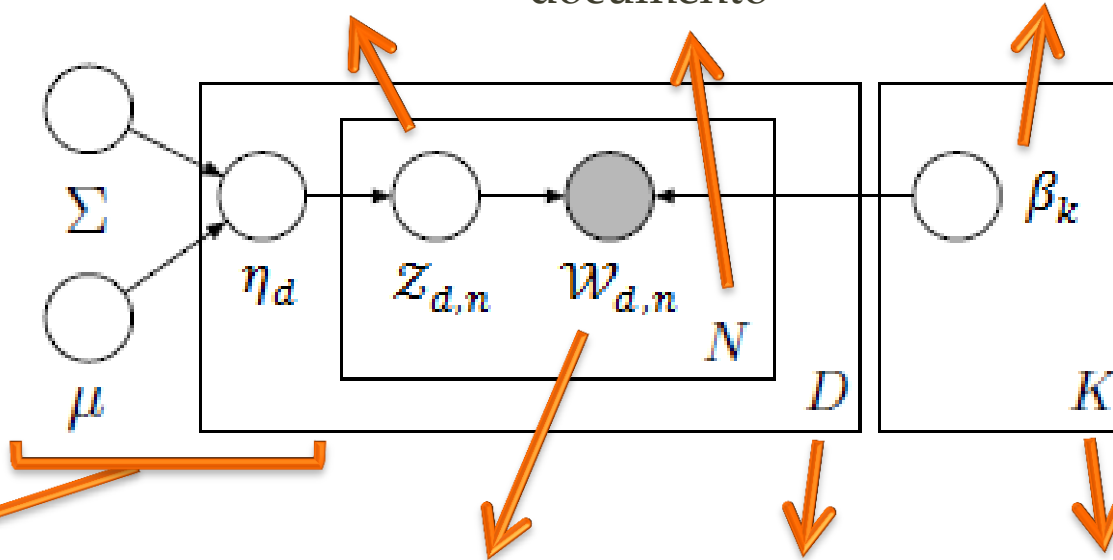
insieme delle  $V$  parole utilizzabili nel corpus

## Modello grafico probabilistico

$z_{d,n}$  realizzazione di  $Mult(\theta_d)$ ; indica l'argomento cui appartiene  $w_{d,n}$

numero di parole del documento

distribuzione dell'argomento  $k$ -esimo sul vocabolario



$\eta_d$  realizzazione di  $\mathcal{N}_{K-1}(\mu, \Sigma)$

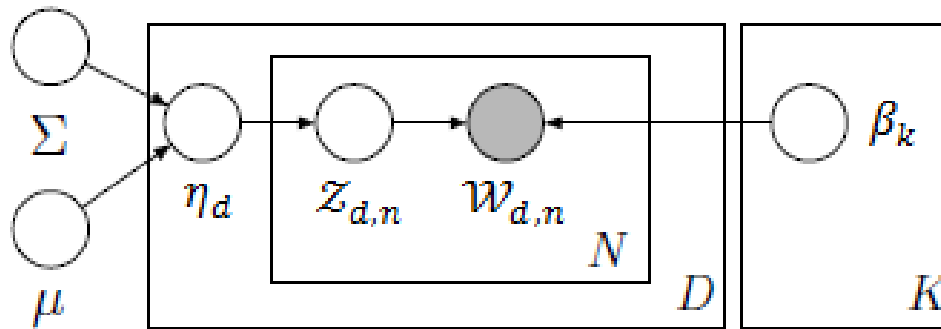
$$\theta_d^* = f(\eta_d) = \frac{e^{\eta_d}}{1 + \sum_{j=1}^{K-1} e^{\eta_d}}$$

$w_{d,n}$  realizzazione di  $Mult(\beta_{z_{d,n}})$  è la  $n$ -esima parola del  $d$ -esimo documento

numero di documenti

numero di argomenti

# Riassumendo...



- 1) Estrazione distribuzione tematica del documento da normale logistica
- 2) Estrazione assegnazione tematica della parola da multinomiale
- 3) Estrazione della parola da multinomiale

Nell'LDA l'estrazione avviene da una Dirichlet

# La Normale Logistica

Nel CTM viene usata al posto della Dirichlet per includere la correlazione tra topics.

