Università degli Studi di Milano Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di laurea in Informatica

Automatic classification of Point of Interest

Relatore: Prof.ssa Vincenzina Messina

Co-relatore: Prof.ssa Elisabetta Fersini

Relazione della prova finale di:
Francesca Pulerà
Matricola 870005

Anno Accademico 2022-2023

Abstract

Il lavoro di questa tesi si è sviluppato nell'ambito del progetto RASTA (Realtà Aumentata e Story-Telling Automatizzato per la valorizzazione di Beni Culturali ed Itinerari). Obiettivo dello stage è stato quello di individuare ed analizzare tramite tecniche di Natural Language Processing i Point Of Interest (POI) e le relative descrizioni testuali per poter estrarre automaticamente i possibili TOI (Topic Of Interest) da utilizzare per il suggerimento di itinerari personalizzati. Come caso di studio sono state prese in considerazione le regioni Umbria e Abruzzo. Durante il lavoro di tesi, sono state acquisite competenze specifiche nel linguaggio di programmazione Python e nell'ambito del Natural Language Processing, con un focus specifico sui Topic Model.

Inizialmente, l'attenzione è stata posta sul recupero dei POI relativi ai beni culturali, partendo da risorse esterne e integrando basi di conoscenza (ad esempio Wikipedia) con basi di dati geografici (ad esempio Open Street Map).

In una seconda fase si sono esplorati differenti approcci (statistici e basati su modelli neurali) per l'estrazione degli argomenti relativi alle descrizioni dei POI precedentemente individuati finalizzati all'identificazione dei possibili TOI.

Gli approcci sono stati valutati studiando differenti metriche e scegliendo il modello migliore dopo un processo di ottimizzazione.

Sommario

Capitolo 1
Introduzione

Comprendere, generare e manipolare il linguaggio umano è uno degli obiettivi che si pone l'Intelligenza Artificiale (AI) grazie ad un insieme di tecniche specifiche per l'elaborazione del linguaggio naturale, dette "Natural Language Processing" (NLP).

Negli ultimi anni l'elaborazione del linguaggio naturale ha assunto sempre più importanza e il suo campo di applicazione si è sviluppato esponenzialmente, permettendo di apportare miglioramenti significativi in diversi settori.

In questa tesi si presentano lo studio e l'applicazione di tecniche di NLP nell'ambito della valorizzazione turistico-culturale, in particolare all'interno del progetto RASTA (Realtà Aumentata e Story-Telling Automatizzato per la valorizzazione di Beni Culturali ed Itinerari).

Nei prossimi capitoli verranno descritti approcci statistici e approcci neurali per classificare automaticamente dei Point of Interest (da cui il titolo "Automatic classification of Point of Interest").

Per fare ciò, è stato realizzato, attraverso l'integrazione di differenti basi di dati, un dataset che include le descrizioni relative ai punti di interesse. In particolare, come caso di studio si sono prese in considerazione le regioni Umbria e Abruzzo.

1.1 Struttura della tesi

Il lavoro di tesi che segue sarà così articolato:

si presenterà nel *Capitolo 2* il dataset e come questo è stato realizzato a partire dalla raccolta di informazioni geolocalizzate di punti di interesse. Inoltre, saranno effettuate anche delle analisi inerenti ad esso, mirate a definire le proprietà che rendono tale dataset fondamentale per i task dei capitoli successivi.

Nel *Capitolo 3* si studierà lo stato dell'arte. Verrà fornita una panoramica del Natural Language Processing (NLP), concentrandosi in particolare sull'analisi e il riconoscimento delle diverse metodologie applicate nel topic modeling. Durante questa fase di studio, si descriveranno le principali tipologie di approcci: il primo di tipo frequentista e il secondo di tipo neurale. Inoltre, si presterà particolare attenzione al framework OCTIS (Optimizing and Comparing Topic models Is Simple), che riveste un ruolo fondamentale nel corso della tesi. Ne verrà effettuata una descrizione esaustiva, seguita da un'analisi delle sue principali caratteristiche e funzionalità, in quanto si tratta di uno strumento di grande importanza per la ricerca e lo sviluppo dei topic model trattati nel contesto della tesi.

Nel capitolo successivo, ovvero il *Capitolo 4,* verrà presentato l'approccio innovativo proposto per l'estrazione dei Topic of Interest (TOI) dal dataset. Saranno fornite dettagliate spiegazioni sui modelli statistici e neurali utilizzati per questo scopo. Verranno descritti i passaggi chiave del processo di addestramento, evidenziando le scelte metodologiche adottate per affrontare le sfide specifiche legate al progetto in questione. Inoltre, saranno discusse le motivazioni che hanno guidato tali scelte.

Infine, il *Capitolo 5* conterrà le conclusioni tratte dall'analisi dei risultati e le proposte di sviluppi futuri.

Si allegano inoltre, in *Appendice A* alcune analisi condotte al fine di studiare meglio il dataset.

Capitolo 2

Dataset

Il processo di creazione del dataset è avvenuto integrando diverse basi di dati al fine di includere le descrizioni dei beni culturali presenti nell'Umbria e nell'Abruzzo. L'obiettivo principale era fornire un ampio insieme di dati che potesse essere utilizzato per sviluppare e applicare algoritmi di NLP.

2.1 Recupero dei POI da OpenStreetMap

Il primo importante step è consistito nel reperire informazioni geolocalizzate per la creazione del dataset di interesse.

A seguito di un’approfondita ricerca sulle fonti e sulla disponibilità dei dati, si è scelto OpenStreetMap come servizio open source da cui recuperare le informazioni geolocalizzate necessarie.

OpenStreetMap (OSM) è un servizio di mappatura globale che consente a chiunque di visualizzare, proporre modifiche e utilizzare i dati geografici in modo libero.

A differenza di altri servizi di mappe online, come Google Maps, OpenStreetMap permette agli utenti di contribuire attivamente alla creazione e all'aggiornamento delle informazioni presenti.

OpenStreetMap associa ad ogni elemento un insieme di etichette (‘tags’). Quest’ultime sono formate da una coppia di dati (chiave-valore), la quale può descrivere sia caratteristiche geografiche, che proprietà peculiari che contraddistinguono ciascun elemento [1].

Ogni elemento presente in questo database geografico è quindi caratterizzato da un insieme di attributi, che aiutano ad organizzare e categorizzare i dati geografici (in OSM). Successivamente ad un'analisi preliminare, si è effettuata una prima scrematura delle chiavi di interesse, selezionando infine le tre principali: *historic*, *amenity* e *tourism*. In *Appendice A* si riporta la tabella comprensiva di tutte le categorie che sono state selezionate successivamente alla prima valutazione.

Nella *Figura 1* è presente un grafico riassuntivo della prima selezione delle etichette di interesse. La legenda, posizionata a destra del grafico, fornisce un elenco delle chiavi, in cui sono evidenziate le tre selezionate durante l’ultima scrematura, mentre il numero associato a ciascuna di esse rappresenta la quantità di valori corrispondenti.

