Università degli Studi di Padova

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di Laurea Triennale in
Scienze Psicologiche dello sviluppo, della personalità e delle relazioni interpersonali

Elaborato Finale

Modelli Latent Dirichlet Allocation ed applicazioni in psicologia

Relatore: Prof. Antonio Calcagnì
Dipartimento di Scienze Psicologiche dello Sviluppo e della Socializzazione

Laureando: Niccolò Cao
Matricola N. 116925

Anno Accademico 2019/2020

Indice

Introduzione **1**

1 Topic Models e LDA **3**

1.1 Terminologia e notazione introduttive 3

1.2 Introduzione ai Topic Models 3

1.2.1 Processo generativo 5

1.3 Latent Dirichlet Allocation 7

1.3.1 Bag-of-Words 7

1.3.2 LDA, il modello 8

1.3.3 La distribuzione Dirichlet 10

1.3.4 Inferenza 11

1.3.5 Metodi di stima dei parametri 13

1.4 Valutazione quantitativa e qualitativa del modello LDA 17

1.4.1 Selezione e valutazione del modello 18

1.4.2 Valutazione del singolo Topic 20

1.4.3 Valutazione qualitativa 21

2 Correlated Topic Models **23**

2.1 CTM, il modello 23

2.2 Stima dei parametri e inferenza 25

3 Biterm Topic Model **27**

3.1 Il modello BTM 27

3.2 Stima dei parametri e inferenza 29

4 Structural Topic Model **31**
4.1 Il modello STM 32
4.1.1 Elementi di notazione 32
4.1.2 Il modello 32
4.2 Stima dei parametri e inferenza 34

5 Applicazione dei Topic Models **37**
5.1 I dati 37
5.2 Implementazione Topic Models 38
5.2.1 Preprocessamento dei dati 38
5.2.2 Implementazione LDA 40
5.2.3 CTM 48
5.2.4 BTM 51
5.2.5 STM 56

6 Conclusioni **61**

Bibliografia **63**

A Codice R utilizzato per i grafici **67**

Introduzione

I *topic models* consistono in un insieme di metodi statistici applicati a dati testuali, che permettono di esplorarne il contenuto. All'interno di un *corpus* di testi questi strumenti statistici individuano degli specifici gruppi di parole, che tendono a formare degli "argomenti", i *topics*. Possiamo pensare a un topic come una "costellazione di parole" che tendenzialmente condividono una stessa categoria semantica; ovvero, un insieme di parole che appartengono allo stesso argomento. Un esempio di topic, che ha come argomento i colori, potrebbe essere il seguente: "rosso", "blu", "verde" e "giallo".
I topic models contribuiscono alla ricerca, oltre che negli ambiti in cui ebbero i natali, quello informatico (*information retrieval*) e quello dell'apprendimento automatico (*machine learning*), anche in discipline quali: linguistica [Hall, Jurafsky e Manning 2008], scienze politiche [Grimmer 2010], sociologia [McFarland et al. 2013], geografia [Yin et al. 2011], psicologia [Gaut et al. 2015], medicina [Wu et al. 2012], biologia [Wang et al. 2011] e molte altre. Come mostra il precedente elenco, i topic models sono strumenti applicati ad una molteplicità di ambiti, anche molto distanti tra loro, e non solo a dati testuali. Ciò che rende queste tecniche versatili fino a tal punto è il fatto che, modificando gli assunti statistici del topic model "più semplice" [Blei 2012], ovvero la *Latent Dirichlet Allocation* (LDA), si possono ottenere modelli costruiti appositamente per gli scopi della propria ricerca. Nel corso di questa tesi ci occuperemo di presentare quattro topic models e di osservare il loro funzionamento nell'ambito della ricerca in psicologia. I dati sui cui verranno applicati i modelli derivano dai trascritti di due gruppi di parola di bambini e adolescenti incentrati sul tema dell'affido. In particolare il presente documento è articolato come segue: il Capitolo 1 è dedicato a una breve in-

troduzione ai topic model e alla presentazione del modello LDA; nel Capitolo 2 è affrontata una breve trattazione del *Correlated Topic Model*; il Capitolo 3 è incentrato sul *Biterm Topic Model*; con il Capitolo 4 si conclude la parte teorica riguardante i modelli, presentando lo *Structural Topic Model*; nel Capitolo 5, infine, vengono applicati i topic models trattati precedentemente ad un insieme di dati provenienti da un caso studio reale.

Capitolo 1

Topic Models e LDA

1.1 Terminologia e notazione introduttive

Al fine di illustrare il funzionamento dei *topic models* risulta necessario introdurre i seguenti elementi di terminologia e notazione riguardanti i concetti di parola, documento e corpus. Una parola distinta w_v, una *word type*, è un oggetto indicizzato appartenente a un vocabolario $V = \{1, \ldots, V\}$. La v-esima parola può essere rappresentata come un vettore V-dimensionale tale che $w_v = 1$ e $w_u = 0$ per $u \neq v$. Mentre, in un documento, l'occorrenza di una parola, o *word token*, verrà indicata con w_i. Quindi definiamo un documento come una sequenza di N occorrenze di parole $d_i = \mathbf{w}_d = (w_1, w_2, \ldots, w_i, \ldots, w_N)$, ovvero un vettore di N *word tokens*. Un corpus, infine, consiste in un insieme di D documenti, $C = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D\} = \{d_1, d_2, \ldots, d_D\}$ e, solitamente, il vocabolario V si costruisce sull'insieme di parole distinte del corpus.

1.2 Introduzione ai Topic Models

Con *topic models*, o *probabilistic topic models*, si indica una classe ampia e diversificata di tecniche statistiche per l'analisi di dati testuali sviluppate nell'ambito dell'*information retrivial* e del *text mining*. Queste tecniche permettono di estrarre quei temi che caratterizzano maggiormente un dato

corpus di testi, i *topics*. I topic models, per la selezione dei topics, non necessitano di codifiche o classificazioni dei dati definite *a priori* da un intervento umano. Si tratta, quindi, di tecniche di apprendimento non-supervisionato.
L'assunto alla base dei *probabilistic topic models* è che ciascun documento sia una mistura di temi, i topics, e che questi temi siano comuni a tutti i documenti di un corpus, in proporzioni diverse a seconda del documento considerato. Pertanto i topics possono essere immaginati come degli spazi semantici che attraversano il corpus e ne descrivono gli aspetti più salienti. Qualcosa di molto simile a quel che comunemente consideriamo un "argomento".
Ad esempio, se stiamo per leggere il numero di una rivista scientifica di psicologia, potremmo aspettarci di trovare i seguenti argomenti: neuroscienze, psicopatologia, psicologia dello sviluppo e psicologia dinamica. A ciascun argomento, naturalmente, sono associate delle parole più probabili: la parola "neurone" è più probabile che compaia quando si parla di neuroscienze piuttosto che di psicopatologia. Quindi quel corpus di documenti che è una rivista potremmo immaginarcelo come una mescolanza di tutti questi temi. Gli argomenti nei singoli articoli possono rimanere separati tra loro caratterizzando ciascuno un articolo diverso, ma possono anche combinarsi tutti e quattro in misure diverse all'interno di ciascun articolo. L'obiettivo dei topic models è proprio catturare questa variabilità semantica dei documenti in ciascun topic.
Più concretamente, i topics rappresentano dei cluster di parole, che tendono a co-occorrere all'interno dei documenti del corpus, identificati tramite un modello probabilistico. Un topic è, infatti, una distribuzione di probabilità su tutte le parole del vocabolario di un corpus, in cui le parole più probabili sono quelle che ne descrivono meglio il contenuto. possiamo osservare nella figura 1.1, a sinistra, una colonna in cui sono rappresentati quattro topics; le parole sono presentate in ordine decrescente rispetto alla loro probabilità all'interno del topic, cioè in base a quanto sono "rappresentative" del topic.
A livello statistico, però, è necessario fare delle assunzioni per utilizzare una tecnica che ci informi su delle quantità latenti, che in questo caso sono le proporzioni di topics nei documenti e le parole più probabili per topic. L'assunto principale è che tutti i documenti di un corpus provengono da uno specifico

insieme di topics e, avendo questi topics generato a livello teorico il corpus, possiamo risalire ad essi tramite una procedura inferenziale. Possiamo, quindi, definire un *topic model* come un modello probabilistico che spiega la procedura attraverso cui si assume che i documenti siano stati generati stocasticamente da una serie di topics "originari". Tale procedura prende il nome di *processo generativo*. Da questa definizione possiamo comprendere anche da dove derivano le differenze tra topic models. Essi si differenziano per gli assunti statistici che ne determinano l'aspetto modellistico, ovvero per il modo secondo cui sono stati "generati" i dati testuali.

1.2.1 Processo generativo

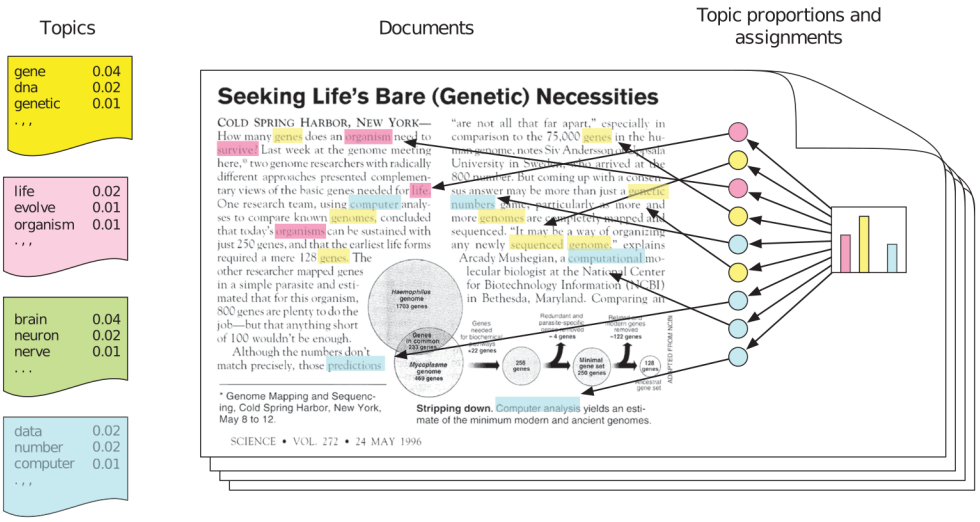


Figura 1.1: Esemplificazione grafica del processo generativo di un *topic model*. Fonte: Blei (2012)

Per comprendere il funzionamento e l'idea alla base dei topic models può essere utile riferirsi alla figura 1.1 per illustrare il processo generativo del *probabilistic topic model* più semplice e più usato, la *Latent Dirichlet Allocation* (LDA), che approfondiremo dettagliatamente in seguito, servendoci in parte della notazione esemplificativa adottata da Steyvers e Griffiths (2007). I topic models, come già accennato, ipotizzano che un dato corpus di documenti sia il risultato di uno specifico processo generativo, secondo cui, controintuitiva-

mente, si assume che siano una serie di argomenti a produrre un testo; come se un autore scegliesse prima i temi di cui vuole parlare e poi li combinasse tra loro in misure diverse per la creazione di un insieme di documenti.
Innanzitutto indicheremo un topic generico con z_j, in cui l'indice j si riferisce al j-esimo topic, e indicheremo ciascuna parola (*word token*) di un dato vocabolario V con w_i, in cui l'indice i è riferito all'i-esima parola di un documento. Nell'ottica generativa, per ottenere un singolo documento come quello presentato nell'immagine, è necessario compiere tre passaggi [Blei 2012], iterabili per ogni d-esimo documento del corpus:

1. assegnare casualmente una probabilità a ciascun topic $\Pr(z_j)$, costruendo la distribuzione di probabilità sui topics, che vale per un solo documento.
2. ottenere le parole dei documenti, per cui bisogna:
   (a) selezionare casualmente un topic z_j dalla distribuzione sui topics $\Pr(z)$
   (b) poi, condizionatamente al topic z_j scelto, selezionare una parola w_i sulla base della probabilità che una parola i sia estratta dato il topic j scelto: $\Pr(w_i \mid z_j = j)$
3. Il secondo passo è ripetuto tante volte quanto è il totale delle occorrenze delle parole N nel documento.

Quindi la distribuzione di probabilità di una parola generica w_i in un documento è dato da:

$$\Pr(w_i) = \sum_{j=1}^{K} \Pr(w_i \mid z_j = j) \Pr(z_j = j) \tag{1.1}$$

in cui definiamo K come il numero dei topics e $\Pr(z_j = j)$ come la probabilità che il topic j sia selezionato per la parola i scelta. Tale formula mostra che la probabilità di una parola in un documento, $\Pr(w_i)$, è data dalla somma, per tutti i K topics, del prodotto tra la probabilità che un topic j compaia nel testo per la parola i e la probabilità di trovare la parola w_i condizionatamente al topic j estratto.
Il primo passo descritto è rappresentato nell'istogramma a destra della figura

1.1, ogni barretta colorata rappresenta la probabilità scelta per ciascun topic di comparire nel singolo documento, $\Pr(z)$. Il secondo passo è raffigurato nei cerchietti colorati: per ciascuna parola viene estratto un cerchietto colorato (un topic) e sulla base della probabilità che questo assegna a ciascuna parola i, ovvero $\Pr(w_i \mid z_j = i)$, come mostrano le seconde frecce nere da sinistra verso destra, si estrae la i-esima parola.
Ovviamente i topic models non servono per generare nuovi testi, ma per determinare, tramite un procedimento inferenziale, quali topics hanno generato più verosimilmente un certo corpus di testi di nostro interesse; il che significa dover invertire la direzione del processo appena descritto, ovvero "iniziare dalla fine": comporre un corpus e, poi, ottenere ciò che teoricamente si trova all'origine del testo, i topics.

1.3 Latent Dirichlet Allocation

La *Latent Dirichlet Allocation* è un modello gerarchico bayesiano presentato per la prima volta da Blei, Ng e Jordan [Blei, Ng e Jordan 2003] che prende le mosse dai lavori riguardanti il *Latent Semantic Indexing* (LSI) [Deerwester et al. 1990] e il *probabilistic Latent Semantic Indexing* (pLSI) [Hofmann 1999].

1.3.1 Bag-of-Words

Il modello LDA, per essere applicato ad un corpus di documenti, assume che questo abbia una determinata rappresentazione statistica. Il corpus è, infatti, rappresentato secondo il *Bag-of-Words assumption*, in linea con l'assunto della scambiabilità di parole e documenti. Il modello LDA, quindi, considera un documento come un vettore di occorrenze di parole che ad esso appartengono, ovvero come una *bag of words*, in cui l'ordine delle parole viene ignorato, così come l'informazione grammaticale. Secondo questo assunto, una collezione di documenti può essere rappresentata da una *Document-Term Matrix* (DTM), laddove ciascuna riga della matrice corrisponde a uno specifico documento indicizzato e ciascuna colonna a una *word type* presente nel vocabolario del corpus, come si vede nell'esempio della tabella 1.1. Nella

DTM i totali di riga restituiscono il numero di occorrenze di ciascuna parola del vocabolario dato un documento; mentre i totali di colonna mostrano il numero totale di occorrenze di ciascuna parola all'interno dell'intero corpus. Questo metodo di rappresentazione dei dati testuali risulta molto diffuso nel campo dell'analisi automatica dei testi e del text mining.

Documenti x Parole	cinema	oggi	vado	al	...
Documento_1	1	0	3	0	...
Documento_2	5	2	1	2	...
Documento_3	0	1	4	3	...