**Def:** sia  $v$  realizzazione di  $\mathcal{N}_{K-1}(\mu, \Sigma)$ .

Allora  $u$  è realizzazione di una normale logistica  $\mathcal{U} \sim \mathcal{L}_{K-1}(\mu, \Sigma)$  se e solo se

$$u = f(v) = \frac{e^v}{1 + \sum_{j=1}^{K-1} e^{v_j}},$$

$$\text{cioè } v = f^{-1}(u) = \ln \frac{u}{1 - \sum_{j=1}^{K-1} u_j}.$$

Il supporto di  $\mathcal{U}$  è

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

# Perché $u$ ha dimensione $K-1$ ?

$u$  deve rappresentare uno spazio di probabilità.

La  $K$ -esima probabilità si ottiene con

$$u_K = 1 - \sum_{j=1}^{K-1} u_j$$

Quindi si definisce

$$\theta = u^* = [u^T \quad u_K] = [u_1 \quad u_2 \quad \dots \quad u_{K-1} \quad u_K],$$

che varia in

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

# Quali parametri ci interessano?

I parametri che vogliamo stimare sono  $\mu$ ,  $\Sigma$  e  $\beta$ .

Sono i parametri sufficienti (e necessari) per ripetere il procedimento.

**N.B.**

*Le uniche variabili osservabili (cioè non latenti) sono le parole  $w$ .*

# Stima dei Parametri

→ **Massima verosimiglianza?**

**no**, perché ci sono variabili latenti.

Si usa allora il metodo

## **VARIATIONAL EXPECTATION-MAXIMIZATION**

1) **E-step**

calcolo approssimazione della distribuzione a posteriori delle variabili latenti condizionandoli a dati e parametri correnti

2) **M-step**

calcolo stima di massima verosimiglianza dei parametri condizionandoli a dati e distribuzione delle variabili latenti calcolata al punto 1)

# Stima dei Parametri

- I due passi vengono ripetuti iterativamente fino a raggiungere la convergenza.

- Si ottengono

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

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

$$\hat{\boldsymbol{\beta}}_i \propto \sum_d [\phi_d]_i m_d .$$

$[\phi_d]_i$  elemento  $i$ -esimo di  $\boldsymbol{\phi}_d = E(\mathbf{Z}_d)$

$\mathbf{m}_d$  vettore  $V$ -variato, conta quante volte ciascuna parola compare nel documento  $d$

$\lambda_d, v_d^2$  parametri variazionali



il caso

*[www.openpolis.it](http://www.openpolis.it)*



# COS'È OPENPOLIS?

- Associazione senza fini di lucro.
- Raccoglie le informazioni su tutti i politici italiani. Ad ognuno è dedicata una scheda dove viene ricostruito il profilo e dove vengono raccolte le sue dichiarazioni pubbliche.
- Il database è enorme e i 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 CORPUS

- In questo caso il corpus è costituito dalle dichiarazioni politiche presenti in **openpolis**.
- Una dichiarazione è contenuta in un file di testo, la cui struttura è:

titolo



data



autore



testo



## Esempio

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 miliardi di euro. E non declassa il Campidoglio. Lo spiega [...]

# FASE 1. PREPARAZIONE DEL CORPUS

- Eliminazione delle dichiarazioni uguali
- Risoluzione problemi di codifica
- Eliminazione di tag HTML
- Individuazione **poliformi**
- Trasformazione in minuscolo
- Correzione errori ortografici
- ...

## **POLIFORME:**

Segmento di testo con significato diverso da quello dei singoli termini che lo compongono.

*Es. "patata bollente", "dare carta bianca"*

<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 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.</b><br />

<br />

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 PD: "Non c'è nessuna città d'Europa che ha lasciato un deficit di 16 mila miliardi di vecchie lire". [...]