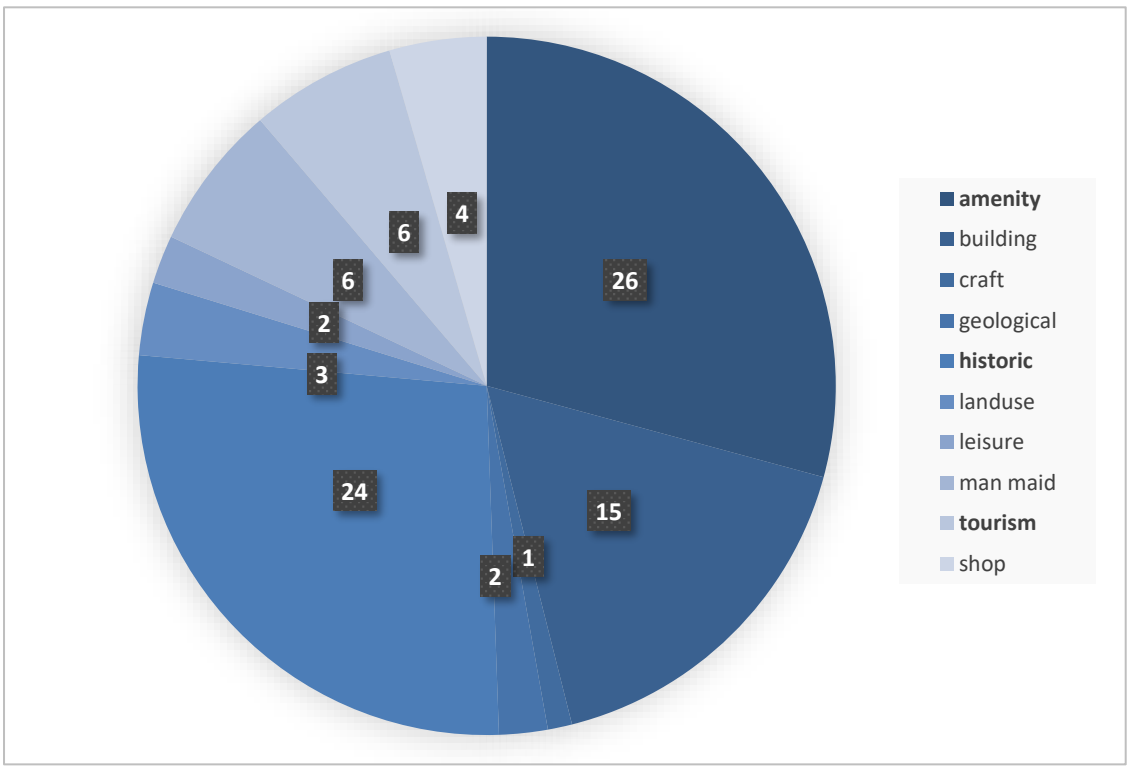

Figura 1: Grafico a torta delle chiavi risultanti dopo la prima scrematura. A ciascuna di esse è associato un numero corrispondente alla quantità di valori relativi

Nella *Tabella 1* sono evidenziate le tre chiavi principali con i relativi valori:

amenity		tourism	historic	
library	grave_yard	artwork	fort	aircraft
research_institute	monastery	attraction	gallows	aqueduct
training	place_of_worship	gallery	highwater_mark	archaeological_site
music_school	city_wall	museum	locomotive	battlefield_site
university	national_park	viewpoint	memorial	bomb_crater
arts_centre	protected_area	zoo	mine	building
fountain	school		milestone	cannon
community_centre	college		monastery	castel
cinema	language_school		monument	castell_wall
music_venue	courthouse		pillory	charcoal_pile
planetarium	townhall		manor	church
public_bookcase	ditch			city_gate
theatre	studio			citywalls

Tabella 1: Rappresenta le tre chiavi selezionate in ultimo e i relativi valori

2.2 Analisi dei dati

Durante la fase iniziale di creazione del dataset, sono quindi stati scaricati tutti i dati relativi all'Umbria e all'Abruzzo che rientravano nelle macrocategorie *amenity*, *historic* e *tourism*.

In seguito, sono state condotte delle analisi sui dati scaricati per valutare la qualità e la quantità di informazioni presenti in OSM.

Nella *Tabella 2*, riguardante un'analisi sui dati estratti relativi alla città di Perugia, e nella *Tabella 3*, concernente un'analisi sui dati estratti relativi alla città dell'Aquila, sono presenti due esempi significativi che mettono in luce la necessità di ricorrere ad ulteriori fonti di informazione, come Wikipedia, per ottenere descrizioni complete ed esaustive dei punti di interesse precedentemente selezionati. Infatti, risulta evidente che solo una quantità esigua dei punti di interesse dispone di una descrizione associata.

PERUGIA	amenity	historic	tourism
Numero totale di POI presenti su OSM	1774	924	438
Numero di POI con relative descrizioni presenti su OSM	64	16	59

Tabella 2: Illustra un confronto tra il totale dei POI relativi a Perugia presenti su OSM e il sottoinsieme dei POI di tale città aventi una descrizione su OSM

L'AQUILA	amenity	historic	tourism
Numero totale di POI presenti su OSM	1290	853	935
Numero di POI con relative descrizioni presenti su OSM	15	19	12

Tabella 3: Illustra un confronto tra il totale dei POI relativi a L'Aquila presenti su OSM e il sottoinsieme dei POI di tale città aventi una descrizione su OSM

Wikipedia rappresenta una risorsa preziosa in quanto fornisce informazioni dettagliate e descrittive sui beni culturali presenti nelle aree di interesse.

2.3 Collegamento dei POI alle relative descrizioni

Grazie alle coordinate geografiche di longitudine ("lon") e latitudine ("lat") è stato possibile stabilire una relazione tra i singoli dati estratti da OSM, che rappresentavano i punti di interesse delle due regioni, e le rispettive pagine su Wikipedia.

Per farlo, è stata utilizzata la libreria Python "*wikipediaapi*", utile per l'accesso e l'estrazione di informazioni da Wikipedia. Tramite l'utilizzo di questa libreria si è reso possibile effettuare la ricerca delle pagine Wikipedia relative ai punti di interesse, accedere al testo delle pagine stesse, recuperare le informazioni

dettagliate da associare a ciascun POI e, infine, integrarle al dataset in questione. L'operazione di collegamento tra il POI e la relativa descrizione, comunemente nota come "linking", è stata effettuata tramite una ricerca delle pagine di Wikipedia che contenevano informazioni inerenti alle coordinate geografiche fornite come input di ricerca. Tale procedura è stata implementata utilizzando la libreria Python menzionata in precedenza.

Studiando le descrizioni ottenute dei POI, si evidenzia che la media della lunghezza dei documenti che compongono il dataset è di circa 800 parole.

I grafici sottostanti riassumono i risultati del calcolo delle lunghezze delle descrizioni estratte da Wikipedia. La *Figura 2* mostra la lunghezza dei documenti relativi ai punti di interesse umbri, mentre la *Figura 3* mostra la lunghezza dei documenti relativi ai punti di interesse abruzzesi.

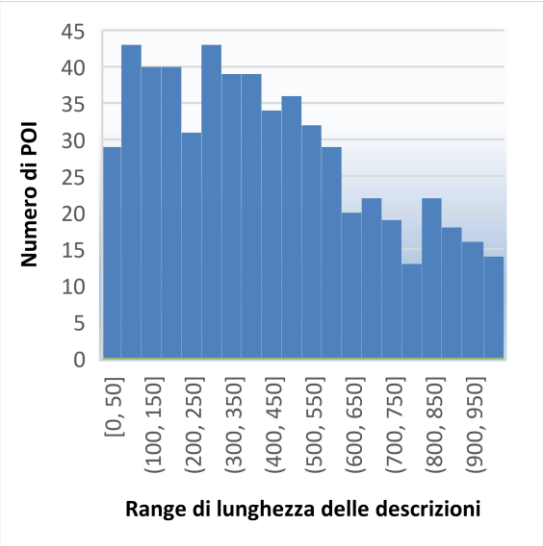

Figura 2: Grafico che rappresenta la lunghezza dei documenti relativi ai POI Umbri. Media=712 parole

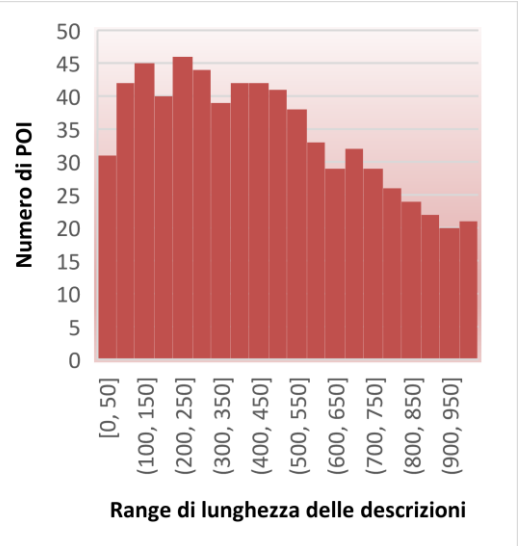

Figura 3: Grafico che rappresenta la lunghezza dei documenti relativi ai POI Abruzzesi. Media=926 parole

Dopo aver associato i POI alle relative pagine Wikipedia, è stato ottenuto il dataset utilizzato successivamente per poter procedere all'addestramento, l'analisi e il confronto dei Topic Model.

Nella *Figura 4*, viene presentata una sezione della mappa dell'Umbria ottenuta da OpenStreetMap, in cui sono stati individuati e segnalati i punti di interesse presenti nel dataset.