Tabella 1.1: Esempio di Document-Term Matrix

1.3.2 LDA, il modello

La LDA, come tutti i *probabilistic topic models*, è un modello generatore di un corpus di documenti, che a partire da tre distribuzioni di probabilità estrae le parole che creeranno ciascun documento. Le tre distribuzioni sono: una Poisson con parametro λ, che governa la numerosità delle parole nei documenti (N_d); una distribuzione Multinomiale θ_d per ogni documento, che rappresenta le distribuzioni di probabilità dei K topics sui D documenti, ovvero il vettore di probabilità che indica in quale misura un certo topic k-esimo sarà presente nel documento d-esimo; e, infine, β_k indica la distribuzione di probabilità di un topic sulle V parole del vocabolario, anch'essa prende la forma di una Multinomiale, una per ogni topic. Le distribuzioni di probabilità θ_d e β_k derivano dall'estrazione di due distinti vettori di probabilità da una distribuzione Dirichlet: il primo proviene da una Dirichlet simmetrica K-dimensionale con il vettore di parametri α_k con $\alpha_k > 0$ e il secondo da una Dirichlet simmetrica V-dimensionale con parametro η_v con $\eta_v > 0$.
Il processo generativo del modello LDA per un documento d è il seguente:

1. Per ciascun topic k-esimo per K fissato e noto,
   (a) si estrae $\beta_k \sim Dirichlet_V(\eta)$
2. Per ciascun documento d-esimo,

   (a) si estrae $N_d \sim Poisson(\lambda)$
   (b) si estrae $\theta_d \sim Dirichlet_K(\alpha)$
3. Per ciascuna parola n-esima in ciascun documento d-esimo,
   (a) si estrae $z_{k,d,n} \sim Multinomiale(\theta_d)$, in cui $z_{k,d,n} \in \{1, \ldots, K\}$, l'assegnazione di una parola a un topic
   (b) si estrae $w_{d,n} \sim Multinomiale(\beta_k \mid z_{k,d,n})$, in cui $w_{d,n} \in \{1, \ldots, V\}$.

Nella figura 1.2 possiamo osservare il modello grafico di questo processo. Con $z_{k,d,n}$ si identifica una variabile-indicatore latente che mostra l'assegnazione del topic k a una parola $w_{d,n}$ sulla base delle probabilità di tutti i topics in un dato documento d. Condizionatamente all'assegnazione del topic $z_{k,d,n}$, si genera la parola $w_{d,n}$ tramite la distribuzione delle parole del topic β_k.
Il processo generativo del modello LDA è descritto dalla seguente distribuzione congiunta di tutte le variabili latenti e osservate dati gli iperparametri α e η:

$$\begin{aligned}\Pr(\beta, \theta, \mathbf{z}, \mathbf{w} \mid \alpha, \eta) = &\prod_{k=1}^{K} \Pr(\beta_k \mid \eta_k) \prod_{d=1}^{D} \Pr(\theta_d \mid \alpha_k) \times \\ &\times \left(\prod_{n=1}^{N_d} \Pr(z_{d,n} \mid \theta_d) \Pr(w_{d,n} \mid z_{d,n}, \beta) \right)\end{aligned} \tag{1.2}$$

Dati gli iperparametri, la probabilità di un corpus C di documenti, si ottiene dalla distribuzione marginale dei singoli documenti:

$$\begin{aligned}\Pr(C \mid \alpha, \eta) = &\int \prod_{k=1}^{K} \Pr(\beta_k \mid \eta) \prod_{d=1}^{D} \int \Pr(\theta_d \mid \alpha) \times \\ &\times \left(\prod_{n=1}^{N_d} \sum_{n=z_n}^{z_N} \Pr(z_{d,n} \mid \theta_d) \Pr(w_{d,n} \mid z_{d,n}, \beta) \right) d\theta d\beta\end{aligned} \tag{1.3}$$

Con questa formula si chiarisce anche formalmente la presenza dei tre livelli descritti in figura 1.2: topics (K), documenti (D) e parole (N). I parametri α e η sono estratti una volta per ogni C, le variabili latenti θ_d sono estratte una volta per ogni documento e le variabili $\mathbf{z_{d,n}}$ e $\mathbf{w_{d,n}}$ sono estratte una volta per ogni parola.

Bisogna, infine, puntualizzare che le distribuzioni di θ_d e β_k sono due Multinomiali con una sola prova e che tale distribuzione è nota anche come distribuzione Categoriale.

1.3.3 La distribuzione Dirichlet

In questo contesto la scelta di una distribuzione Dirichlet come *a priori* è molto conveniente, in quanto rappresenta una distribuzione congiunta *a priori* (*conjugate prior*) della Multinomiale, come della Categoriale. In questo modo durante la fase di stima dei parametri e di inferenza si evitano le difficoltà dovute all'eventuale incompatibilità tra distribuzione di probabilità *a priori* e distribuzione *a posteriori*.
Data una distribuzione Dirichlet, questa ha un supporto su un generico vettore $\mathbf{x}$, J-dimensionale, con $x_i \in \{x_1, \ldots, x_J\}$ tale che $x_i \in (0, 1)$ e $\sum_{i=1}^{J} x_i = 1$. Ciò permette di interpretare qualunque realizzazione di una Dirichlet come un vettore di probabilità associabile a un insieme J-dimensionale di eventi discreti, che possono essere distribuiti secondo una Multinomiale o una Categoriale. Una distribuzione Dirichlet *a priori* K-dimensionale su una distribuzione Multinomiale, $p = (p_1, \ldots, p_K)$, avrà una densità di probabilità data da [Steyvers e Griffiths 2007]:

$$f(p_1, \ldots, p_K \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{K} p_i^{\alpha_i - 1} \tag{1.4}$$

in cui Γ rappresenta la funzione Gamma.
Nel caso della LDA viene solitamente utilizzata una distribuzione Dirichlet *simmetrica*, in cui il vettore di parametri α è uguale per tutti i K elementi $\alpha = (\alpha_1, \ldots, \alpha_K) = \alpha_i$, in modo da modellare l'assenza di una conoscenza *a priori* sulla probabilità dei topics. Allo stesso modo anche η consiste in un vettore di parametri equivalenti.
Inoltre α, siccome rappresenta un vettore di "parametri di concentrazione" sulle Multinomiali θ, controlla la forma delle singole probabilità dei topics su tutti i documenti. Solitamente α assume valori molto vicini allo zero, $\alpha < 1$, in modo che le probabilità tendano a concentrarsi su un numero ristretto di topics, ovvero a concentrare alte densità di probabilità su pochi topics all'in-

terno dei documenti. Ciò produce la "sparsità" di θ, ovvero la tendenza del modello a scegliere per ciascun documento pochi topics più probabili.
Anche η è un vettore di parametri di concentrazione e concentra le densità di probabilità delle parole di ciascun *topic*. La sparsità di ogni β_k, controllata appunto da η, riguarda il grado di similarità che le parole devono avere (il grado di co-occorrenza all'interno del corpus) per essere assegnate allo stesso topic. Per valori alti di η si ottengono topics caratterizzati da più parole con alte probabilità, suggerendo la scelta di un numero basso di K topics [Griffiths e Steyvers 2004]. Invece, per valori più bassi è preferibile la scelta di un maggior numero di topics, poiché, in questo modo, si ottengono meno parole con alte probabilità.
Possiamo quindi considerare α come un indicatore della similarità dei contenuti latenti tra i documenti, mentre η come un'indicatore del minimo grado di co-occorrenza tra parole nel corpus per inferire un topic [Heinrich 2009].
Spesso i ricercatori tendono a favorire la sparsità per entrambi gli iperparametri. Ciò tende a creare dei topics più distinti tra loro e a individuare, al loro interno, solo le parole più numericamente rappresentative.

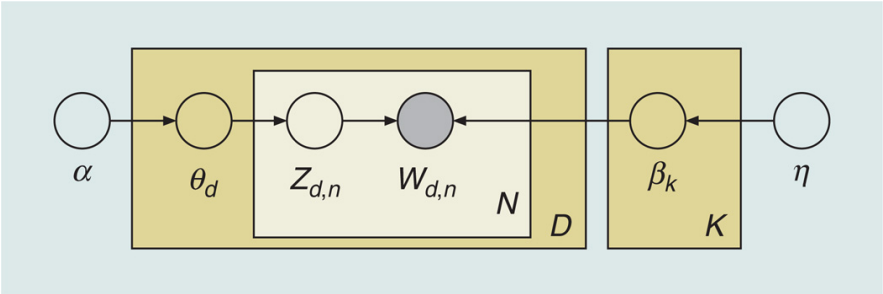


Figura 1.2: Modello grafico LDA. I rettangoli rappresentano la ripetizione (N, D o K) di ciascuna variabile in essi inclusa; l'unica variabile osservata è $w_{d,n}$, colorata in grigio, mentre le restanti sono variabili latenti [Blei, Carin e Dunson 2010].

1.3.4 Inferenza

Finora abbiamo concepito il modello LDA solo come modello generativo, e, come osservato precedentemente, ciò permette di costruire la distribuzione congiunta delle variabili osservate e delle variabili latenti, ovvero di costruire

lo spazio parametrico del modello supposto alla base di ciascun documento di un corpus. Ciononostante il modello LDA è non stato progettato per "generare" nuovi corpora, ma al fine di stimare per un dato corpus: i topics, le proporzioni dei topics nei documenti, le assegnazioni dei topics sulle parole e le parole più probabili dato ciascun topic. Conviene, pertanto, esplicitare formalmente la distribuzione *a posteriori* delle quantità latenti che ci interessa stimare caratterizzanti un generico corpus C, fissati gli iperparametri α e η:

$$\Pr(\beta, \theta, \mathbf{z} \mid \eta, \alpha, C) = \frac{\Pr(\beta, \theta, \mathbf{z}, C \mid \alpha, \eta)}{\Pr(C \mid \alpha, \eta)}. \tag{1.5}$$

In questo modo tutte le variabili latenti sono condizionate ai dati testuali osservati C e agli iperparametri che assumiamo già noti, in modo tale da far emergere la struttura probabilistico-tematica propria del corpus. Mentre il numeratore di 1.5 può essere calcolato senza difficoltà, il denominatore, che rappresenta la probabilità marginale del corpus osservato, risulta intrattabile dal punto di vista computazionale e, per questo motivo, deve essere approssimato. Il problema computazionale, più nello specifico, è dato dal dover assegnare, in tutti i modi possibili, tutte le parole osservate di ogni documento a ciascun topic [Blei, Carin e Dunson 2010], che possiamo meglio esplicitare nella distribuzione di probabilità marginale del corpus C nei termini dei parametri del modello:

$$\begin{aligned}\Pr(C \mid \alpha, \eta) = \int \int \sum_{i=1}^{Z} \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k + n_{d,k,\cdot} - 1} \right) \times \\ \left(\prod_{k=1}^{K} \frac{\Gamma(\sum_{v=1}^{V} \eta_{k,v})}{\prod_{v=1}^{V} \Gamma(\eta_{k,v})} \prod_{v=1}^{V} \beta_{k,v}^{\eta_{k,v} + n_{\cdot,k,v} - 1} d \right) d\theta d\beta\end{aligned} \tag{1.6}$$

La prima parentesi contiene le proporzioni dei topics per ciascun documento θ_d e l'assegnazione di un topic $\mathbf{z}_k$ a ogni parola in un documento, il cui conteggio è dato da $n_{d,k,\cdot}$; nella seconda parentesi osserviamo la distribuzione delle parole all'interno dei topics e la probabilità del corpus dati i topics β_k e le assegnazioni $\mathbf{z}$ per ciascuna parola distinta dell'intero corpus (*word type*), conteggiate da $n_{\cdot,k,v}$ [Ponweiser 2012]. Possiamo, quindi, individuare nella grandezza dei conteggi $n_{d,k,\cdot}$ e $n_{\cdot,k,v}$ il maggior ostacolo per il calcolo esatto

della somma su tutte le possibili assegnazioni $\mathbf{z}$ a tutte le parole. Di qui la necessità di dover approssimare anche $\Pr(\beta, \theta, \mathbf{z} \mid \eta, \alpha, C)$, in quanto dipende da $\Pr(C \mid \alpha, \eta)$. Pertanto l'inferenza sulle quantità latenti associate ai topics avviene tramite la scelta di un metodo di stima dei parametri, che consenta di calcolare le quantità latenti, in modo che i valori scelti rappresentino le stime di *massima verosimiglianza* (ML) per ciascun parametro del modello LDA selezionato.

1.3.5 Metodi di stima dei parametri

Per risolvere il problema inferenziale della corretta stima dei parametri latenti sono stati proposti diversi metodi di approssimazione della distribuzione *posterior* (Eq. 1.5), riconducibili, nella maggioranza dei casi, a due categorie: metodi variazionali bayesiani (*Variational Bayes methods*) e metodi di simulazione basati su algoritmi *Markov-chain-Monte-Carlo* (MCMC).

Variational Expectation-Maximization

Tra i Variational Bayesian methods ci concentreremo brevemente sull'algoritmo *Variational Expectation-Maximization* (VEM), che è il metodo di stima proposto nel primo articolo sul modello LDA [Blei, Ng e Jordan 2003]. L'obiettivo del VEM è trovare le stime di massima verosimiglianza dei parametri del modello tramite i valori di α e η, che massimizzino la log-verosimiglianza del corpus:

$$\ell(\alpha, \eta) = \sum_{d=1}^{D} \log p(C \mid \alpha, \eta) \tag{1.7}$$

Per la massimizzazione di questa quantità rispetto ai parametri da stimare, il VEM utilizza una distribuzione più semplice e calcolabile della *a posteriori* vera, che è detta *variazionale*, all'interno della quale le variabili β, θ e $\mathbf{z}$ sono indipendenti tra loro:

$$q(\beta, \theta, \mathbf{z} \mid \lambda, \gamma, \phi) = \prod_{k=1}^{K} q(\beta_k \mid \lambda_k) \prod_{d=1}^{D} \left(q(\theta_d \mid \gamma_d) \prod_{n=1}^{N} q(\mathbf{z}_{d,n} \mid \phi_{d,n}) \right) \tag{1.8}$$

in cui λ (distribuzione Dirichlet V-dimensionale), γ (distribuzione Dirichlet K-dimensionale) e ϕ (distribuzione Multinomiale K-dimensionale) sono i *pa-*

rametri variazionali che governano la distribuzione variazionale. Al fine di ottenere un'approssimazione il più vicina possibile alla distribuzione a posteriori vera $p(\beta, \theta, \mathbf{z} \mid \eta, \alpha, C)$, si cercano quei parametri variazionali che minimizzano la divergenza di Kullback-Leiber (KL) tra le due distribuzioni:

$$\min_{(\gamma,\phi)} KL(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \eta)) \tag{1.9}$$

I parametri sono, infatti, ottimizzati iterativamente alternando una fase di Expectation (o E-step), in cui vengono trovati i valori ottimali dei parametri γ e ϕ e massimizzato il limite inferiore della log-verosimiglianza dei dati (1.7), e una fase di Maximization (o M-step), in cui viene massimizzato il limite inferiore della verosimiglianza trovato nel passo precedente rispetto ai parametri α e η [Grün e Hornik 2011]. Quindi il procedimento *variational EM*, quando converge, fornisce le stime di massima verosimiglianza di β, θ e $\mathbf{z}$.
Il metodo VEM si configura quindi come un algoritmo deterministico, che trasforma il problema inferenziale della distribuzione *a posteriori* vera in un problema di ottimizzazione della distribuzione variazionale [Blei 2012].