PRIMA

DOPO

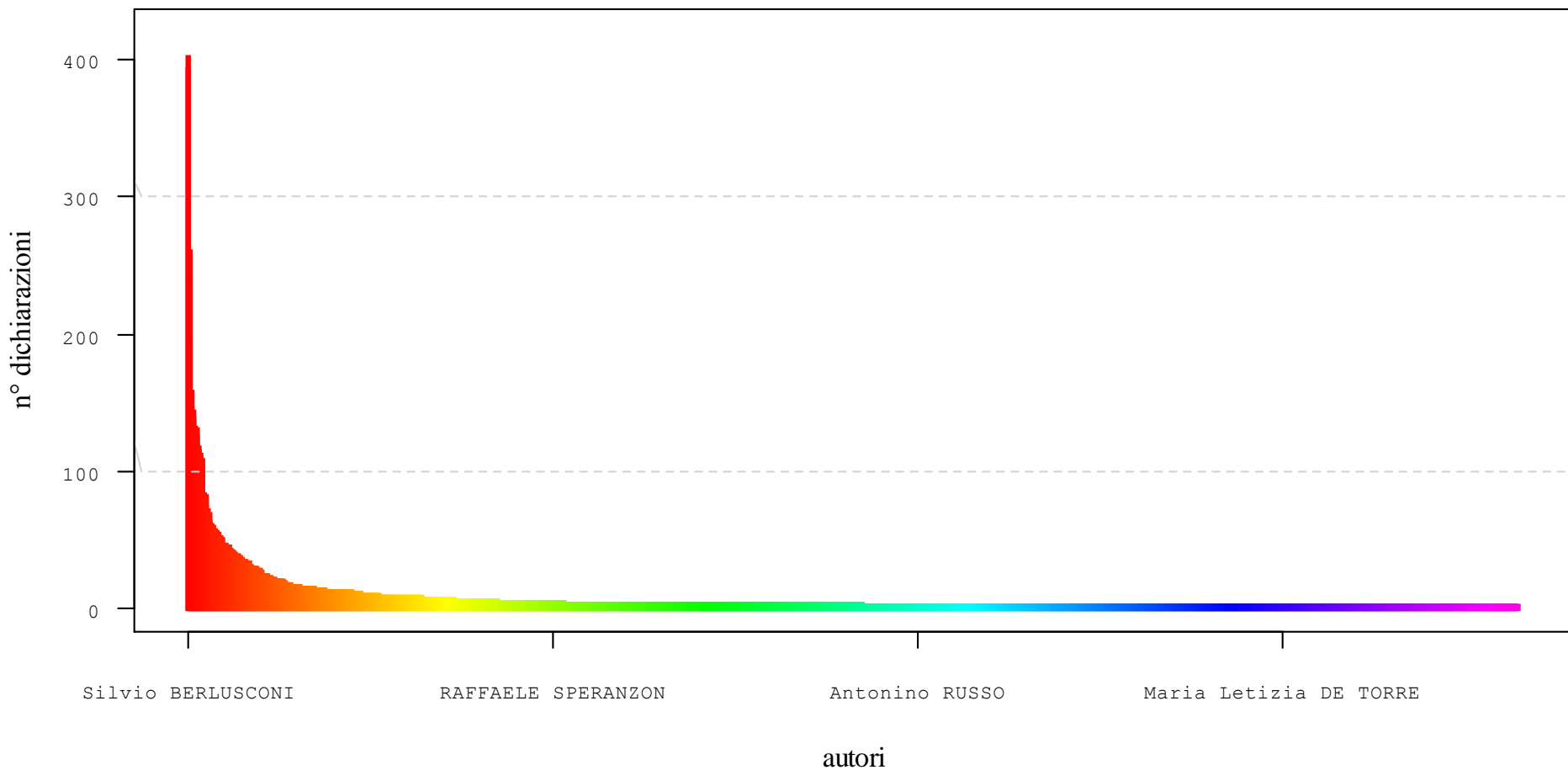
standard and poor' s smentisce alemanno e berlusconi sul debito del comune di roma. non 10 miliardi come ripetuto in un mantra dalle destre, ma 6, 9 miliardi di euro. e non declassa il campidoglio. lo spiega in un'intervista a la stampa myriam fernandez de heredia, responsabile per standard and poor' 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. [...]