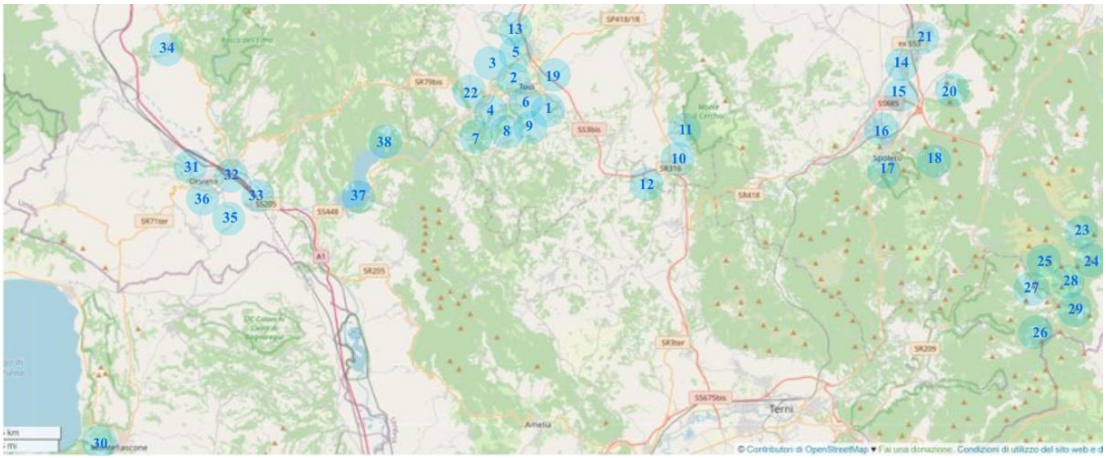

Figura 4: Cartina di OSM in cui sono evidenziati alcuni dei POI estratti

1	Chiesa di San Fortunato	**20**	Stazione di Cortaccione
2	Palazzo dei Priori	**21**	Chiesa della Madonna della Bianca
3	Palazzo del Popolo	**22**	Convento di Montesanto
4	Palazzo del Capitano	**23**	Rifugio Giovanni Giacomini
5	Concattedrale della Santissima Annunziata	**24**	Sito archeologico Forca di Ancarano
6	Museo civico di Todi	**25**	Basilica di San Benedetto
7	Torre Caetani	**26**	Concattedrale di Santa Maria Argentea
8	Tempio di Santa Maria della Consolazione	**27**	Stazione di Villa di Serravalle
9	Chiesa di Santa Maria in Camuccia	**28**	Basilica inferiore di Santa Rita da Cascia
10	Chiesa di San Felice	**29**	Santuario di Santa Rita da Cascia
11	Santuario della Madonna della Pace	**30**	Circuito internazionale di Viterbo
12	Santuario dell'Amore Misericordioso	**31**	Stazione di Orvieto
13	Stazione di Todi Ponte Rio	**32**	Tempio del Belvedere
14	Stazione di San Giacomo di Spoleto	**33**	Fortezza Albornoz
15	Chiesa Parrocchiale di San Sabino	**34**	Castello della Sala
16	Stazione di Spoleto	**35**	Chiesa di Santa Maria dei Servi
17	Teodelapio (Alexander Calder)	**36**	Cappella Petrucci
18	Villa Redenta	**37**	Lago di Corbara
19	Chiesa del Santissimo Crocifisso	**38**	Parco fluviale del Tevere

Tabella 4: Mostra i POI evidenziati sulla mappa riportata nella Figura 4

Capitolo 3

Stato dell’arte

Il Natural Language Processing (NLP) è una disciplina che si pone a metà tra l’informatica e la linguistica. Si occupa di estrarre informazioni da documenti scritti in linguaggio naturale, con l’obiettivo di automatizzare alcune attività svolte degli esseri umani per delegarle alle macchine.

Il ruolo fondamentale del NLP è quello di mediare tra linguaggio umano (indefinito e ambiguo) e linguaggio macchina (definito e con regole formali).

3.1 Le applicazioni di NLP

Ai giorni nostri il Natural Language Processing è in continua espansione e, grazie alle nuove tecnologie informatiche, ricopre un ruolo sempre più fondamentale in differenti aree di applicazione, come, ad esempio, per quanto riguarda i motori di ricerca (Google, Bing, Yahoo, …) e le traduzioni automatiche (Google Translate, DeepL, …). Inoltre, l'aumento significativo della potenza di calcolo dei computer più recenti ha notevolmente agevolato l'esecuzione e l'elaborazione di algoritmi anche molto complessi e computazionalmente onerosi.

3.1.1 Task dell'NLP

L'elaborazione del linguaggio naturale comprende numerosi task che coinvolgono l'analisi, la comprensione e la generazione del linguaggio umano mediante sistemi informatici. Tra i principali task dell'NLP si annoverano [2]:

- Text Analysis:
  Analisi di un testo e, laddove richiesto, individuazione di elementi chiave (es. argomenti, persone, date);
- Text Classification:
  Interpretazione di un testo per classificarlo in una categoria predefinita (es. spam);
- Sentiment Analysis:
  Rilevamento dell'umore all'interno di un testo (es. recensione positiva/negativa);
- Intent Monitoring:
  Comprensione del testo per prevedere comportamenti futuri (es. la volontà di acquisto da parte di un cliente);
- Smart Search:
  Ricerca, all'interno di archivi, dei documenti che meglio corrispondono ad un'interrogazione posta in linguaggio naturale;
- Text Generation:
  Generazione automatica di un testo;
- Automatic Summarization:
  Produzione di una versione sintetica di uno o più documenti testuali;
- Language Translation:
  Traduzione di testi scegliendo, volta per volta, il significato migliore a seconda del contesto.

Questi molteplici compiti sono spesso combinati per conseguire task più complessi in ottica di una comprensione più approfondita del linguaggio naturale in vari contesti di analisi testuale.

I task di interesse di questa tesi sono *Text analysis* e *Text classification*.

3.2 Topic Modeling

Con *Topic Modeling* si fa riferimento ad uno specifico task del NLP che permette di identificare gli argomenti presenti in un insieme di testi non etichettati e, successivamente, raggruppare quest'ultimi.
Si tratta di un task che segue un approccio non supervisionato, in quanto si basa su un dataset non etichettato di cui non si conosce a priori l'insieme degli argomenti presenti.

Nella rappresentazione visiva relativa alla *Figura 5* è illustrato un esempio dell'applicazione di algortimi di topic modeling al fine di individuare 9 principali topic (indicati con riquadri di dimensione maggiore) in un insieme di testi. Inoltre, si può notare la possibilità di associare ad uno o più topic le parole presenti nei documenti.

