Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in

Scienze Statistiche

Supervised or structural topic modelling: un'analisi di podcast su Spotify

Relatore: Prof. Emanuele Aliverti

Dipartimento di Scienze Statistiche

Laureando: Gianluca Ivo Tori

Matricola N. 2055680

Anno Accademico 2022/2023

Alla mia famiglia.

Indice

Introduzione

Nel corso degli ultimi decenni, il progressivo accrescimento dell'interazione tra essere umano e digitalizzazione, dovuto all'amplificazione dei mezzi di comunicazione come *social media* e piattaforme *streaming*, ha sottolineato il bisogno di sviluppare strumenti avanzati capaci di comprendere la complessità della comunicazione umana. A tal proposito, nasce l'*Elaborazione del Linguaggio Naturale* (NLP, dall'acronimo inglese *Natural Language Processing*), una disciplina che si pone come obiettivo quello di interpretare, emulare e sintetizzare il linguaggio umano. La maggior parte dei consumatori odierni vi ha probabilmente interagito senza esserne a conoscenza. Ad esempio, essa è alla base del funzionamento di assistenti virtuali come Siri o Alexa: quando un utente pone una domanda a questi assistenti virtuali, le tecniche di NLP permettono non solo di comprenderne la richiesta, ma anche di formulare una risposta adatta. Le tecniche di NLP sono anche applicate al fine di automatizzare la sintesi di una vasta raccolta di documenti di testo quali report aziendali, articoli di notizie o documenti scientifici.

Queste tecniche vengono spesso affiancate dal *Text Mining*. Quest'ultimo tratta dell'applicazione di tecniche di analisi dei dati per estrarre informazioni o conoscenza da grandi quantità di testo non strutturato, mentre l'NLP si concentra sulla comprensione del linguaggio, il *Text Mining* si focalizza sull'identificazione di schemi o relazioni nei dati testuali. Ad esempio, il *Text Mining* può essere utilizzato per identificare argomenti comuni in un grande insieme di documenti o per estrarre relazioni tra entità menzionate in un testo. In base a quanto detto, si potrebbe sostenere che il *Text Mining* sia un'applicazione o un sottoinsieme dell'NLP, poiché spesso utilizza tecniche di NLP per elaborare e analizzare i dati. Tuttavia, il *Text Mining* può anche avvalersi di altre tecniche, come quelle di *machine learning* o analisi statistica, che vanno oltre il tradizionale ambito dell'NLP.

Nel seguente elaborato vengono presentate tecniche di NLP e *Text Mining* applicate a dati relativi a podcast di Spotify, in modo da ottenerne una catalogazione automatica. I podcast

rappresentano una vasta raccolta di file audio digitale, di natura seriale o a episodi, distribuiti sul web e fruibili da dispositivi portatili in modalità online o offline. La recente riscoperta di questa forma di intrattenimento ha reso le raccolte di podcast un insieme di documenti in continua espansione e di notevole complessità: rispetto ad altri dati di tipo audio provenienti da trasmissioni TV o audiolibri, i podcast risultano più variegati nello stile, formato, genere e tematiche trattate. Questi vengono poi trascritti tramite riconoscimento vocale automatico: in particolare, per l'estrapolazione dei dati qui presentati è stato utilizzato *Google's Cloud Speech-to-Text API* che converte dati audio in dati testuali trascrivendo in modo automatico conversazioni registrate fornendo in aggiunta una specifica formattazione e punteggiatura. Si tratta, dunque, di una raccolta di documenti affascinante ma rumorosa e complessa. Congiuntamente alle trascrizioni, vengono utilizzati riassunti forniti dagli autori stessi dei podcast, in modo da integrarli nella ricerca di argomenti latenti all'interno delle diverse puntate di podcast analizzate. Si vuole quindi analizzare il comportamento dei diversi episodi a disposizione, con particolare attenzione agli argomenti che essi trattano, utilizzando sia le informazioni presenti nelle trascrizioni, sia quelle contenute nei rispettivi riassunti.

Le metodologie presentate fanno parte dei cosiddetti *Topic Model*, ossia modelli probabilistici che attraverso l'analisi delle parole caratterizzanti nei testi, individuano gli argomenti trattati e le loro connessioni; questo tipo di modelli non necessita di alcuna annotazione manuale, i temi emergono direttamente dall'analisi permettendo di archiviare e classificare ciò che sarebbe umanamente impossibile. Tali modelli possono essere utilizzati in diverse forme, supervisionati nei quali viene introdotta una variabile risposta, e strutturati nei quali vengono introdotte delle variabili esplicative. Particolare attenzione sarà rivolta alla struttura probabilistica sottostante, definendo le dipendenze statistiche tra le variabili in gioco cercando di evidenziare il legame tra le trascrizioni dei diversi episodi e il loro riassunto.

Capitolo 1

Definizione di analisi testuale e problema in analisi

L'interesse nei confronti dei dati testuali ha subito una crescita significativa nel ventunesimo secolo. Ciò è dovuto all'incremento della facile reperibilità di tali dati. Essi presentano interessanti spunti di ricerca, tra i quali ottenere delle sintetizzazioni o catalogazioni automatiche. L'analisi di questo tipo di dato complesso presenta una vasta gamma di approcci, uno dei quali è la caratterizzazione in argomenti, che però dev'essere attuata su un insieme di dati che porti con sé la maggior quantità di segnale possibile.

I dati testuali non detengono uno schema costante di elaborazione preliminare, visualizzazione e modellazione. L'elaborazione preliminare risulta di fondamentale importanza per le analisi successive. Infatti la frequenza delle unità di base, ovvero le parole, all'interno delle osservazioni, i documenti, rappresenta una delle caratteristiche più importanti. La rimozione di parole troppo frequenti che non rappresentano argomenti sottostanti i documenti, ad esempio, permette di risaltare le parole più esplicative per un determinato argomento. Nel presente contesto vi è anche un problema per ciò che riguarda la visualizzazione dei dati, dovuto alla loro elevata dimensionalità. Le osservazioni possono essere visualizzate in uno spazio latente, definendo delle opportune trasformazioni che permettano la loro proiezione in una dimensione ridotta.

In questo capitolo viene fornita una contestualizzazione rispetto all'analisi di dati testuali, e del problema in analisi. Si presentano i dati utilizzati, come questi vengono ottenuti e le procedure di pulizia preliminare effettuate. In seguito, si mostra un'analisi preliminare al fine di valutare, tramite un approccio esplorativo, possibili aspetti descrittivi che categorizzino le

diverse osservazioni, sia indipendentemente sia condizionatamente ad alcuni raggruppamenti preliminari di cui si dispone. Si definisce poi una notazione generale con la quale si vuole descrivere tutte le quantità a disposizione. Infine, si presenta una visualizzazione dei dati a disposizione tramite tecniche di *embedding*, le quali permetteranno di proiettare su dimensioni latenti i vari episodi a disposizione.

1.1 Contestualizzazione dell'analisi su dati testuali

L'analisi di dati testuali è un insieme di tecniche linguistiche e statistiche che modellano e strutturano il contenuto informativo di testi. Questa consiste nell'estrazione di informazioni, tramite analisi lessicale volta a esaminare la frequenza delle parole per schematizzare, etichettare o sintetizzare un ampio insieme di documenti, tramite tecniche di *data mining*, come il raggruppamento di documenti simili, la visualizzazione di documenti in uno spazio appropriato e la previsione di tendenze future. In letteratura si ha una prima citazione di analisi di dati testuali intorno agli anni 80' (Walker, 1982). Con analisi di dati testuali ci si riferisce all'applicazione di analisi del testo volte a rispondere a problemi, indipendentemente o in combinazione con l'analisi di altri tipi di dati. La maggior parte delle informazioni al giorno d'oggi hanno origine in forma non strutturata, tra cui quelle testuali. Queste tecniche permettono di evidenziare e scoprire informazioni, che altrimenti rimarrebbero nascoste all'interno di documenti scritti. Per affrontare un'analisi su dati testuali si ricorre prevalentemente a due tipi di tecniche: il *Natural Language Processing* e il *Text Mining*.

Con *Natural Language Processing* si fa riferimento a quella branca dell'intelligenza artificiale che si concentra sulla comprensione, interpretazione e generazione del linguaggio umano da parte delle macchine. L'NLP si occupa di problemi come la traduzione automatica, il riconoscimento del discorso e la generazione di testo, il cui obiettivo è consentire alle macchine di comprendere e rispondere al linguaggio umano in modo che possano eseguire compiti specifici o interagire con gli esseri umani in modo più naturale.

Il *Text Mining* si riferisce al processo di estrazione di informazioni da testi, le quali vengono generalmente ottenute tramite la stima di modelli statistici appropriati. Il *Text Mining* utilizza diverse tecniche, che spaziano dall'analisi del linguaggio naturale alla statistica, per identificare e studiare schemi e relazioni in dati testuali, permettendo così di estrarre informazioni significative

o conoscenze da collezioni di testi non strutturati. Il *Text Mining* è pertanto l'applicazione di algoritmi volti alla scoperta di schemi significativi in un campione di dati testuali (Broder, 1997). Le operazioni più comuni che fanno riferimento al *Text Mining* riguardano la classificazione dei testi, il raggruppamento di contenuti, l'identificazione di concetti latenti, la *sentiment analysis* e la sintesi di documenti.

1.2 Dataset Podcast Spotify

I dati utilizzati nel presente elaborato sono stati presentati in Clifton et al. (2020) e rappresentano le trascrizioni di alcune puntate di podcast su Spotify, ottenute tramite *Google's Cloud Speech-to-Text API*. A questi dati testuali si aggiungono anche dei metadati, i quali includono il nome dell'episodio, il nome dello spettacolo, la descrizione dell'episodio, l'autore, la durata e l'intestazione RSS. Di queste informazioni si pone maggiore attenzione alle descrizioni degli episodi, anch'esse di tipo testuale, le quali rappresentano un riassunto fornito dall'autore del podcast stesso, per ogni episodio trascritto. Una prima importante distinzione risiede nei due diversi testi a disposizione: le trascrizioni e i riassunti. Le prime risultano essere dati testuali rumorosi in quanto trascrizioni di dati audio effettuate da un *API*. Essi possono, inoltre, presentare un rumore aggiuntivo all'interno delle conversazioni stesse, come: frasi colloquiali, saluti, presentazioni e molte altre parti di una puntata nella quale il focus non è prettamente sull'argomento trattato. Il secondo insieme di dati a disposizione, i riassunti, sono forniti dall'autore direttamente in formato testuale, presumendo che risultino più diretti e concisi rispetto all'argomento trattato. Ci si aspetta inoltre che episodi appartenenti al medesimo *show* mostrino strutture e terminologie simili tra loro. Nel caso presente si hanno a disposizione 114 episodi relativi a 28 differenti *show* ognuno dei quali presenta almeno 3 episodi.

1.2.1 Natural Language Processing

I dati a disposizione necessitano di una fase di elaborazione preliminare. Data la diversa natura dei due dati testuali a disposizione, riassunti e trascrizioni, si è deciso di operare diversamente a seconda del contesto. Si procede quindi utilizzando alcune tecniche di *Natural Language Processing*, al fine di estrarre la maggior quantità di segnale possibile dai dati a disposizione.

Per prima cosa, sono state eliminate le *stopwords* presenti nella lingua inglese, le quali sono uguali per entrambi i testi. Le *stopwords* comprendono tutti quei termini utilizzati comunemente nel linguaggio e presenti in tutti i testi, come articoli, congiunzioni e pronomi. Inoltre, si rimuovono nomi propri di persona, che in alcuni contesti possono essere associati all'autore dello *show*, anche questa operazione è analoga in entrambi i testi. Vengono poi fatte delle considerazioni specifiche sulla rimozione di termini a seconda del testo considerato. Per i riassunti, si vogliono rimuovere le varie sponsorizzazioni dei siti web appartenenti ai vari *show*, o altre parole che riassumono lo *show* in generale. In alcuni casi sono stati rimossi anche i nomi specifici degli *show*, come "*chompers*" o "*afterbuzztv*". Dalle trascrizioni vengono invece rimosse tutte quelle parole che possono essere esclamazioni, come ad esempio "*yeah*" o "*oh*", che sono comunque state trascritte. Questi termini sono stati eliminati perchè non aggiungono nessuna informazione riguardo l'argomento trattato in un episodio, contrariamente porterebbero notevoli problemi di stima vista l'alta frequenza con cui vengono utilizzati. Poiché le trascrizioni presentano un errore dovuto alla modalità con cui esse vengono ottenute, infatti per quanto sofisticata l'API utilizzata presenta un margine di errore nel trasformare un dato audio in uno testuale, si decide di non modificare ulteriormente la semantica al loro interno.

	Parole	**Parole uniche**
Trascrizioni	123529	14802
Riassunti	2177	1069

Tabella 1.1: Numero di parole presenti all'interno dei due insiemi.

Contrariamente per i riassunti, che rappresentano una sintetizzazione delle trascrizioni stesse, si decide di attuare delle ulteriori operazioni di elaborazione, al fine di renderli il più esplicativi possibili. Si attua quindi ai soli riassunti la lemmatizzazione di ogni parola, la quale consiste nell'identificazione del lemma di una parola in base al significato che essa assume in una frase. In linguistica un lemma rappresenta la parte principale di una parola. Questa tecnica è molto legata allo *stemming*, che però opera su una singola parola senza conoscerne il contesto, la lemmatizzazione opera sulla parola valutandola rispetto al contesto in cui questa compare. Infine, si cerca di identificare quei bigrammi maggiormente diffusi all'interno dei riassunti. Un bigramma è rappresentato da una coppia di parole che compaiono l'una vicina all'altra. L'identificazione dei bigrammi viene attuata in base alla loro frequenza, nel caso presente, sono

stati identificati quei bigrammi con una frequenza minima pari a 5. Nel seguito dell'elaborato nell'indicare una parola si fa riferimento anche ai bigrammi identificati.

Si ottengono così due diversi insiemi di dati testuali. Come riportato in Tabella 1.1, i riassunti dei diversi episodi presentano un numero considerevolmente basso di parole se confronati con le relative trascrizioni. Questa caratteristica risulta naturale, in quanto i riassunti dovrebbero rappresentare una sintetizzazione degli episodi, quindi delle trascrizioni.

1.3 Analisi preliminari

Vengono presentate delle analisi preliminari, nelle quali si valutano le frequenze delle parole nei due diversi testi a disposizione. Queste operazioni vogliono evidenziare eventuali analogie e differenze per i due tipi di testi considerati.

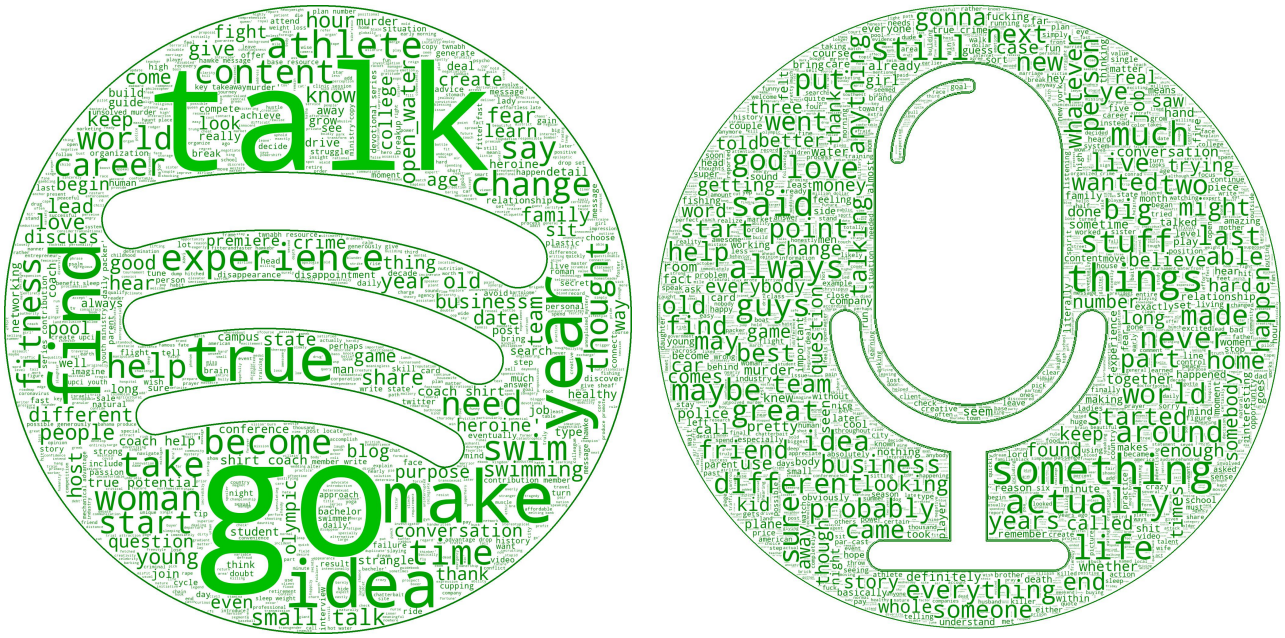

Figura 1.1: *Worldcloud*, (Sinistra) dei riassunti (Destra) delle trascrizioni.

Si vogliono quindi valutare quali possano essere alcuni degli argomenti trattati, in Figura 1.1 vengono mostrate le *worldcloud* ottenute dopo la fase di elaborazione: a sinistra quella relativa ai riassunti, mentre a destra quella riguardante le trascrizioni. Si può notare come la *worldcloud* inerente ai riassunti presenti tematiche come: sport ("*athlete*", "*fitness*"), dibattiti ("*talk*", "*content*") o imprenditoriali ("*business*", "*career*"). La *worldcloud* associata alle trascrizioni sembra invece evidenziare tematiche più generiche come la condivisione di esperienze

che possano essere in qualche modo d'aiuto ad altri ("*life*", "*love*", "*help*"). Da questa prima visualizzazione le trascrizioni sembrerebbero essere più vaghe rispetto ai riassunti.

Si distingue una maggior densità di parole nella *worldcloud* pertinente alle trascrizioni, questo è dovuto alla maggior quantità di parole che esse contengono. Questo si riscontra anche in Tabella 1.2, la quale mostra la frequenza di parole presenti all'interno dei diversi testi. Si evidenzia da subito una notevole differenza tra le il numero di parole medie presenti nelle trascrizioni e quelle presenti nei riassunti, anche per il numero minimo di parole vi è una spiccata differenza, in quanto il riassunto più corto, dopo le varie procedure di elaborazione preliminare, presenta solo una parola. Infatti nella fase di elaborazione preliminare, avviene l'eliminazione delle diverse sponsorizzazioni, e l'individuazione di bigrammi frequenti, portando ad ottenere un numero molto ridotto di parole.

	Minimo	**1o Qu.**	**Mediana**	**Media**	**3o Qu.**	**Massimo**
Trascrizioni	71.0	326.2	707.5	1083.6	1646.0	4078.0
Riassunti	1.0	6.0	11.5	19.1	18.7	147.0

Tabella 1.2: Conteggio delle parole per ogni episodio.

Queste rappresentazioni preliminari mostrano come i testi considerati siano considerevolmente diversi tra loro, in particolare come i riassunti siano molti brevi, ma mirati all'argomento trattato, mentre le trascrizioni siano più lunghe e dispersive. Nella fase di modellazione si cercherà di tenere conto di queste caratteristiche dei dati, proponendo modelli che possano incorporare la sintetizzazione presente nei riassunti e la gran quantità di parole presenti nelle trascrizioni.