Gibbs sampler

Il *Gibbs sampler* (o *Campionamento di Gibbs*) è un algoritmo MCMC (*Markov chain Monte Carlo*) introdotto da Geman e Geman [Geman e Geman 1984]. Il Gibbs sampler è stato ampiamente utilizzato nel contesto dei modelli LDA in quanto offre una semplice e valida soluzione per problemi computazionali derivanti da distribuzioni multivariate di grandi dimensioni. Questo algoritmo consente, infatti, di generare la distribuzione stazionaria di una *Markov chain* in modo che approssimi una distribuzione di probabilità di nostro interesse, ovvero di simularla. L'aspetto essenziale di questo metodo di stima è dato dal fatto che la distribuzione stazionaria viene costruita campionando sequenzialmente nuovi valori da una serie di distribuzioni univariate condizionate (*full conditional distribution*), in cui ciascun valore dipende dal precedente. Ciò significa che ogni valore estratto deriva da una distribuzione in cui tutte le variabili aleatorie presenti sono fissate eccetto una, rendendo il campionamento più semplice. Si ripete il processo di campionamento

sequenziale finché non si ottiene un'approssimazione alla distribuzione vera adeguata.
Per introdurre più formalmente il Gibbs sampler possiamo ricorrere al caso più semplice [Ghosh, Delampady e Samanta 2007]: poniamo che X e Y siano due variabili aleatorie, di cui ci interessa calcolare la distribuzione bivariata $\Pr(X, Y)$. Costruiamo, quindi, le distribuzioni condizionate: per ciascuna realizzazione y, sia $\Pr(X \mid y)$ la distribuzione condizionata univariata di X dato $Y = y$ e per ogni x, sia $\Pr(Y \mid x)$ la distribuzione univariata di Y dato $X = x$. Partendo da un valore casuale $X_0 = x_0$ si costruisce la distribuzione $\Pr(Y_0 \mid x_0)$, da cui si estrae un'osservazione $Y_0 = y_0$. Poi da $\Pr(X_1 \mid y_0)$ si genera un'osservazione $X_1 = x_1$. Il Gibbs sampler segue la seguente procedura:

$$x_i^{(t_{n+1})} \sim \Pr(X_i \mid Y_{i-1} = y_{i-1})$$
$$y_i^{(t_{n+1})} \sim \Pr(Y_i \mid X_i = x_i)$$

In questo modo si crea una catena di Markov bivariata $\mathbf{Z}_n = (\mathbf{X}_n, \mathbf{Y}_n)$ per n ripetizioni del processo, ovvero n vettori aleatori bivariati simulati. Per n sufficientemente grande possiamo considerare $\mathbf{Z}_n$ come un campione estratto da una distribuzione che approssima $\Pr(\mathbf{X}, \mathbf{Y})$.
Il modello LDA necessita però di fare inferenza su una distribuzione multivariata. Nel caso multivariato il Gibbs sampler funziona come descritto nel seguente esempio: sia $\Pr(X_1, X_2, \ldots, X_k)$ la distribuzione di probabilità del vettore aleatorio k-dimensionale $\mathbf{X}$. Sia $\mathbf{X}_{-i} = (X_1, X_2, X_{i-1}, X_{i+1}, \ldots, X_k) = \mathbf{x}_{-i}$ il vettore precedente a cui è stata rimossa una variabile i-esima con $\mathbf{x}_{k-1}$ realizzazioni. La distribuzione condizionata univariata di X_i è data da $\Pr(X_i \mid \mathbf{X}_{-i} = \mathbf{x}_{-i})$, ovvero la distribuzione full conditional. Quindi si inizia da un primo vettore casuale $\mathbf{X}_0 = (x_{0,1}, x_{0,2}, \ldots, x_{0,k})$, a partire dal quale si inizia con il generare il vettore $\mathbf{X}_1 = (X_{1,1}, X_{1,2}, \ldots, X_{1,k})$, nel seguente modo:

$$x_{1,1}^{(t_1)} \sim \Pr(X_{1,1} \mid \mathbf{x}_{0_{-1}})$$
$$x_{1,2}^{(t_1)} \sim \Pr(X_{1,2} \mid X_{1,1}, x_{0,3}, x_{0,4}, \ldots, x_{0,k})$$

fino a produrre, prima il vettore $\mathbf{X}_1$, poi $\mathbf{X}_k = (X_{k,1}, X_{k,2}, \ldots, X_{k,k})$, condizionatamente a tutte le variabili stimate. A questo punto si itera lo stesso procedimento finché non si raggiunge la distribuzione stazionaria di $\Pr(\mathbf{X})$, aggiornando i valori delle realizzazioni di ciascuna variabile del vettore $\mathbf{X}$ per t volte.
Nel modello LDA le quantità da stimare sono θ_d, la proporzione dei topics nei documenti, e β_k, la distribuzione delle parole sui topics. Ciononstante non conviene stimare direttamente θ_d e β_k [Griffiths e Steyvers 2004], ma inferire le due quantità dalla stima della distribuzione a posteriori di $\Pr(\mathbf{z} \mid \mathbf{w})$, marginalizzando rispetto ad esse. Il Gibbs sampler viene, quindi, applicato solo alla probabilità che il topic k sia assegnato a una parola w_i $(z_{i,k})$, fissate tutte le altre assegnazioni dei topics alle restanti parole $\mathbf{z}_{-i,k}$ [Carpenter 2010]. Dato che integriamo i parametri multinomiali e campioniamo solo le assegnazioni, utilizziamo la variante del Gibbs sampler denominata: "Collapsed Gibbs sampling" [Darling 2011]. La probabilità full conditional dell'assegnazione di una parola a un topic è data da [Griffiths e Steyvers 2004]:

$$p(z_{i,k} \mid \mathbf{z}_{-i,k}, \mathbf{w}, \alpha, \eta) \propto \frac{n_{-i,k}^{V_v} + \eta}{\sum_{v=1}^{V} n_{-i,k}^{V} + V\eta} \frac{n_{-i,k}^{d_m} + \alpha}{\sum_{k=1}^{K} n_{d,k}^{d_m} + K\alpha} \tag{1.10}$$

in cui la prima frazione indica la probabilità della parola w_i dato il topic k, mentre la seconda mostra la probabilità del topic k nel documento d_m. Più dettagliatamente $n_{-i,k}^{V_v}$ rappresenta il conteggio delle assegnazioni della v-esima *word type* del vocabolario V al topic k-esimo, esclusa la i-esima parola. Invece $n_{-i,k}^{d_m}$ rappresenta il conteggio delle assegnazioni della parola all'interno del documento m-esimo al topic k-esimo, esclusa la parola i-esima. A questo punto l'algoritmo del Gibbs sampler può procedere con la stima [Steyvers e Griffiths 2007]:

1. Per ciascuna iterazione t:
   1.1 Per ogni *word token* w_i:
      (a) si sceglie casualmente un'assegnazione z_{i,k_j} a un topic $\{1, \ldots, K\}$.
   1.2 Per ciascuna assegnazione $z_{i,k}$:

(a) si rimuove l'assegnazione già ottenuta, z_{i,k_j}, dai conteggi $n_{i,k}^{V_v}$ e $n_{i,k}^{d_m}$
(b) si estrae una nuova assegnazione a un topic $z_{i,k_{j+1}}$ con una probabilità proporzionale a 1.10;
(c) si aggiornano le variabili conteggio aggiungendo la nuova assegnazione tematica.

Dopo un numero sufficiente di iterazioni, si rimuovono le iterazioni iniziali (burn-in) in modo da evitare il rumore derivante dalla casualità dei primi passi dell'algoritmo.
Un'altra eventuale problematica può dipendere dalla presenza di correlazioni tra i campioni estratti dalla catena di Markov. Una soluzione, spesso utilizzata, consiste nell'accettare un campione ogni n campioni scartati.
Infine possiamo derivare dalle variabili conteggio le stime di β e θ:

$$\hat{\beta}_k^V = \frac{n_k^V + \eta}{\sum_{v=1}^{V} n_k^V + V\eta}$$
$$\hat{\theta}_k^D = \frac{n_k^D + \alpha}{\sum_{k=1}^{K} n_k^D + K\alpha}$$

1.4 Valutazione quantitativa e qualitativa del modello LDA

Nell'implementazione del modello LDA è raro emergano fin da subito dei topics di buona qualità: che abbiano un'ottima coerenza interna, non presentino "intrusi", siano facilmente interpretabili da un punto di vista semantico e rispecchino adeguatamente il contenuto di uno o più documenti.
Per costruire un topic model che presenti un buon adattamento a un corpus, nella letteratura, si tende coniugare metodi e metriche basati sulla *likelihood* per la stima dei parametri α, η e K a una valutazione *qualitativa* dell'interpretabilità dei topics in sé e rispetto ai documenti. In questo contesto ci concentreremo in particolare sulla stima del miglior valore di K tenendo α e η fissati a valori generalmente considerati ottimali.
Inoltre bisogna sottolineare che è fondamentale scegliere il metodo, o i me-

todi, di valutazione di un topic model in funzione del suo utilizzo, in quanto non è scontato che un metodo di valutazione sia utile e informativo per lo scopo della ricerca. Infatti Chang et al. (2009) mostra come valori più alti di verosimiglianza predittiva, misura quantitativa, siano associati a minor interpretabilità dei topics da parte di un campione di soggetti, misura qualitativa. In questo scritto utilizzeremo l'approccio quantitativo più comune in letteratura, associato a una serie di indici di fit. L'obiettivo è ottenere dei topics capaci di informarci sull'effettivo contenuto dei documenti scelti, non creare un modello dotato di un'ottima capacità predittiva.

1.4.1 Selezione e valutazione del modello

La selezione di un modello LDA verte attorno a due punti critici: la selezione del numero di topics e la scelta dei parametri di concentrazione. Per quanto riguarda i valori α e η (spesso, in questo contesto denominata β), non ci soffermeremo nello specifico per una serie di motivi: abbiamo già spiegato il loro ruolo nel modello LDA; non è un argomento sufficientemente trattato in letteratura e solitamente i valori ritenuti "ottimali" sono $\alpha = 50/K$ e $\eta = 0.01$ [Steyvers e Griffiths 2007]; in generale il metodo migliore per trovare il miglior valore di questi due parametri è modificarli a seconda delle esigenze del modello che si intende creare e dei dati a disposizione [Boyd-Graber, Mimno e Newman 2014]. Invece per la scelta del valore K, come vedremo, si tende ad utilizzare il metodo *held-out* e una serie di metriche specifiche.

Held-out method

Nel contesto dei topic models, una volta implementato il modello, non abbiamo il vantaggio tipico delle tecniche supervisionate, per cui, avendo definito a priori delle categorie, possiamo verificare immediatamente la capacità del modello di imparare a gestire i dati correttamente. In aggiunta non sappiamo, a priori, qual è il numero di topics più adeguato per un insieme di dati che il modello LDA deve individuare. Una soluzione a quest'ultimo problema consiste nel confrontare la qualità di una serie di modelli con K diversi dal punto di vista della capacità di generalizzazione del modello. La tecnica più diffusa è quella dell'*held-out method*, ovvero una validazione [Wallach et

al. 2009;Blei 2012]. Questo metodo prevede una divisione randomizzata del corpus in due o più campioni (*k-folds*). Tra questi insiemi di dati uno è il *test set*, ovvero il campione su cui sarà testato il modello, che chiameremo corpus-test; mentre gli altri sono quelli su cui il modello sarà stato adattato precedentemente (*training set*), il corpus-training.
Generalmente il criterio per individuare il modello più performante consiste nella misurazione della *predictive-perplexity* di tutti i modelli implementati. La predictive-perplexity consiste nella verosimiglianza di un insieme di documenti dato il modello da valutare, ovvero la capacità del modello $\mathcal{M}$ di predire dati diversi da quelli del *training*, il corpus-test (C_{test}):

$$perplexity(C_{test} \mid \mathcal{M}) = \exp\left\{ - \frac{\sum_{d=1}^{D} \log p(\mathbf{w}_d) \mid \mathcal{M}}{\sum_{d=1}^{D} N_d} \right\} \tag{1.11}$$

La perplexity è definita come il reciproco della media geometrica della verosimiglianza di una *word token* nel C_{test}. Più è basso il valore della perplexity più il modello tende a non rimanere "perpelsso" rispetto ai dati del test, ma ad aspettarseli. La scelta di questo metodo è suggerita anche da Blei e Lafferty (2009), secondo cui è utile per fini qualitativi, come l'esplorazione di un corpus. Bisogna però considerare, come accennato precedentemente, che una buona capacità di generalizzazione del modello, che tende ad aumentarne la complessità, può facilmente entrare in conflitto con la produzione di topics comprensibili, proprio a causa del fatto che la complessità dei topics generati può superare la capacità umana di interpretarli.

Indici di fit

Il pacchetto *ldatuning* [Nikita 2016] implementato in R [R Core Team 2020] fornisce quattro strumenti per individuare il miglior valore di K:

1. **Griffith2004**: il metodo, tipicamente bayesiano, si basa sul computo, tramite Gibbs sampling, della probabilità a posteriori di un insieme di modelli rispetto a tutte le parole del corpus, ovvero della verosimiglianza $\Pr(\mathbf{w} \mid K)$. Tale verosimiglianza viene approssimata determinando la media armonica di $\Pr(\mathbf{w} \mid \mathbf{z}, K)$ per mezzo di $\mathbf{z}$ campionato da

$\Pr(\mathbf{z} \mid \mathbf{w}, K)$. Più è alto il valore, maggiore è la verosimiglianza del numero di topics [Griffiths e Steyvers 2004].

2. **CaoJuan2009**: la tecnica seleziona il modello migliore sulla base della densità dei clusters (topics). Si calcola per ciascun modello la distanza tra i topics tramite la similarità media del coseno ($\mathbf{r}$), che ci permette di definire la densità di un topic rispetto agli altri. Il valore migliore di K è raggiunto quando $\mathbf{r}$ tra topics è al minimo, ovvero quando la distanza è massima [Cao et al. 2009].
3. **Arun2010**: questa misura calcola la divergenza Kullback-Leibler simmetrica tra le distribuzioni dei valori singolari della matrice Topic-Word e le distribuzioni dei valori singolari della matrice Document-Topic di un topic. Il miglior numero di topics è identificato quando il valore della divergenza tra queste due misure in ciascun topic è al minimo [Arun et al. 2010].
4. **Deveaud2014**: la metrica proposta sfrutta la divergenza Jensen-Shannon per determinare quel valore di K che massimizza la distanza tra ciascuna coppia di topics [Deveaud, SanJuan e Bellot 2014].

1.4.2 Valutazione del singolo Topic

In Boyd-Graber, Mimno e Newman (2014) e in Mimno et al. (2011) sono state elaborate delle metriche basate sulle statistiche delle singole parole presenti in un topic per valutarne la qualità:

- **Topic size**: un numero basso di assegnazioni di *word token* a un topic, via Gibbs sampler, è spesso indice di bassa qualità del topic.
- **mean-token-length**: la lunghezza media delle parole più probabili di un topic può indicare la tendenza di un topic ad essere più specifico, per valori maggiori, o più generico, per valori minori.
- **dist-from-corpus**: utilizzando la distribuzione delle frequenze normalizzate della parole di un corpus come distribuzione di probabilità del corpus sulle parole, possiamo calcolare la distanza (Jensen-Shannon o Hellinger) dalla distribuzione di probabilità di un topic, ottenendo una misura della generalità del topic. Topics con parole molto frequenti hanno una minor distanza dal corpus e tendono ad essere poco

informativi.

- **tf-df-dist**: distanza calcolata tra le frequenze di assegnazione di ciascuna parola a un topic e le frequenze di assegnazione di ciascun documento al medesimo topic. Questa metrica consente di individuare i topics influenzati dalla *burstiness*. La burstiness consiste nella presenza di parole rare all'interno del corpus, ma frequenti a livello locale di uno o più documenti.
- **doc-prominence**: per distinguere topics che descrivono documenti specifici rispetto a topics che riguardano un numero maggiore di documenti, si possono misurare le proporzioni di documenti assegnati a ciascun topic.
- **topic-coherence**: una metrica molto semplice, che indica quanto spesso le parole più probabili di ciascun topic compaiono insieme nello stesso documento [Mimno et al. 2011].
- **topic-exclusivity**: una misura di quanto le parole più probabili di un topic sono "esclusivamente" assegnate a quel topic rispetto agli altri.