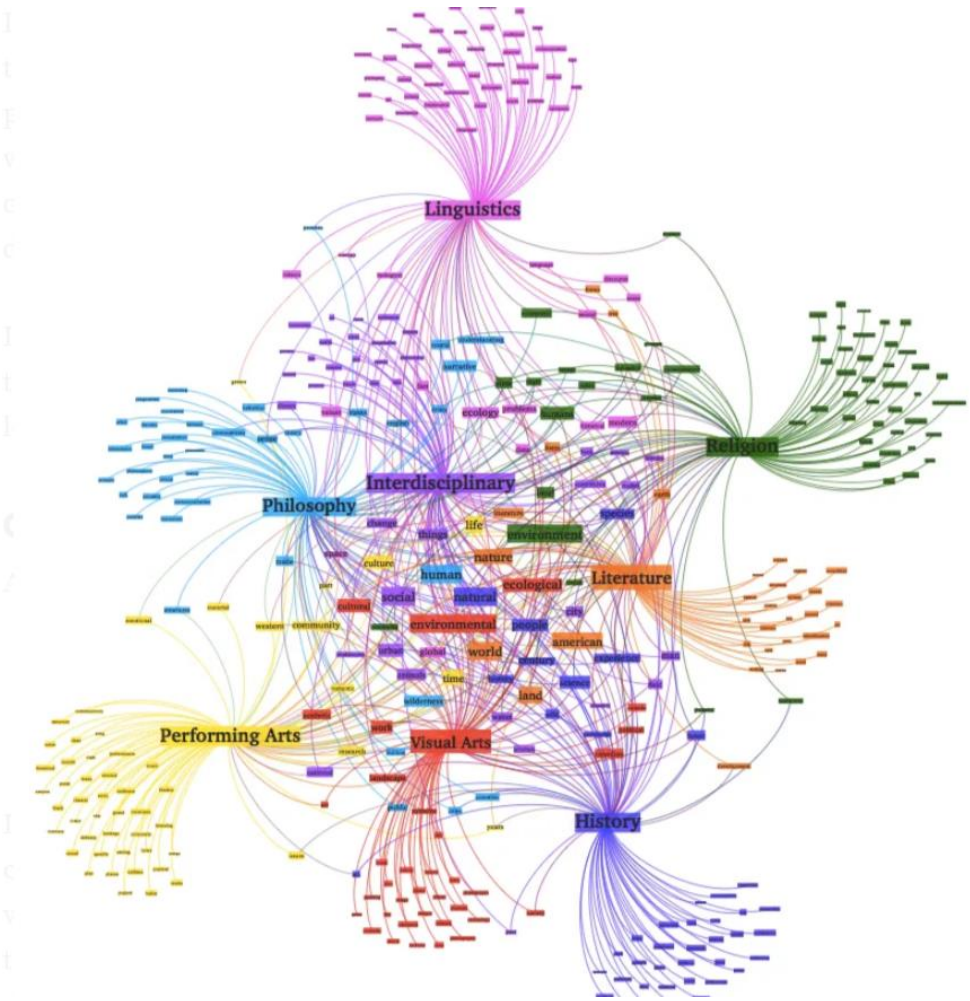

Figura 5: Topic Modeling

Esistono diversi approcci di topic modeling, il cui scopo è assegnare un topic (o argomento) ad ogni documento della collezione: quest'informazione permette di avere una sintesi facilmente interpretabile del documento di riferimento.

La *Figura 6* fornisce un'idea del concetto di topic modeling: i documenti vengono dati in input al modello e questo si occupa di estrarre i topic o argomenti principali.

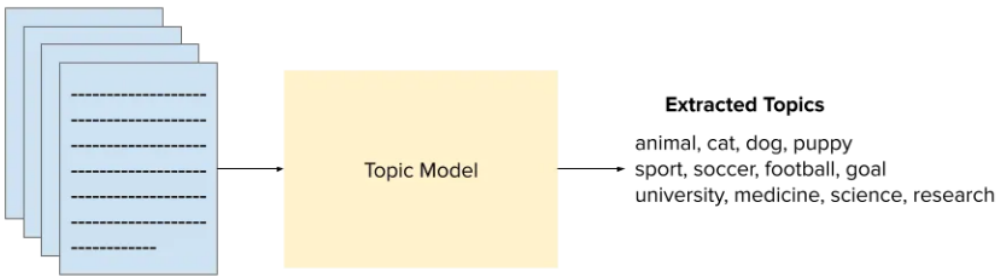

Figura 6: concetto generale di "topic modeling"

I due principali tipi di approcci utilizzati nel topic modeling sono quello neurale e quello frequentista. Entrambi mirano ad individuare le strutture tematiche nascoste nei testi per rivelare informazioni rilevanti e scoprire i temi sottostanti, al fine di facilitarne la comprensione. Nel proseguo verranno approfondite le principali differenze sui due tipi di approcci.

Ad oggi, uno dei modelli più popolari di topic modeling è il Latent Dirichlet Allocation (LDA), introdotto nel 2003 da David Blei, Andrew Ng e Michael Jordan.

Prima di intraprendere un'analisi più approfondita sul funzionamento della Latent Dirichlet Allocation e degli altri modelli, è opportuno fornire una definizione chiara di due termini fondamentali che verranno frequentemente utilizzati nel contesto di questa tesi:

- Un *corpus* rappresenta una collezione di testi che vengono utilizzati come base di dati per condurre ricerche linguistiche o per addestrare modelli di apprendimento automatico.
- Un *topic model* è un modello il cui obiettivo è trovare i topic (o argomenti) astratti contenuti in un insieme di documenti (*corpus*).

I topic non sono noti a priori, ma vengono identificati autonomamente dall'algoritmo in base alla frequenza e al numero di occorrenze delle parole nei vari testi. Ad esempio, le parole "croce" e "chiostro" compariranno molto di frequente nei testi relativi alle chiese, mentre parole come "parlamento" e "legislatura" saranno univocamente associate a testi relativi alla politica.

Sfruttando statistiche di questo tipo un *topic model* sarà quindi in grado di individuare un numero arbitrario di tematiche generali (i cosiddetti *topic*) presenti nei vari testi ed assegnare correttamente ad ogni testo il suo rispettivo topic semantico.

LDA è un modello generativo che assegna ai documenti di un corpus una combinazione di topic, dove ciascun argomento è rappresentato come una distribuzione di parole.

Negli ultimi anni, sono state introdotte numerose varianti di LDA che hanno mantenuto una rilevanza significativa sia nella ricerca accademica che nell'applicazione pratica. Ad esempio, sono state introdotte versioni più complesse dell'algoritmo di Latent Dirichlet Allocation, come la Supervised LDA (sLDA) che incorpora informazioni supervisionate nella modellazione dei temi.

Un'ulteriore tendenza nell'ambito del topic modeling consiste nell'integrare congiuntamente innovative tecniche di elaborazione del linguaggio naturale, quali la rappresentazione compatta delle parole nei documenti (word embeddings) e l'uso di nuove architetture, come le reti neurali. Negli ultimi anni quest'ultime sono infatti state ampiamente adottate nel NLP grazie alla loro capacità di apprendimento automatico e alla loro flessibilità nel modellare relazioni complesse presenti nei testi.

Le reti neurali trovano applicazione in diverse aree del Natural Languge Processing, tra cui:

- Classificazione del testo:

  Le reti neurali possono essere addestrate per classificare i testi in categorie o etichette specifiche. Ad esempio, possono essere impiegate per analizzare il sentiment di un testo (determinando se è positivo, negativo o neutrale),

classificare notizie in categorie specifiche, o identificare le intenzioni degli utenti in un sistema di chatbot.

- Generazione del testo:
  Le reti neurali possono essere utilizzate per generare testo coerente e significativo. Attraverso l'addestramento, possono generare automaticamente descrizioni di immagini, creare dialoghi o produrre testi creativi.
- Traduzione automatica:
  Le reti neurali possono essere impiegate per la traduzione automatica di testi da una lingua all'altra. I modelli di traduzione automatica basati su un approccio di tipo neurale, come l'architettura transformer, si sono rivelati molto promettenti.
- Riconoscimento e generazione del linguaggio naturale:
  Le reti neurali possono essere utilizzate per compiti come il riconoscimento del parlato (convertendo l'audio in testo) e la generazione di parlato (convertendo il testo in audio).
- Elaborazione del linguaggio naturale basata sul contesto:
  Le reti neurali possono essere addestrate per comprendere e generare testo considerando il contesto circostante. Ad esempio, i modelli di linguaggio basati su transformer (come GPT) sono in grado di generare testo coerente tenendo conto del contesto precedente.

Questi esempi illustrano solo alcune delle molteplici applicazioni delle reti neurali nell'NLP.

La combinazione di metodi statistici e di apprendimento automatico basati su reti neurali ha contribuito ad importanti progressi nel campo del Natural Language Processing negli ultimi anni.

È inoltre importante sottolineare che la ricerca nell'ambito del topic modeling è ancora in corso e continuano ad emergere nuovi metodi e miglioramenti.

Nella *Figura 7* viene illustrato un esempio di come una piccola parte di testo di un singolo documento è stata classificata tramite il topic modeling: le figure colorate a sinistra del diagramma descrivono i topic individuati dal modello e le parole in ciascun riquadro sono quelle che compaiono più frequentemente all'interno di ogni topic [3].