1.3.1 Terminologia

Per analizzare dati testuali si fa riferimento a parole, documenti e *corpus*, nel caso presente con il termine documento si definisce una trascrizione o un riassunto di un episodio.

- Una *parola* è l'unità di base, definita come un elemento di un vocabolario, indicizzato da $\{1, ..., V\}$. Queste vengono rappresentate da vettori i quali contengono un elemento uguale a uno e tutti gli altri uguali a zero. La v-esima parola nel vocabolario è rappresentata da un vettore V-dimensionale w tale per cui $w_v = 1$ e $w_u = 0$ per $u \neq v$.

- Un *documento* è una sequenza di N_d parole denotate da $\mathbf{w}_d = (w_1, w_2, ..., w_{N_d})$, dove w_n è l'n-esima parola nella sequenza.

- Un *corpus* è una collezione di M documenti denotato da $\mathbf{D} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$.

Come detto precedentemente si dispone di due diversi tipi di testo per ogni episodio. Pertanto, si considerano due diversi corpora, uno relativo alle trascrizioni ed un altro relativo ai riassunti, entrambi di dimensione $M = 114$. La dimensione N_d di ogni documento varia anche all'interno del *corpus* stesso, tale distribuzione nei due corpora è mostrata in Tabella 1.2. Infine la dimensione dei due vocabolari utilizzati V è riconducibile al numero di parole uniche presentate in Tabella 1.1, pertanto, il vocabolario relativo ai riassunti presenta una dimensione $V = 1153$, mentre il vocabolario relativo alle trascrizioni presenta una dimensione $V = 15090$.

Viene inoltre presentata la misura di *term frequency-inverse document frequency*, con la quale si definisce la matrice di *tf-idf*. Questa trasformazione deriva da un'equazione che combina due quantità: la frequenza con cui una parola viene usata in un documento (tf), e la frequenza con cui tale parola appare in tutto il *corpus* (idf). La formulazione dell'indice *tf-idf* risulta la seguente

$$tf - idf(t, d) = tf(t, d) \times idf(t, d)$$

Nella quale *tf(t,d)* rappresenta la frequenza relativa della parola w_t nel documento $\mathbf{w}_d$, mentre il termine $idf(t, d)$ è esprimibile come

$$idf(t,d) = \log \frac{M}{1 + df(d,t)}$$

in cui M è il numero di documenti all'interno del *corpus* e *df(t,d)* è il numero di documenti d che contengono il termine t. Un'alternativa frequentemente utilizzata è quella di definire l'*idf* nel seguente modo

$$idf(t, d) = \log \frac{1 + M}{1 + df(d, t)},$$

e ridefinire la trasformazione di *tf-idf* come

$$tf - idf(t,d) = tf(t,d) \times (idf(t,d) + 1).$$

Nel caso presente viene applicata questa seconda formulazione, la quale aiuta ad assegnare un peso pari a zero ai termini che possono ricorrere in tutto il documento. Una volta trovati i *tf-idf* con la precedente procedura, può risultare opportuna una loro normalizzazione.

$$tf - idf(t,d)_{norm} = \frac{tf - idf(t,d)}{\|tf - idf(i,d)\|} = \frac{tf - idf(t,d)}{(\sum_{i=1}^{N_d}(tf - idf(i,d))^2)^{(1/2)}}.$$

La matrice risultante da questa trasformazione rappresenta la matrice di $tf - idf$ normalizzata che viene indicata con X_{tf-idf}, in cui in riga vengono rappresentati i documenti ed in colonna le parole, dunque l'elemento relativo all'i-esima riga della j-esima colonna di tale matrice risulta essere $x_{i,j} = tf - idf(j,i)_{norm}$.

1.3.2 Analisi esplorativa

Si vuole ora cercare di confrontare i due corpora a disposizione. Per farlo si ricorre alle frequenze relative delle parole all'interno di ciascun documento. In particolare, si rappresenta con N_d la dimensione di un documento, si confrontano quindi le seguenti quantità $Freq_d = N_d / \sum_{j=1}^{M} N_j$ di entrambi i corpora. Questa operazione preliminare viene attuata al fine di evidenziare, per via esplorativa, eventuali proporzionalità tra i due corpora. In un primo momento si valutano direttamente i documenti dei due corpora ed in seguito condizionatamente agli *show*. Quest'ultimi rappresentano una categorizzazione preliminare dei documenti, durante le analisi si assume però che per un medesimo *show* vi possano appartenere molteplici argomenti, come più *show* possano trattatare lo stesso argomento.

La statistica $Freq$ permette quindi di confrontare i due corpora, nonostante la distribuzione delle parole nei documenti risulti essere nettamente diversa, come mostrato in Tabella 1.2. Si vuole valutare se la lunghezza delle trascrizioni influisca sulla stesura dei riassunti da parte degli autori, una volta registrata una puntata. Inoltre, si cerca di comprendere se tutti i documenti di un medesimo *show* abbiano la stessa dimensione all'interno dei diversi corpora. Per confrontare le $Freq$ dei corpora si ricorre ad un grafico di dispersione, il quale proietta ogni unità nello

spazio identificato dalle *Freq* del relativo *corpus* di appartenenza. Il grafico di dispersione mette a confronto le *Freq* su entrambi i corpora considerati, ogni documento è poi colorato in base allo *show* di appartenenza.

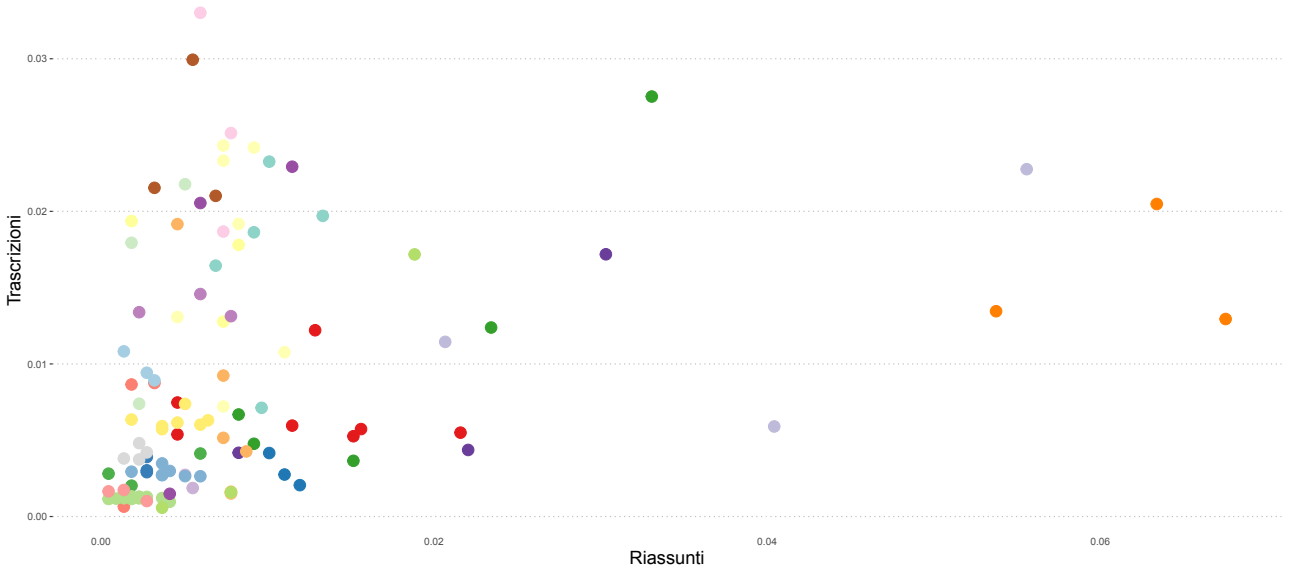

Figura 1.2: Frequenza relativa di termini all'interno dei diversi episodi, colorati per *show*.

Come mostrato in Figura 1.2, non si evidenzia una costante proporzionalità tra i due corpora. In certe situazioni per una singola unità i documenti hanno lo stesso peso all'interno del *corpus*. In altre circostanze presentano, una maggior *Freq* i documenti appartenenti al *corpus* dei riassunti ed in altri a quello delle trascrizioni. Questa diversa proporzionalità mostra una prima importante differenza tra i due corpora, ossia che gli autori pur conducendo episodi particolarmente brevi o lunghi, non tengano conto della lunghezza della puntata nella stesura del riassunto. In altre circostanze, però, gli autori risultano essere prolissi nella stesura di una sintetizzazione, che per quanto debba essere breve, si presenta in proporzione anche maggiore alle stesse trascrizioni.

Si vuole ora considerare il numero di parole condizionatamente allo *show* di appartenenza. Poiché per ogni *show* si hanno a disposizione un minimo di 3 episodi, le *Freq* vengono rappresentate tramite dei punti, allineati per *show* di appartenenza. Come mostrato in Figura 1.3, si possono visualizzare le *Freq* condizionatamente ai vari *show*, in verde dei riassunti, in nero delle trascrizioni. Si sottolinea come le trascrizioni presentino una maggior variabilità rispetto ai relativi riassunti. Questo potrebbe derivare anche dalle diverse lunghezze degli episodi asso-

ciati al medesimo *show*. Pertanto, i riassunti sembrerebbero contenere un'informazione meno variabile, condizionatamente allo *show* di appartenenza.

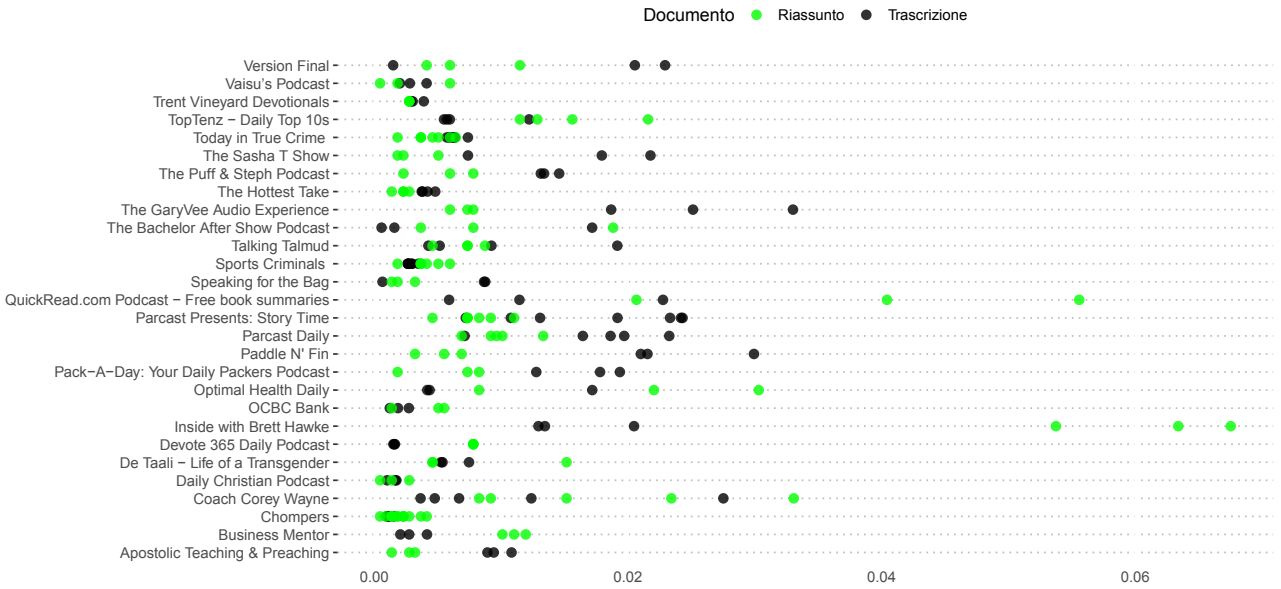

Figura 1.3: Frequenza relativa di termini all'interno dei diversi *show*.

Questa diversa proporzionalità mostra una prima importante differenza tra i due corpora, in particolare sottolinea un effetto autore, secondo cui alcuni tendano a fare riassunti molto prolissi ed altri invece siano più coincisi. Si nota una scarsa variabilità tra i riassunti appartenenti allo stesso *show*, come ad esempio "*Chompers*","*Bussines Mentor*", mentre *show* con un elevata variabilità, sono rappresentati da lunghezze molto elevate.

Per quanto riguarda le *Freq* del *corpus* dei riassunti, si nota come vi siano alcuni *show* anomali, in cui si evidenziano chiaramente alcune osservazioni molto estreme, come ad esempio "*Inside with Brett Hawke*". Si potrebbe pensare che un autore tenda a fornire riassunti simili tra loro. Contrariamente la gestione delle parole trascritte non è sempre costante. Questo potrebbe essere dovuto anche alla presenza di eventuali ospiti, o interventi esterni.

Dai nomi dei diversi *show* si possono inoltre evidenziare alcune sotto aree tematiche tra cui quella religiosa ("*Daily Christian Podcast*", "*Apostolic Teaching & Preaching*"), storie criminali ("*Today in True Crime*"), attualità ("*Top Tenz - Daily Top 10s*"), lettura di libri ("*QuickRead.com Podcast - Free book summaries*"), sport ("*Paddle N'Fin*"), diritti ("*De Taali - Life of a Transgender*") e strategie di mercato ("*Bussines Mentor*, "*OCBC Bank*").

Infine viene proposta una visualizzazione delle matrici X_{tf-idf} per entrambi i corpora a disposizione, con l'intento di valutarne la sparsità, tramite un'analisi grafica nella quale vengono confrontate le relative *heatmap*.

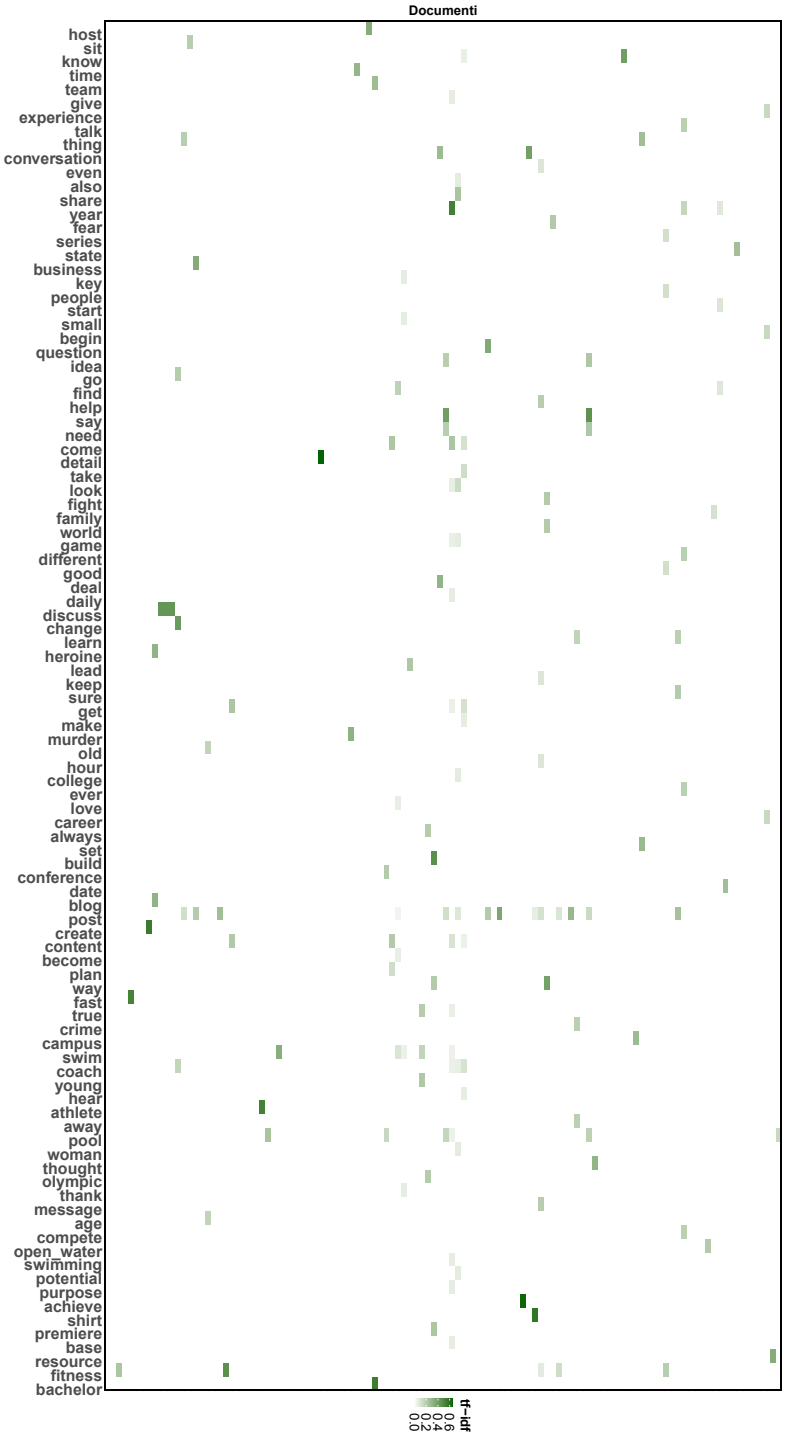

Figura 1.4: *Heatmap* relativa alla matrice *tf-idf* del *corpus* dei riassunti.

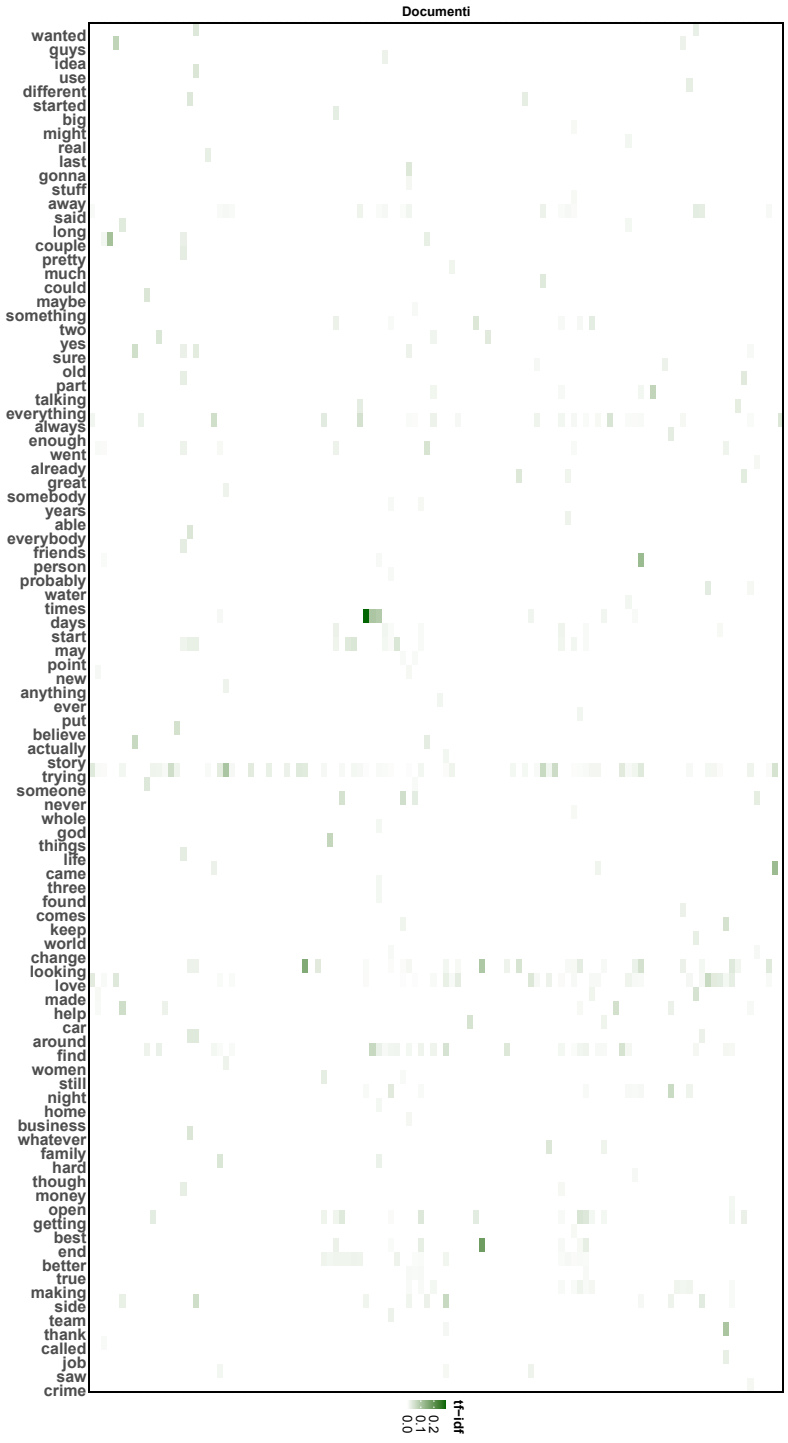

Figura 1.5: *Heatmap* relativa alla matrice *tf-idf* del *corpus* delle trascrizioni.