1.4.3 Valutazione qualitativa

Per una prima valutazione, piuttosto semplice, possiamo verificare la presenza di topics appartenenti alle seguenti categorie [Boyd-Graber, Mimno e Newman 2014;Mimno et al. 2011]:

- Topics con parole casuali o non interpretabili
- Topics con parole troppo generiche o troppo specifiche
- Topics con sottoinsiemi di parole incoerenti tra loro (es.:"olio, oliva, frantoio, gatto, cane, topo")
- Topics con "intrusi", ovvero una o più parole non associate a quelle che caratterizzano il topic
- Topics con parole identiche a quelle di un altro topic

I topics sono spazi semantici qualitativi, che ci possono informare nella misura in cui sono comprensibili, ben definiti e rappresentativi dell'insieme dei dati. Pertanto bisogna comprendere se sono validi a livello individuale, se i documenti che descrivono sono appropriati e se, infine l'aspetto più impor-

tante, sono utili all'esplorazione dei dati testuali. Quando il modello genera topics dotati di significato, infatti, il momento della valutazione qualitativa consiste nel giudicare l'effettiva capacità del modello di produrre ulteriore conoscenza sull'oggetto di studio.

Capitolo 2

Correlated Topic Models

Il modello LDA, assumendo che la distribuzione sui topics derivi da una Dirichlet, considera i topics pressoché indipendenti l'uno dall'altro [Blei e Lafferty 2009]. Anzi, come abbiamo osservato, in precedenza, per i modelli LDA sono stati elaborati indici di fit basati sulla massimizzazione delle distanze tra le distribuzioni di probabilità dei topics [Cao et al. 2009].
D'altra parte, può essere molto utile, e più realistico, introdurre in un topic model la possibilità di modellare la forza delle associazioni tra topics. Per questo motivo è stato proposto il *Correlated Topic Model* (CTM), che stima anche le correlazioni tra topics [Blei e Lafferty 2006].
Di seguito provvederemo ad una sintetica presentazione del modello.

2.1 CTM, il modello

Nel CTM si assume che la distribuzione di probabilità sui topics sia generata da una variabile aleatoria Normale-Logistica, anziché da una Dirichlet. Tale variabile, generando le frequenze dei topics in ciascun un documento, considera anche come tali frequenze covariano tra loro all'interno della matrice di covarianza.
Una variabile aleatoria Normale-Logistica deriva da una trasformazione logistica applicata ad una distribuzione Normale. Questa trasformazione permette di avere come supporto di una variabile Normale $(K-1)$-dimensionale, anziché tutto lo spazio reale $\mathbb{R}^{K-1}$, solo il simplesso $\mathbb{S}^{K-1}$ tale che: tutti i

punti presenti nel simplesso siano minori di 1 e sia 1 la somma dei valori associati a ciascun punto nel simplesso [Atchison e Shen 1980]. In questo modo si ottengono i vettori di probabilità che definiscono la distribuzione Multinomiale dei topics per ogni documento θ_d.
Al posto di α, come si osserva in figura, il CTM utilizza i due parametri μ e $\mathbf{\Sigma}$.

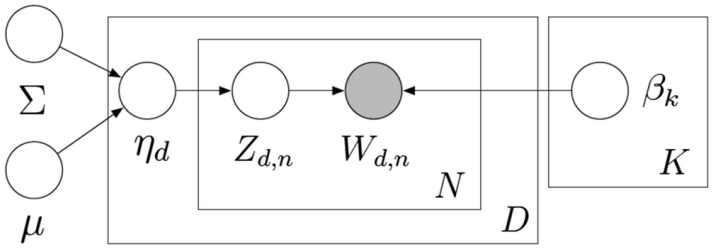


Figura 2.1

Questi rappresentano rispettivamente la media, vettore $(K-1)$-dimensionale, e la matrice di covarianza, $K-1 \times K-1$, della variabile Normale $(K-1)$-variata. Da questa variabile si estrae η_d[1], che consiste nel vettore di frequenza dei topics in un documento. La distribuzione Multinomiale θ_d si ricava quindi, dalla variabile η_d, che, in questo contesto, rappresenta un vettore di valori appartenenti ad una distribuzione normale multivariata, associati al simplesso tramite l'equazione [Blei, Lafferty et al. 2007]:

$$\theta_{d,k} = f(\eta_{d,k}) = \frac{\exp \eta_{d,k}}{\sum_{i=1}^{K} \exp \eta_{d,i}} \tag{2.1}$$

Infine nel CTM la distribuzione Multinomiale delle parole sui topics β_K non è fornita di una prior per controllare la sparsità delle parole, come nella LDA. Il processo generativo del CTM è il seguente:

1. Per ogni documento d:
   (a) Si estrae $\eta_d \sim \mathcal{N}_{K-1}(\mu_d, \mathbf{\Sigma})$
2. Per ogni i-esima parola in ciascun documento d:
   (a) si estrae un'assegnazione $\mathbf{z}_{k,d,n} \sim Multinomiale(f(\eta_d))$;
   (b) si estrae una parola $w_i \sim Multinomiale(\beta_{z_{k,d,n}})$

[1] Attenzione η_d del CTM non va confusa con η del modello LDA.

2.2 Stima dei parametri e inferenza

A questo punto emerge un problema, già affrontato nella trattazione del modello LDA: la *posterior* $\Pr(\eta, \mathbf{z} \mid \mathbf{w}_d, \beta, \mu, \mathbf{\Sigma})$ non è calcolabile. Il metodo più utilizzato per approssimare la *posterior* e ottenere la stima di massima verosimiglianza dei parametri latenti è la variational Expectation-Maximization[2]. L'obiettivo è stimare η e $\mathbf{z}$ per ciascun documento utilizzando la seguente distribuzione a posteriori:

$$\Pr(\eta, \mathbf{z} \mid \mathbf{w}_d, \beta, \mu, \mathbf{\Sigma}) = \frac{\Pr(\eta \mid \mu, \mathbf{\Sigma}) \prod_{n=1}^{N} \Pr(z_n \mid \eta) \Pr(w_n \mid z_n, \beta)}{\int \Pr(\eta \mid \mu, \mathbf{\Sigma}) \prod_{n=1}^{N} \sum_{z_n=1}^{K} \Pr(z_n \mid \eta) \Pr(w_n \mid z_n, \beta) d\eta} \tag{2.2}$$

Per ottenere le stime di massima verosimiglianza dei parametri η e $\mathbf{z}$ si ricorre a una distribuzione variazionale. Si massimizza la verosimiglianza del corpus rispetto a β e a $\mathcal{N}_{K-1}(\mu, \mathbf{\Sigma})$ per minimizzare la divergenza KL tra la distribuzione variazionale approssimata, i cui parametri sono stimati in funzione di questa minimizzazione, e quella vera. Nel caso del CTM la verosimiglianza di un documento da massimizzare è la seguente:

$$\ell(\mu, \mathbf{\Sigma}, \beta) = \log \Pr(\mathbf{w} \mid \mu, \mathbf{\Sigma}, \beta) \tag{2.3}$$

[2]Si veda p. 13, *infra*.

Capitolo 3

Biterm Topic Model

Il *Biterm Topic Model* (BTM) è stato proposto come una soluzione al problema dell'applicazione dei topic models a corpus, in cui ciascun documento presenta un numero molto basso di parole [Yan et al. 2013]. I dati testuali provenienti, ad esempio, da social media, da conversazioni di gruppo o da siti web, spesso formano corpus di testi dalla lunghezza molto ridotta. Il modello LDA o il CTM, applicati a questo genere di dati, si trovano a dover utilizzare numeri molto bassi di *word tokens*, in quanto le occorrenze a livello del singolo documento sono scarse. Di conseguenza il modello ha maggior difficoltà a individuare dei pattern distinti e coerenti di co-occorrenze di parole.

3.1 Il modello BTM

Il BTM modella la generazione di coppie di parole non-ordinate (*biterms*) direttamente all'interno del corpus senza la mediazione del raggruppamento delle parole per documento. Applicando il BTM, quindi, si ottengono solo la distribuzione dei topics nel corpus e delle parole nei topics, ma, implementato il modello, è possibile identificare anche le distribuzioni dei topics sui documenti. Il BTM, infatti, assegna un topic a ciascun biterm.
Nel BTM, inoltre, le co-occorrenze delle parole in un corpus sono modellate integralmente, in quanto in ciascun biterm si utilizza una stessa parola

i-esima in seconda posizione e poi in prima posizione (es.: "colgo un", "un frutto", "frutto da").

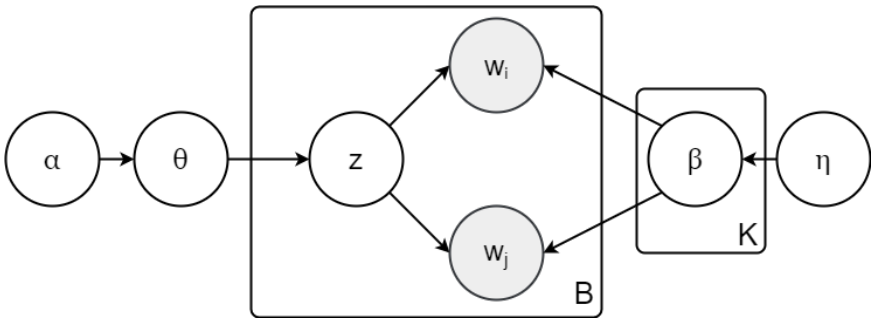


Figura 3.1: Modello grafico del Biterm Topic Model

Il processo generativo del BTM è descritto in questo modo:

1. Per il corpus C:
   (a) si estrae una distribuzione sui topics $\theta \sim Dirichlet(\alpha)$
2. Per ciascuna assegnazione tematica a un biterm $z_{b,k}$:
   (a) si estrae una specifica distribuzione delle parole sui topic $\beta_{z_{b,k}} \sim Dirichlet(\eta)$.
3. Per ciascun biterm b nell'insieme di biterms del corpus B:
   (a) si estrae un'assegnazione tematica $z_{b,k} \sim Multinomile(\theta)$
   (b) si estraggono due parole: $w_i, w_j \sim Multinomiale(\beta_{z_{b,k}})$

Nel BTM θ rappresenta la proporzione dei topics rispetto all'intero corpus, mentre θ_d del modello LDA indica le proporzioni dei topics rispetto a un singolo documento. Infatti il BTM utilizza direttamente le co-occorrenze dei biterms contenute nel corpus. In questo modo risolve il problema della carenza di un numero sufficiente di conteggi parole all'interno di ogni documento. Un altro vantaggio del BTM sta nel fatto che la modellazione di coppie di parole rende più chiaro il tessuto semantico a cui si riferisce ciascun topic.

3.2 Stima dei parametri e inferenza

Come nel caso del modello LDA e del CTM, anche nell'implementazione del modello BTM i parametri possono essere stimati mediante l'approssimazione tramite *Gibbs sampler*. La tecnica di stima è applicata in modo molto simile a quanto avviene per il modello LDA: campioniamo le assegnazioni tematiche per ciascun biterm z_b dalla distribuzione *full conditional* delle restanti variabili:

$$\Pr(z_{b,k} \mid \mathbf{z}_{-b,k}, \mathbf{B}, \alpha, \eta) \propto (n_{z_{b,k}} + \alpha) \frac{(n_{w_i|z_{b,k}} + \beta)(n_{w_j|z_{b,k}} + \beta)}{(\sum_{v=1}^{V} n_{w_v|z_k} + M\beta)} \tag{3.1}$$

in cui $n_{z_{b,k}}$ è il conteggio dei biterms b assegnati al topic k-esimo, $n_{w_i|z_k}$ è il conteggio delle assegnazioni della *word type* w_v al topic k e, infine, $n_{w_i|z_{b,k}}$ è il conteggio della parola i-esima dato il topic e il biterm. Bisogna notare che le assegnazioni $z_{b,k}$ prevedono l'attribuzione dello stesso topic a entrambe le parole del biterm. Ottenuti i conteggi delle assegnazioni tematiche dei biterm e delle parole possiamo stimare β e θ:

$$\beta_{z_{w|k}} = \frac{n_{w_v|k} + \beta}{\sum_{v=1}^{V} n_{w_v|k} + M\beta} \tag{3.2}$$

$$\theta_{z_{b|k}} = \frac{n_{z_{b,k}} + \alpha}{\mid B \mid + K\alpha} \tag{3.3}$$

In ultimo, per ottenere la proporzione dei topics in ciascun documento si assume che ogni θ_d è uguale al valore atteso delle proporzioni dei topics di biterms generati dal d-esimo documento:

$$\Pr(z \mid d) = \sum_{b=1}^{B} \Pr(z \mid b) \Pr(b \mid d) \tag{3.4}$$

Capitolo 4

Structural Topic Model

Lo *Structural Topic Model* (STM) consiste in un'estensione del modello LDA per introdurre i topic models nel mondo della ricerca nelle scienze sociali e nelle scienze del comportamento, non solo in qualità di strumenti esplorativi, ma anche sperimentali [Roberts et al. 2013]. STM, infatti, non si limita a inferire i topics sulla base del modello che propone, ma permette anche di modellare delle informazioni aggiuntive rispetto ai soli dati testuali, cioè dei metadati. Questi metadati sono introdotti nel modello come covariate a livello dei documenti e possono influenzare due elementi del modello: la prevalenza tematica (*topic prevalence*) θ_d, ovvero la proporzione di ciascun documento riguardante ciascun topic, e il contenuto tematico (*topical content*) $\beta_{v,k,d}$, le parole più probabili in un topic. Pertanto, il modello STM ci permette di trovare le differenze o le somiglianze nel linguaggio nei termini di frequenza nell'uso dei topics e del vocabolario utilizzato per descrivere i topics, ad esempio, tra persone appartenenti a categorie diverse, definite nei metadati.
In particolare lo Structural Topic Model è stato originariamente realizzato per analizzare dati testuali derivanti da sondaggi a risposta aperta [Roberts et al. 2014], ma si è rivelato uno strumento valido anche nel contesto dell'analisi del contenuto [Grajzl e Murrell 2019].

4.1 Il modello STM

4.1.1 Elementi di notazione

Per presentare il modello, oltre alla notazione definita inizialmente, bisogna considerare la presenza di due matrici, una per la prevalenza tematica, $\mathbf{X}$, e una per il contenuto tematico, $\mathbf{Y}$. Entrambe le matrici presentano per ogni riga un vettore di tutte le covariate per ciascun documento, La matrice $\mathbf{X}$ ha dimensione $D \times P$ e la matrice $\mathbf{Y}$ ha dimensione $D \times A$. Le righe di queste matrici sono indicate con $\mathbf{x}_d$ e $\mathbf{y}_d$.

4.1.2 Il modello

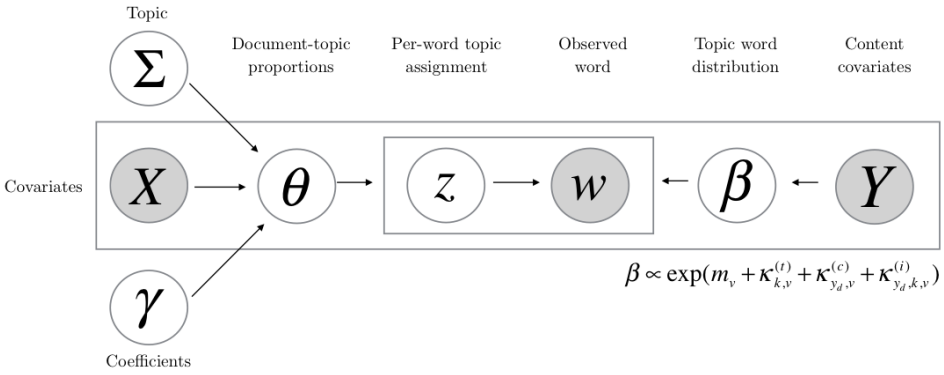


Figura 4.1: Modello grafico del STM, in cui il rettangolo più grande indica che le variabili contenute in esso sono diverse per ogni documento e quello più piccolo per ogni parola.