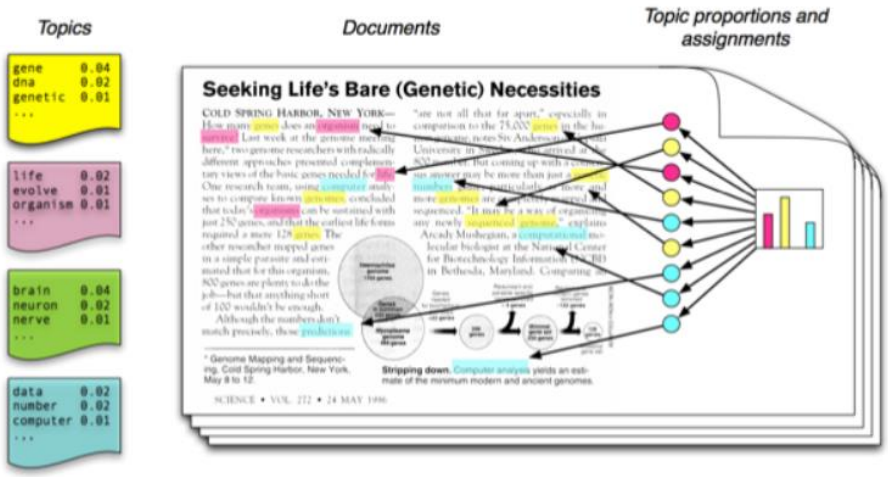

Figura 7: Processo per estrarre i topic dai documenti in modo non supervisionato

3.2.1 Approcci statistici vs approcci neurali

I modelli di topic model, sia statistici che neurali, rappresentano due approcci distinti per identificare ed analizzare gli argomenti presenti in un insieme di testi. Di seguito sono riportate le principali differenze tra i due:

- Rappresentazione dei testi:
  Nei modelli statistici, i documenti vengono solitamente rappresentati mediante matrici di conteggio delle parole, in cui ogni riga rappresenta un documento e ogni colonna rappresenta una parola. Nei modelli neurali, invece, si fa spesso uso di rappresentazioni compatte dei documenti, come i word embeddings, che catturano le relazioni semantiche tra le parole.
- Capacità di apprendimento:
  I modelli statistici di topic model sono in grado di identificare i temi basandosi sulle co-occorrenze statistiche delle parole nei testi. Tuttavia, possono presentare limiti nel rilevare sfumature e complessità presenti nei

dati testuali. Al contrario, i modelli neurali, grazie alla loro capacità di apprendimento più flessibile, possono individuare relazioni semantiche più complesse tra le parole e catturare informazioni contestuali.

- Interpretabilità:
  Un vantaggio dei modelli statistici di topic model risiede nella loro interpretabilità. I temi estratti sono spesso rappresentati come distribuzioni di parole, agevolando la comprensione delle parole associate a ciascun tema. Nei modelli neurali, invece, la rappresentazione interna dei temi può risultare meno interpretabile, poiché le reti neurali apprendono rappresentazioni implicite dei temi.
- Dimensione dei dati:
  I modelli statistici di topic model sono generalmente più adatti per dataset di piccole e medie dimensioni, mentre i modelli neurali possono gestire anche dataset di dimensioni maggiori grazie alla loro scalabilità e alle tecniche di addestramento su larga scala.

In generale, i modelli statistici di topic modeling sono stati ampiamente utilizzati garantendo una buona interpretabilità dei risultati. I modelli neurali di topic modeling, invece, rappresentano un campo di ricerca più recente che potenzialmente offre prestazioni superiori su dati complessi, ma richiede tipicamente maggiori risorse computazionali e può presentare una minore interpretabilità. La scelta tra i due approcci dipende dalle specifiche esigenze dell'applicazione e dalla natura dei dati testuali da analizzare.

3.3 OCTIS

Nel contesto di questa tesi, è stato utilizzato OCTIS (Optimizing and Comparing Topic models Is Simple), un framework open-source per l'addestramento, l'analisi e il confronto di Topic Model i cui hyper-parametri ottimali sono stimati utilizzando un approccio di Ottimizzazione Bayesiana.

OCTIS [4] consente a ricercatori e professionisti di avere un confronto equo tra i diversi modelli statistici e neurali che propone, ed una visualizzazione interattiva dei risultati per comprendere ciascun modello.

3.3.1 Principali contributi

Di seguito elencati i principali contributi del framework proposto [4]:

- Diversi topic model sono stati integrati in un framework unificato, fornendo un'interfaccia comune che consente agli utenti di sperimentare facilmente i topic model;
- È stato integrato un approccio di ottimizzazione bayesiana a singolo obiettivo per determinare i valori ottimali degli hyper-parametri di ciascun modello;
- Una visualizzazione interattiva dei risultati per ispezionare i dettagli dei modelli;

3.2.2 Progettazione e architettura del sistema

Le funzionalità più rilevanti di OCTIS sono relative alla pre-elaborazione (*preprocessing*) dell'insieme di dati, all'addestramento dei topic model, alla stima delle metriche di valutazione, all'ottimizzazione degli hyper-parametri (*hyper-parameter optimization*) e alla visualizzazione interattiva della web dashboard.

Nella *Figura 8* è riportato il diagramma dell'interazione tra i differenti moduli di cui OCTIS si compone:

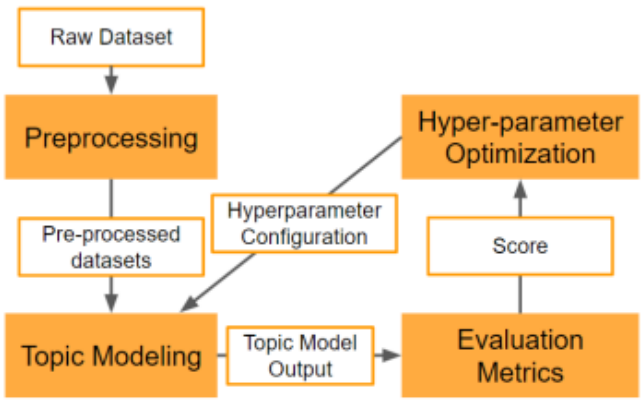

Figura 8: Flusso di lavoro del framework OCTIS [4]

- **Preprocessing**:
  Il primo passaggio della pipeline di modellazione dei topic implica una serie di operazioni (che verranno approfondite successivamente), tra cui la

conversione del testo in minuscolo, la rimozione della punteggiatura, la lemmatizzazione, la rimozione delle stopwords, la rimozione delle parole più e meno frequenti e la rimozione dei documenti con un numero ridotto di parole. Tuttavia, è importante notare che alcune di queste funzionalità potrebbero non essere adatte per lingue o domini specifici. Pertanto, OCTIS offre la flessibilità di disabilitare, abilitare o personalizzare i parametri di tali operazioni in base alle esigenze specifiche dell'utente. Inoltre, sebbene OCTIS offra già alcuni dataset predefiniti, gli utenti hanno la possibilità di caricare e pre-elaborare qualsiasi dataset di loro interesse utilizzando la libreria Python fornita.

- **Topic Modeling**:

  OCTIS presenta un'integrazione di modelli di topic sia statistici che neurali, tra cui il Latent Dirichlet Allocation (LDA), il Product-of-Experts LDA (ProdLDA) e i Contextualized Topic Models (CTM). Questi sono i tre modelli specificamente utilizzati durante il corso del lavoro svolto.

  OCTIS consente di utilizzare i differenti topic model come una "scatola nera", ossia un sistema in cui, dati degli input, si osservano solo gli output, senza poter vedere il suo funzionamento interno. Questa black-box riceve in input un dataset e un insieme di hyper-perparametri, un insieme di valori che caratterizzano la struttura del modello. Il modello restituisce la distribuzione dei topic nei vari documenti e le prime k parole, denominate "*top-k parole*", che rappresentano ciascun topic identificato. Le *top-k parole* indicano le parole più rilevanti o indicative associate a ciascun topic identificato dal modello. Queste parole rappresentano i termini più significativi e distintivi all'interno di un determinato argomento.

  Il valore k ed il numero di topic sono parametri arbitrari e possono essere specificati in base alle necessità del contesto.

- **Evaluation Metrics**:

  Le performance di un topic model possono essere valutate analizzando diversi aspetti, secondo le seguenti misure di valutazione messe a disposizione da OCTIS:

  - Topic coherence

    Misura quanto le *top-k parole* di un topic sono correlate fra loro
  - Topic significance

Valuta l'importanza o la rilevanza di un determinato argomento all'interno di un dataset o di un contesto specifico

- Topic diversity

  Valuta quanto gli argomenti all'interno di un insieme di dati siano diversificati e vari

- Topic classification

  Viene impiegata per valutare l'abilità di un modello nell'assegnare correttamente i documenti o le istanze ai rispettivi argomenti di riferimento.