Rispettivamente in Figura 1.4, viene mostrata la *heatmap* relativa al *corpus* dei riassunti, mentre in Figura 1.5 quella relativa al *corpus* delle trascrizioni. Vengono riportate le 100 parole più frequenti all'interno dei corpora. In ogni colonna sono rappresentati i documenti, mentre in riga le parole.

La *heatmap* dei riassunti mostra una sparsità più elevata, evidenziata dalla maggior presenza di celle completamente bianche. Questa contiene valori di $tf - idf$ notevolmente alti, con un massimo superiore a 0.6, mentre la *heatmap* relativa al *corpus* delle trascrizioni, presenta un massimo di $tf - idf$ di appena 0.2. Questo potrebbe essere dovuto alla dimensione ridotta dei documenti, N_d, a cui fa riferimento il *corpus* relativo ai riassunti. Infatti l'indice $tf - idf$ tiene conto della dimensione dei documenti per calcolare l'importanza relativa di una parola all'interno di una frase.

I due corpora in esame presentano tra loro alcune differenze. In particolare, il *corpus* dei riassunti non risulta sempre proporzionale al corrispettivo delle trascrizioni. In alcune circostanze una trascrizione più breve non implica una sintetizzazione breve, o al contrario una trascrizione più lunga non implica una sintetizzazione lunga. Nei riassunti gli autori vogliono racchiudere le parti principali dell'episodio, senza dare troppe anticipazioni su tutti i dettagli, ma in certi casi risultano più che esaustivi. Infine, si evidenzia come vi sia una maggior sparsità nei documenti del *corpus* relativo ai riassunti rispetto a quello pertinente alle trascrizioni, in particolare come alcune parole in questo *corpus* assumano valori di $tf - idf$ notevolmente alti, mentre nel *corpus* delle trascrizioni vi siano valori più bassi, mentre si riduce la sparsità.

1.4 Embedding

I dati testuali a disposizione, nonostante le operazioni di pulizia, risultano ancora difficili da visualizzare. La difficoltà più grande risiede nell'elevata dimensionalità di questi: il numero di termini utilizzati è decisamente troppo grande per riuscire a visualizzare i documenti analizzati. Si presentano due tecniche di *embedding*, allo scopo di individuare eventuali raggruppamenti latenti dei diversi episodi considerati. Con *embedding* si intende una mappatura in uno spazio vettoriale latente. Nel caso presente si vogliono visualizzare i vari episodi in uno spazio bidimensionale. Vengono presentate tecniche che operano sulla matrice X_{tf-idf}, che come visto precedentemente risulta molto sparsa per entrambi i corpora.

1.4.1 Componenti principali

Un metodo frequentemente utilizzato è quello delle componenti principali: una trasformazione ortogonale lineare del sistema di coordinate, basata sulla varianza delle osservazioni. Più for-

malmente, le componenti principali di un insieme di dati in $\mathbb{R}^V$ forniscono una sequenza dei migliori approssimatori lineari, per tutti i ranghi $q \leq V$ (Friedman et al., 2001). Sia $x_1, ..., x_M$ un insieme di unità, il modello lineare di rango q che le rappresenta è definito

$$f(\lambda) = \mu + P_q \lambda$$

con μ vettore di posizione in $\mathbb{R}^V$, P_q matrice $V \times q$ a colonne ortogonali e λ vettore q-dimensionale di parametri. La matrice P_q è ottenibile da

$$\min_{P_q} \sum_{i=1} \|(x_i - \bar{x}) - P_q P_q^T (x_i - \bar{x})\|^2$$

(per convenienza si può assumere $\bar{x} = 0$). La matrice $H_q = P_q P_q^T$ è una matrice di proiezione $V \times V$, che mappa ogni elemento x_i nel sottospazio delle colonne di P_q, definendo la proiezione $H_q x_i$. La stessa soluzione è ottenibile attraverso la scomposizione in valori singolari di X_{tf-idf}:

$$X_{tf-idf} = UDP^T.$$

Per ogni rango q, la soluzione per Pq consiste nelle prime q colonne di P; le colonne di UD sono definite componenti principali di X_{tf-idf}.

Vengono quindi mostrati i diversi documenti proiettati sullo spazio latente individuato dalle prime due componenti principali. Nello specifico in Figura 1.6 sono rappresentate le prime due componenti principali, dei riassunti, mentre in Figura 1.7 delle trascrizioni. In questa fase preliminare non si è ritenuto di ottimizzare il numero di componenti da utilizzare in quanto il fine del loro utilizzo fosse prettamente grafico. La scelta delle prime due componenti risiede nel fatto che queste presentano la maggior variabilità all'interno dei dati, anche se, come si può vedere dagli assi nei quali è riportata anche la percentuale di variabilità spiegata, ogni componente non contenga piu del 4% della variabilità totale.

Si evidenzia come la dispersione apportata dalle componenti principali relative alle trascrizioni permetta di identificare, alcuni degli *show* di appartenenza di ciascun documento, mentre le prime due componenti principali relative al *corpus* dei riassunti tendano a schiacciare notevolmente la maggior parte dei documenti verso l'origine degli assi.

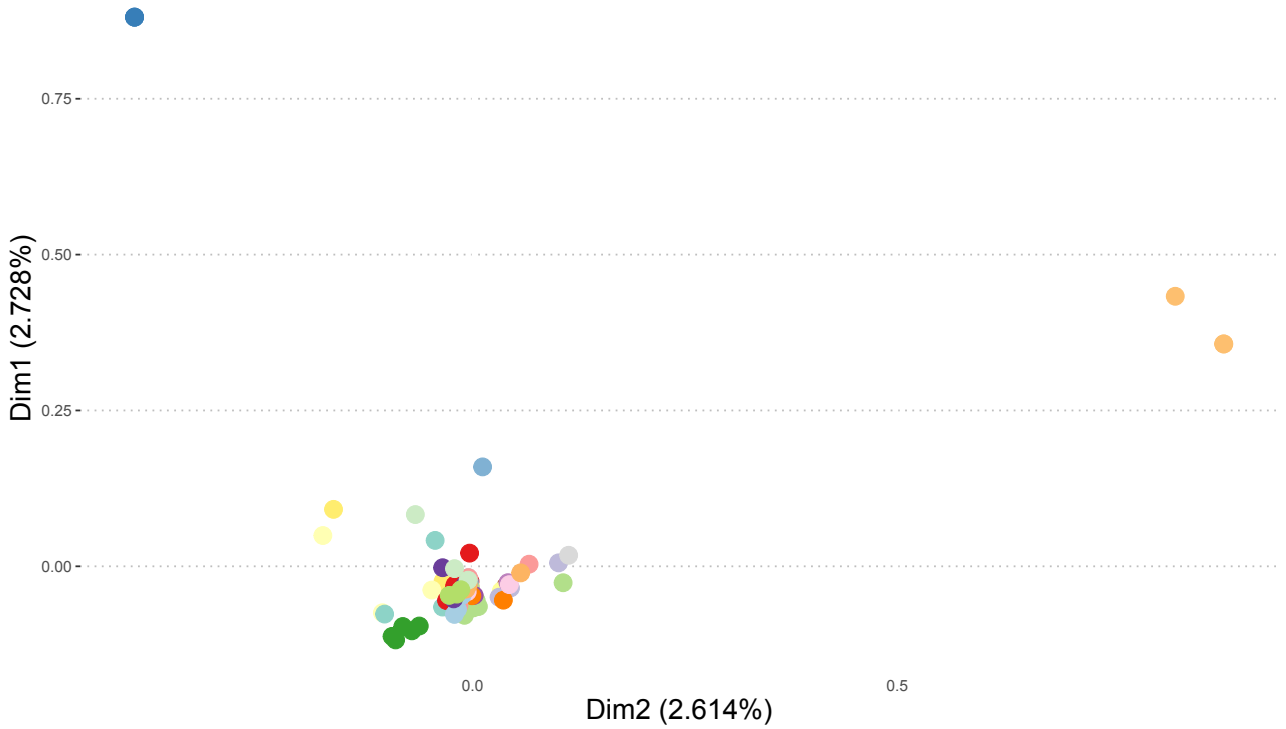

Figura 1.6: Prime due componenti principali calcolate sulla matrice *tf-idf* relativa al *corpus* dei riassunti.

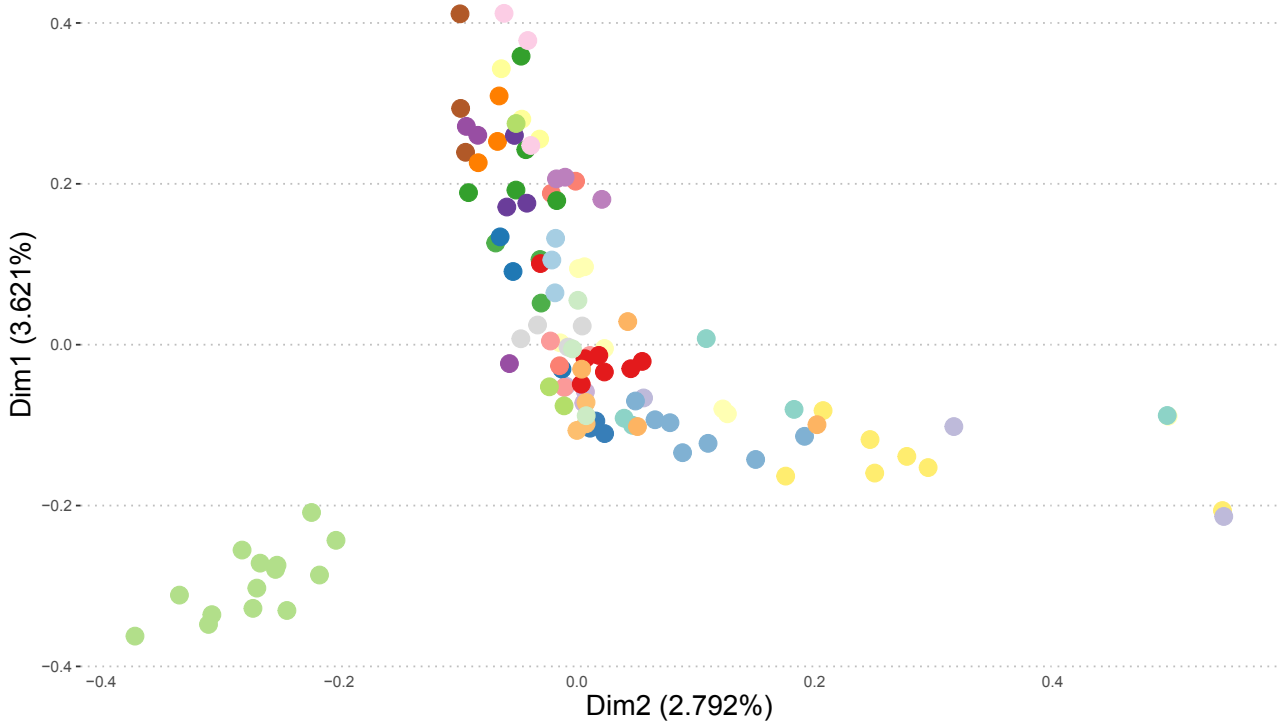

Figura 1.7: Prime due componenti principali calcolate sulla matrice *tf-idf* relativa al *corpus* delle trascrizioni.

1.4.2 T-distributed stochastic neighbor embedding

Un metodo alternativo è il t-SNE (*t-distributed stochastic neighbor embedding*), (Maaten & Geoffrey, 2008); lo scopo è sempre quello di definire un omeomorfismo $g : \mathbb{R}^V \longrightarrow \mathbb{R}^p$, con $p \ll V$, nel caso presente $p = 2$. Questa tecnica non-lineare consiste in due fasi:

- Si definisce una distribuzione sulle coppie di unità V-dimensionali, assegnando probabilità alta alle unità più simili, e viceversa probabilità bassa alle unità più diverse;
- Si definisce in modo analogo una distribuzione sui punti dello spazio p-dimensionale, e si minimizza la divergenza di Kullback-Leibler tra le due distribuzioni.

Siano $x_1, ..., x_V$ un insieme di unità V dimensionali. Si definisce somiglianza tra x_i e x_j come la probabilità condizionata, $p_{j|i}$, che x_i abbia x_j come punto più vicino, sapendo che x_j è stato estratto con probabilità proporzionale alla densità di una variabile casuale Gaussiana centrata in x_i. Per punti vicini, la probabilità $p_{j|i}$ è relativamente alta, mentre per punti distanti la probabilità è infinitesimale (assumendo di avere valori di σ_i ragionevoli). Matematicamente, la probabilità condizionata $p_{j|i}$ è definita come

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}$$

dove σ_i è la varianza della variabile casuale Gaussiana centrata in x_i. È possibile dunque specificare la funzione di probabilità congiunta di tutto lo spazio come

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2V}.$$

Ciò garantisce che per tutti i punti x_i $\sum_j p_{ij} > 1/2V$, di conseguenza ciascun punto x_i apporta un contributo significativo alla funzione di obbiettivo. Per contrapporre i punti nello spazio a bassa dimensionalità ϵ_i e ϵ_j con i punti ad elevata dimensionalità x_i e x_j, è possibile calcolare, in modo analogo, una probabilità condizionata che viene indicata con $q_{j|i}$. Pertanto, siano $\epsilon_1, ..., \epsilon_p$ un insieme di unità p-dimensionali; si definisce la somiglianza tra le unità ϵ_i e ϵ_j come probabilità congiunta

$$q_{ij} = \frac{(1 + \|\epsilon_i - \epsilon_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|\epsilon_k - \epsilon_l\|^2)^{-1}}.$$

Importante notare la differenza nella definizione della probabilità nei due spazi: la prima fa riferimento ad una densità Gaussiana, la seconda da una t di Student con un grado di libertà. Questa particolare scelta permette di evitare il cosiddetto *crowding*, cioè la tendenza dei punti dello spazio p-dimensionale ad accorparsi in una sfera; le code più pensanti della t di Student permettono di avere più repulsione tra punti maggiormente distanti nello spazio V-dimensionale. Una volta definite le due funzioni di somiglianza nei due spazi, è necessario che queste ultime definiscano le stesse distribuzioni di probabilità. Una scelta naturale per misurare la somiglianza in ambito probabilistico è la divergenza di Kullback-Leibler. La funzione da minimizzare, rispetto ai punti ϵ_i, risulta dunque

$$C = KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

dove P risulta essere la distribuzione di probabilità congiunta, nello spazio ad elevata dimensionalità, mentre Q rappresenta la distribuzione di probabilità congiunta nello spazio a bassa dimensionalità.

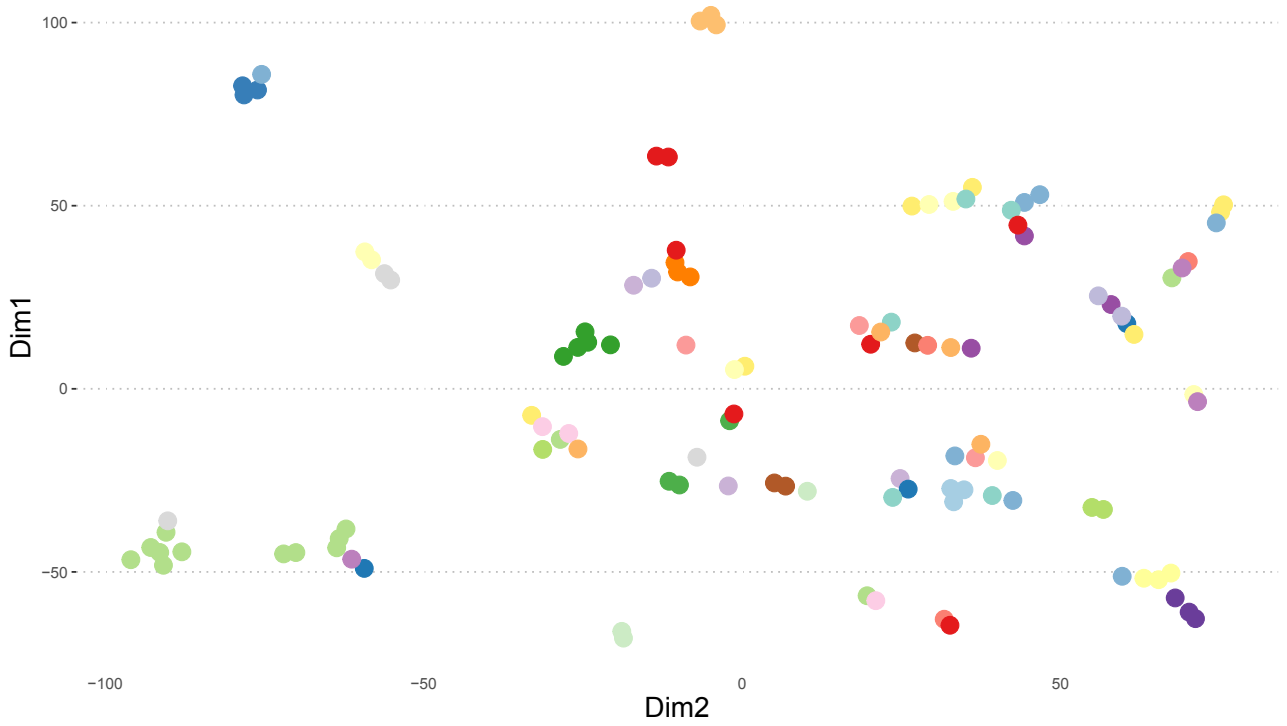

Figura 1.8: Dimensioni ottenute tramite tsne calcolate sulla matrice *tf-idf* relativa al *corpus* dei riassunti.