# QUALCHE NUMERO...

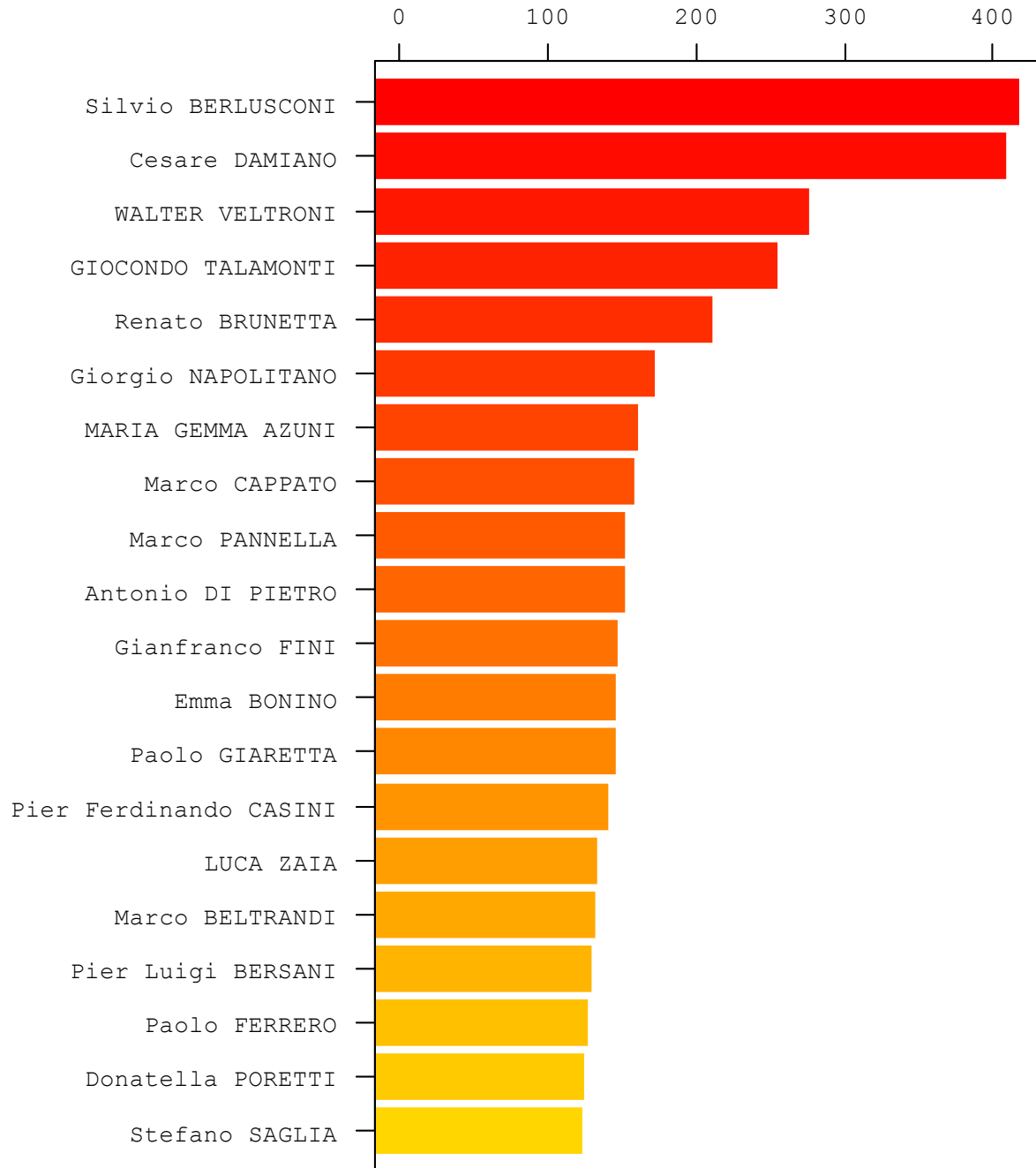
- **12.807**: dichiarazioni disponibili inizialmente
- **12.575**: dichiarazioni rimaste dopo la fase di preparazione
- **3.970.979**: numero complessivo di termini del corpus
- **80.475**: numero di parole distinte del corpus (dimensione del vocabolario)
- **1.823**: politici presenti nel database (al 01/12/2010)
- **20/10/1984**: data della dichiarazione meno recente
- **01/12/2010**: data delle dichiarazioni più recenti