- **Hyper-parameter Optimization**:

  OCTIS sfrutta l'ottimizzazione bayesiana come metodo per individuare gli hyper-perparametri ottimali, adottando un approccio efficace ed efficiente per determinare la configurazione migliore. L'ottimizzazione bayesiana esplora diverse combinazioni di valori per gli hyper-parametri, identificando l'insieme più promettente tra di essi.

  Per valutare questo insieme di hyper-parametri, vengono adottate metriche di riferimento come la *topic diversity* e la *topic coherence*.

Capitolo 4

Approccio proposto

Il caso di studio in oggetto prende in considerazione i documenti relativi alle descrizioni di ciascun Point of Interest delle regioni Umbria e Abruzzo. Come sottolineato in precedenza, i documenti utilizzati sono stati estratti da pagine Wikipedia relative agli specifici POI. Pertanto, questi documenti contengono tutte le informazioni tipiche che normalmente si possono trovare sul sito Web in questione, come ad esempio la storia, l'architettura, gli scavi, le note e la bibliografia associate al POI.

L'obiettivo è riuscire ad estrarre dei topic dall'insieme di documenti (*corpus*) presenti all'interno del dataset creato, i quali saranno i Topic of Interest candidati da utilizzare per il suggerimento di itinerari personalizzati all'interno del progetto RASTA.

Per il perseguimento dello scopo di questa tesi si è deciso quindi di utilizzare tre diversi topic model (uno di tipo statistico e due di tipo neurale) presenti in OCTIS, il framework introdotto nel capitolo precedente. Verranno qui mostrate le motivazioni di questa scelta, l'evoluzione del metodo sviluppato e varie considerazioni sui risultati ottenuti.

4.1 LDA

Il modello *Latent Dirichlet Allocation* (LDA) rappresenta il primo dei modelli di topic modeling che sono stati impiegati come parte integrante del presente progetto. LDA è un modello frequentista che permette di identificare un numero predefinito di topic ed attribuire ad ogni documento una probabilità che in esso sia presente un determinato topic. Il modello associa: ad ogni documento una distribuzione di probabilità di appartenere ai differenti topic (Topic-document distribution) e ad ogni parola la distribuzione di probabilità che sia associabile ad un certo topic (Topic-word distribution).

LDA deve il suo nome al tipo di distribuzione che utilizza per modellare i parametri interni del modello, nota come distribuzione di Dirichlet. Essa è una distribuzione di probabilità continua multivariata utilizzata comunemente nell'ambito dell'analisi statistica e del machine learning, soprattutto nei contesti legati ai modelli generativi, come, per l'appunto, i topic model.

Per applicare il modello LDA ad un corpus, è necessario specificare due valori: k, che rappresenta il numero desiderato di *top-k parole* da considerare, e il numero di topic desiderati.

Nel contesto di questa tesi, si è deciso di valutare un intervallo di numeri di topic compreso tra 5 e 15, al fine di trovare un compromesso tra un adeguato numero di topic e le differenti metriche utilizzate per ottimizzare gli hyper-parametri.

Ogni parola che compare nel corpus viene assegnata, tramite una distribuzione di Dirichlet, ad uno dei k topic.

Di seguito si riporta la *Figura 8* relativa alla architettura del modello LDA, in cui è possibile osservare la presenza di due hyper-parametri, α (alpha) e β (beta), che regolano rispettivamente la Document-Topic distribution e la Word-Topic distribution. Inoltre, le variabili colorate in grigio sono le variabili osservate, quindi che sono note e conosciute, mentre le variabili colorate in bianco sono quelle nascoste.

Nella figura *Figura 9* le parole osservate (*Observed Word)* sono identificate con la lettera *w*. La variabile T corrisponde al numero di topic da individuare, mentre β è un hyper-parametro che regola la distribuzione word-topic ψ. La *Document-Topic distribution* è indicata come θ, ed è regolata dall'hyper-parametro α. La variabile z è il topic latente e, infine, N_d corrisponde al numero di parole nel documento D.

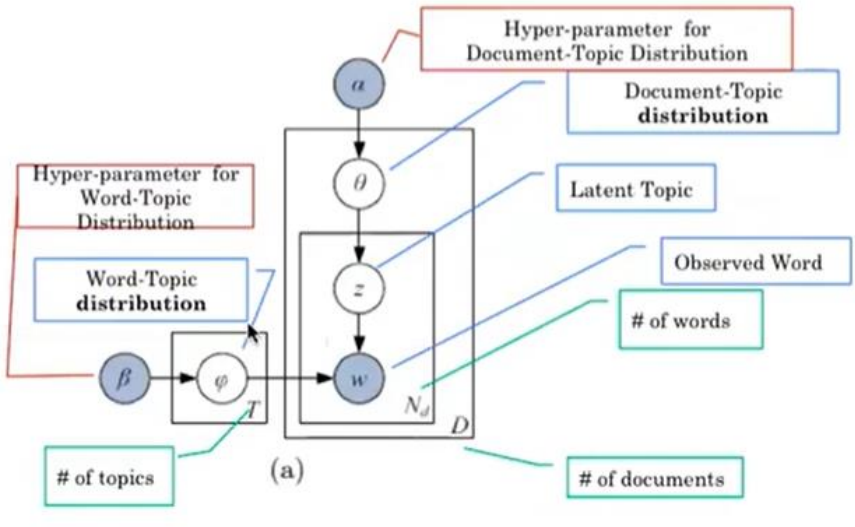

Figura 9: Struttura interna del modello generativo LDA

Si procede assegnando quindi a ciascun documento il topic più probabile, sulla base delle probabilità attribuite. È possibile, inoltre, stabilire una soglia di probabilità al di sopra della quale un documento contenga uno o più dei topic specificati.

Questo approccio consente di ottenere una rappresentazione dei documenti in termini dei topic specificati, che può essere utilizzata per l'analisi, l'organizzazione e la categorizzazione del corpus. [5]

4.1.1 Preprocessing

I dati di tipo testuale richiedono una fase di preparazione prima di poter essere elaborati. Questa prima revisione dei dati è essenziale per poter eliminare dai documenti delle parole che non danno alcun contributo utile al fine di elaborazioni successive. [6]

Prima di procedere alla modellazione dei testi è dunque necessaria una fase di preparazione e pulizia dei dati; le operazioni più importanti sono la rimozione della punteggiatura, la *lemmatizzazione* e la rimozione delle *stopword*:

- La prima pulizia importante riguarda i segni di punteggiatura. Segni come i punti e le virgole sono utili nel linguaggio per capire quando finisce una frase o per dare la giusta intonazione di lettura; tuttavia, LDA si basa esclusivamente sulla frequenza delle parole, per cui in prima istanza si procede con il rimuovere tutti questi segni dai documenti. Si vuole comunque sottolineare che non è un passo obbligatorio nell'ambito dell'elaborazione del linguaggio naturale. Ad esempio, esistono modelli molto recenti nell'ambito del Natural Language Processing che cercano di catturare schemi all'interno del documento, basandosi proprio sui segni di punteggiatura per carpire l'ironia delle frasi, come ad esempio CTM.
- Le *stopwords* sono un insieme di parole utilizzate frequentemente nel linguaggio e presenti in tutti i testi, quali articoli, congiunzioni, pronomi etc... Questi non aggiungono nessuna informazione riguardo l'argomento del quale si parla in un documento, anzi, porterebbero problemi di stima vista l'alta frequenza con cui vengono utilizzati, dunque vengono rimossi dal *corpus*. Nel caso della lingua italiana alcune stopwords comuni sono ad esempio: "quando", "allora", "e" "anche", etc. Nel contesto specifico del nostro caso, sono state considerate le stopwords della lingua italiana poiché i documenti presenti nel dataset erano scritti in italiano.
- La lemmatizzazione, invece, ha lo scopo di trovare una radice comune, il quale rappresenti un insieme di parole che hanno approssimativamente lo stesso significato. Questo procedimento non si basa però semplicemente sul troncare la parola, ma su un meccanismo molto più complesso. Ad esempio, lemmatizzare la serie di parole correre, corro, corriamo e correremo, produrrebbe il lemma "correre". Questa pratica non è particolarmente complessa quando si tratta della lingua inglese, con la sua grammatica relativamente semplice e poche forme irregolari. Tuttavia, diventa più impegnativa quando si considera la lingua italiana, che presenta una grammatica più articolata. Esistono librerie Python atte ad affrontare questa sfida, come ad esempio *spacy* o *nltk*. Esse offrono funzionalità e risorse

linguistiche specifiche per la lingua italiana, semplificando così l'elaborazione del testo.