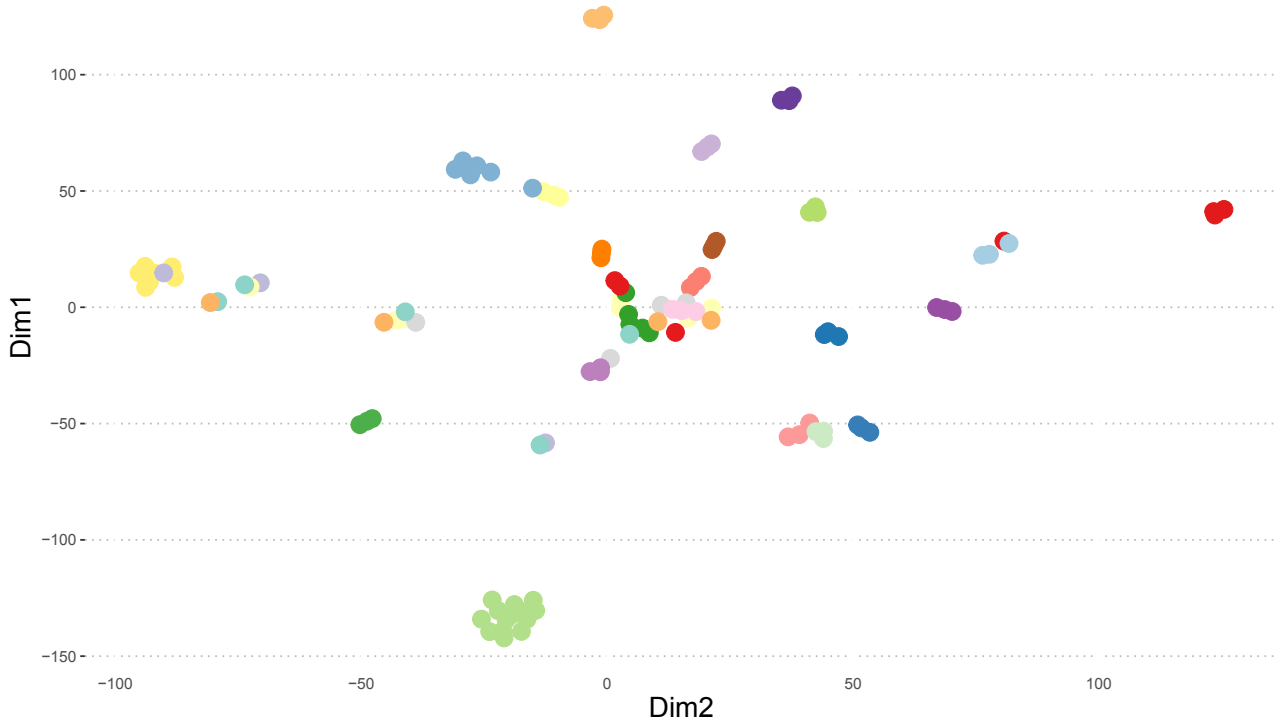

Figura 1.9: Dimensioni ottenute tramite tsne calcolate sulla matrice *tf-idf* relativa al *corpus* delle trascrizioni.

Vengono quindi riportate le proiezioni sulle dimensioni latenti identificate dalla tecnica di *t*-SNE. Nello specifico in Figura 1.8 e 1.9 vengono proiettati i diversi documenti in relazione al *corpus* di appartenenza, rispettivamente di riassunti e trascrizioni. Il secondo di questi permette di discriminare alcuni tra gli *show* a disposizione, indentificandoli in delineate zone del piano latente. Per ciò che riguarda lo spazio latente identificato sul *corpus* dei riassunti, sembra essere particolarmente difficile distinguere i diversi *show*. Questo conferma ancora una volta che l'elevata dimensionalità presente nel *corpus* delle trascrizioni possa contenere un'informazione più discriminante in ottica di modellazione degli argomenti.

Capitolo 2

Topic Model

L'obiettivo dell'elaborato consiste nell'estrapolazione di tematiche latenti sottostanti i podcast per facilitarne la catalogazione, utilizzando entrambi i corpora a disposizione. Si ricorre dunque ai *Topic Model*, i quali permettono di raggruppare documenti in base alla frequenza delle parole che li compongono. I *Topic Model* sono metodi statistici che analizzano le parole del testo per scoprire i temi che compaiono al loro interno, (Blei, 2012). L'idea di base di tali modelli risiede nel valutare quali documenti condividano la stessa semantica; infatti, se vi è un insieme di parole che presentano un'elevata frequenza in più documenti, questi metodi tendono a raggruppare tali documenti in dei gruppi tematici quali i *topic*. I *Topic Model* assumono distribuzioni su raccolte di documenti in cui ciascun documento è rappresentato come un insieme di variabili casuali, le quali sono parole. Nei *Topic Model* si trattano le parole di un documento come derivanti da un insieme di *topic* latenti, cioè da un insieme di distribuzioni sconosciute all'interno del vocabolario. I documenti in un *corpus* condividono lo stesso insieme di argomenti, ma ogni documento risulta essere una mistura di *topic*. In questo capitolo vengono presentati i *Topic Model* utilizzati durante l'analisi: il primo di questi è *Latent Dirichlet Allocation*, (Blei et al., 2003), il quale viene utilizzata sui due corpora separatamente. I risultati di questo primo modello vengono poi integrati con altri *Topic Model*, per ottenere dei *topic* che racchiudano le informazioni di entrambi i corpora. Successivamente viene presentata la *Supervised Latent Dirichlet Allocation*, (Blei & Mcauliffe, 2007), che è un *Topic Model* supervisionato, per il quale nel caso presente si utilizza come variabile risposta le etichette di *topic* identificati precedentemente dalla *Latent Dirichlet Allocation* sul *corpus* dei riassunti. Infine si presenta lo *Structural Topic*

Model, (Roberts et al., 2016), il quale è un modello che consente l'utilizzo di covariate a livello di *topic*, in particolare nelle analisi svolte queste covariate risultano essere le proporzioni di *topic* per ogni documento, stimate dalla *Latent Dirichlet Allocation* sul *corpus* delle trascrizioni.

Come detto in precedenza, il fine delle analisi è rivolto all'identificazione di aree tematiche in cui raggruppare i diversi episodi. Per farlo si vuole utilizzare l'informazione presente in entrambi gli insiemi di dati testuali di cui si dispone. La suddivisione delle analisi in due passi permette di utilizzare nella seconda parte dei *Topic Model* più complessi che ammettono l'utilizzo di informazioni esterne oltre ad un *corpus* di riferimento dal quale far emergere tematiche oggettive. I documenti a disposizione non risultano essere etichettati, per tale motivo si presenta la necessità di utilizzare un approccio in due fasi per la sLDA, in cui in una prima fase si individuano possibili gruppi latenti tramite la LDA, ed in seguito si utilizzano le etichettature derivati da questo metodo di raggruppamento come supervisione per sLDA. Questa scelta è giustificabile in quanto le uniche etichette disponibili sarebbero i titoli degli *show* di appartenenza dei diversi documenti, portando però ad un vasto numero di livelli della variabile risposta, che per ogni modalità presenterebbero un numero esiguo di documenti al loro interno. Tale procedura renderebbe, inoltre, più complesso il raggruppamento di *show* con nomi diversi, ma tematiche simili. In modo analogo lo *Structural Topic Model* richiede una procedura in due fasi, infatti questo modello consente l'utilizzo di covariate a livello di documento, ma l'elevata dimensionalità del vocabolario delle trascrizioni renderebbe intrattabile l'inversione della matrice del disegno in fase di stima. Per cui si effettua il primo passo sul *corpus* relativo alle trascrizioni, da cui se ne ricava una matrice di proporzioni, la quale viene utilizzata come matrice del disegno per l'inserimento di covariate a livello di documento. L'analisi viene, pertanto, divisa in due passi al fine di rendere possibile l'utilizzo delle informazioni presenti in entrambi i corpora, senza limitarsi all'utilizzo di solo uno di questi, sfruttando la struttura dei *Topic Model* più complessi stimati nel secondo passo.

2.1 Approccio all'analisi

Vi sono quindi due importanti fasi che caratterizzano la modellazione. La prima di queste consiste nella stima ed ottimizzazione della *Latent Dirichlet Allocation*, la quale fornisce delle informazioni che verranno utilizzate per le analisi successive. Queste riguardano, appunto, il

numero ottimale di *topic* latenti per ciascun *corpus* e, a seconda del contesto, una possibile etichettatura dei documenti, o una loro rappresentazione quantitativa. Queste quantità sono parte integrante della seconda fase. In particolare le etichette vengono riprese nella stima della *Supervised Latent Dirichlet Allocation*, mentre la rappresentazione quantitativa dei documenti, ovvero la rappresentazione delle proporzioni di *topic* all'interno dei documenti, viene utilizzata nella stima dello *Structural Topic Model*. Le procedure proposte, come detto precedentemente, implicano quindi due diverse fasi di modellazione. Nella prima viene applicata la *Latent Dirichlet Allocation*, i cui risultati vengono utilizzati nella seconda fase. Viene naturale sottolineare che, per quanto sia un approccio sensato, i dati ottenuti nella prima fase non sono realmente osservati, ma derivanti dalla stima di un primo modello.

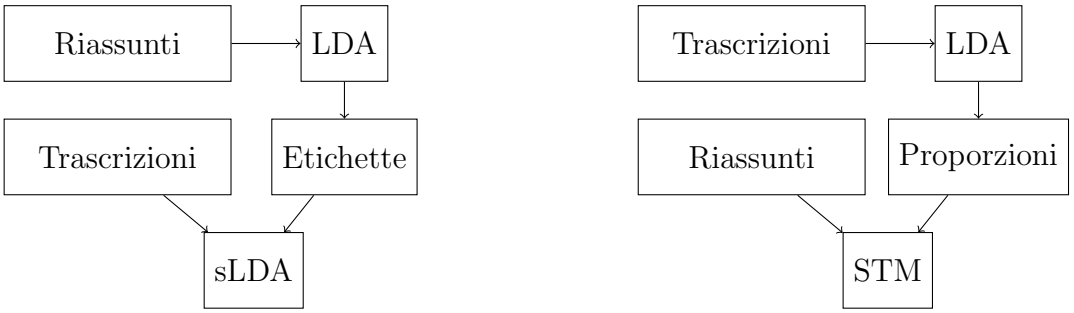

Figura 2.1: (Sinistra) Primo approccio di analisi utilizzando un modello di *topic* supervisionato. (Destra) Secondo approccio di analisi utilizzando un modello di *topic* strutturato.

Le motivazioni che hanno portato ad un tale approccio risiedono nel voler utilizzare le informazioni presenti nei due corpora, senza limitarsi all'utilizzo di solo uno di questi, sfruttando le strutture di *Topic Model* che oltre ad un *corpus* di riferimento permettano l'inserimento di altre quantità. In particolare la *Latent Dirichlet Allocation* è un modello per l'identificazione di gruppi latenti sottostanti i dati, questi gruppi però non sono confermati in nessun modo. Quindi si vuole valutare se le etichette per la *Supervised Lantent Dirichlet Allocation*, e le proporzioni per lo *Structural Topic Model*, possano in qualche modo migliorare i risultati ottenuti nella prima fase. Le fasi delle analisi svolte sono schematizzate in Figura 2.1, la quale mostra come i corpora entrino in gioco per i diversi modelli. In particolare, a sinistra, si può osservare il primo approccio all'analisi nel quale si stimano le etichette dei *topic* sul *corpus* dei riassunti, e con queste si stima la *Supervised Latent Dirichlet Allocation*, che le utilizza come variabile risposta. A destra invece, è visibile il secondo approccio utilizzato in analisi per il quale si evidenzia come in un primo momento vengano stimati i *topic* sul *corpus* delle trascrizioni, i quali presentano

le relative proporzioni per ogni documento. Questa matrice di proporzioni viene passata allo *Structural Topic Model* insieme al *corpus* dei riassunti. Come esplicitato precedentemente, il fine dell'analisi è quello di caratterizzare dei *topic* latenti all'interno dei podcast incorporando le informazioni presenti in entrambi i corpora a disposizione.

2.2 Contestualizzazione sui Topic Model

Nell'analisi testuale è comune modellare l'effetto dei diversi argomenti, o *topic*, all'interno di un documento considerando ogni *topic* come una distribuzione di probabilità sulle parole. In questo modo, un documento viene visto come una mistura di *topic*. Sia $p(z)$ la distribuzione di probabilità sui *topic* in un documento, e $p(w|z)$ la distribuzione di probabilità di una parola dato un *topic* z. Di conseguenza $p(z_i = j)$ è la probabilità che il j-esimo *topic* sia estratto per l'i-esima parola, mentre $p(w_i|z_i = j)$ rappresenta la probabilità della parola w_i sotto il *topic* j. Se vi sono K *topic* latenti, è possibile esprimere la probabilità della i-esima parola in un documento come:

$$p(w_i) = \sum_{j=1}^{K} p(w_i|z_i = j)p(z_i = j).$$

La distribuzione $p(w|z)$ può essere interpretata come le parole che sono rilevanti per un determinato argomento, mentre $p(z)$ rappresenta la prevalenza degli argomenti all'interno di un documento. Nei *Topic Model* i documenti vengono generati scegliendo prima una distribuzione sui *topic* $p(z)$, che determina le parole in quel documento. La stima del modello generativo diventa quindi un problema di massimizzazione rispetto ad una certa distribuzione, poichè si vuole stimare la distribuzione $p(w|z)$, ma la sua soluzione esatta è complessa e si ricorre spesso ad approssimazioni che fanno ricorso all'inferenza variazionale.

Si denoti con π l'insieme dei parametri e delle variabili latenti e con $\mathbf{O}$ i dati osservati. L'obiettivo dell'inferenza variazionale è quello di approssimare la distribuzione a posteriori $p(\pi|\mathbf{O})$. A tale scopo, si definisce una distribuzione variazionale $q(\pi)$ che minimizza la divergenza di Kullback-Leibler con la distribuzione a posteriori esatta. Si definisce allora una famiglia di densità ristrette per le quali se non ci sono restrizioni su $q(\pi)$ il minimo è raggiunto quando $q(\pi) = p(\pi|\mathbf{O})$, ma specificando una famiglia che contiene la distribuzione a posteriori si arriva ad un problema intrattabile, che viene risolto con la specificazione di una famiglia di

distribuzioni più semplice. Per ulteriori approfondimenti in merito alle procedure di inferenza variazionale si rimanda a Blei et al. (2017).

2.3 Latent Dirichlet Allocation

La *Latent Dirichlet Allocation* è un modello probabilistico generativo per insiemi di dati come quelli testuali. LDA presenta una struttura gerarchica a tre livelli, nei quali ogni documento è rappresentato come una mistura di *topic*, ognuno di questi ha una distribuzione Multinomiale. Nello specifico si assume il seguente processo generativo per ogni documento $\mathbf{w}_d$ in un *corpus* $\mathbf{D}$:

1. $\theta_d \sim Dirichlet(\alpha)$, con $d \in \{1, ..., M\}$
2. $\varphi_k \sim Dirichlet(\beta)$, con $k \in \{1, ..., K\}$
3. Per ogni parola w_i del documento d, con $i \in \{1, ..., N_d\}$:
   (a) Si estrae un *topic* $z_{i,d} \sim Multinomial(\theta_d)$.
   (b) Si estrae una parola $w_{i,d} \sim Multinomial(\varphi_{z_{i,d}})$.

Nel presente contesto la distribuzione di Dirichlet risulta essere una distribuzione a priori molto conveniente, in quanto risulta essere la coniugata della distribuzione Multinomiale. Sia dunque ω una variabile di Dirichlet k-dimensionale che assume valori nel simplesso $(k-1)$-dimensionale (un vettore k-dimensionale ω sta nel simplesso $(k-1)$-dimensionale se $\omega_i \geq 0$, $\sum_{i=1}^{k} \omega_i = 1$), con la seguente densità di probabilità in questo simplesso:

$$p(\omega|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \omega_i^{\alpha_i - 1}$$

in cui il parametro α è un vettore k-dimensionale con componenti $\alpha_i > 0$, e dove $\Gamma(x)$ è la funzione Gamma. La dimensionalità k della distribuzione di Dirichlet è assunta nota e fissata, come il parametro α, si pone $(\alpha_1, ..., \alpha_k) = \alpha$.

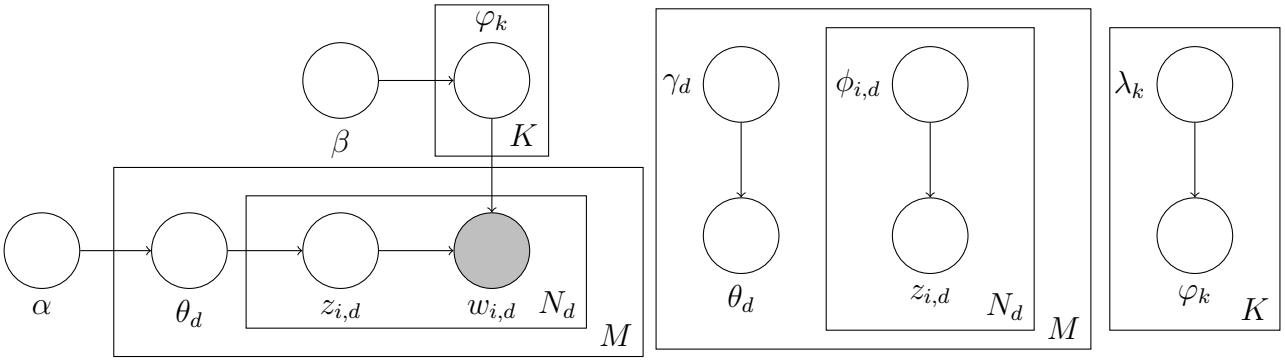

Figura 2.2: (Sinistra) Rappresentazione grafica della Latent Dirichlet Allocation. (Destra) Rappresentazione grafica della distribuzione variazionale utilizzata per approssimare la posteriori nel LDA.

Il modello mistura finale per la generazione di un *corpus* composto da M documenti ognuno dei quali formato da N_d parole ciascuno, in cui compaiono K diversi *topic*, risulta essere il seguente:

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{k=1}^{K} p(\varphi_k; \beta) \prod_{d=1}^{M} p(\theta_d; \alpha) \prod_{i=1}^{N_d} p(z_{i,d}|\theta_d) p(w_{d,i}|\varphi_{z_{d,i}}).$$

La struttura gerarchica della *Latent Dirichlet Allocation* è rappresentabile come una struttura probabilistica grafica, mostrata in Figura 2.2, sinistra, nel quale si nota la rappresentazione a tre livelli di tale modello generativo. I parametri φ_k sono variabili a livello del *corpus*, si assume che vengano campionati una volta nel processo di generazione di un *corpus*. Le variabili θ_d sono variabili a livello del documento, campionati una volta per documento. Le variabili $z_{i,d}$ e $w_{i,d}$ sono variabili a livello della parola e sono campionate una volta per ogni parola in ogni documento.