Il modello STM può essere suddiviso in tre parti e si può osservare nella figura 4.1. La prima rappresenta il modello generatore della prevalenza tematica, la seconda il modello generatore del contenuto tematico, mentre la terza combina i prodotti delle precedenti per la generazione delle singole parole di un documento.
Partendo dalla prima componente del modello STM: per governare la prevalenza tematica di un documento θ_d si utilizza una distribuzione Normale-

Logistica, che possiede un vettore dei valori della media parametrizzato in funzione delle covariate $\mathbf{X}$. In questo modo si assegna a ciascun documento una specifica distribuzione *a priori* sui topics in base alle modalità che questo presenta nelle covariate. A questo punto ci interessa capire come ottenere questi valori del vettore della media. Innanzitutto si estraggono i coefficienti $\gamma_{p,k}$ di ciascuna covariata per ciascun topic da una Normale P-variata con media zero e con varianza controllata da una variabile Gamma-Inversa. Generato il vettore-media, come già osservato nel CTM, possiamo ricavare una Normale-logistica da una Normale $(K-1)$-variata definendone un simplesso come supporto tramite l'equazione: $\theta_{d,k} = \exp(\eta_{d,k})/(\sum_{i=1}^{K} \exp(\eta_{d,i}))$; per rendere il modello identificabile $\eta_{d,k}$ è fissato a zero. Più nello specifico, questa variabile aleatoria Normale-logistica, che ha una matrice di covarianza $K-1 \times K-1$ e i valori del vettore-media dati da $\mu_d = \gamma_K \mathbf{x}_d$, consente di differenziare i valori di θ_d per ciascun documento sulla base delle covariate, definite in $\mathbf{x}_d$. La prevalenza tematica emerge, quindi, da un modello lineare normale multivariato con una sola matrice di covarianza condivisa da tutti i θ_d. Pertanto, un modello STM senza le covariate $\mathbf{X}$ risulta equivalente al CTM.

Per quanto riguarda il contenuto tematico $\beta_{\cdot,k,d}$, invece, ciascun topic ha una distribuzione sulle parole che può variare a seconda della modalità della covariata $\mathbf{Y}$. Tali coviariate non sono incluse nel modello, come in precedenza, tramite una distribuzione *a priori* su $\beta_{\cdot,k,d}$, ma agiscono direttamente sulla parametrizzazione della distribuzione dei topics sulle parole. In primo luogo si parametrizza la distribuzione Multinomiale delle occorrenze di ciascuna parola m_v come il logaritmo delle quote di deviazione rispetto alla distribuzione-base di tutte le parole del corpus $\mathbf{m}$. Queste quote di deviazione possono essere specificate come una funzione dei topics, delle covariate osservate e delle interazioni tra topic e covariata e sono indicate con la lettera κ. Ogni κ si differenzia tramite gli apici e i pedici: la quota di deviazone di un topic è $\kappa_{k,v}^{(t)}$, di una covariata $\kappa_{y_d,v}^{(c)}$ e dell'interazione topic-covariata $\kappa_{y_d,k,v}^{(i)}$. Come si evince dal pedice, $\kappa_{k,v}^{(t)}$ identifica il logaritmo della quota di deviazione per il topic k e la parola v dal logaritmo del contributo della stessa parola v nella distribuzione del corpus. Più in generale $\kappa^{(t)}$ è una matrice $K \times V$, le cui deviazioni sono condivise per ciascuna modalità A delle covariate $\mathbf{Y}_d$.

Per ogni covariata Y_d e per ogni parola v si ha un logaritmo della quota di deviazione dal logaritmo della quota della parola v nel corpus, che rappresenta un'istanza della matrice $\kappa^{(c)}$ di dimensione $A \times V$. In conclusione, la matrice che raccoglie gli effetti delle interazioni topic-covariata di dimensione $A \times K \times V$ è $\kappa^{(i)}$. Pertanto, il contenuto tematico di ciascun topic si può considerare come derivante da una regressione Multinomiale-Logistica, in cui le covariate sono le assegnazioni dei topics alle parole ($\mathbf{z}_{k,n}$), i metadati ($\mathbf{Y}$) e le loro interazioni.
La terza componente del STM descrive la combinazione tra la prevalenza tematica e il contenuto tematico. Come già visto in precedenza da θ_d si estrae un'assegnazione tematica $z_{d,n}$ che permette l'estrazione di una parola condizionatamente al contenuto tematico $\beta_{z_{d,n}}$.
Il processo generatore è quindi il seguente:

1. $\gamma_k \sim Normale_P(0, \sigma_k^2 I_P)$
2. $\theta_d \sim Normale\text{-}Logistica_{K-1}(\gamma \mathbf{x}_d, \mathbf{\Sigma})$
3. $\mathbf{z}_{d,n} \sim Multinomiale_K(\theta_d)$
4. $\mathbf{w}_{d,n} \sim Multinomiale_V(\beta_{z_{d,n}})$
5. $\beta_{v,k,d} = \frac{\exp(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)})}{\sum_{v=1}^{V} \exp(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)})}$

in cui $\sigma_k^2 I_P$ definisce la presenza di una distribuzione prior con forma di Gamma-Inversa sulla varianza σ_k^2, con parametri fissati a $a = 1$ e $b = 1$.

4.2 Stima dei parametri e inferenza

La distribuzione *a posteriori* da stimare è la seguente:

$$\Pr(\eta, \mathbf{z}, \kappa, \gamma, \mathbf{\Sigma} \mid \mathbf{w}, \mathbf{X}, \mathbf{Y}) \propto \left(\prod_{d=1}^{D} (\eta_d \mid \mathbf{X}_d \gamma, \mathbf{\Sigma}) \left(\prod_{n=1}^{N} (z_{d,n} \mid \theta_d)(w_n \mid \beta_{k,d=z_{d,n}}) \right) \right) \times \prod \Pr(\kappa) \prod \Pr(\gamma) \quad (4.1)$$

in cui η consiste nel vettore estratto per ciascun documento dalla Normale-Logistica poi trasformato in θ. Come nei topic models visti precedentemente, anche in questo caso la distribuzione deve essere approssimata. Gli autori del STM propongono l'algoritmo della Variational Expectation-Maximization

utilizzando un'approssimazione alla distribuzione Laplace per gestire il fatto che la Normale-logistica non è congiunta alla Multinomiale. L'inferenza consiste, pertanto, in un E-step, in cui si ottimizzano i parametri della distribuzione *a posteriori* variazionale per le proporzioni dei topics in ciascun documento, e in un M-step, in cui si stimano i coefficienti della prevalenza tematica, del contenuto tematico e si aggiorna la matrice di covarianza.

Capitolo 5

Applicazione dei Topic Models

5.1 I dati

Il campione di dati analizzato consiste in un insieme di dati testuali relativi ai trascritti degli scambi verbali avvenuti in due gruppi di parola[1]. Nel nostro caso entrambi i gruppi di parola sono composti da bambini e adolescenti che condividono l'esperienza dell'affido. Il primo gruppo si è svolto nei mesi di Giugno e Luglio del 2018 in 4 incontri di circa un'ora e mezza. Ad esso hanno partecipato i figli biologici della famiglie affidatarie: 6 soggetti, 2 femmine e 4 maschi, dai 9 ai 15 anni (Milo, Olivia, Francesco, Elisa, Nicola e Davide). Il secondo gruppo di parola ha avuto luogo nei mesi di Novembre e Dicembre 2019 in 4 incontri di circa un'ora e mezza, i cui partecipanti sono stati degli adolescenti in affido. Hanno partecipato 8 soggetti, 5 femmine e 3 maschi, dai 15 ai 17 anni (Alice, Anita, Giuditta, Linda, Lucrezia, Jacopo, Zeno e Dario)[2]. Si noti come Elisa e Francesco del primo gruppo siano fratelli ed hanno Alice, del secondo gruppo, in affido nella loro famiglia. Analogamente, Milo e Olivia sono fratelli ed hanno in affido, nella propria famiglia,

[1]Un gruppo di parola è un gruppo finalizzato al supporto e al confronto tra partecipanti su certi temi principali. Lo scopo è diminuire l'impatto dei fattori di rischio eventualmente derivanti dalla situazione di vita su cui si concentra il gruppo e, al contempo, promuovere attraverso le varie attività di gruppo lo sviluppo di fattori protettivi.

[2]Si noti come ai partecipanti di entrambi i gruppi sono stati assegnati nomi fittizi

Anita. Infine, Carlo e Adele sono i nomi dei due conduttori in entrambi i gruppi.

5.2 Implementazione Topic Models

I topic ci consentono, in sostanza, di osservare delle costellazioni di parole (o spazi latenti) che possono suggerire la presenza di un certo contenuto semantico, indicandoci modi anche alternativi di intendere dei concetti all'interno di uno o più testi. La principale utilità di questo strumento riguarda la possibilità di accostarci qualitativamente al contenuto latente di un insieme di documenti, che può offrirci un primo approccio ad esso o fornirci nuove prospettive per la ricerca. Pertanto lo scopo applicativo di questa tesi è primariamente indagare come i topic models si comportano nell'elaborazione del linguaggio naturale, nello specifico in un contesto di ricerca di ambito psicologico.
I topic models sono stati implementati tramite R [R Core Team 2020], software *open-source* di analisi statistica dei dati, all'interno dell'ambiente di sviluppo integrato (IDE) *RStudio*.

5.2.1 Preprocessamento dei dati

Si importano i dati in formato txt:

```
g1 <- readLines("~/Tesi/gruppo1.txt", encoding = "UTF-8")
g2 <- readLines("~/Tesi/gruppo2.txt", encoding = "UTF-8")
```

In questo modo *g1* e *g2* rappresentano due array in cui ciascun elemento corrisponde all'intervento di una persona. Bisogna, quindi, creare un corpus suddiviso in documenti, nello specifico scegliere come strutturare il corpus: per intervento, per incontro, per partecipanti o per attività. Si è preferito utilizzare le attività, in quanto sono queste a influenzare in maniera più determinante il linguaggio e gli argomenti affrontati in ciascun incontro. Gli incontri, in questo contesto, sono un'insieme di attività separate legate dal tema principale che è l'affido. Questo formato, inoltre, permette di avere un

numero sufficiente di documenti e un sufficiente equilibrio tra eterogeneità e omogeneità tra i diversi documenti. Considerare come documento tutta la produzione verbale di ciascun partecipante non è una scelta ottimale, poiché alcuni tendono ad essere più laconici, contribuendo in minor misura all'emersione dei topic affrontati nel complesso delle interazioni. L'ultima possibilità per suddividere il corpus sarebbe quella di considerare ciascun intervento come un documento a sé, ma, eccetto i conduttori del gruppo, le frasi sono molto brevi, composte di poche parole. Quindi, dopo aver suddiviso entrambi i vettori per attività, procediamo creando un *data.frame*:

```
gA1<-data.frame(documenti=c("a1","a2","a3","a4","a5","a6","a7","a8","a9","
    a10","a11","a12"),testi=g1)
gA2<-data.frame(documenti=c("a1","a2","a3","a4","a5","a6","a7","a8","a9"),
    testi=g2)
```

Il passo successivo è la normalizzazione del testo, nel nostro caso: la lessicalizzazione di parole che formano insieme un concetto altrimenti non riconoscibile ai fini dell'analisi del testo (es.: da "mamma affidataria" a "mamma_affidataria"); riduzione del testo a caratteri minuscoli; rimozione della punteggiatura, dei numeri e delle *stopwords*. Le stopwords sono parole molto frequenti nel testo, ma irrilevanti per le analisi statistiche. Generalmente le stopwords da eliminare provengono da una lista costruita *ad hoc* per i dati a disposizione [Wallach, Mimno e McCallum 2009]. Per queste operazioni utilizziamo il pacchetto *quanteda* [Benoit et al. 2018], iniziando con il caricamento del *file* contenete le stopwords e con la creazione del corpus[3]:

```
stopwords<-readLines("~/Tesi/stopwords.txt", encoding = "UTF-8")

library(quanteda)

corpus1<-corpus_reshape(corpus(gA1$testi),to="documents")
docnames(corpus1)<-gA1$documenti

corpus_norm <- corpus1 %>% gsub(pattern = "'", replacement = " ") %>%
tokens(remove_numbers = TRUE,
```

[3]Nel seguito, quando le operazioni sono identiche per entrambi i gruppi, sono riportati, per semplicità, solo i passaggi destinati al primo gruppo.

```
remove_punct = TRUE,
remove_symbols = TRUE) %>%
tokens_tolower() %>%
tokens_select(pattern = stopwords,
selection = "remove",
verbose = TRUE
valuetype="fixed")
```

L'ultimo passaggio consiste nella creazione della *Document-Feature Matrix* (DFM), o Document-Term Matrix, sempre tramite il pacchetto *quanteda*, rimuovendo le parole con frequenza inferiore a 2:

```
dfm1<-dfm(corpus1, stem = FALSE)
dfm1<-dfm_trim(dfm1, min_termfreq = 2)
```

Al termine del preprocessamento otteniamo, per il primo gruppo, un DFM con 887 *word types* e 2067 *word tokens* e, per il secondo gruppo, un DFM con 920 *word types* e 1854 *word tokens*.

5.2.2 Implementazione LDA

ldatuning

Per applicare il modello LDA ai dati bisogna fissare il parametro K numero di topic e per identificarne, inizialmente, il valore ottimale ci affideremo alle metriche proposte nel pacchetto *ldatuning* per entrambi i gruppi:

```
library(ldatuning)

metriche1 <- FindtopicNumber(dfm1,
topic = c(2:50),
metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
method = "Gibbs",
control = list(seed = 503),
mc.cores = NA,
verbose = TRUE)

FindtopicNumber_plot(metriche1)
```

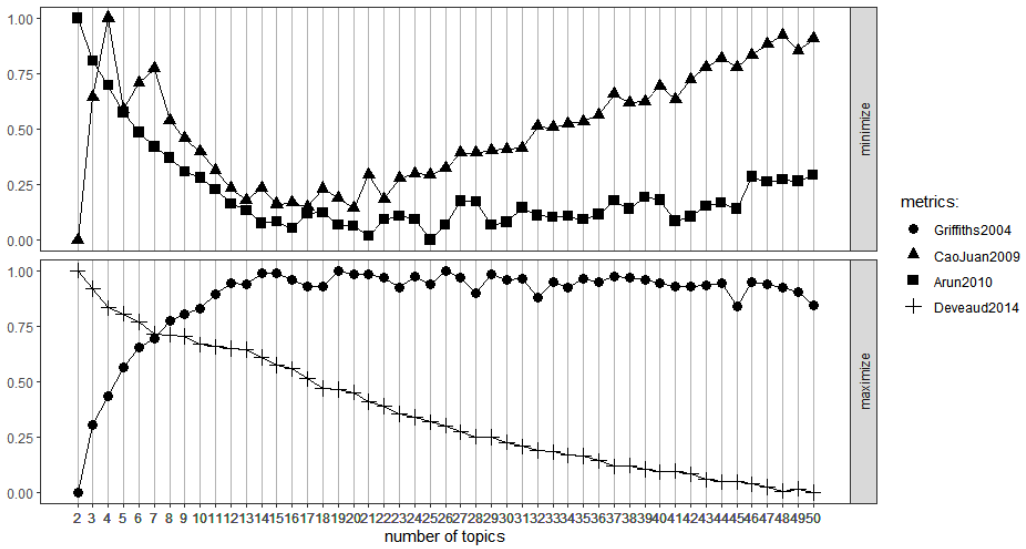


Figura 5.1: Metriche per il valore ottimale di K per il gruppo uno.