La conversione del testo in minuscolo, la rimozione della punteggiatura, la lemmatizzazione, la rimozione di parole con meno di tre caratteri e la rimozione delle stopword sono tutte operazioni presenti nel preprocessing di OCTIS.

La *Figura 10* rappresenta visivamente il corpus nella sua forma originale, prima di eseguire il preprocessing dei dati. D'altra parte, nella *Figura 11*, viene presentato il corpus dopo aver completato il processo di preprocessing.

Figura 10: Corpus prima del preprocessing

Figura 11: Corpus dopo il preprocessing

Prima di passare alla fase successiva della pipeline riguardante l'*Optimization*, è stata effettuata un'altra operazione importante durante il *Preprocessing*: lo studio della *legge di Zipf* sulle parole del *corpus*.

La *legge di Zipf* fornisce una stima della probabilità di incontrare una specifica parola all'interno del corpus, tenendo conto del suo rango di frequenza. In sostanza, essa quantifica la probabilità che una parola si verifichi in base alla sua posizione nella classifica delle frequenze delle parole nel corpus.

La *legge di Zipf* afferma che la frequenza di una parola è inversamente proporzionale al suo rango di frequenza. Ciò significa che poche parole molto comuni, come articoli e preposizioni, hanno una frequenza molto più alta rispetto alle parole meno comuni. Nel Natural Language Processing viene utilizzata per modellare la distribuzione di probabilità discreta che descrive la frequenza delle parole in un determinato corpus.

La *Figura 12* mostra un grafico inerente alla distribuzione delle parole del corpus, da cui si evince l'applicazione della legge di Zipf.

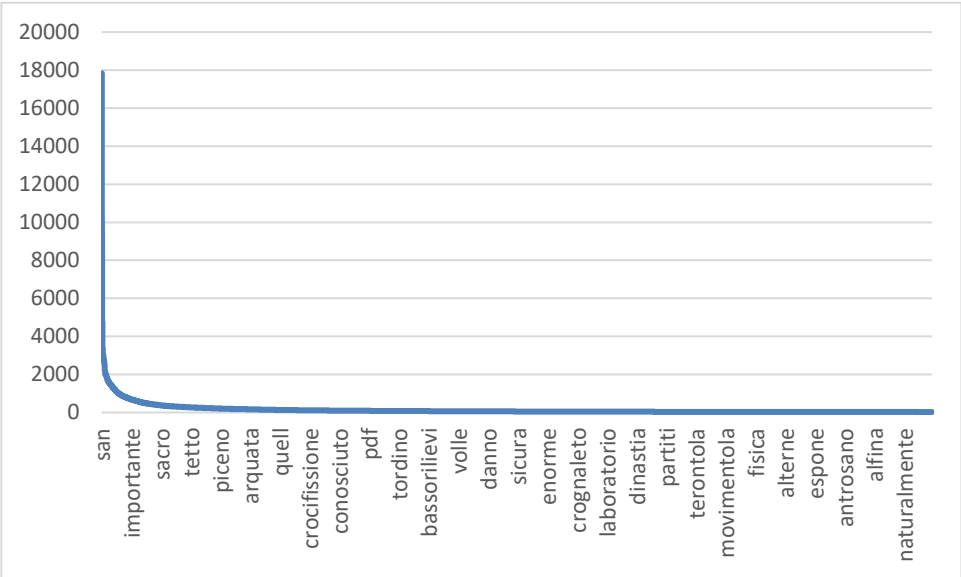

Figura 12: Distribuzione delle parole del corpus

Nel contesto del topic modeling, in particolare con l'utilizzo di modelli come LDA, le parole più comuni possono dominare la rappresentazione dei topic, rendendo difficile la scoperta di argomenti più specifici o meno frequenti. Effettuando un taglio alla *legge di Zipf*, si rimuovono le parole più comuni, solitamente considerando *stopwords* anch'esse, per mitigare l'influenza eccessiva che queste parole possono avere sulla modellazione dei topic.

Il taglio sullo studio della legge di Zipf consente di concentrarsi su parole più informative e specifiche, migliorando l'interpretazione e la scoperta di topic più rilevanti all'interno del *corpus*. Tuttavia, la scelta di quali parole includere o escludere nel taglio dipende dal contesto dell'applicazione e dalle specifiche esigenze dell'analisi dei testi.

Nella *Tabella 5*, vengono presentate le 10 parole più comuni all'interno del *corpus*, il quale è composto da un totale di 82.123 parole. Ogni parola viene accompagnata dalla sua frequenza di occorrenza nel corpus. Al contrario, la *Tabella 16* mostra alcune delle 10 parole meno frequenti.

PAROLA	FREQUENZA
san	17840
chiesa	10685
secolo	8407
comune	4766
parte	4747
città	3460
paese	3313
maria	3309
castello	3230
centro	3164

Tabella 5: Parole più comuni

PAROLA	FREQUENZA
liberalizzazioni	1
spiritosi	1
giocondo	1
miscuglio	1
stilizzare	1
moschee	1
infondono	1
puntandogli	1
rodio	1
suddividendo	1

Tabella 6: Parole meno comuni

Al fine di migliorare la qualità dei topic, è stato quindi deciso di applicare un taglio alla distribuzione delle frequenze delle parole, utilizzando la legge di Zipf. In particolare, le prime sei parole più comuni (san, chiesa, secolo, comune, parte, città), che dominavano la rappresentazione dei topic (si veda *Tabella 7*), sono state aggiunte alla lista delle *stopwords*.

1	**san**	**secolo**	**chiesa**	**parte**	castello	**comune**	**città**	stazione	paese	monte
2	**san**	**chiesa**	**secolo**	madonna	**parte**	**comune**	anni	palazzo	interno	trova
3	**san**	**chiesa**	**secolo**	**parte**	paese	**comune**	venne	via	stato	anni
4	**san**	**chiesa**	**secolo**	**comune**	maria	valle	palazzo	**parte**	**città**	madonna
5	**san**	**chiesa**	**secolo**	**parte**	**comune**	centro	paese	territorio	monte	anni

Tabella 7: Un esempio di 5 topic estratti prima dell'applicazione del taglio della legge di Zipf

4.1.2 Optimization

Dopo aver completato la fase di *preprocessing* del dataset, si è proceduto all'ottimizzazione sia della diversità degli argomenti, misurata attraverso la *topic diversity*, che della coerenza degli argomenti, valutata mediante la *topic coherence*.

L'ottimizzazione consiste nel selezionare la configurazione ottimale dei valori assegnabili agli hyper-parametri del modello. Per far ciò si valutano, per ogni

hyper-parametro, i differenti valori nel relativo range e, valutando una metrica appropriata (come in questo caso la topic diversity e la coherence), si scelgono quei valori che minimizzano o massimizzano la precedentemente presentata funzione obiettivo.

Solitamente, quando si parla di coerenza, si fa riferimento ad una caratteristica di cooperazione tra gli argomenti. Un insieme di argomenti è considerato coerente se vi è una reciproca conferma tra di essi. La metrica di *topic coherence* valuta quanto bene un argomento è "supportato" da un *corpus*. Utilizzando statistiche e probabilità tratte dal corpus, concentrandosi in particolare sul contesto delle parole, viene assegnato un punteggio di coerenza a ciascun argomento. Questo aspetto riveste grande importanza poiché dimostra che la coerenza di un topic non dipende solamente dalle parole che lo compongono, ma anche dal corpus di riferimento. L'obiettivo è quindi generare topic che presentino un alto valore di coerenza, con un intervallo di valutazione compreso tra 0 e 1.

La *topic diversity* si riferisce alla varietà e all'eterogeneità dei topic individuati dal modello. Maggiore è la diversità degli argomenti, maggiore sarà la ricchezza e l'ampiezza delle tematiche trattate dal modello.

Un buon modello di LDA dovrebbe cercare di bilanciare l'obiettivo di scoprire topic coerenti e distinti, evitando l'eccessiva sovrapposizione o la concentrazione su un numero limitato di argomenti.