Viene quindi presentato un approccio di stima dei parametri basato su estensioni dell'algoritmo EM (*Online Variational Bayes*), presentato in Hoffman et al. (2010). Risulta possibile, dunque, fare inferenza sulla distribuzione a posteriori, approssimando quest'ultima tramite una semplice distribuzione $q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi})$, detta distribuzione variazionale. Si considera $p(\mathbf{w}|\alpha, \beta)$ la vera distribuzione a posteriori del modello, poichè risulta intrattabile per un'inferenza esatta si ricorre ad un approccio di inferenza variazionale, con il quale si ricava il sequente *Evidence Lower BOund* (ELBO):

$$\log p(\mathbf{w}|\alpha, \beta) \geq \mathcal{L}(\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = E_q[\log p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\alpha, \beta)] - E_q[\log q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi})]. \tag{2.1}$$

Massimizzare l'Equazione 2.1 è equivalente a minimizzare la divergenza di Kullback-Leibler tra $q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi})$ e la posteriori $p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{w}, \alpha, \beta)$. Si definisce $q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi})$ in modo che sia fattorizzabile rispetto alle variabili latenti $\{\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi}\}$. La distribuzione avrà dunque la seguente forma:

$$q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi}) = q(\mathbf{z}|\phi)q(\boldsymbol{\theta}|\gamma)q(\boldsymbol{\varphi}|\lambda).$$

La scelta delle distribuzioni può apportare significative semplificazioni analitiche, infatti queste sono scelte in modo da appartenere alla stessa famiglia esponenziale delle corrispettive distribuzioni a priori. Le distribuzioni risultano essere:

$$q(\mathbf{z}) = Multinomial(\mathbf{z}|\phi), \qquad q(\boldsymbol{\theta}) = Dirichlet(\boldsymbol{\theta}|\gamma), \qquad q(\boldsymbol{\varphi}) = Dirichlet(\boldsymbol{\varphi}|\lambda).$$

A questo punto è possibile, con un algoritmo di tipo *coordinate ascent*, la massimizzazione del limite inferiore. Gli aggiornamenti dei parametri, analogamente alla procedura *Expectation-Maximization (EM)*, avvengono in modo alternato: durante la fase di *Expectation* si stimano γ e ϕ, tenendo λ fissato; durante la fase di *Maximization* si stima λ fissando ϕ. Più formalmente, le formule di aggiornamento dei parametri sono le seguenti:

$$\phi_{dwk} \propto \exp\left(E_q(\log \theta_{dk}) + E_q(\log \phi_{kw})\right)$$
$$\gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk}$$

I valori attesi rispetto a $\log\theta$ e $\log\phi$ sono calcolabili come segue:

$$E_q(\log \theta_{dk}) = \Psi(\gamma_{dk}) - \Psi\left(\sum_{i=1}^{K} \gamma_{di}\right)$$
$$E_q(\log \phi_{kw}) = \Psi(\lambda_{kw}) - \Psi\left(\sum_{i=1}^{W} \lambda_{ki}\right)$$

dove Ψ indica la funzione Digamma. È possibile sfruttare la scomposizione di $\mathcal{L}$, definita come somma di contributi di ogni singolo documento:

$$\mathcal{L}(n, \lambda) = \sum_d l\left(n_d, \gamma\left(n_d, \lambda\right), \phi\left(n_d, \lambda\right), \lambda\right)$$

dove $\gamma(n_d, \lambda)$ e $\phi(n_d, \lambda)$ sono le stime dei parametri ottenute nella fase di *Expectation*. A questo punto, è possibile suddividere le osservazioni in sottocampioni, detti *mini-batch*. Si definisce con S il numero di osservazioni per ogni *mini-batch*, il passo di *Maximization*, definito come segue:

$$\tilde{\lambda}_{kw} = \eta + \frac{M}{S} \sum_{s} n_{tsk} \phi_{tskw}$$

$$\lambda = (1 - \rho_t)\lambda + \rho_t \tilde{\lambda}_t$$

dove t è il numero di iterazioni. L'aggiornamento è fatto in due passi: il primo incorpora il peso $\frac{M}{S}$; nel secondo passo avviene l'effettivo aggiornamento delle stime, utilizzando una media mobile esponenziale con peso $\rho_t = (\tau_0 + t)^{-\kappa}$, dove $\kappa \in (0.5, 1]$ è il parametro di *smoothing* e $\tau_0 \geq 0$ limita l'influenza delle stime ottenute nelle prime iterazioni.

2.4 Supervised Latent Dirichlet Allocation

La *Supervised Latent Dirichlet Allocation* (sLDA) è un modello statistico per documenti etichettati, tale modello viene proposto per modellare le informazioni ottenute da un *Topic Model* precedente stimato sul *corpus* dei riassunti, mentre il *corpus* di documenti testuali utilizzato per la stima di questo modello, risulta essere quello relativo alle trascrizioni. Questa struttura permette quindi l'utilizzo di entrambi i corpora a disposizione, in particolare, l'informazione presente nei riassunti viene utilizzata come variabile risposta, mentre il *corpus* relativo alle trascrizioni viene utilizzato per l'identificazione dei *topic*. In questa parte dell'elaborato si concentra l'attenzione sulla variabile risposta, esterna alle parole presenti nelle trascrizioni, ma che si presume contenga un'informazione relativa alle tematiche latenti all'interno dei podcast analizzati. I *Topic Model* supervisionati hanno l'obiettivo di dedurre *topic* latenti che siano predittivi per la risposta, che nel caso presente non rappresenta un'etichettatura tematica fornita dagli autori, o dalla piattaforma di streaming, in quanto non se ne dispone. Data l'assenza di questa informazione si ricorre ad una procedura in due passi, in cui nel primo si individuano delle possibili tematiche latenti per ogni episodio considerato, ed in seguito con tali etichettature si stima un *Topic Model* supervisionato come sLDA. Si decide di utilizzare le etichette derivanti da un *Topic Model* precedentemente stimato sul *corpus* dei riassunti, in quanto questi vengono presentati dai diversi autori successivamente alla registrazione degli episodi, ovvero le

trascrizioni. Risulta importante sottolineare che questa successione rispetta gli assunti base del modello secondo cui prima viene generato il testo, ed in seguito la variabile risposta.

Nella *Supervised Latent Dirichlet Allocation*, si aggiunge al modello LDA una variabile risposta connessa a ciascun documento, la cui stima si basa nuovamente su metodi variazionali, per gestire la posteriori che altrimenti risulterebbe intrattabile. Vengono quindi modellati congiuntamente documenti ed etichette associate, al fine di trovare argomenti latenti che possano aumentare le capacità previsive di futuri documenti non etichettati. sLDA fa uso dello stesso meccanicismo probabilistico di un modello lineare generalizzato per diversi tipi di riposta. Si specificano dunque i parametri del modello: K argomenti rappresentati da un vettore di probabilità β, un parametro θ corrispondente alle proporzioni di *topic* per ogni documento derivante da una distribuzione di Dirichlet di parametro α, come per LDA, e i parametri della variabile risposta η e δ. Secondo un modello sLDA, ogni documento e relativa risposta derivano dal seguente processo generativo:

1. $\theta_d \sim Dirichlet(\alpha)$, con $d \in \{1, ..., M\}$
2. Per ogni parola w_i del documento d, con $i \in \{1, ..., N_d\}$:
   (a) Si estrae un *topic* $z_{i,d} \sim Multinomial(\theta_d)$.
   (b) Si estrae una parola $w_{i,d} \sim Multinomial(\beta_{z_{i,d}})$.
3. Generare la variabile risposta $y \sim GLM(\bar{z}, \eta, \delta)$, dove si definisce

$$\bar{z} = (1/N_d) \sum_{i=1}^{N_d} z_i. \tag{2.2}$$

La distribuzione della risposta è rappresentabile come un modello lineare generalizzato

$$p(y) = h(y, \delta) \exp\left\{ \frac{(\eta^T \bar{z})y - A(\eta^T \bar{z})}{\delta} \right\}. \tag{2.3}$$

Due sono gli elementi principali in un modello lineare generalizzato (GLM): la componente casuale e la componente sistematica. Per la componente casuale, si assume che la distribuzione della risposta sia una famiglia di dispersione esponenziale con parametro naturale $\eta^T \bar{z}$ e parametro di dispersione δ. Il parametro di dispersione fornisce ulteriore flessibilità nel modellare

la varianza di y. Per la componente sistematica nei GLM, si mette in relazione il parametro della famiglia esponenziale relativo alla componente casuale nel predittore lineare, tramite una combinazione lineare di variabili esplicative. La struttura dei GLM fornisce la flessibilità di modellare qualsiasi tipo di variabile risposta la cui distribuzione può essere espressa in forma di famiglia esponenziale come nell'Equazione 2.3.

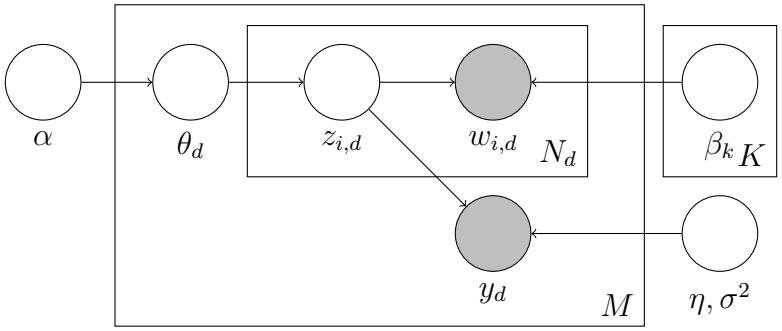

Figura 2.3: Rappresentazione grafica della Supervised Latent Dirichlet Allocation.

La differenza tra sLDA e un classico GLM risiede nelle covariate, che rappresentano le frequenze empiriche non osservate dei *topic* in ciascun documento. Nel processo generativo, mostrato in Figura 2.3, tali variabili latenti sono responsabili della scrittura delle parole del documento, pertanto, la risposta e le parole sono legate. I coefficienti su tali frequenze compongono il predittore lineare η. Si noti che un GLM solitamente include un termine di intercetta, il che equivale ad aggiungere una covariata che è sempre uguale ad 1. Nel caso presente tale termine è ridondante, poiché le componenti di $\bar{z}$ sommano sempre ad uno, come mostrato in Equazione 2.2. Si regredisce la risposta sulle frequenze empiriche dei *topic*, la risposta viene trattata come non scambiabile con le parole. Il documento viene generato per primo, in piena scambiabilità delle parole; poi, in base al documento, viene generata la variabile di risposta. La formulazione scelta sembra sensata: la risposta dipende dalle frequenze dei *topic* effettivamente presenti nel documento, piuttosto che dalla media della distribuzione che genera i *topic*, così come le tematiche presenti nei riassunti si presume dipendano dagli argomenti affrontati durante la registrazione di una puntata. La stima di tale modello, completamente scambiabile con un numero sufficiente di argomenti consente di utilizzare i *topic* latenti per spiegare le variabili risposta. In altre parole, nel presente modello le variabili latenti che governano la risposta sono le stesse variabili latenti che governano le parole presenti nei documenti.

La sLDA si contraddistingue dalla classica LDA per la presenza della distribuzione relativa

alla variabile risposta nel modello, presentando due ulteriori parametri che si riferiscono a quest'ultima distribuzione. Si ha dunque che il modello mistura per la generazione di un *corpus* composto da M documenti di lunghezza N_d ciascuno, risulta essere:

$$p(\mathbf{w}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta}; \alpha, \beta, \eta, \sigma^2) = \prod_{d=1}^{M} p(\theta_d; \alpha) p(y_d; \eta, \sigma^2) \prod_{i=1}^{N_d} p(z_{i,d}|\theta_d) p(w_{d,i}|\beta_{z_{d,i}}).$$

Risulta evidente che in contrapposizione alla *Latent Dirichlet Allocation* non supervisionata descritta precedentemente, vi è una dipendenza aggiuntiva rispetto ai parametri relativi alla variabile risposta. Tali parametri influenzeranno l'esplicazione dei *topic* latenti da parte del modello. Si passa dunque a delineare la procedura di stima dei parametri, esplicitando l'*Evidence Lower BOund* (ELBO) come segue:

$$\begin{aligned} \log p(\mathbf{w}, \mathbf{y}|\alpha, \beta, \eta, \sigma^2) &\geq \mathcal{L}(\mathbf{w}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma}) \\ &\geq E_q[\log p(\mathbf{w}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta}|\alpha, \beta, \eta, \sigma^2)] - E_q[\log q(\mathbf{z}, \boldsymbol{\theta})]. \end{aligned}$$

Il primo termine è il valore atteso del logaritmo della probabilità congiunta delle variabili latenti e di quelle osservate; il secondo termine è l'entropia della distribuzione variazionale. Nella sua forma estesa, l'ELBO per sLDA è

$$\begin{aligned} \mathcal{L}(\mathbf{w}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma}) = E_q[\log p(\theta; \alpha)] + \sum_{i=1}^{N_d} E_q[\log p(z_i; \theta)] \\ + \sum_{i=1}^{N_d} E_q[\log p(w_i; z_i, \theta, \beta)] + E_q[\log p(y; z, \eta, \delta)] + H(q). \end{aligned} \tag{2.4}$$

L'ELBO rappresentato in Equazione 2.4, come spiegato precedentemente, si restringe quando $q(\mathbf{z}, \boldsymbol{\theta})$ è la posteriori, ma specificando una famiglia che contenga la distribuzione a posteriori si arriva ad un problema di ottimizzazione intrattabile, viene specificata quindi una famiglia approssimata più semplice. Stringere l'ELBO, è equivalente a trovare la distribuzione che è più vicina, secondo la divergenza di Kullback-Leibler, alla posteriori. Si sceglie una distribuzione completamente fattorizzata

$$q(\mathbf{z}, \boldsymbol{\theta}|\gamma, \phi) = q(\theta|\gamma) \prod_{i=1}^{N_d} q(z_i|\phi_i),$$

dove γ è un vettore di parametri della distribuzione di Dirichlet K-dimensionale, ogni ϕ_i parametrizza una distribuzione Multinomiale di K elementi. Ne segue che $E[z_i] = q(z_i) = \phi_i$, il quale rappresenta l'assegnazione dei *topic* latenti z, che può essere visto come un vettore di indicatrici K-dimensionale. Pertanto, data una coppia documento-risposta, si massimizza l'Equazione 2.4 rispetto a ϕ e γ per ottenere una stima della distribuzione a posteriori. I primi tre termini e l'entropia della distribuzione variazionale hanno forma equivalente ai corrispondenti termini nell'ELBO per LDA non supervisionata, mentre si aggiunge il seguente termine relativo alla variabile risposta:

$$E_q[\log p(y; z, \eta, \delta)] = \log h(y, \delta) + \frac{1}{\delta}[\eta^T (E[\bar{z}]y) - E[A(\eta^T \bar{z})]],$$

dove il primo valore atteso, relativo alla proporzione di *topic* stimati per un singolo documento d, corrisponde alla seguente formulazione:

$$E[\bar{z}_d] = \bar{\phi}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} \phi_i,$$

questo deriva dalla struttura di z_d, la quale risulta essere un vettore di indicatrici. Il secondo rappresenta il valore atteso del predittore lineare specifcato per il GLM relativo alla variabile risposta y.

I dati sono coppie documento-risposta $\{\mathbf{w}_d, y_d\}_{d=1}^M$, mentre i parametri di sLDA sono i K *topic* β, gli iperparametri della distribuzione di Dirichlet α, i coefficienti η e il parametro di dispersione δ del GLM. Questi parametri si stimano con un *Variational Expetaction Maximization* (VEM). Viene utilizzata una procedura di *block coordinate-ascent variational inference*, la quale permette la massimizzazione dell'Equazione 2.4 rispetto a ciascun vettore di parametri variazionali. I termini variazionali γ non coinvolgono la variabile di risposta y, contrariamente gli aggiornamenti del parametro relativo alla distribuzione variazionale Multinomiale dipendono dalla forma di $E[A(\eta^T \bar{Z})]$. L'inferenza variazionale procede aggiornando iterativamente i parametri variazionali $\{\gamma, \phi\}$. Infine la distribuzione variazionale risultante viene utilizzata come approssimazione per la posteriori. Nella parte di *E-step*, viene stimata la distribuzione a posteriori approssimata per ciascuna coppia documento-risposta. Nella parte di *M-step*, viene massimizzato l'ELBO a livello di *corpus* rispetto ai parametri del modello. Il *Variational Expe-*

tation Maximization trova un ottimo locale dell'equazione 2.4 eseguendo l'iterazione tra questi passaggi. Durante la parte di *M-step* gli aggiornamenti degli argomenti β tengono in considerazione che la probabilità di una parola sotto un *topic* sia proporzionale al numero atteso di volte a cui è stata assegnata a quel *topic*. In pratica, la procedura alterna l'aggiornamento dei parametri variazionali con la stima dei parametri del modello. Gli aggiornamenti dei parametri variazionali risultano essere

$$\gamma^{new} = \alpha + \sum_{i=1}^{N_d} \phi_i,$$
$$\phi_i^{new} \propto \exp\{E_q[\log\theta] + E_q[\log p(w|\beta)] + \left(\frac{y}{N_d\delta}\right)\eta - \left(\frac{1}{\delta}\right)\frac{\partial}{\partial\phi_i}\left\{E[A(\eta^T\bar{Z})]\right\}\}.$$

Gli aggiornamenti, durante la parte di *M-step*, dei parametri β, sono rappresentati dalla probabilità di una parola sotto un *topic* che risulta essere proporzionale al numero atteso di volte a cui è stata assegnata a quel *topic*

$$\hat{\beta}_{k,w}^{new} \propto \sum_{d=1}^{M}\sum_{i=1}^{N_d} \mathbb{1}(w_{d,i} = w)\phi_{d,i}^{k},$$

dove ogni $\hat{\beta}_k^{new}$ è normalizzato per sommare ad uno.

Nel caso presente si specifica per la variabile risposta y una distribuzione Multinomiale. Sia pertanto $Y = (y_1, ..., y_K)$ un insieme di variabili casuali. In particolare, y_h rappresenta il conteggio del numero di volte in cui si verifica l'evento h, su M prove indipendenti. Sia π_h la probabilità che si verifichi l'h-esimo evento. Allora si può specificare la distribuzione di y come famiglia esponenziale seguendo i seguenti passi:

$$\begin{aligned} p(y,\pi) &= \frac{M!}{y_1!...y_K!}\prod_{i=1}^{K}\pi_i^{y_i} \\ &= \frac{M!}{y_1!...y_K!}\exp\left\{\sum_{i=1}^{K} y_i\log\pi_i\right\}. \end{aligned}$$

Si concentra particolare attenzione sull'argomento dell'esponenziale, si può riscrivere il modello

tenendo una categoria di riferimento, nel seguente modo:

$$\begin{aligned} p(y,\pi) &= \exp\left\{\sum_{i=1}^{K} y_i \log \pi_i\right\} \\ &= \exp\left\{\sum_{i=1}^{K-1} y_i \log \pi_i + \left(1 - \sum_{i=1}^{K-1} y_i\right) \log\left(1 - \sum_{i=1}^{K-1} \pi_i\right)\right\} \\ &= \exp\left\{\sum_{i=1}^{K-1} \log\left(\frac{\pi_i}{1 - \sum_{i=1}^{K-1} \pi_i}\right) y_i + \log\left(1 - \sum_{i=1}^{K-1} \pi_i\right)\right\}, \end{aligned}$$

riprendendo l'Equazione 2.3 si specificano le diverse quantità come segue:

$$\begin{aligned} \eta_k^T \bar{z} &= \log\left(\frac{\pi_k}{1 - \sum_{i=1}^{K-1} \pi_i}\right) = \log\left(\frac{\pi_k}{\pi_K}\right) \\ \pi_k &= \frac{e^{\eta_k \bar{z}}}{\sum_{i=1}^{K} e^{\eta_i^T \bar{z}}}. \end{aligned}$$

Infine, l'ultimo termine di riferimento con una generica famiglia esponenziale

$$A(\eta^T \bar{z}) = -\log\left(1 - \sum_{i=1}^{K-1} \pi_i\right) = \log\left(\sum_{i=1}^{K} e^{\eta_i^T \bar{z}}\right).$$

Si noti come in questa circostanza il parametro di dispersione $\delta = 1$. Il predittore lineare viene inizializzato per poi essere aggiornato in modo iterativo secondo la proporzione di *topic* presenti nei documenti.