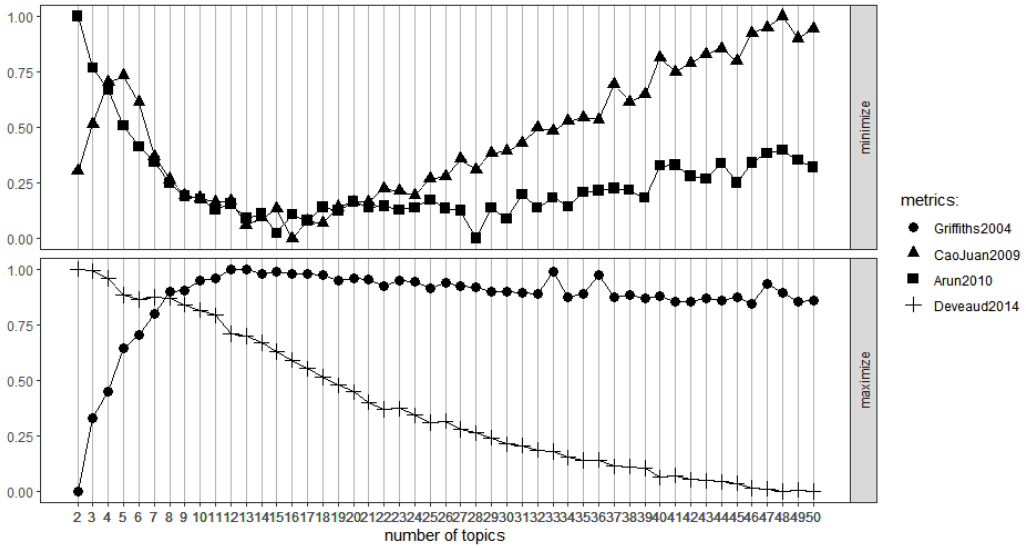
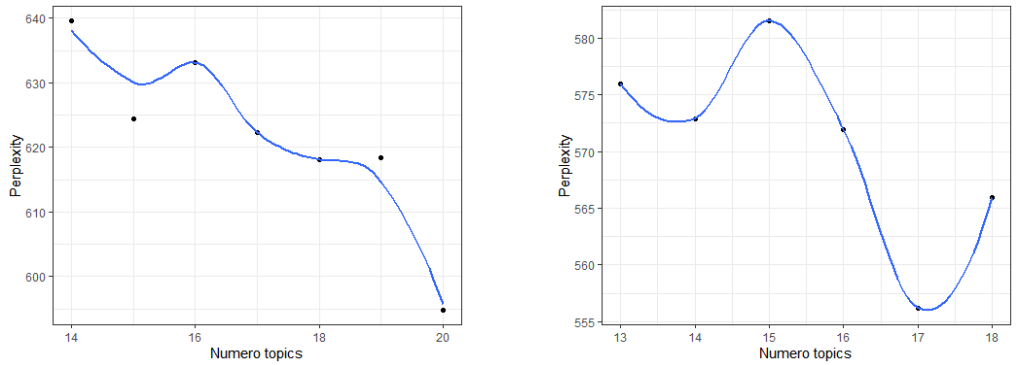


Figura 5.2: Metriche per il valore ottimale di K per il gruppo due.

Il grafico 5.1 sembra suggerire che il numero ottimale di topic per il primo gruppo si trovi nell'intervallo tra 14 e 20. Mentre, per il secondo gruppo, il grafico 5.2 indica l'intervallo, all'incirca, 13-18.
Per valutare più nello specifico il valore migliore all'interno degli intervalli

individuati utilizziamo il metodo *held-out* si veda p. 18, *infra*. Dividiamo i dati in due sottogruppi:

```
n1<-12  #numero documenti
splitter<-sample(1:n1, round(n1 * 0.75))

train_set<-dfm1[splitter, ]
valid_set<-dfm1[-splitter, ]

dfmT1<-convert(train_set, to = "topicmodels")
dfmV1<-convert(valid_set, to = "topicmodels")

library(topicmodels)

k<-c(13:17)
perplexT<-vector(length = length(k))
for (i in 1:length(k)){
lda_T<-LDA(dfmT1, method = "Gibbs", k = k[i], control = list(burnin=1000))

perplexT[i]<-perplexity(lda_T, newdata = dfmV1)
}
```



(a) Gruppo 1 **(b)** Gruppo 2

Dal grafico (5.3a)[4] osserviamo che il modello con 20 topic ha una performance di adattamento a nuovi dati migliore rispetto a gli altri modelli. Per il secondo

[4]Codice in appendice (A.1)

gruppo (5.3b) il valore preferibile è invece 17.

Implementazione

Avendo implementato ciascun modello nell'intervallo precedentemente individuato, si è riscontrato che il modello più consistente e interpretabile è quello con $K = 15$ per il primo gruppo e $K = 14$ per il secondo. I risultati del *held-out method* per entrambi i gruppi non sono stati informativi.
Per l'implementazione del modello LDA si è utilizzato il pacchetto *topicmodels*[5] [Grün e Hornik 2011]:

```
library(topicmodels)

dfmTM1<-convert(dfm1, to="topicmodels")
lda_model1<-LDA(dfmTM1,method = "Gibbs",
k=14,
control= list(alpha=50/k, delta=0.1,
burnin=1000, seed=50312))

get_terms(lda_model1, 10)

dfmTM2<-convert(dfm2, to="topicmodels")
lda_model2<-LDA(dfmTM2,method = "gibbs",
k=14,
control= list(alpha=50/k,delta=0.1,
burnin=1000, seed=50312))

get_terms(lda_model2, 10)
```

[5]Nel contesto del pacchetto *topicmodels* δ equivale a η

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	scuola	acqua	affido	futuro	genitori
2	caratteristica	sostanza	domande	idea	alice
3	sport	gelatina	domanda	idee	fratello
4	età	legno	fratello_affidatario	mare	carte
5	nome	sole	linea	desiderio	strategia
6	affido	ghiaccio	ragazzi	affido	consigli
7	anni	cuoio	intervistatore	bar	pazienza
8	preferito	palla	rai_affido	pensa	ricatto
9	alice	occhi	risposte	anello	parlare
10	calcio	pallone	gruppo	felici	amico

	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	insieme	affido	pensare	mamma	stanza
2	esempio	fratello	scotch	colore	ricordo
3	ragazzi	grande	proposito	arrabbiato	affido
4	rai_affido	famiglia	idea	stagno	giorno
5	parlare	famiglie	giocare	pagina	pagine
6	pensare	fratelli	tempo	rubato	nome
7	libro	ragazzi	comodo	momento	bambino
8	possibilità	esperienza	gioco	tv	anello
9	calcio	bambini	rabbia	stefano	settimana
10	compagno	difficile	organizzare	fratelli_biologici	gruppo

	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
1	genitori	calzini	inizio	anita	anita
2	sorella	giochi	pazineza	casa	mamma_origine
3	ascoltare	fratelli	regole	anni	andrea
4	giochi	vacanza	maschio	genitori	ciuccetti
5	giocare	vestiti	comunità	camera	casa
6	pronti	italiano	fratelli	papà	settimana
7	cartellone	sveglia	rapporto	bambino	gruppo
8	femmine	foto	arrabbiare	mamma	sera
9	compenso	regalo	paziente	sorella	bar
10	merenda	fratello_affidatario	età	amici	alice

Tabella 5.1: topic del gruppo 1.

	Topic 1	Topic 2	Topic 3	Topic 4
1	assistente_sociale	gioco	ragazzi	affido
2	silvia	possibilità	famiglia	tribunale
3	edoardo	cambiare	ragazzo	anni
4	famiglie	periodi	affido	domande
5	comunità	campo	percorso	rispsote
6	nomi	capo	affidatari	assistente_sociale
7	momento	risposta	famiglie	cellulare
8	affidi	ragazzi	anni	psicologo
9	casa	famglia_affidataria	genitori	giudiziale
10	scelta	buono	contenuti	consensuale

	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
1	periodo	ragazzi	gruppo	vita	regola
2	mare	buio	pronto_affidi	parlare	domande
3	ricordo	affidatari	whatsapp	tranquilla	gruppo
4	lettera	sciare	affidi	mamma	destra
5	scuola	momento	idee	importanza	maschio
6	mamma_origine	descrivere	nome	blu	femmina
7	amico	parlare	descrizione	scritta	domanda
8	buio	ricordi	gioco	famiglia	rispondere
9	squadra	oca	pronti	famiglie	giro
10	famiglia	turno	avatar	riferimento	stanza

	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
1	affido	parlare	affido	gruppo	like
2	insieme	condividere	anni	famiglia	casa
3	famiglia_affidataria	famiglia	bagno	immagine	diciotto
4	pazienza	pensare	famiglia_affidataria	sorella	affidatari
5	vivere	decidere	origine	mamma_origine	affido
6	interno	merenda	adozione	profilo	tempo
7	aspettative	pensiero	ragazzi	messaggio	regole
8	relazione	gruppo	cellulare	partecipanti	esempio
9	sé	cerchio	frase	foto	mamma_origine
10	famigliare	problemi	genitori	descrizione	papà

Tabella 5.2: topic del gruppo 2.

Per quanto riguarda il primo gruppo, i topic del modello LDA presentano in generale una discreta qualità, ovvero ci permettono di "immergerci" nel testo e osservare alcuni punti salienti delle discussioni del gruppo. Il topic 1 riguarda il tema delle presentazione di sé al gruppo. Il topic 2 individua una specifica attività di immaginazione, in cui bisognava pensare di essere una "sostanza" con attenzione alle qualità e alla forma di questa "sostanza". Il topic 3 ci parla dell'attività di "giornalismo" sul tema dell'affido. Il topic 4

è molto interessante perché tratta il tema del futuro e dei desideri, mostrando un rgomento di riflessione del gruppo. Il topic 5 è piuttosto confuso, ma sembra riferirsi a un'attività in cui l'obiettivo era trovare le strategie rispetto alle problematiche nelle relazioni interpersonali. Il topic 7 distingue il tema dell'affido rispetto alla famiglia e come esperienza. Il topic 10 sembra mostrare un'altro oggetto di riflessione di gruppo riguardo al tema del ricordo rispetto al primo incontro e al periodo iniziale con la sorella o del fratello in affido. Mentre il topic 13 tende a trattare il tema del rapporto con il minore in affido più in generale. Il topic 11 sembra riferirsi alle attività di gioco svolte nel gruppo. il topic 14 e il topic 15 sembrano riferirsi a ciò che pensano i due fratelli, Olivia e Milo, nei confronti di "anita", la minore in affido nella loro famiglia. Il topic 14 si riferisce ad Anita nel contesto famigliare e neri rapporti con i genitori affidatari, il topic 15 invece parla di Anita ma più in relazione alla famiglia di origine.
Il modello LDA applicato al secondo gruppo si comporta discretamente anche in questo caso, individuando dei topic piuttosto informativi. Il topic 1 è connesso al tema delle figure professionali che si occupano del mondo dell'affido, in cui i nomi comparsi sono di professionisti del settore. Il topic 3 tratta il tema dell'affido a livello famigliare e come "percorso". Il topic 4 è incentrato sull'attività di informazione svolta riguardante l'affido in una prospettiva giurisprudenziale. Il topic 5 e il topic 6 sembrano indicare un momento di riflessione sui ricordi dei partecipanti. Il topic 14 è piuttosto interessante perché, pur apparendo molto confuso, si riferisce molto precisamente a una specifica attività, che prevedeva un gioco in cui una persona, rientrata nella stanza, doveva scoprire con delle domande la "regola" scelta dal gruppo durante la sua assenza. Il topic 10 sembra trattare il tema dell'affido più nel concreto delle proprie vite da un punto di vista emotivo, ma anche riflessivo. il topic 11 raccoglie delle parole riguardanti probabilmente l'agire dei partecipanti, ma le parole di cui è composto sono molto generiche. Il topic 14 sembra riguardare il tema dei diciotto anni in relazione alla famiglia affidataria.
Conoscere la distribuzione dei topic sui documenti (θ) ci può aiutare nell'interpretazione dei topic, come strumento di validazione delle proprie ipotesi sul significato di un topic. Naturalmente i topic che hanno maggiori proba-

bilità di comparire in più documenti come topic principali, tenderanno ad avere un contenuto più generale e ad essere meno informativi sul singolo documento.

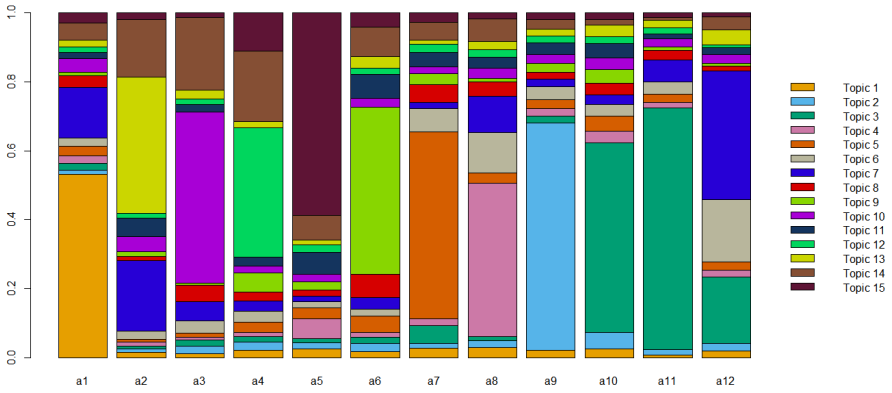


Figura 5.3: Proporzioni dei topic per documento, gruppo 1.

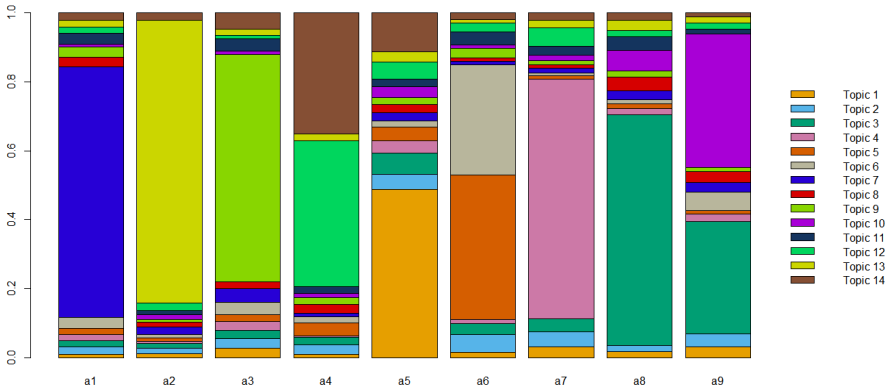


Figura 5.4: Proporzioni dei topic per documento, gruppo 2.

Entrambi i grafici mostrano che tendenzialmente tutte le attività hanno un topic principale che ne descrive il contenuto e che questo topic è esclusivamente dedicato a ciascuna attività, eccetto che il topic 3 per il primo gruppo e sempre il topic 3 per il secondo gruppo.