# ... E QUALCHE GRAFICO DESCRITTIVO

Frequenza dichiarazioni per politico



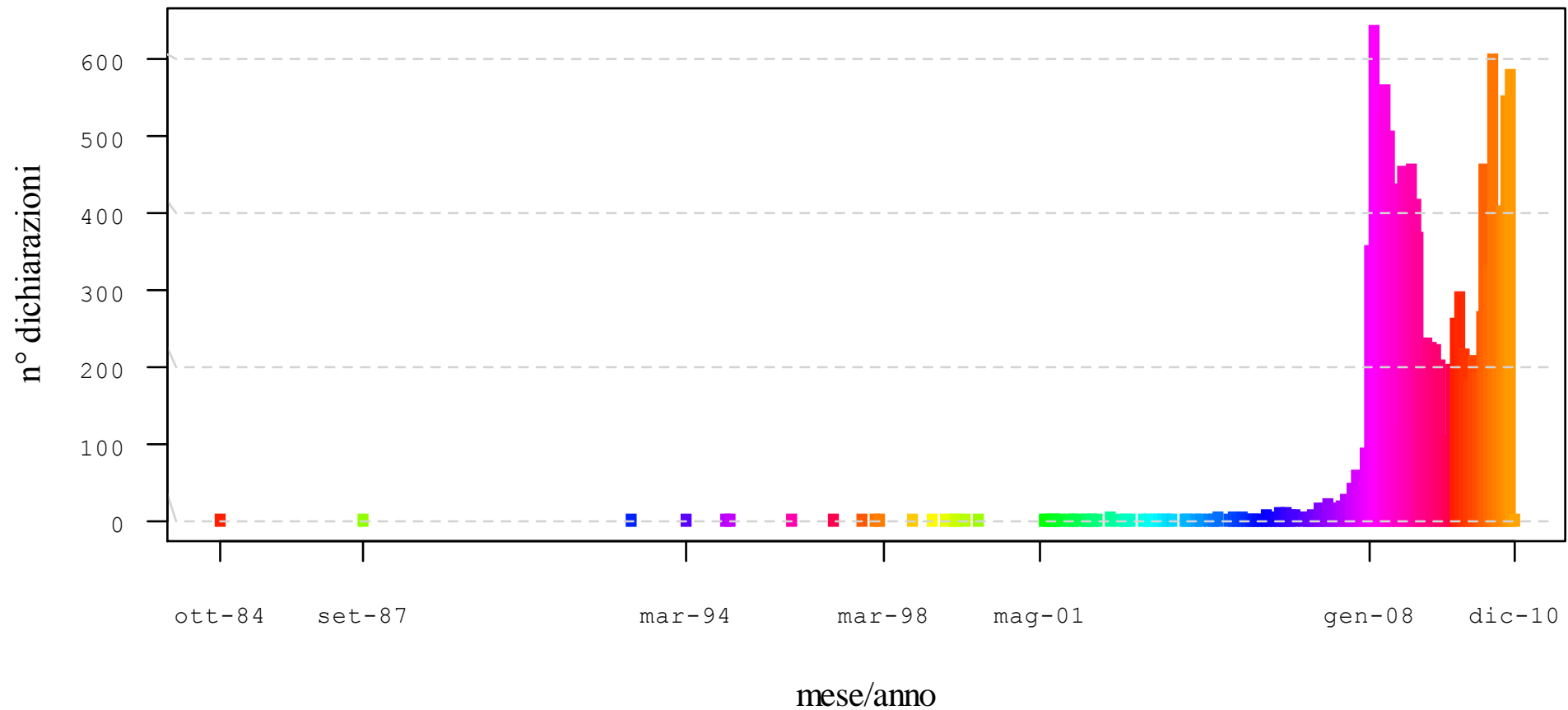
n° dichiarazioni



## Frequenza dichiarazioni per politico

limitata ai  
politici più  
presenti in  
*openpolis*  
(# dichiarazioni > 100)