2.5 Structural Topic Model

In questa sezione viene presentato lo *Structural Topic Model* (Roberts et al., 2019), una rivisitazione del *Correlated Topic Model*, (Lafferty & Blei, 2005), dove le proporzioni di *topic* presentano una correlazione tramite la distribuzione Logistica-Normale, *Logistic-Normal Topic Admixture Model* (LoNTAM) (Ahmed & Xing, 2007), la quale permette di lavorare su un simplesso continuo. Questo modello può rappresentare la correlazione tra *topic* tramite l'utilizzo di tale distribuzione, al fine di modellare la variabilità nella mistura di *topic* sottostante i documenti, oltre a permettere l'inserimento di covariate a livello di documento, le quali rappresentano delle proporzioni di *topic* precedente stimate. Tale struttura presenta uno svantaggio dato dal fatto

che la coniugata di una distribuzione Multinonomiale non risulti essere la Logistica-Normale, rendendo l'inferenza a posteriori e la stima dei parametri differente rispetto ad i precedenti approcci nei quali veniva proposta una distribuzione di Dirichlet. L'utilizzo di covariate che corrispondono a proporzioni stimate precedentemente ha il fine di rafforzare la struttura che mira a rappresentare la correlazione sottostante i *topic* che vengono identificati. La grande dimensione dei vocabolari relativi ai corpora a disposizione non permette l'utilizzo di una matrice del disegno che possa tener conto di tutte le parole, in quanto la stima del predittore lineare avviene come in una classica regressione lineare. Più formalmente, per generare un documento $\mathbf{w}_d$, con un vocabolario di grandezza V, per uno *Structural Topic Model* con K *topic*, si procede come segue:

1. $\theta_d \sim LogisticNormal(\mu = x_d\Delta, \Sigma)$, con $d \in \{1, ..., M\}$ e x_d un vettore di covariate a livello di documento

2. Per ogni parola w_i del documento d, con $i \in \{1, ..., N_d\}$:

   (a) Si estrae un *topic* $z_{i,d} \sim Multinomial(\theta_d)$

   (b) Si estrae una parola $w_{i,d} \sim Multinomial(\beta_{z_{i,d}})$

La relazione tra una distribuzione Logistica-Normale ed una distribuzione Normale può essere formalizzata nel seguente modo: dato $\gamma \sim Normal(\mu, \Sigma)$ con $\gamma \in \mathbb{R}^{K-1}$, se $\theta = logit(\gamma)$ risulta:

$$\theta = \left[\frac{\exp\{\gamma_1\}}{1 + \sum_{i=1}^{K-1} \exp\{\gamma_i\}}, ..., \frac{\exp\{\gamma_{K-1}\}}{1 + \sum_{i=1}^{K-1} \exp\{\gamma_i\}}, \frac{1}{1 + \sum_{i=1}^{K-1} \exp\{\gamma_i\}} \right],$$

quindi si ha che $\theta \sim LogisticNormal(\mu, \Sigma)$, la cui funzione di densità risulta essere:

$$p(\theta, \mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \frac{1}{\prod_{i=1}^{K}(\theta_i(1-\theta_i))} \exp\left\{ -\frac{1}{2} \left(\log\left(\frac{\theta}{1-\theta}\right) - \mu \right)^T \Sigma^{-1} \left(\log\left(\frac{\theta}{1-\theta}\right) - \mu \right) \right\}.$$

Come nei precedenti modelli viene dunque riportato il modello mistura per la generazione di un *corpus* composto da M documenti ognuno dei quali formato da N_d parole ciascuno:

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}; \beta, \mu, \Sigma) = \prod_{d=1}^{M} p(\theta_d; \mu, \Sigma) \prod_{i=1}^{N_d} p(z_{i,d}|\theta_d) p(w_{d,i}|\beta_{z_{d,i}})$$

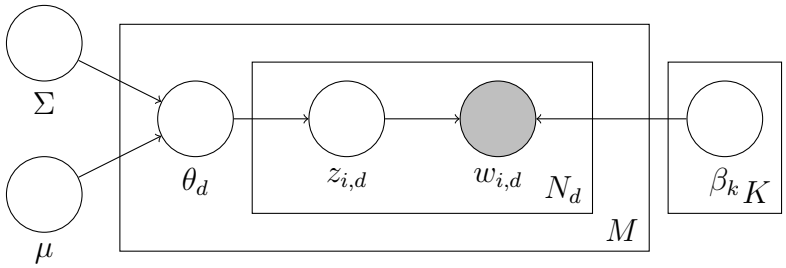

Figura 2.4: Rappresentazione grafica del Strucural Topic Model.

Come illustrato in Figura 2.4 negli *Structural Topic Model*, ogni *topic* è rappresentato da due parametri, μ che viene considerato il predittore lineare e che permette l'inserimento di covariate a livello di documento, e Σ che fornisce una struttura di correlazione tra i *topic*. Ogni parola è rappresentata a sua volta da un vettore di proporzioni β il quale dipende dai *topic* estratti per il documento corrente. Sostanzialmente, le proporzioni di *topic* θ vengono campionate da una distribuzione Logistica-Normale, i quali poi vengono utilizzati come parametro di una distribuzione Multinomiale per campionare i *topic* z nel documento corrente, successivamente le parole sono campionate basandosi su questo vettore di argomenti.

Si vuole ora dare una breve illustrazione della procedura di stima di tale modello generativo. In primo luogo, ci si concentra sulla variabile θ, la quale può essere parametrizzata secondo una distribuzione Normale, come mostrato in precedenza. Sia dunque $\gamma_d \sim Normal(\mu, \Sigma)$, una proiezione nel simplesso tramite la seguente trasformazione logistica:

$$\theta_{d,k} = \exp\{\gamma_{d,k} - C(\gamma_d)\},$$

dove $C(\gamma) = \log(\sum_{k=1}^{K} \exp(\gamma_{d,k}))$, rappresenta la constante di normalizzazione. A causa del vincolo di normalizzazione sui parametri della distribuzione multinomiale, θ presenta solo $K-1$ gradi di libertà. Si necessita pertanto di scrivere le prime $K-1$ componenti di γ da una Normale multivariata $(K-1)$-dimensionale, infine si pone $\gamma_{d,K} = 0$ per rendere il modello identificabile. Si modella quindi il vettore media della distribuzione Logistica-Normale come un semplice modello lineare. Come fatto precedentemente, si ricorre all'algoritmo *Variational Expectation Maximization* (VEM), nel quale si considera $p(\mathbf{w}|\mu, \Sigma, \beta)$ come la vera distribuzione a posteriori del modello, la quale nuovamente risulterebbe intrattabile per un'inferenza esatta. Si ricava

dunque il seguente *Evidence Lower BOund* (ELBO):

$$\log p(\mathbf{w}|\mu, \Sigma, \beta) \geq E_q[\log p(\gamma|\mu, \Sigma)] + \sum_{i=1}^{N_d}(E_q[\log p(z_i|\gamma)] + E_q[\log p(w_i|z_i, \beta)]) + H(q)$$

il valore atteso è ottenuto rispetto alla distribuzione variazionale delle variabili latenti, $H(q)$ denota ancora una volta l'entropia della distribuzione variazionale. Si definisce nuovamente una distribuzione fattorizzata:

$$q(\boldsymbol{\gamma}, \mathbf{z}|\lambda, \nu, \phi) = \prod_{k=1}^{K} q(\gamma_k|\lambda_k, \nu_k^2) \prod_{i=1}^{N_d} q(z_i|\phi_i)$$

Le distribuzioni variazionali riferite alle variabili discrete z sono specificate dai parametri Multinomiali ϕ. La distribuzione variazionale delle variabili continue γ_k risultano invece K Normali indipendenti univariate parametrizzate da $\{\lambda_k, \nu_k^2\}$. La non coniugazione della distribuzione Logistica-Normale rende difficile il calcolo del valore atteso del logaritmo della distribuzione di assegnazione dei *topic*

$$E_q[\log p(z|\gamma)] = E_q[\gamma^T z] - E_q\left[\log\left(\sum_{k=1}^{K}\exp\{\gamma_k\}\right)\right].$$

Al fine di preservare il limite inferiore sulla trasformazione logaritmica di tale distribuzione, si ricorre ad un espasione di Taylor come segue:

$$E_q\left[\log\left(\sum_{k=1}^{K}\exp\{\gamma_k\}\right)\right] \leq \zeta^{-1}\left(\sum_{k=1}^{K}E_q[\exp\{\gamma_k\}]\right) - 1 + \log(\zeta)$$

in cui si introduce un quarto parametro variazionale ζ. Il valore atteso $E_q[\exp\{\gamma_k\}]$ risulta, dunque, essere la media della distribuzione log-Normale con media e varianza ottenuti dai parametri variazionali $\{\lambda_k, \nu_k^2\}$, ne risulta che $E_q[\exp\{\gamma_k\}] = \exp\{\lambda_k + \nu_k^2/2\}$.

Gli aggiornamenti della procedura di *Expectation Maximization* avvengono tramite l'usuale procedura alternata in cui: durante l'*E-step* si stimano i parametri variazionali $\{\phi, \zeta, \lambda, \nu\}$ massimizzando l'ELBO e tenendo fissati gli altri parametri, mentre nel *M-step* la massimizzazione avviene rispetto ai parametri del modello, con le stime trovate nell'*E-step*. Si ottengono quindi

i seguenti aggiornamenti per i primi due parametri:

$$\zeta = \sum_{k=1}^{K} \exp\{\lambda_k + \nu_k^2/2\}$$
$$\phi_{n,i} \propto \exp\{\lambda_i\}\beta_{i,w_n}$$

Rimane quindi la massimizzazione rispetto agli ultimi due parametri variazionali i quali, però, non ammettono una soluzione analitica. Infatti, questi vengono stimati con un metodo del gradiente coniugato, comunemente applicato per risolvere sistemi di equazioni lineari, in particolare quando la matrice del sistema è simmetrica e definita positiva, in cui le derivate risultano:

$$\frac{\partial}{\partial\lambda} = -\Sigma^{-1}(\lambda - \mu) + \sum_{i=1}^{N_d} \phi_i - (N_d/\zeta)\exp\{\lambda + \nu^2/2\}$$
$$\frac{\partial}{\partial\nu^2} = -\Sigma_{ii}^2/2 - N_d/2\zeta \exp\lambda + \nu_i^2/2 + 1/(2\nu_i^2)$$

Gli aggiornamenti dei rimanenti parametri Δ e Σ avvengono come in una classica regressione lineare, utilizzando le proporzioni di *topic* stimate precedentemente nell'*E-step*.

Capitolo 3

Applicazione ai dati

Nel presente capitolo, vengono illustrati i procedimenti attuati durante l'analisi, oltre ai risultati ottenuti mediante l'applicazione dei vari modelli proposti. Si introducono anche tre diverse metriche di coerenza, le quali forniscono un criterio di confronto tra i risultati emersi nelle due fasi di stima. In particolare, una di queste metriche è stata adottata per ottimizzare i parametri della *Latent Dirichlet Allocation*. L'obiettivo di tale ottimizzazione è l'identificazione di un numero ideale di *topic* per ciascun *corpus*.

Le tecniche di ottimizzazione adottate nella fase iniziale puntano alla massimizzazione di una delle tre metriche di coerenza proposte. Infine, si valuta se l'approccio in due fasi, il quale permette l'estrapolazione di informazioni da entrambi i corpora, migliori le prestazioni in relazione all'individuazione dei *topic*, rispetto a modelli che considerano le informazioni presenti in un solo *corpus* come quelli stimati nella prima fase. In questa ottica, si procede con la valutazione dei *topic* stimati attraverso il confronto di tutte le metriche di coerenza proposte, mirando a delineare tematiche concrete dai termini più rappresentativi dei vari *topic*. Si riporta anche una valutazione qualitativa dei *topic* individuati, al fine di delineare se questi presentino tematiche oggettive date dall'insieme di più parole.

3.1 Metriche di ottimizzazione e valutazione

Nell'ambito dell'analisi testuale i risultati ottenuti, solitamente, vengono valutati da linguisti specializzati, i quali redigono una scala di gradimento per valutare i risultati delle analisi, che

varia da "pessime", per un insieme di parole che non rispecchiano minimamente il contenuto dei documenti che rappresentano, ad "eccellenti", che invece evidenziano parole che riescono a riassumerne perfettamente il contenuto, questo tipo di approccio viene utilizzanto anche in Clifton et al. (2020). Tale procedura di valutazione può però risultare onerosa e soggetta a disaccordi.

Per superare queste problematiche, sono state introdotte diverse metriche che cercano di quantificare quanto bene i *topic* identificati riflettano le strutture semantiche presenti nei dati, (Stevens et al., 2012). Una di queste metriche è la *coherence*, che misura la co-occorrenza delle parole all'interno dei documenti del *corpus*, ovvero la frequenza con cui due o più termini appaiono insieme nei documenti. Queste valutano le parole all'interno di un dato argomento misurandone la similarità semantica tra di loro, per cui un elevato valore indica una forte correlazione semantica tra le parole dell'argomento. Si definisce $D(w_i)$ come la frequenza di documenti che contengono la parola w_i, e $D(w_i, w_j)$ la frequenza di documenti che contengono sia la parola w_i che la parola w_j, la formulazione generale della *coherence* per un dato *topic* k è data da:

$$C(w_i^{(k)}, w_j^{(k)}) = f(D(w_i^{(k)}, w_j^{(k)}), D(w_i^{(k)}), D(w_j^{(k)}), \epsilon),$$

dove $f(.)$ rappresenta una funzione non lineare che combina le frequenze $D(.)$, ed un termine di lisciamento ϵ. Si vuole quindi esplicitare la formulazione di tale metrica per ogni modello considerato indipendentemente dal numero di *topic* stimati risulta essere:

$$C = \frac{1}{K} \sum_{k=1}^{K} \sum_{m=2}^{L} \sum_{l=1}^{m-1} C(w_i^{(k)}, w_j^{(k)}),$$

dove L sono il numero di parole più probabili per il *topic* k; si pone $L = 20$ per ogni modello stimato. La *coherence* si basa esclusivamente sulla co-occorrenza delle parole raccolte dal *corpus* modellato, evitando la necessità di un *corpus* di riferimento esterno. Vengono esplicitate tre diverse varianti di tale metrica: *coherence UMass* (C_{UMass}), *coherence UCI* (C_{UCI}) e *coherence NPMI* (C_{NPMI}), dove la prima viene utilizzata non solo per un confronto finale tra tutti i modelli, ma anche come criterio di ottimizzazione nella prima fase dell'analisi.

La *coherence UMass* (Mimno et al., 2011), viene utilizzata per ottimizzare il numero di *topic* presenti nei testi. Essa viene definita tramite un punteggio basato sulla co-occorrenza dei

documenti come segue:

$$C_{UMass}(w_i^{(k)}, w_j^{(k)}) = \log \frac{D(w_i^{(k)}, w_j^{(k)}) + \epsilon}{D(w_i^{(k)})}.$$

Questa, infatti, stima il grado di coerenza tra parole all'interno di un dato *topic* calcolando il logaritmo del rapporto delle frequenze relative alle co-occorrenze nel *corpus* di documenti. C_{UMass} conta il numero di volte in cui una coppia di parole co-occorre in un dato *corpus* e lo confronta con il numero di co-occorrenze di parole distribuite in tutto il *corpus*. Più formalmente, C_{UMass} calcola il logaritmo della frequenza della parola $w_i^{(k)}$ e $w_j^{(k)}$ divisa per la frequenza della parola $w_i^{(k)}$.

La seconda metrica di coerenza proposta per la comparazione dei modelli risulta essere C_{UCI} (Mahanty et al., 2019). Essa misura la forza dell'associazione tra coppie di parole basata sulla loro co-occorrenza, tenendo conto anche delle frequenze individuali delle due parole considerate. In questo contesto la coerenza dell'argomento è definita come il logaritmo in base 2 del rapporto della frequenza di co-occorrenza della parola $w_i^{(k)}$ e $w_j^{(k)}$ all'interno di un dato argomento k. Viene, pertanto, definita nel seguente modo:

$$C_{UCI}(w_i^{(k)}, w_j^{(k)}) = \log_2 \left(\frac{D(w_i^{(k)}, w_j^{(k)}) + \epsilon}{D(w_i^{(k)}) D(w_j^{(k)})} \right).$$

Questa è calcolata prendendo il rapporto tra la frequenza congiunta di due parole $D(w_i^{(k)}, w_j^{(k)})$ che appaiono insieme, e le frequenze individuali delle parole $D(w_i^{(k)})$ e $D(w_j^{(k)})$ che compaiono separatamente.

La terza ed ultima metrica proposta risulta essere C_{NPMI} (Aletras & Stevenson, 2013), la quale tiene conto del fatto che alcune parole sono più comuni di altre e regola la frequenza delle singole parole di conseguenza. Questa viene definita come la metrica C_{UCI} fratto il logaritmo in base 2, di segno negativo, della frequenza congiunta di due parole, nello specifico:

$$C_{NPMI}(w_i^{(k)}, w_j^{(k)}) = \frac{\log_2 \left(\frac{D(w_i^{(k)}, w_j^{(k)}) + \epsilon}{D(w_i^{(k)}) D(w_j^{(k)})} \right)}{-\log_2 (D(w_i^{(k)}, w_j^{(k)}) + \epsilon)}.$$

C_{NPMI} è, pertanto, una normalizzazione della metrica C_{UCI} la quale assume valori che variano

tra -1 e 1, risultando generalmente più sensibile a parole rare. Per tutte le metriche proposte si pone $\epsilon = 1$, al fine di evitare logaritmi di zero.

3.2 Ottimizzazione LDA

Vengono quindi presentati gli andamenti della metrica di *coherence UMass*, introdotta precedentemente, per ottimizzare i parametri della *Latent Dirichlet Allocation*. Si vuole andare a selezionare un numero di *topic* K che massimizzi tale metrica, congiuntamente ai parametri delle distribuzioni a priori α e β. Tali parametri vengono ottimizzati indipendentemente sui due corpora. Si propongono tre griglie identiche per entrambi i corpora, di cui si riportano i valori dei parametri α e β in Tabella 3.1, mentre i *topic* variano da un minimo di 2 fino ad un massimo di 11. Si decide di esplorare un numero ridotto di *topic* al fine di rispettare ciò che è emerso durante le analisi esplorative, presentate nel primo capitolo, sottolineando la condivisione di tematiche da parte di molteplici *show*, dove ognuno di questi presenta almeno tre episodi. Viene quindi stimata la *Latent Dirichlet Allocation* su tutte le possibili combinazioni di queste griglie.

Ci si potrebbe aspettare che il numero di argomenti latenti sia unico indipendentemente dal *corpus* sul quale viene stimato il modello. Nella seconda fase di analisi si terranno in considerazione entrambi i risultati ottenuti, rispettivamente per stimare la *Supervised Latent Dirichlet Allocation* si utilizza il numero di *topic* identificati sul *corpus* dei riassunti, mentre per stimare lo *Structural Topic Model* si utilizza il numero di *topic* identificati sul *corpus* delle trascrizioni. Tale procedura permette di stimare dei modelli che possano incorporare ulteriormente informazioni da entrambi i corpora, poiché oltre al numero di *topic* si utilizzano le quantità descritte precedentemente, ovvero: un'etichettatura per la *Supervised Latent Dirichlet Allocation* ed una distribuzione di proporzioni per lo *Structural Topic Model*.

$\boldsymbol{\alpha}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\boldsymbol{\beta}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

Tabella 3.1: Grigilia di valori valutati per la *Latent Dirichlet Allocation*.