5.2.3 CTM

In questa sezione ci occupiamo del Correlated Topic Model al fine di valutare la qualità dei topic prodotti rispetto al modello LDA. Il CTM tende ad avere migliori valori di *perplexity*, ma minor interpretabilità dei topic rispetto al modello LDA [Chang et al. 2009].
Il pacchetto *topicmodels* è fornito anche di una funzione per l'implementazione del Correlated Topic Model e richiede lo stesso fromato di DFM del modello LDA:

```
ctm1<-CTM(dfmT1, k = 15, method = "VEM")
ctm2<-CTM(dfmT2, k = 14, method = "VEM")
```

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	affido	affido	anita	affido	affido
2	domande	ragazzi	calzini	futuro	ricordo
3	linea	fratellli	papà	famiglia	stanza
4	ragazzi	fratello	anni	genitori	casa
5	risposte	famiglie	giochi	idea	gelatina
6	rai_affido	esperienza	carte	desiderio	sostanza
7	intervistatore	pagine	casa	bar	acqua
8	esperienza	possibilità	vacanza	pensare	anita
9	intervista	famigla	alice	mare	legno
10	studio	grande	palla	anello	pagine

	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	affido	ragazzi	arrabbiato	genitori	affido
2	grande	fratello	pagina	consigli	grande
3	alice	tempo	stagno	fratello	comunità
4	famiglia	affido	mamma	strategia	compenso
5	genitori	casa	stefano	carte	inizio
6	famiglie	famiglia	anita	alice	fratello_affidatario
7	ragazzi	problema	affido	mamma	esperienza
8	sorella	genitori	gioco	ricatto	fratelli
9	fratello_affidatario	fratello_affidatario	colore	anita	anni
10	pensare	anni	retino	parlare	regole

	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
1	scuola	anita	affido	mamma_origine	grande
2	caratteristica	casa	domande	anita	genitori
3	affido	famiglia	domanda	casa	ragazzi
4	sport	inizio	conduttore	ciuccetti	casa
5	età	anni	intervistatore	andrea	affido
6	anni	fratello	casa	settimana	anita
7	anita	grande	fratello_affidatario	bar	famiglia
8	preferito	fratelli	intervistato	amica	anni
9	nome	maschio	foglio	comprare	sorella
10	calcio	bambino	risposte	scarpe	papà

Tabella 5.3: CTM del gruppo 1.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	ragazzi	periodo	famiglia	ragazzi	gruppo
2	affido	buio	ragazzi	buio	ragazzi
3	insieme	mare	ragazzo	ricordo	parlare
4	pazienza	affido	affido	parlare	nome
5	famiglia	gioco	anni	domanda	gioco
6	famiglia_affidataria	lettera	genitori	affidatari	momento
7	percorso	squadra	contenuti	famiglia	tempo
8	famigile	sciare	famiglie	mamma_origine	mano
9	ragazzo	ragazzi	gruppo	parola	domanda
10	relazione	risposta	attività	domande	casa

	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	affido	affido	affidatarie	assistente_sociale	gruppo
2	like	tribunale	partita	silvia	pronto_affidi
3	anni	anni	ricordi	edoardo	whatsapp
4	famiglia	assistente_sociale	ragazzi	affido	idee
5	casa	risposte	possibilità	affidi	affidi
6	diciotto	domanda	famiglie	like	nome
7	affiatari	giudiziale	domande	comunità	pronti
8	ragazzi	consensuale	contenuti	casa	descrizione
9	adozione	tutore	massiccia	sonia	avatar
10	tempo	sposati	necessari	educatore	canzone

	Topic 11	Topic 12	Topic 13	Topic 14
1	regola	ragazzi	gruppo	gruppo
2	domande	periodo	famiglia	affido
3	gruppo	affido	immagine	affidatari
4	maschio	buio	sorella	ragazzi
5	domanda	affidatari	mamma_origine	genitori
6	destra	amico	profilo	tempo
7	femmina	mare	messaggio	possibilità
8	stanza	gruppo	partecipanti	bisogno
9	donna	perso	descrizione	parlare
10	giro	parlare	foto	momento

Tabella 5.4: CTM del gruppo 2.

I topic ottenuti con gli stessi valori di K del modello LDA dal CTM tendono ad essere meno interpretabili e a presentare più "parole-intruso". Ciononostante l'interpretabilità dei topic del secondo gruppo si dimostra migliore rispetto a quella dei topic del primo gruppo.

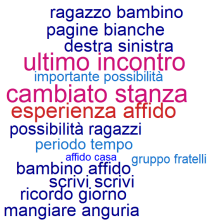
5.2.4 BTM

Come abbiamo visto all'inizio di questo capitolo, i dati a disposizione non devono essere necessariamente divisi in documenti che rappresentino il singolo intervento verbale di ciascun partecipante. Quindi nel nostro caso non sussiste il problema di dover applicare un topic model a documenti eccessivamente poveri di parole. Tuttavia possiamo testare le capacità del Biterm Topic Model, confrontando la qualità dei topic prodotti da questo con la qualità dei topic generati dal modello LDA applicato ai singoli interventi.
Per l'implementazioe del Biterm Topic Model è disponibile il pacchetto dedicato *BTM* [«BTM: Biterm Topic Models for Short Text. R package»]. La funzione *BTM* richiede un oggetto *data.frame* contenente, come prima colonna, una stringa identificativa del documento di appartenenza di ciascun biterm e, come seconda colonna, i biterms; i due *data.frame* sono rispettivamente *bitg1* e *bitg2* per il primo e il secondo gruppo.
Per conservare una cera qualità interpretativa dei risultati, si è provveduto ad utilizzare un numero non elevato di topic ($K = 15$).

```
library(BTM)

btm1<-BTM(bitg1, k = 15)
btm2<-BTM(bitg2, k = 15)
```

(a) Topic 1

(b) Topic 2



(c) Topic 3

(d) Topic 4

(e) Topic 5

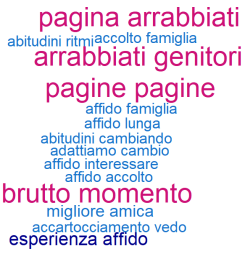
(f) Topic 6

(g) Topic 7

(h) Topic 8

(i) Topic 9

(j) Topic 10



(k) Topic 11

(l) Topic 12

(m) Topic 13

(n) Topic 14

(o) Topic 15

Figura 5.5: Modello BTM del gruppo 1.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	francesco	fratello	milo	casa	caratteristica
2	nome	persone	chiedo	stanza	senso
3	grande	grande	età	camera	giusto
4	papà	basta	sport	bravo	fratello_affidatario
5	teniamo	esempio	vacanza	letto	finito
6	dico	pagine	settimana	maschi	famiglia
7	aspetta	fratello_affidatario	assieme	vicino	dobbiamo
8	pazienza	domande	voleva	anno	affidatari
9	scriviamo	calzini	vado	trovare	fratelli
10	giochi	dice	arrivata	papà	scriviamo

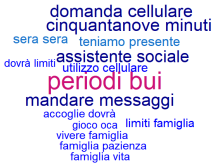
	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	milo	domande	elisa	importante	casa
2	famiglie	olivia	davide	famiglia	persone
3	anni	esperienza	secondo	mamma_origine	cambiato
4	vuoi	insieme	inizio	volete	fratello
5	mente	ora	volete	cinque	giocare
6	settimana	maschio	pagina	andrea	mare
7	aspetta	problema	sorella	liceo	qualcuno
8	bambino	bella	regalato	giusto	arrabbiare
9	cominciamo	male	sentita	disponibile	anni
10	futuro	difficile	sostanza	ricordate	olivia

	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
1	persona	affido	nicola	fratelli	mamma
2	gruppo	ragazzi	giusto	genitori	ricordo
3	potete	scuola	tempo	giorno	bella
4	anni	genitori	pensare	figli	subito
5	palla	attimo	parlare	bar	carte
6	poco	chiede	calcio	secondo	gelatina
7	abbastanza	amico	scrivete	tv	idea
8	ricordi	vive	arrabbia	parlare	vedi
9	andava	nuovo	inizio	prendere	aiuta
10	genitori	stanza	capire	ricordo	acqua

Tabella 5.5: LDA applicata sui singoli interventi del primo gruppo.

(a) Topic 1

(b) Topic 2



(c) Topic 3

(d) Topic 4

(e) Topic 5

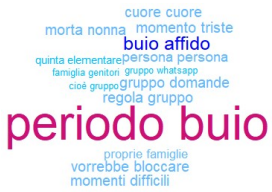
(f) Topic 6

(g) Topic 7

(h) Topic 8

(i) Topic 9

(j) Topic 10



(k) Topic 11

(l) Topic 12

(m) Topic 13

(n) Topic 14

(o) Topic 15

Figura 5.6: Modello BTM del gruppo 2.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	risposta	casa	ragazzi	domande	famiglia
2	parlare	silvia	partecipanti	rispondere	esempio
3	ricordo	immagine	papà	scelta	affido
4	aspetto	diventare	frase	esempio	descrizione
5	risposte	sciare	fratello	donna	femmina
6	capelli	gruppo	origine	sorella	differenza
7	diciotto	incontro	profilo	piedi	pronti
8	significato	ragazze	assistente_sociale	bellissimo	giocare
9	profilo	gente	disposto	proseguire	casa
10	origine	indovina	cerchio	famiglia_affidataria	futuro

	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11
1	affido	ragazzi	gruppo	scuola	periodo	nome
2	anni	famiglie	affidi	idee	genitori	pronto_affidi
3	ragazzo	insieme	papà	bambini	buio	mare
4	gioco	possibilità	amico	momento	whatsapp	attività
5	diciotto	percorso	affidatario	film	adozione	foto
6	tribunale	pazienza	opportunità	whatsapp	periodi	conoscere
7	vivere	affidatari	anno	cuore	doccia	figlio
8	giudiaziale	affidatarie	dispiace	ricordo	sposati	squadra
9	chiaro	cominciare	capo	famigliare	affidatari	pronti
10	ragazzi	messaggio	parlare	squadre	piscina	sennò

	Topic 12	Topic 13	Topic 14	Topic 15
1	mamma_origine	affidatari	like	cellulare
2	regola	tempo	domanda	casa
3	genitori	ricordo	famiglia_affidataria	gruppi
4	immagine	idea	assistente_sociale	amici
5	nomi	capelli	accordo	interno
6	forza	giorno	momento	ricordi
7	regole	cena	luisa	tempo
8	arrabbiare	rabbia	mamma	descrizione
9	storia	famiglia_affidataria	fromazione	esperienza
10	lettera	rappresentare	entrare	sinistra

Tabella 5.6: LDA applicata sui singoli interventi del secondo gruppo.

I topic del BTM sono presentati come *wordcloud* in cui le parole più probabili hanno grandezza maggiore e sono di colore rosa scuro e rosso, mentre una probabilità più bassa è indicata dal blu scuro e dall'azzurro.
Rispetto al primo gruppo, il topic 2 sembra trattare il tema della presentazione. Il topic 3, pur essendo in parte confuso, sembrerebbe riguardare l'esperienza dell'affio in relzione ai primi ricordi. Il topic 4, il 5 e, in parte, il 10 pare che facciano emergere un nuovo tema: l'attvità di riflessione sulle qualità e i comportamenti del fratello affidatario. Il topic 7 riguarda i

"ciuccetti", che sono un tipo caramelle. Nonostante alcune "parole-intruso", il topic 9 indica l'attività di giornalismo sull'affido insieme al topic 13. Il topic 11 distingue l'attività in cui i bambini venivano invitati utilizzare una pagina di un diario per descrivere un momento di rabbia in relazione alle esperienze in famiglia legate all'affido. Il topic 14 riguarda il tema del sesso femminile e maschile. Il topic 15 sembra alluderea al tema del primo periodo dell'affido.
Anche i topic relativi al secondo gruppo, oltre a quelli del primo, offrono un buon livello qualitativo. Il topic 1 è legato al tema della famiglia affidataria e di origine. Il topic 2 è riferito all'affido da un punto di vista più formale e in relazione al compimento dei diciotto anni. Il topic 3 tende a parlare del uso del cellulare e delle limitazioni in famiglia, pur essendo abbastanza confuso. Il topic 5 descrive, nonostante le "parole-intrusi", inieme al topic 11 il tema dei "periodi bui". Il topic 6 verte sul tema di come funzionano o dovrebbero funzionare le relazioni nella famiglia affidataria tra famiglia e minore in affido. Il topic 7 riguarda Benedetta Parodi e il programma televisivo "Bake Off". Il topic 8 e il 13 parlano dell'invenzione del nome "pronto affidi" per il gruppo Whatsapp dei minori in affido. Il topic 9 sembra connesso all'affido e all'*iter* attraverso cui una famiglia diventa una famiglia affidataria. Il topic 10 riguarda la fiura dell'assistente sociale. Il topic 12 riguarda il tema della famiglia connessa all'attività del gruppo di Whatsapp. Il topic 14 affronta l'affido dal punto di vista giurisprudenziale.
I topic generati dal BTM sia per il primo che per il secondo gruppo sono nettamente migliori di quelli del modello LDA, che sono pressoché tutti non interpretabili.

5.2.5 STM

L'ultimo topic model applicato ai trascritti è lo Structural Topic Model tramite il pacchetto *stm* [Roberts, Stewart, Tingley et al. 2014]. Questo topic model, come già visto, permette di identificare eventuali differenze nella proporzione di topic all'interno dei documenti e nella distribuzione delle parole all'interno dei topic tra sottogruppi di dati, aventi metadati diversi. Si è scelto di utilizzare questo modello per esplorare il rapporto tra i dati appartenenti al primo gruppo e quelli del secondo in termini di contenuto tematico. Il *da-*

ta.frame G12 contiene i testi suddivisi per attività e l'informazione relativa all'appartenenza di questi al primo o al secondo gruppo.
Il preprocessamento dei dati prevede di applicare gli passaggi svolti in precedenza tramite *quanteda* fino alla conversione del DFM nel formato compatibile con il pacchetto *stm*:

```
library(stm)

#preprocessamento
dfm_stm<-convert(dfmG12,to="stm")
```

Per individuare il miglior valore di K è possibile utilizzare la funzione *searchK* specificando *attivita* come predittore della prevalenza tematica dei topic, in quanto è proprio per attività che sono divisi i documenti, e *gruppo* come predittore del contenuto tematico dei topic:

```
fitK <- searchK(documents = dfm_stm$documents, vocab = dfm_stm$vocab, data
     = dfm_stm$meta, K = c(10:30), prevalence=\sim attivita, content=\sim
    gruppo, N = 2 , proportion = 0.5, M = 10, init.type = "Spectral",
    heldout.seed = 23154, cores = 1)

plot.searchK(fitK)
```

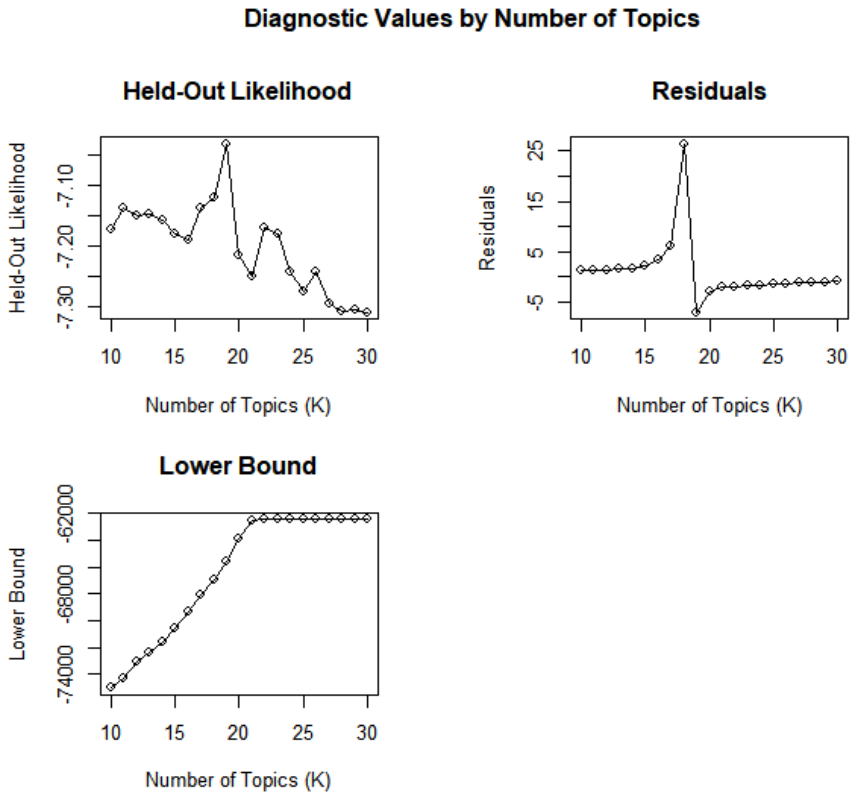


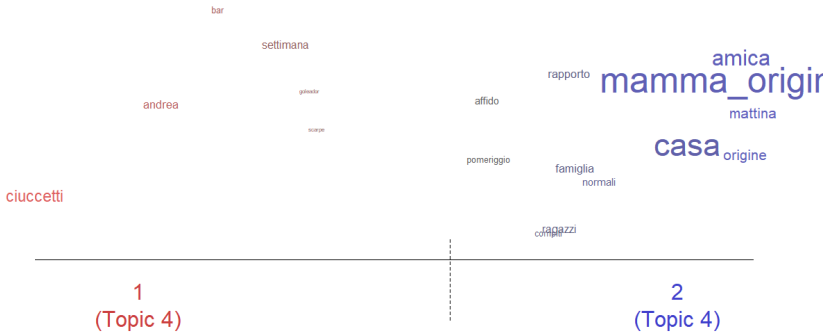
Figura 5.7: Metriche per la valutazione della bontà del modello.