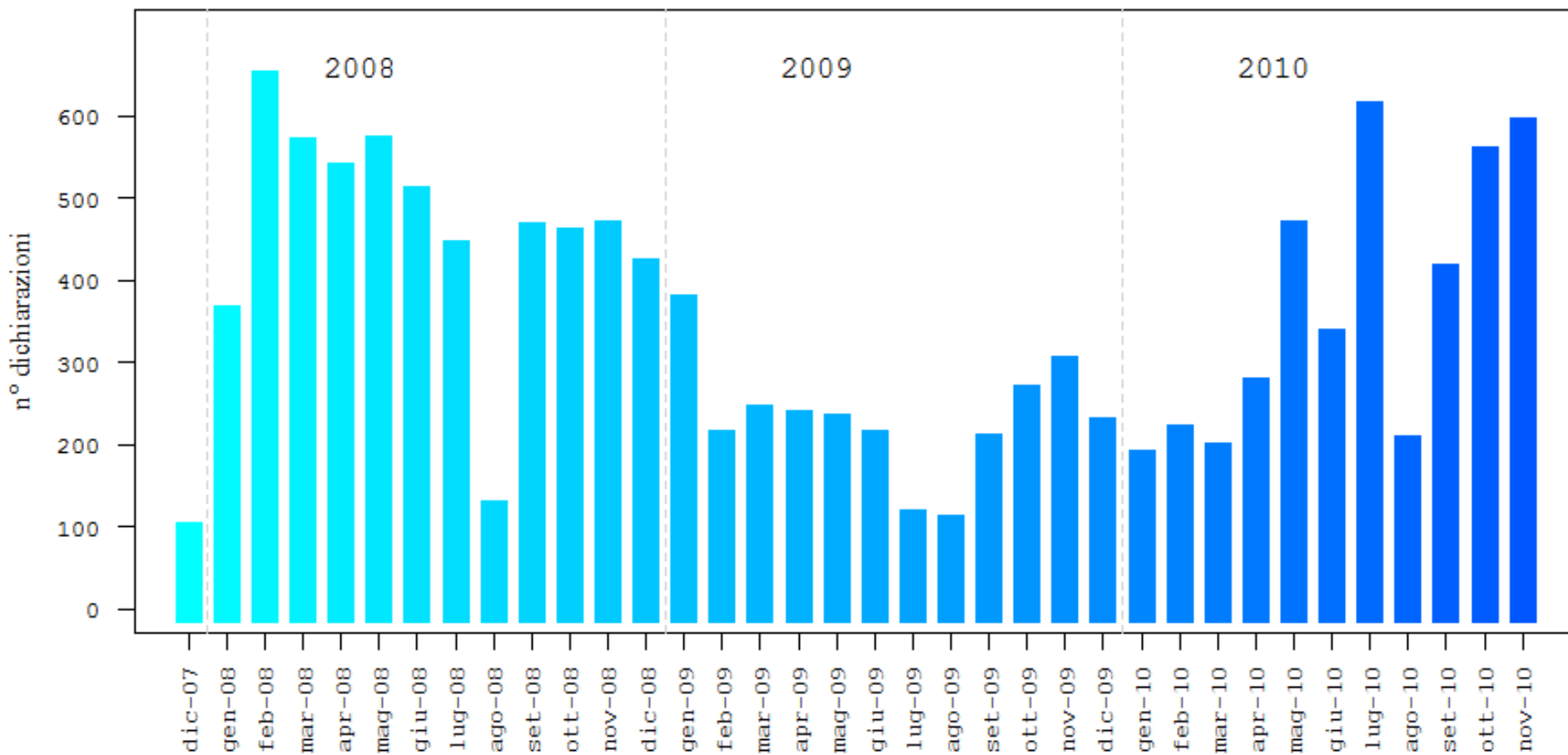
# Frequenza dichiarazioni nel tempo



# Frequenza dichiarazioni nel tempo

limitata all'intervallo 2008 - 2010

**Frequenza dichiarazioni da fine 2007 a fine 2010**



## FASE 2. APPLICAZIONE DEL MODELLO

- Usiamo il comando `CTM()` del package `topicmodels` di R.
- Numero di topics ( $k$ ) = 20
- Ignoriamo le parole estremamente frequenti e quelle estremamente rare.

Per individuarle assegniamo a ogni parola dei pesi tramite la matrice **Term Frequency – Inverse Document Frequency**

## FASE 3. ANALISI DEL RISULTATO

- Il modello è in un oggetto di classe "CTM, *topic model*".
- Contiene:
- $\hat{\mu}, \hat{\Sigma}$  (stime dei parametri della normale logistica)
- il vocabolario
- $\log(\hat{\beta}_i) \forall i$  ( $\beta_i$  è la distribuzione sul vocabolario per il *topic i*)
- ...