Vengono quindi riportati gli andamenti della *coherence UMass* utilizzata per la prima fase di ottimizzazione, al variare dei tre parametri da ottimizzare.

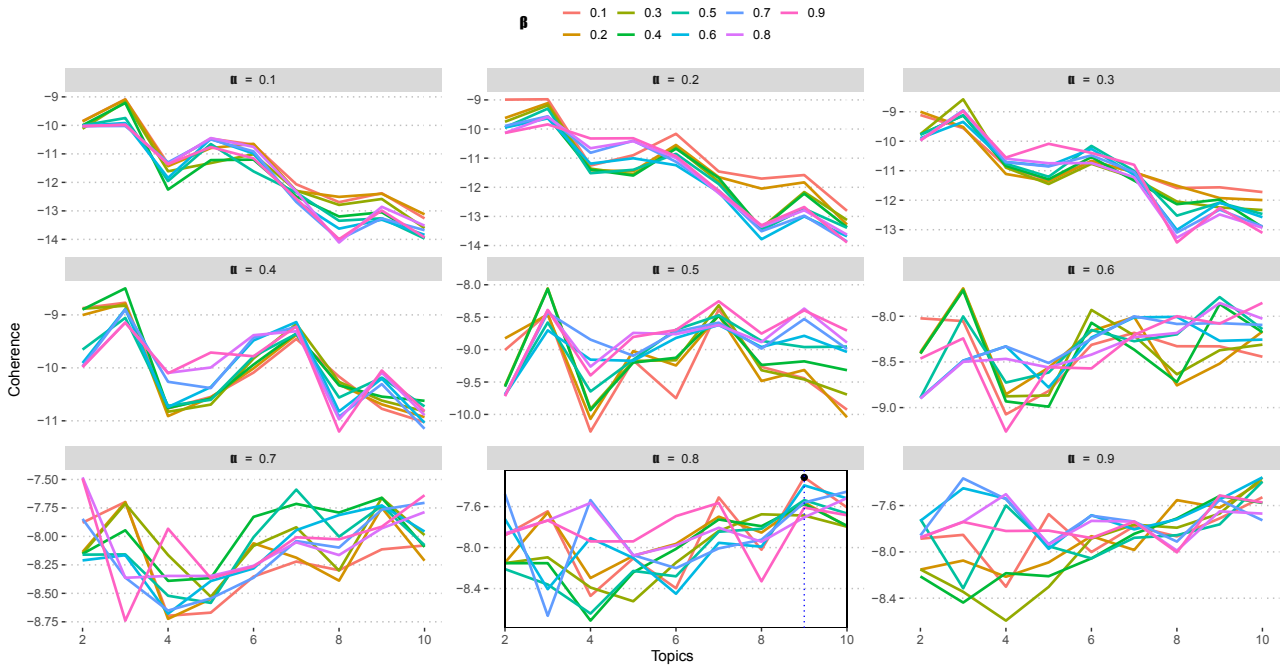

Figura 3.1: Andamento coherence UMass al variare del numero di topic e dei parametri α e β, nel corpus dei riassunti.

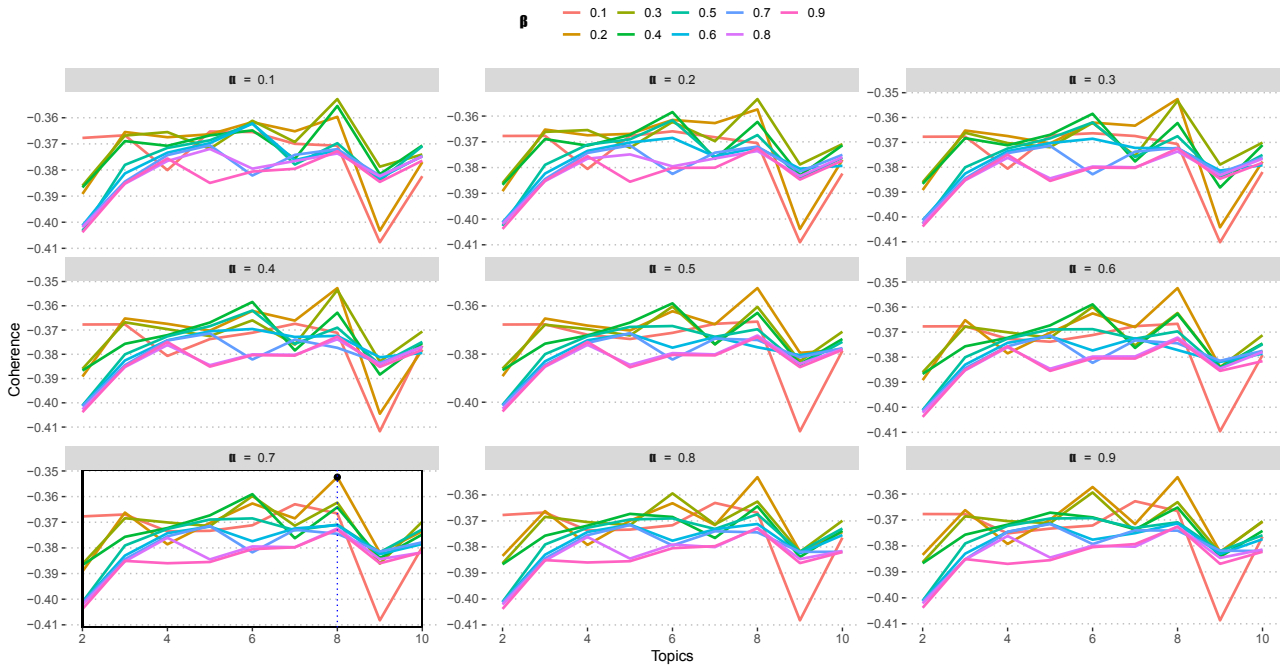

Figura 3.2: Andamento coherence UMass al variare del numero di topic e dei parametri α e β, nel corpus delle trascrizioni.

In particolare in Figura 3.1 viene mostrato l'andamento di tale metrica sul *corpus* dei riassunti, mentre in Figura 3.2 si riportano i risultati ottenuti sul *corpus* delle trascrizioni. I

colori nelle diverse figure discriminano i vari valori di β, mentre i quadranti rappresentano i diversi valori di α. Sull'asse delle ascisse si trova il numero di *topic* esaminati, mentre sull'asse delle ordinate si trova la metrica *coherence UMass*.

La combinazione di parametri che massimizza tale metrica viene evidenziata dal quadrante nero, si può inoltre visualizzare il numero ottimo di *topic* selezionato in corrispondenza del punto di massimo ottenuto. Si nota per prima cosa come vi sia un effetto *corpus*, questo viene caratterizzato dal fatto che in Figura 3.2 per ogni combinazione dei parametri α e β la *coherence* si massimizzi per un numero di *topic* pari ad 8. Contrariamente in Figura 3.1 si nota come al variare delle combinazioni dei parametri α e β vi sia un andamento sostanzialmente diverso, in alcune circostanze descresce in altre cresce al variare del numero di *topic*.

	Numero di Topic (K)	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	**Coherence UMass**
LDA-Trascrizioni	8	0.7	0.2	-0.3524
LDA-Riassunti	9	0.8	0.1	-7.3146

Tabella 3.2: Valori dei parametri ottimizzati per la *Latent Dirichlet Allocation*.

In Tabella 3.2 vengono riportati i valori dei parametri selezionati ottimizzando la metrica di coerenza proposta. Nonostante il numero di *topic* identificati sui due corpora differisca, si osserva una notevole vicinanza tra questi, suggerendo che i risultati tra i due corpora possano essere molto simili tra loro in ottica di modellazione.

Per quanto l'ottimizzazione di questi avvenga secondo la metrica *coherence UMass*, si vogliono ulteriormente esaminare qualtativamente i gruppi tematici identificati. Per farlo si valutano sia le parole con la maggior frequenza all'interno di ciascun gruppo, sia tramite le distribuzioni delle proporzioni di *topic* all'interno dei documenti. Pertanto, si propone anche una valutazione qualitativa, con il fine di individuare a che tematica possano appartenere l'insieme delle parole più frequenti di ogni *topic*. Vengono dunque riportarti i risultati dei modelli LDA stimati sui due corpora.

In Tabella 3.3 si possono visualizzare le parole in ordine decrescente di frequenza, all'interno di ogni *topic* stimato, sul *corpus* dei riassunti. Questi presentano molte parole tra loro ripetute, come "*talk*", "*go*", "*true*", che non sono riconducibili a specifiche tematiche. Invece in altre circostanze si evidenzia la presenza di parole che potrebbero discriminare maggiormente alcune tematiche sottostanti i dati, come: "*athlete*", "*swim*" o "*talk*", "*experience*", queste però oltre

a comparire in diversi *topic*, non risultano essere le parole che presentano la frequenza maggiore all'interno del *topic* di riferimento.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
talk	talk	coach	go	talk	true	talk	coach	true
go	year	year	true	go	talk	year	athlete	help
year	content	athlete	talk	year	coach	go	year	coach
coach	small	go	make	give	go	find	talk	fear
true	idea	talk	year	coach	help	make	make	year
help	go	true	message	make	small	message	career	talk
make	make	idea	coach	find	year	swim	message	idea
give	help	find	find	swim	find	help	find	go
purpose	murder	make	give	true	woman	need	true	become
idea	create	swim	help	ever	experience	give	experience	find
daily	find	woman	love	help	idea	experience	time	experience
fitness	give	help	murder	experience	become	heroine	swim	say
find	conversation	message	resource	time	make	take	take	heroine
say	networking	post	ever	message	know	become	murder	create
take	change	time	also	take	start	idea	idea	also
experience	start	content	become	thought	also	young	go	need
athlete	daily	become	series	swimming	host	swimming	say	make
time	woman	state	idea	become	people	coach	thought	ever
message	business	thought	woman	young	give	share	help	give
change	time	team	change	change	time	true	give	start

Tabella 3.3: Parole con maggior frequenza identificate dalla *Latent Dirichlet Allocation* sul *corpus* dei riassunti.

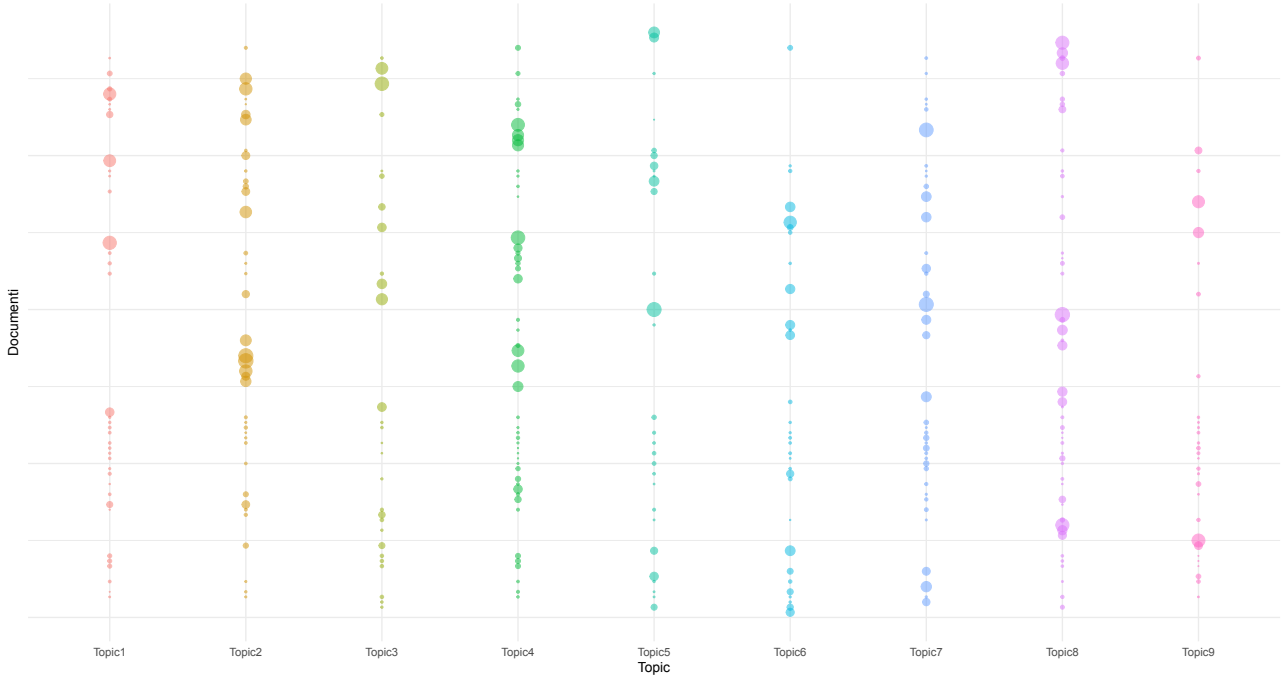

Figura 3.3: Distribuzione delle proporzioni di *topic* identificati dalla *Latent Dirichlet Allocation* sul *corpus* dei riassunti.

Nel caso del *topic 2* si potrebbe pensare ad una tematica imprenditoriale data la presenza di parole come: "*business*", "*networking*", anche in tale circostanza queste non risultano essere tra

le parole di maggior rilevanza all'interno de *topic* considerato. La distribuzione delle proporzioni di *topic* emersi nel *corpus* dei riassunti è visualizzabile in Figura 3.3, dalla quale si può notare come ogni *topic* individuato in questo *corpus* tenda a presentare una proporzione nella maggior parte dei documenti. Infatti, o un *topic* risulta completamente assente all'interno del *corpus*, oppure vi sono documenti che presentano un'equidistribuzione di più *topic*. Sicuramente si sottolinea come il *topic 4* e il *topic 2* siano quelli più diffusi all'interno del *corpus*.

Per identificare le etichette da utilizzare nella fase successiva viene selezionato il *topic* con la proporzione maggiore all'interno del documento. Chiaramente i riassunti, per quanto presumibilmente contengano un'informazione più diretta sulle tematiche affrontate, sono un *corpus* aggiuntivo ai dati testuali sui quali cercare delle tematiche sottostanti, in cui vi sia un vocabolario di una grande dimensione. Si passa dunque alla valutazione dei risultati ottenuti dalla *Latent Dirichlet Allocation* stimata sul *corpus* delle trascrizioni, per la quale tramite la Tabella 3.4 si visualizzano i diversi insiemi di parole maggiormente frequenti, per ogni *topic* stimato.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
could	actually	something	something	things	could	things	things
much	life	could	could	something	two	something	something
something	things	said	years	could	life	actually	much
things	something	much	things	always	something	could	could
two	could	years	actually	life	god	new	god
years	said	stuff	still	never	years	life	life
said	love	things	two	much	things	years	years
stuff	much	love	much	years	still	find	two
never	years	god	guys	said	world	much	actually
life	still	life	new	around	new	said	love
love	god	two	around	still	around	love	said
new	guys	guys	stuff	two	never	maybe	always
always	two	still	never	love	old	different	person
around	around	around	maybe	two	much	world	never
still	great	different	said	new	love	always	next
big	sure	actually	next	new	love	idea	stuff
maybe	always	sure	three	world	said	experience	great
sure	different	new	world	maybe	last	still	stuff
actually	never	always	life	big	actually	still	great
great	new	maybe	made	next	story	two	stuff

Tabella 3.4: Parole con maggior frequenza identificate dalla *Latent Dirichlet Allocation* sul *corpus* delle trascrizioni.

Figura 3.4: Distribuzione delle proporzioni di *topic* identificati dalla *Latent Dirichlet Allocation* sul *corpus* delle trascrizioni.

Contrariamente alle parole con la maggior frequenza identificate nei *topic* sul *corpus* dei riassunti, le parole evidenziate nel *corpus* delle trascrizioni dalla *Latent Dirichlet Allocation* risultano essere parole molto comuni che anche se valutate singolarmente non sembrerebbero rappresentare particolari tematiche oggettive. In Figura 3.4 si nota inoltre come la distribuzione di proporzioni all'interno del *corpus* sia notevolmente diversa dalla precedente. Nel caso presente, i documenti non sono rappresentati dall'insieme di più *topic* ciascuno, infatti per ogni documento vi è un unico *topic* con una grande proporzione al suo interno. In particolare i *topic 5,6,7* presentano una proporzione di rilevanza in pochi documenti, mentre i *topic 2* o *3* risultano prevalenti in circa la metà dei documenti. La distribuzione dei *topic* all'interno del *corpus* rappresentata in Figura 3.4 mostra la matrice di proporzioni utilizzata nella stima dello *Structural Topic Model*. La *Latent Dirichlet Allocation* viene ottimizzata secondo la metrica di *coherence UMass*, ma questa non garantisce che i *topic* identificati rappresentino delle tematiche ben distinte tra loro.

Poichè si dispone di due diversi corpora, risulta naturale provare ad incorporare i risultati ottenuti in questa prima fase di analisi con una seconda. In seguito oltre a rappresentare in maniera analoga a come fatto finora i risultati, i modelli che incorporano le informazioni di

entrambi i corpora vengono messi a confronto con quelli stimati in questa prima fase, tramite la comparazione delle ulteriori metriche di coerenza proposte precedentemente.

3.3 Stima sLDA e STM

Si presentano quindi le stime dei modelli appartenenti alla seconda fase di modellazione, nella quale si integrano ai corpora i risultati ottenuti durante la prima fase. Questi, infatti, presentano informazioni aggiuntive ai dati testuali di cui si dispone. Inoltre, sfruttando la struttura dei modelli considerati in questa fase, tali informazioni risultano utilizzabili in modo opportuno. Pertanto, in questa seconda fase di modellazione si utilizzano:

- le etichette dei *topic* identificate sul *corpus* dei riassunti, che rappresentano una quantità che può essere prevista, come una variabile risposta, utilizzata nella *Supervised Latent Dirichlet Allocation*;

- le proporzioni dei *topic* identificate sul *corpus* delle trascrizioni, queste possono rappresentare una quantità che apporti delle indicazioni preliminari per l'individuazione delle tematiche latenti, utilizzate nello *Structural Topic Model*.

L'idea di base rimane legata alla sequenzialità dei due diversi corpora, nella realtà prima un episodio viene registrato, pertanto, prima vi sono le trascrizioni, poi l'autore del podcast rilascia una breve sintetizzazione; quindi, i riassunti arrivano in un secondo momento. Secondo questa logica e la gerarchia dei modelli proposti, per la seconda fase di modellazione si è ritenuto opportuno tener fede a questo ordinamento temporale.

Si presentano quindi i risultati relativi alla *Supervised Latent Dirichlet Allocation* utilizzando come variabile risposta le etichette identificate dalla *Latent Dirichlet Allocation* sul *corpus* dei riassunti, come spiegato precedentemente nel secondo capitolo. Per quanto le etichette non rispecchino delle categorie specifiche di podcast, si vuole valutare se queste possano comunque migliorare l'identificazione di *topic* sul *corpus* delle trascrizioni. Come numero ottimale di *topic* da stimare in questo contesto, si fa uso delle informazioni derivanti dalla modellazione preliminare effettuata sul *corpus* dei riassunti, nel caso specifico quindi si andranno a stimare 9 diversi gruppi.

	Coherence UMass	Coherence UCI	Coherence NPMI
LDA	-0.3524	1.8936	0.101
sLDA	-0.8388	3.7449	0.135

Tabella 3.5: Metriche di coerenza relative al *corpus* delle trascrizioni.