Il grafico sembra tendenzialmente indicare 20 come numero migliore di topic.
Il modello STM viene stimato nel modo seguente:

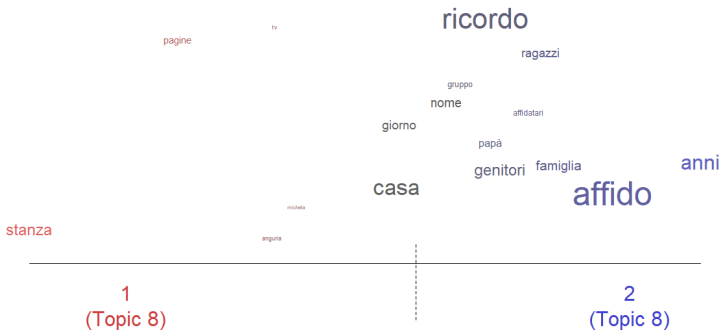
```
stm_contenuto <- stm(documents = dfm_stm$documents, vocab = dfm_stm$vocab,
    data = dfm_stm$meta, K = 20, prevalence=\sim attivita, content=\sim
   gruppo, init.type = "Spectral", seed = 23154)
```

Le differenze più interessanti tra i due gruppi nell'utilizzo del linguaggio all'interno dei topic stimati sono raffigurate nelle seguenti figure, in cui la dimendione della parola indica il suo peso nel topic, mentre la distanza orizzontale dalla linea centrale tratteggiata indica l'intensità con cui una data parola appartiene a una covariata[6]:

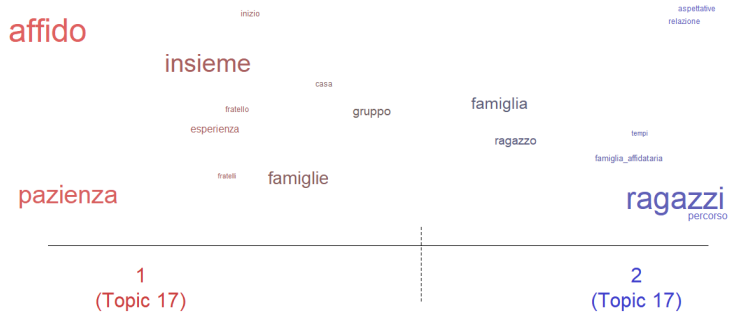
[6] Codice in appendice (A.3)



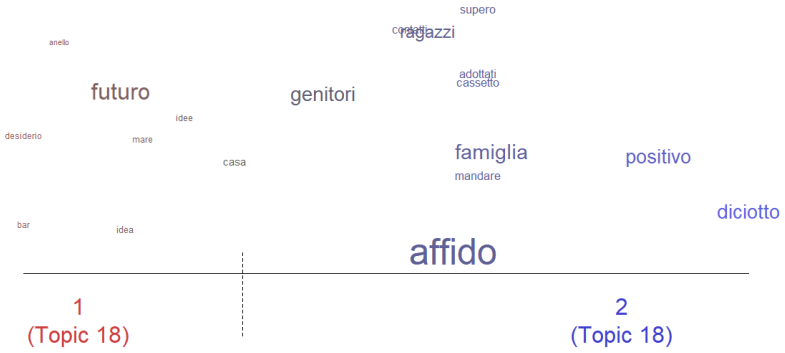
(a) Topic 4



(b) Topic 8



(c) Topic 17



(d) Topic 18

Figura 5.8: topic a confronto tra i due gruppi per contenuto tematico.

Nel topic 4 l'argomento è la "mamma_origine", che nel primo gruppo compare principalmente in relazione all'acquisto di dolciumi (i "ciuccetti" e le "goleador"), nel secondo invece la frequenza della parola è di gran lunga maggiore e più legata a temi diversi. Nel topic 8 il tema che emerge è quello del ricordo rispetto all'affido: il primo gruppo si concentra più in particolare sul tema della stanza del minore in affido, mentre il secondo gruppo parla più approfonditamente del ricordo in relazione, più in generale, alla famiglia affidataria. Il topic 17 sembra mostrare la differenza tra il primo e il secondo gruppo nell'affrontare il tema di una riflessione sui vissuti dell'esperienza dell'affido in famiglia: il primo gruppo parla di "pazienza","fratelli", "esperienza" e "insieme", mentre il secondo di "percorso", "relazione", "aspettative" e "tempi". Nel topic 18, infine, si parla del futuro: nel primo gruppo considerando più i desideri, nel secondo trattando il tema in maniera più approfondita rispetto ai diciotto anni e ai legami con la famiglia affidataria. Il modello STM, pertanto, si dimostra capace di fornire risultati interpretabili e informativi anche per delineare differenze di contenuto tematico tra corpus diversi.

Capitolo 6

Conclusioni

In questa sezione verranno fornite alcune considerazioni conclusive rispetto a quanto è emerso dall'applicazione dei topic models ai dati testuali.
L'utilità di un topic model, come già evidenziato, dipende, in prima istanza, dalla possibilità di riconoscere in ciascun topic generato un tema distinto e ben definito. I topics emersi dai trascritti di entrambi i gruppi, invece, possono sembrare piuttosto confusi e, certuni, del tutto insensati, soprattutto agli occhi di chi non conosce il contenuto di questi testi. L'interpretazione dei topics proposta, infatti, è stata possibile principalmente grazie a una lettura integrale dei dati testuali a disposizione, ovvero grazie a una conoscenza approfondita dei trascritti. Senza questa "validazione qualitativa" sarebbe stato molto più arduo formulare ipotesi interpretative rispetto a buona parte dei topics ottenuti e probabilmente la maggior parte di essi sarebbe stata ritenuta incomprensibile. La difficoltà interpretativa potrebbe essere dovuta al fatto che questi topics raramente si riferiscono a un tema con parole che minimizzano lo sforzo di comprensione dell'essere umano. Spesso i topics qui inferiti, infatti, più che delineare argomenti collegati a un unico concetto, come ad esempio il concetto di "famiglia", indicano dei veri e propri "argomenti di discussione", ovvero dei temi più complessi e articolati, che fanno riferimento a più concetti e, di conseguenza, possono essere descritti da parole apparentemente molto distanti tra loro, complicandone l'interpretazione. Quindi i topic models ottenuti dai dati, in positivo, sembrano capaci di poter individuare, piuttosto discretamente, degli autentici "spazi di significato" al-

l'interno del corpus di testi. In questo modo sembrano emergere dai trascritti dei *clusters* di parole semanticamente coerenti, che ne mettono in luce alcuni contenuti caratterizzanti e piuttosto specifici. Tuttavia più i topics diventano specifici e concettualmente complessi, più una conoscenza approfondita del corpus diventa necessaria per la loro comprensione. Bisogna, però, rilevare il fatto che il corpus a disposizione è di piccole dimensioni e che ruota attorno ad un unico tema centrale, l'affido, il che potrebbe aver influenzato i risultati. Un eventuale ricerca a questo riguardo potrebbe prevedere l'implementazione di topic models su dati di dimensioni maggiori, ad esempio provenienti da trascritti di gruppi più longevi di quelli considerati in questa sede e che non abbiano uno specifico tema di riferimento. Un'altra linea di ricerca potrebbe riguardare, invece, lo sviluppo di topic models capaci di fornire a ciascun topic un contesto semantico per migliorarne l'interpretazione, come, ad esempio, parole che aiutino la disambiguazione del significato di ciascuna parola presente nel topic, cosicché si riduca il problema interpretativo a carico del ricercatore.

Bibliografia

Arun, Rajkumar et al. (2010). «On finding the natural number of topics with latent dirichlet allocation: Some observations». In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer, pp. 391–402.

Atchison, J e Sheng M Shen (1980). «Logistic-normal distributions: Some properties and uses». In: *Biometrika* 67.2, pp. 261–272.

Benoit, Kenneth et al. (2018). «quanteda: An R package for the quantitative analysis of textual data». In: *Journal of Open Source Software* 3.30, p. 774. DOI: 10.21105/joss.00774. URL: https://quanteda.io.

Blei, David, Lawrence Carin e David Dunson (2010). «Probabilistic topic models». In: *IEEE signal processing magazine* 27.6, pp. 55–65.

Blei, David e John Lafferty (2006). «Correlated topic models». In: *Advances in neural information processing systems* 18, p. 147.

Blei, David M (2012). «Probabilistic topic models». In: *Communications of the ACM* 55.4, pp. 77–84.

Blei, David M e John D Lafferty (2009). «Topic models». In: *Text mining: classification, clustering, and applications* 10.71, p. 34.

Blei, David M, John D Lafferty et al. (2007). «A correlated topic model of science». In: *The Annals of Applied Statistics* 1.1, pp. 17–35.

Blei, David M, Andrew Y Ng e Michael I Jordan (2003). «Latent dirichlet allocation». In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Boyd-Graber, Jordan, David Mimno e David Newman (2014). «Care and feeding of topic models: Problems, diagnostics, and improvements». In: *Handbook of mixed membership models and their applications* 225255.

Cao, Juan et al. (2009). «A density-based method for adaptive LDA model selection». In: *Neurocomputing* 72.7-9, pp. 1775–1781.

Carpenter, Bob (2010). «Integrating out multinomial parameters in latent Dirichlet allocation and naive Bayes for collapsed Gibbs sampling». In: *Rapport Technique* 4, p. 464.

Chang, Jonathan et al. (2009). «Reading tea leaves: How humans interpret topic models». In: *Advances in neural information processing systems*, pp. 288–296.

Darling, William M (2011). «A theoretical and practical implementation tutorial on topic modeling and gibbs sampling». In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 642–647.

Deerwester, Scott et al. (1990). «Indexing by latent semantic analysis». In: *Journal of the American society for information science* 41.6, pp. 391–407.

Deveaud, Romain, Eric SanJuan e Patrice Bellot (2014). «Accurate and effective latent concept modeling for ad hoc information retrieval». In: *Document numérique* 17.1, pp. 61–84.

Gaut, Garren et al. (2015). «Content coding of psychotherapy transcripts using labeled topic models». In: *IEEE journal of biomedical and health informatics* 21.2, pp. 476–487.

Geman, Stuart e Donald Geman (1984). «Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images». In: *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 721–741.

Ghosh, Jayanta K, Mohan Delampady e Tapas Samanta (2007). *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media.

Grajzl, Peter e Peter Murrell (2019). «Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas». In: *Journal of Comparative Economics* 47.1, pp. 111–135.

Griffiths, Thomas L e Mark Steyvers (2004). «Finding scientific topics». In: *Proceedings of the National academy of Sciences* 101.suppl 1, pp. 5228–5235.

Grün, Bettina e Kurt Hornik (2011). «topicmodels: An R Package for Fitting Topic Models». In: *Journal of Statistical Software* 40.13, pp. 1–30. DOI: 10.18637/jss.v040.i13.

Hall, David, Dan Jurafsky e Christopher D Manning (2008). «Studying the history of ideas using topic models». In: *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 363–371.

Heinrich, Gregor (2009). *Parameter estimation for text analysis*. Rapp. tecn. Technical report.

Hofmann, Thomas (1999). «Probabilistic latent semantic indexing». In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57.

McFarland, Daniel A et al. (2013). «Differentiating language usage through topic models». In: *Poetics* 41.6, pp. 607–625.

Mimno, David et al. (2011). «Optimizing semantic coherence in topic models». In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 262–272.

Nikita, Murzintcev (2016). *ldatuning*. Ver. 1.0.2. URL: https://CRAN.R-project.org/package=ldatuning (visitato il 09/08/2020).

Ponweiser, Martin (2012). «Latent Dirichlet allocation in R». In:

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley et al. (2014). «stm: R package for structural topic models». In: *Journal of Statistical Software* 10.2, pp. 1–40.

Roberts, Margaret E et al. (2013). «The structural topic model and applied social science». In: *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. Vol. 4. Harrahs e Harveys, Lake Tahoe.

Roberts, Margaret E et al. (2014). «Structural topic models for open-ended survey responses». In: *American Journal of Political Science* 58.4, pp. 1064–1082.

Steyvers, Mark e Tom Griffiths (2007). «Probabilistic topic models». In: *Handbook of latent semantic analysis* 427.7, pp. 424–440.

Wallach, Hanna M., David M. Mimno e Andrew McCallum (2009). «Rethinking LDA: Why Priors Matter». In: *Advances in Neural Information Processing Systems 22*. A cura di Y. Bengio et al. Curran Associates, Inc., pp. 1973–1981. URL: http://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter.pdf.

Wallach, Hanna M et al. (2009). «Evaluation methods for topic models». In: *Proceedings of the 26th annual international conference on machine learning*, pp. 1105–1112.

Wang, Huijun et al. (2011). «Finding complex biological relationships in recent PubMed articles using Bio-LDA». In: *PloS one* 6.3.

Wijffels, Jan e Xiaohui Yan. «BTM: Biterm Topic Models for Short Text. R package». Ver. 0.3.1. In: ().

Wu, Yonghui et al. (2012). «Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation». In: *Biocomputing 2012*. World Scientific, pp. 422–433.

Yan, Xiaohui et al. (2013). «A biterm topic model for short texts». In: *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445–1456.

Yin, Zhijun et al. (2011). «Geographical topic discovery and comparison». In: *Proceedings of the 20th international conference on World wide web*, pp. 247–256.

Appendice A

Codice R utilizzato per i grafici

Codice A.1: Grafico perplexity

```
perpl<-data.frame("k"=k,"perplexity"=perplexT, stringsAsFactors = F)

library(ggplot)

ggplot(perpl, aes(x = k, y = perplexity)) +
geom_point() +
geom_smooth(se = FALSE) +
labs(x = "Numero topics", y = "Perplexity")+
theme_bw()
```

Codice A.2: Grafico proporzioni topics nei documenti

```
#gruppo 1
colori <- c("#E69F00", "#56B4E9", "#009E73", "#CC79A7", "#D55E00", "#
    b8b69c","#2700d6","#d60000","#88d600","#a800d6","#14345e","#00d65d","#
    cbd600","#854f34","#5e1435")
legenda<-c("Topic 1","Topic 2","Topic 3","Topic 4","Topic 5","Topic 6","
    Topic 7","Topic 8","Topic 9","Topic 10","Topic 11","Topic 12","Topic
    13","Topic 14","Topic 15")

gamma1<-lda_model1@gamma #matrice documenti x topics
rownames(gamma)<-lda_model1@documents
```

```
par(mar=c(5.1,4.1,4.1,10.1))
barplot(t(gamma),col = colori)
legend("right", legend = legenda, fill = colori, inset = c(-0.27,0), xpd =
    TRUE, mar(c(7,7,7,7)), bty = "n")

#gruppo 2
colori <- c("#E69F00", "#56B4E9", "#009E73", "#CC79A7", "#D55E00", "#
    b8b69c","#2700d6","#d60000","#88d600","#a800d6","#14345e","#00d65d","#
    cbd600","#854f34")
legenda<-c("Topic 1","Topic 2","Topic 3","Topic 4","Topic 5","Topic 6","
    Topic 7","Topic 8","Topic 9","Topic 10","Topic 11","Topic 12","Topic
    13","Topic 14")

gamma2<-lda_model2@gamma #matrice documenti x topics
rownames(gamma)<-lda_model2@documents

par(mar=c(5.1,4.1,4.1,10.1))
barplot(t(gamma),col = colori)
legend("right", legend = legenda, fill = colori, inset=c(-0.27,0), xpd=
    TRUE, mar(c(7,7,7,7)), bty="n")
```

Codice A.3: Grafico confronto contenuto tematico

```
plot.STM(stm_contenuto, type = "perspectives", topics = c(4), text.cex =
    1.5)
plot.STM(stm_contenuto, type = "perspectives", topics = c(8), text.cex =
    1.5)
plot.STM(stm_contenuto, type = "perspectives", topics = c(17), text.cex =
    1.5)
plot.STM(stm_contenuto, type = "perspectives", topics = c(18), text.cex =
    1.5)
```