# STIMA DI $\mu$

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

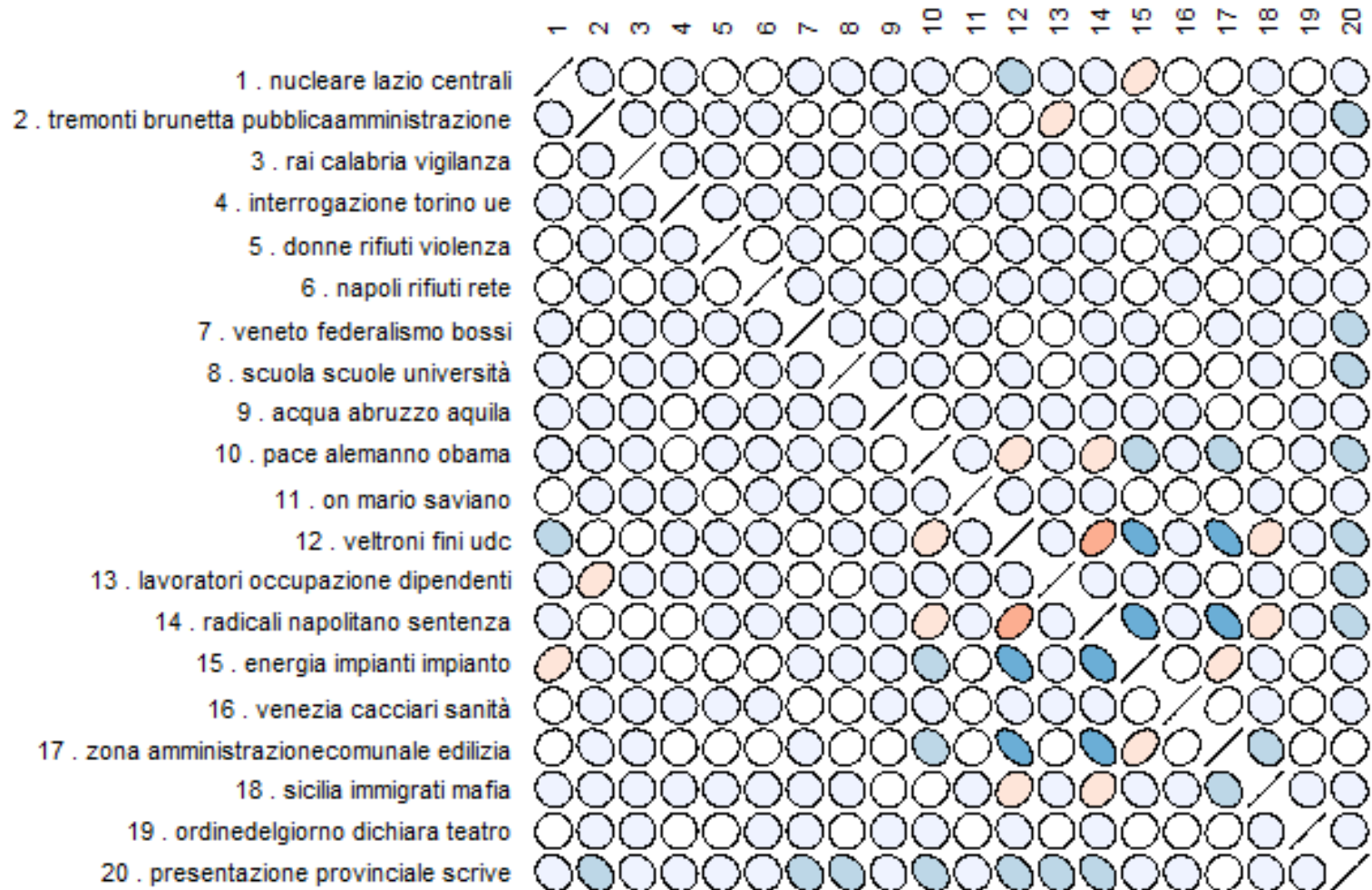
$$\hat{\theta}^* = \frac{e^{\hat{\mu}}}{1 + \sum_{i=1}^{19} e^{\hat{\mu}_i}} ; \quad \hat{\theta} = [\hat{\theta}^*, 1 - \sum_{i=1}^{19} \hat{\theta}_i^*]$$

$$\hat{\theta} = \begin{bmatrix} 0.035 & 0.044 & 0.046 & 0.061 & 0.054 & 0.058 & 0.043 \\ 0.024 & 0.056 & 0.035 & 0.056 & 0.083 & 0.041 & 0.042 \\ 0.036 & 0.049 & 0.056 & 0.062 & 0.057 & 0.063 \end{bmatrix}$$

# STIMA DI $\Sigma$

visualizzazione grafica: correlazione tra topics

## Correlazione tra topics





# PAROLE CON FREQUENZA ALTA ESCLUSE DAL VOCABOLARIO

Alcune sono parole vuote, altre sono significative.

Esempio:

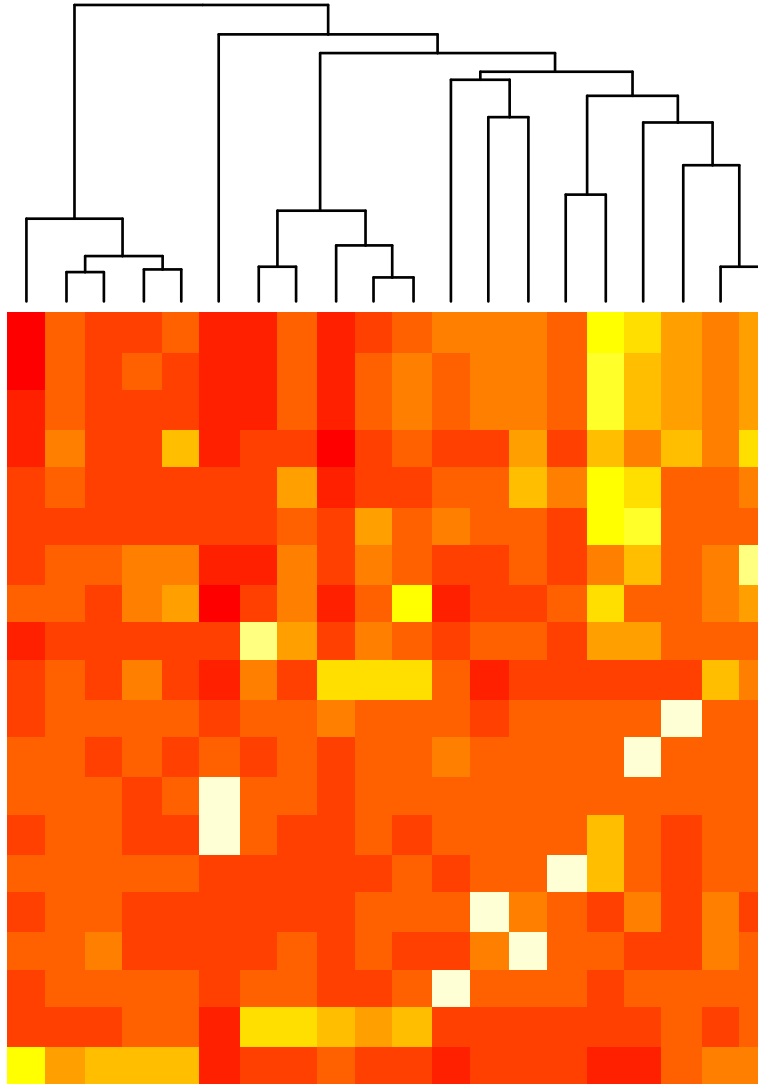
<b>costituzione</b>	credo	<b>crisi</b>	dare	dati	dato
davvero	<b>decisione</b>	<b>decreto</b>	<b>democrazia</b>	<b>destra</b>	detto
devono	dice	dire	<b>diritti</b>	<b>diritto</b>	<b>discussione</b>
dobbiamo	dovrebbe	<b>economia</b>	<b>elezioni</b>	esempio	<b>euro</b>
<b>europa</b>	ex	fa	<b>famiglie</b>	far	fatti
fatto	forse	forte	forza	fronte	fuori

# PAROLE CON FREQUENZA ALTA ESCLUSE DAL VOCABOLARIO

Principali parole **significative** escluse dalla TF-IDF:

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

# QUALCHE GRAFICO PER COMPRENDERE MEGLIO I RISULTATI



- 19 . ordinedelgiorno dichiara teatro
- 11 . on mario saviano
- 20 . presentazione provinciale scrive
- 6 . napoli rifiuti rete
- 16 . venezia cacciari sanità
- 5 . donne rifiuti violenza
- 9 . acqua abruzzo aquila
- 18 . sicilia immigrati mafia
- 4 . interrogazione torino ue
- 10 . pace alemanno obama
- 3 . rai calabria vigilanza
- 17 . zona amministrazionecomunale edilizia
- 1 . nucleare lazio centrali
- 15 . energia impianti impianto
- 7 . veneto federalismo bossi
- 13 . lavoratori occupazione dipendenti
- 2 . tremonti brunetta pubblicaamministrazione
- 8 . scuola scuole università
- 14 . radicali napolitano sentenza
- 12 . veltroni fini udc

er Ferdinando CASINI  
 WALTER VELTRONI  
 Pier Luigi BERSANI  
 Gianfranco FINI  
 Antonio DI PIETRO  
 Stefano SAGLIA  
 Marco CAPPATO  
 Donatella PORETTI  
 Marco PANNELLA  
 Emma BONINO  
 Giorgio NAPOLITANO  
 CONDO TALAMONTI  
 Cesare DAMIANO  
 Renato BRUNETTA  
 Paolo GIARETTA  
 LUCA ZAIA  
 MARIA GEMMA AZUNI  
 Marco BELTRANDI  
 Paolo FERRERO  
 Silvio BERLUSCONI



**PER CONCLUDERE...**

# CORRELATED TOPIC MODEL

## VANTAGGI

- strumento informativo
- applicabile a testi in più lingue
- possibili confronti su sottoinsiemi del corpus
- suggerisce i documenti di interesse (a partire dalla correlazione)

## SVANTAGGI

- molto costoso computazionalmente
- come scegliere  $k$ ?
- restrittivo sugli argomenti