Da subito come mostrato in Tabella 3.5, si confronta questo modello con la *Latent Dirichlet Allocation* stimata sul medesimo *corpus* al passo precedente. Si nota come le considerazioni differiscano al variare della metrica considerata, in particolare la *coherence UMass* viene massimizzata dal modello preliminare, mentre le metriche *coherence UCI* e *coherence NPMI* risultano maggiori per il modello *Supervised Latent Dirichlet Allocation*. Questo ad indicare che, nonostate sia minimo, si evidenzia comunque un miglioramento rispetto al modello stimato al primo passo sullo stesso *corpus*. Si passa quindi ad una valutazione qualitativa dei *topic* stimati dal seguente modello, come fatto precedentemente. Come si può vedere in Tabella 3.6, i *topic* individuati sembrano rappresentare delle categorie tematiche. In particolare, si potrebbero associare le seguenti tematiche ai rispettivi *topic*:

- *topic 1*: Religioso;
- *topic 2*: Sportivo;
- *topic 3*: Racconti personali;
- *topic 4*: Impreditoriale;
- *topic 5*: Cronaca nera;
- *topic 6*: *Gender Equality*;
- *topic 7*: Criminalità organizzata;
- *topic 8*: Cospirazioni sul volo *Malaysia Airlines 370*;
- *topic 9*: Generico.

Vi sono alcune parole all'interno di tali *topic* che possono comunque non far parte dell'area tematica a cui si possano associare. Si ha comunqune un miglioramento, evidenziato in primo luogo dalle metriche di coerenza, in aggiunta dalla suddivisione delle parole chiave per ogni argomento se si confrontano le Tabelle 3.6 e 3.4. Questo confronto potrebbe suggerire che identificare i *topic* sulle trascrizioni, utilizzando una variabile che possa rappresentare delle macrocategorie tematiche, nel caso presente le etichette derivanti dai riassunti, apporti un miglioramento delle rappresentazioni tematiche trovate. Si nota inoltre come i *topic* identificati

secondo tale procedura risultino equamente distribuiti all'interno del *corpus*, come mostrato in Figura 3.5 con una prevalenza associata al *topic 3* quello che si potrebbe definire relativo ai racconti personali e al *topic 9*, riguardante tematiche generali.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
god	something	love	business	police	brush	crime	plane	stuff
spirit	sort	person	content	crime	brushing	organized	flight	guys
fear	world	something	creative	murder	mouth	number	malaysian	pretty
lord	team	someone	industry	home	chrysippus	sleep	government	actually
ladies	athletes	care	idea	murders	side	waterfront	march	money
hallelujah	water	life	company	par	could	criminal	air	everybody
thank	whether	put	brand	killer	bottom	american	search	much
pray	things	better	job	cast	transgender	friendly	ocean	definitely
life	always	call	advertising	true	crime	arrest	missing	probably
said	great	shit	talent	found	source	gangster	malaysia	little
faith	practice	things	companies	brick	version	gang	indian	something
peace	coach	help	action	old	information	organized	passengers	people
word	important	friends	things	killed	switch	weight	conspiracy	big
cupping	training	anything	create	death	history	america	debris	things
verse	open	probably	world	unsolved	julian	mafia	traffic	last
sister	idea	kids	conversation	later	heroines	gangs	southwest	point
chapter	much	use	brands	conrad	key	muscle	court	getting
forgive	swim	trying	whether	college	women	groups	three	literally
house	israel	relationship	service	case	king	billion	lost	around

Tabella 3.6: Parole con maggior frequenza identificate dalla *Supervised Latent Dirichlet Allocation* sul *corpus* delle trascrizioni.

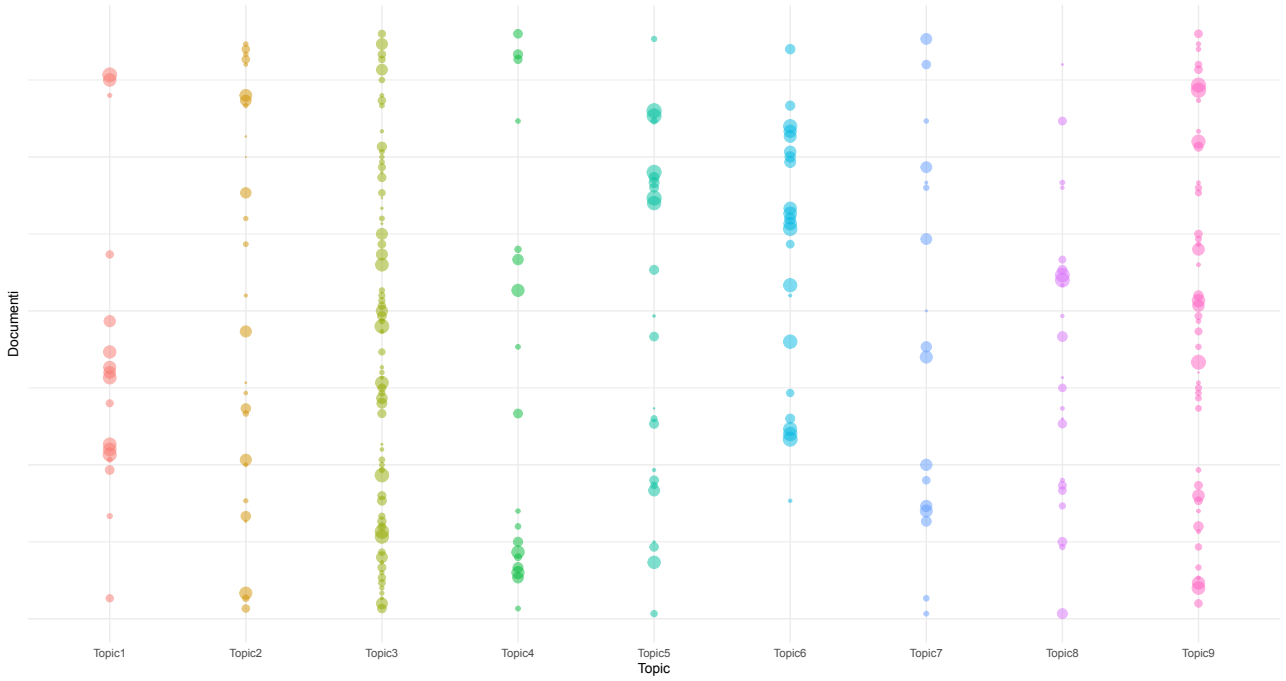

Figura 3.5: Distribuzione delle proporzioni di *topic* identificati dalla *Supervised Latent Dirichlet Allocation* sul *corpus* delle trascrizioni.

Si presenta ora l'ultimo modello, in cui si individuano *topic* latenti sul *corpus* dei riassunti. Questi vengono stimati tenendo conto delle proporzioni di *topic* precedentemente stimati dalla

Latent Dirichlet Allocation sul *corpus* delle trascrizioni. Tali informazioni vengono incorporate nello *Structural Topic Model* come covariate, che rientrano nel predittore lineare della distribuzione Logistica-Normale. Come fatto in precedenza dapprima vengono mostrati i valori delle

	Coherence UMass	**Coherence UCI**	**Coherence NPMI**
LDA	-7.3146	-4.6057	0.0581
STM	-11.4264	-8.0325	0.0180

Tabella 3.7: Metriche di coerenza relative al *corpus* dei riassunti.

metriche di coerenza proposte, in Tabella 3.7, la quale evidenzia come il modello stimato presenti dei livelli di coerenza sempre minori rispetto alla *Latent Dirichlet Allocation* stimata sul medesimo *corpus*. Inoltre in Tabella 3.8 non si discriminano particolari tematiche oggettive. Alcuni di questi *topic* come il terzo possono essere accostati a tematiche sportive, ma in generale non si evidenziano particolari miglioramenti rispetto ai *topic* individuati precedentemente dalla *Latent Dirichlet Allocation* sul medesimo *corpus*. In questa circostanza come suggerito dalla Figura 3.6, la distribuzione della proporzione di *topic* stimata, presenta uno sbilanciamento tra *topic* particolarmente diffusi all'interno del *corpus*, come il *topic 2*, *4*, *6* e *8*, ed altri molto rari come in particolare il *topic 3*.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
murder	resource	swim	come	true	make	talk	fitness
content	bachelor	athlete	natural	coach	series	small	daily
year	fight	find	start	heroine	change	idea	post
message	give	open_water	give	help	help	conversation	set
idea	people	swimming	game	also	ever	networking	roman
hour	make	message	know	woman	devotional	change	advantage
create	fear	make	heroine	potential	job	build	cupping
old	base	thought	host	date	get	avoid	drop
college	copy	young	experience	purpose	write	drug	mechanical
campus	face	career	year	shirt	guide	hit	prayer
strangle	generously	experience	time	love	loss	mind	question
student	ministry	coach	add	discuss	benefit	perhaps	say
know	number	take	fishing	interview	contribution	situation	people
become	sheaf	olympic	israeli	achieve	member	human	much
crime	twnabh	pool	sister	man	mystery	high	locate
blog	upci	time	corporation	card	step	thing	flight
discover	plan	need	introduce	video	weight	year	chatterbait
lady	possible	key	yoga	fear	state	sure	frame
generate	create	compete	share	breakup	fitness	conference	haunt
imagine	youth	disappointment	yoga	break	state	sure	frame

Tabella 3.8: Parole con maggior frequenza identificate dalla *Logistic Normal Topic Admixture Model* sul *corpus* dei riassunti.

Figura 3.6: Distribuzione delle proporzioni di *topic* identificati dallo *Structural Topic Model* sul *corpus* dei riassunti.

3.4 Commenti finali sull'analisi

Dopo aver valutato quantitativamente e qualitativamente i modelli di *topic* stimati è possibile effettuare un confronto tra essi. Non vi è un modello che presenta una netta superiorità in termini di prestazioni relative alle metriche di coerenza, sicuramente il modello *Supervised Latent Dirichlet Allocation* risulta apportare un minimo miglioramento rispetto al modello preliminare per quel che riguarda le metriche *coherence UCI* e *coherence NPMI*. Inoltre, se valutata da un punto di vista qualitativo tale modello presenta tematiche che possono risultare maggiormente oggettive all'essere umano. Contrariamente lo *Structural Topic Model* non migliora in termini quantitativi, secondo le metriche di coerenza, rispetto al modello preliminare.

In entrambi i modelli stimati sul *corpus* dei riassunti si evidenzia come in parte la sintetizzazione fornita dagli autori non racchiuda le parole chiave, anche se le etichette emerse in tale contesto hanno aiutato l'identificazione di gruppi tematici oggettivi come emerso dai risultati della *Supervised Latent Dirichlet Allocation*.

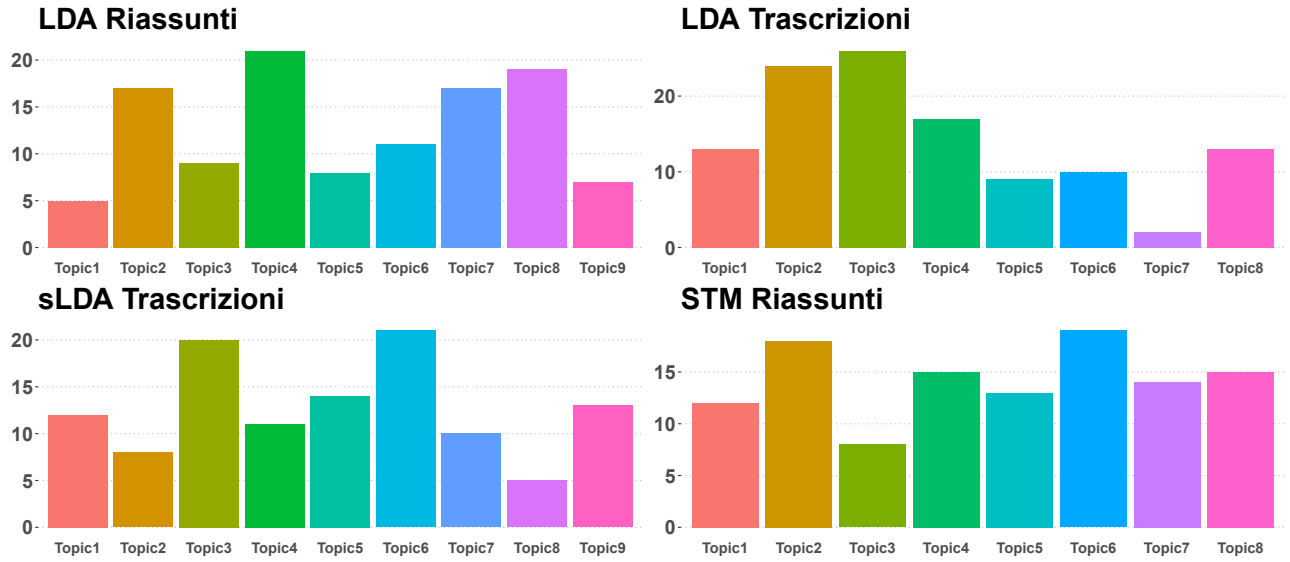

Figura 3.7: Frequenze assolute *topic* stimati: (Alto Sinistra) LDA sul *corpus* dei riassunti, (Alto Destra) LDA sul *corpus* delle trascrizioni, (Basso Sinistra) sLDA sul *corpus* delle trascrizioni, (Basso Destra) STM sul *corpus* dei riassunti.

Si sono valutate le distribuzioni delle proporzioni di *topic* individuate per ogni modello stimato, cercando di confrontare la distribuzione delle etichette stimate dai modelli e come queste variano tra quelli stimati al primo passo e quelli stimati nel secondo. In Figura 3.7 vengono riportati il numero di documenti che fanno parte di un *topic* al variare del modello considerato, si nota come la distribuzione tra il primo grafico in alto a sinistra relativo al modello LDA stimato sul *corpus* dei riassunti, presenti conteggi simili al grafico in basso a sinistra relativo al modello sLDA stimato sul *corpus* delle trascrizioni. Si deve precisare che nei *Topic Model* come in qualsiasi modello di raggruppamento, non si hanno delle etichette per gruppo, quindi il l'i-esimo *topic* non rappresenta la medesima tematica in entrambi i modelli, pertanto, si valuta il numero di documenti del *topic 4* nel grafico in alto a sinistra con il *topic 6* del grafico in basso a sinistra e cosi via. Si nota, quindi una somiglianza nel conteggio di documenti presenti in questi modelli, ad indicare come nel sLDA vengano tenute in considerazione le informazioni aggiuntive derivanti dal primo passo.

Contrariamente il grafico in basso a destra rappresentate il modello STM stimato sul *corpus* dei riassunti, non rispecchia i conteggi individuati dal modello LDA stimato sul *corpus* delle trascrizioni. Questo pone ulteriore attenzione sulla possibilità di sfruttare possibili etichettature per la ricerca di tematiche latenti all'interno dei documenti, infatti il *Topic Model* supervisionato tende a rispettare le proporzionalità imposte dall'etichettatura.

Sembra quindi che attraverso l'utilizzo di un'etichetta rappresenta delle macrocategorie di argomenti, le trascrizioni dei podcast prese in considerazione risultano esplicative riguardo eventuali argomenti latenti permettendo anche una possibile previsione futura, da cui ne potrebbe derivare una catalogazione automatizzabile dei podcast sulla piattaforma di Spotify.

Conclusioni

Nel presente elaborato si è proceduto all'impiego di una serie di *Topic Model* al fine di condurre un'analisi su un corpora di documenti attinente ai podcast disponibili su Spotify, presentato in Clifton et al. (2020). Tale corpora è costituito, per ogni singolo episodio, da due documenti distinti: il primo relativo alle trascrizioni integrali dei dati audio ottenute tramite *Google's Cloud Speech-to-Text API*, il secondo riferito ai riassunti forniti direttamente dall'autore.

L'obiettivo dell'analisi era di catalogare e sintetizzare i diversi episodi attraverso l'utilizzo di entrambi i documenti a disposizione, facendo emergere gli argomenti latenti trattati. Al fine di raggiungere tale obiettivo, è stata adottata una procedura in due passi: inizialmente, è stata effettuata l'estrazione di informazioni preliminari mediante l'utilizzo della *Latent Dirichlet Allocation* (Blei et al., 2003), che ha permesso di ottenere etichette descrittive e proporzioni tematiche utilizzate in seguito. Successivamente, nella seconda fase dell'analisi, sono stati stimati *Topic Model* più complessi. Più specificatamente, è stata applicata la *Supervised Latent Dirichlet Allocation* (Blei & Mcauliffe, 2007) al *corpus* delle trascrizioni, utilizzando come variabile risposta le etichette precedentemente stimate dalla *Latent Dirichlet Allocation* sul *corpus* dei riassunti. In aggiunta, è stato stimato lo *Structural Topic Model* (Roberts et al., 2016) sul *corpus* dei riassunti; quest'ultimo, permettendo l'inclusione di covariate a livello di documento, ha facilitato l'impiego delle proporzioni tematiche individuate nella prima fase sull'insieme delle trascrizioni.

Nonostante le aspettative, l'analisi comparativa tramite le metriche di coerenza proposte non ha rivelato miglioramenti sostanziali nei modelli stimati nella seconda fase rispetto a quelli della prima. Tuttavia, è doveroso riconoscere un incremento, sebbene marginale, della prestazione della *Supervised Latent Dirichlet Allocation*, la quale ha mostrato un miglioramento in termini di coerenza su due delle tre metriche considerate. Questo miglioramento si è manifestato in particolare attraverso la valutazione delle parole con la maggior frequenza nei diversi *topic*,

le quali hanno mostrato di essere più significative e rappresentative per la catalogazione in tematiche latenti oggettive.

Bibliografia

Ahmed, A. & Xing, E. P. (2007). Seeking the truly correlated topic posterior - on tight approximate inference of logistic-normal admixture model. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, M. Meila & X. Shen, eds., vol. 2 of *Proceedings of Machine Learning Research*. San Juan, Puerto Rico: PMLR.

Aletras, N. & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Potsdam, Germany: Association for Computational Linguistics.

Blei, D., Jordan, M. & Ng, A. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3* , 993–1022.

Blei, D. & Mcauliffe, J. (2007). Supervised topic models. In *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer & S. Roweis, eds., vol. 20. Curran Associates, Inc.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* **55**, 77 – 84.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877.

Broder, A. (1997). On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*.

Clifton, A., Reddy, S., Yu, Y., Pappu, A., Rezapour, R., Bonab, H., Eskevich, M., Jones, G., Karlgren, J., Carterette, B. & Jones, R. (2020). 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Friedman, J., Hastie, T. & Tibshirani, R. (2001). . *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin.

Hoffman, M., Bach, F. & Blei, D. (2010). Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel & A. Culotta, eds., vol. 23. Curran Associates, Inc.

Lafferty, J. & Blei, D. (2005). Correlated topic models. In *Advances in Neural Information Processing Systems*, Y. Weiss, B. Schölkopf & J. Platt, eds., vol. 18. MIT Press.

Maaten, L. & Geoffrey, H. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research 9* , 2579–2605.

Mahanty, S., Boons, F., Handl, J. & Batista-Navarro, R. (2019). *Studying the Evolution of the 'Circular Economy' Concept Using Topic Modelling*. pp. 259–270.

Mimno, D., Wallach, H., Talley, E., Leenders, M. & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Roberts, M. E., Stewart, B. M. & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association* **111**, 988–1003.

Roberts, M. E., Stewart, B. M. & Tingley, D. (2019). stm: An r package for structural topic models. *Journal of Statistical Software* **91**, 1–40.

Stevens, K., Kegelmeyer, P., Andrzejewski, D. & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12. USA: Association for Computational Linguistics.

Walker, D. E. (1982). Natural-language-access systems and the organization and use of information. In *International Conference on Computational Linguistics*.