Dopo aver impostato un intervallo di numeri di topic compreso tra 5 e 15, si sono ottenuti i seguenti risultati come conseguenza del processo di ottimizzazione:

MISURA	NUMERO DI TOPIC	VALORE DELLA FUNZIONE
Topic coherence	5	0,00809153
	12	-0,01601109
	15	0,036178
Topic diversity	5	0,5
	12	0,55833333
	15	0,213333

Tabella 8: Risultati ottenuti a seguito delle ottimizzazioni con LDA

1	stazione	territorio	sito	stato	maria	castello	circa	palazzo	museo	monte
2	abruzzo	monte	maria	dopo	territorio	stato	castello	stata	area	nome
3	territorio	monte	valle	maria	abruzzo	castello	madonna	provincia	giovanni	tre
4	madonna	stato	maria	porta	francesco	facciata	mentre	torre	tre	palazzo
5	castello	palazzo	dopo	maria	madonna	fino	stato	poi	territorio	facciata

Tabella 9: I 5 topic estratti mediante l'ottimizzazione di Topic Diversity con LDA

1	valle	stato	roma	castello	circa	territorio	dopo	frazione	poi	nome
2	territorio	madonna	palazzo	castello	sito	provincia	valle	stato	abruzzo	maria
3	territorio	palazzo	italia	castello	dopo	fino	stato	roma	nome	circa
4	stazione	palazzo	maria	madonna	tre	spoleto	castello	dopo	contiene	giovanni
5	museo	madonna	francesco	maria	giovanni	castello	opera	stato	abruzzo	altare
6	castello	maria	torre	fino	madonna	valle	monte	stato	nome	dopo
7	monte	territorio	castello	lago	fino	maria	torre	nome	dopo	borgo
8	maria	monte	abruzzo	palazzo	territorio	comuni	castello	provincia	valle	stato
9	maria	castello	provincia	stato	francesco	monte	dopo	madonna	territorio	sito
10	monte	stazione	territorio	pescara	frazione	nord	sud	fino	valle	abruzzo
11	teramo	abruzzo	palazzo	maria	dopo	stato	stata	facciata	fino	piazza
12	lago	stato	maria	territorio	abruzzo	madonna	area	grande	oggi	dopo

Tabella 10: I 12 topic estratti mediante l'ottimizzazione di Topic Diversity con LDA

1	territorio	stato	madonna	abruzzo	maria	fino	roma	secondo	circa	castello
2	stato	castello	spoleto	monte	nome	territorio	maria	madonna	fino	contiene
3	valle	maria	territorio	castello	stato	abruzzo	torre	palazzo	provincia	francesco
4	madonna	dopo	castello	stato	monte	territorio	abruzzo	maria	fino	torre
5	palazzo	castello	territorio	monte	stato	provincia	valle	maria	fino	abruzzo
6	maria	stato	monte	castello	madonna	dopo	tre	territorio	facciata	abruzzo
7	monte	palazzo	madonna	provincia	territorio	maria	castello	valle	contiene	abruzzo
8	maria	stato	abruzzo	castello	francesco	facciata	provincia	provincia	territorio	porta
9	maria	valle	dopo	francesco	abruzzo	provincia	stato	oggi	tre	madonna
10	maria	territorio	castello	url	stata	fino	monte	sito	nome	madonna
11	madonna	castello	dopo	palazzo	territorio	monte	valle	maria	abruzzo	tre
12	maria	castello	stato	territorio	monte	borgo	stata	facciata	fino	contiene
13	maria	palazzo	stato	castello	monte	presso	fino	stazione	dopo	abruzzo
14	territorio	maria	madonna	monte	abruzzo	mentre	dopo	stato	castello	sito
15	maria	territorio	mentre	valle	madonna	monte	fino	giovanni	piazza	dopo

Tabella 11: I 15 topic estratti mediante l'ottimizzazione di Topic Diversity con LDA

Come si può notare dalle tabelle soprastanti, i risultati ottenuti non sono ottimali: risulta complesso riuscire ad attribuire un'etichetta a ciascun topic estratto. Si è quindi deciso di procedere all'addestramento di modelli neurali, quali ProdLDA e CTM.

4.2 ProdLDA

Productive LDA (ProdLDA) è una variante del modello di topic modeling Latent Dirichlet Allocation che introduce delle modifiche per affrontare alcune limitazioni di LDA. [7]

LDA considera che ogni documento sia composto da una combinazione di argomenti e che ogni argomento sia caratterizzato da una distribuzione di parole. Tuttavia, LDA non considera l'aspetto della produttività differenziale degli argomenti, cioè il fatto che alcuni argomenti possono essere associati a un numero maggiore di parole rispetto ad altri all'interno del corpus. Ad esempio, in un corpus di notizie, il topic "politica" potrebbe essere associato a molte più parole rispetto al topic "meteo".

ProdLDA introduce una modifica a LDA includendo una distribuzione di "produttività" degli argomenti (propria di ogni topic), la quale rappresenta la probabilità che l'argomento sia associato ad una parola nel corpus. In questo modo, ProdLDA può catturare la differenza nella produttività degli argomenti e ottenere una migliore rappresentazione dei topic all'interno del corpus.

L'aggiunta di questa distribuzione di produttività rende ProdLDA un modello più flessibile e adattabile a diversi tipi di corpus.

In sintesi, ProdLDA è un modello neurale che nasce per migliorare la rappresentazione dei topic all'interno di un corpus introducendo una distribuzione di produttività dei topic, la quale tiene conto della differenza nel numero di parole associate a ciascun topic.

ProdLDA produce in modo coerente argomenti migliori rispetto a LDA standard. Inoltre, poiché effettuiamo inferenza probabilistica utilizzando una rete neurale,

possiamo addestrare un topic model su circa un milione di documenti in meno di 80 minuti su una singola GPU.

Di seguito verrà fatto un piccolo approfondimento (basato sulla *Figura 13*) di come funziona l'architettura del modello ProdLDA con l'ausilio di un Variational AutoEncoder (VAE). Per poter procedere, però, è necessario chiarire brevemente che cosa sia un VariationalAutoEncoder: consiste in una rete neurale che sfrutta una tecnica di apprendimento automatico per le rappresentazioni latenti dei dati. Essa mira a rappresentare i dati in uno spazio di dimensioni ridotte, chiamato spazio latente, in modo da catturare le caratteristiche essenziali dei dati originali.

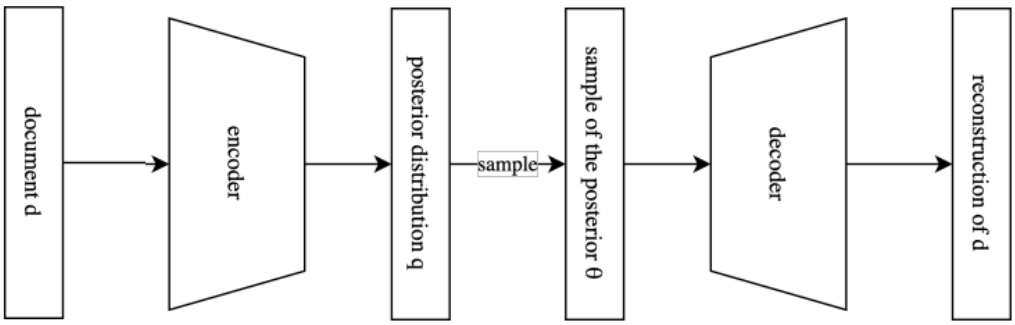

Figura 13: Struttura interna del modello ProdLDA basato su un VariationalAutoEncoder [10]

Nella *Figura 13*, si possono notare due parti principali di cui si compone il modello: *encoder* e *decoder*.

Il vettore sulla sinistra rappresenta i dati di input (ad esempio, un documento) e viene fornito all'encoder. L'encoder è responsabile di trasformare il vettore di input in una rappresentazione latente, cioè un vettore più piccolo che cattura le caratteristiche essenziali dei dati.

Il vettore più piccolo ottenuto dall'encoder è la rappresentazione latente dei dati. Questa rappresentazione contiene informazioni essenziali sui dati di input e si trova nello spazio latente.

Sul lato destro dell'immagine è presente il decoder. Esso prende la rappresentazione latente e la trasforma nuovamente nel vettore originale, cercando di riprodurre i dati di input il più fedelmente possibile.

Il processo di addestramento di un VAE avviene attraverso l'ottimizzazione di una funzione obiettivo che cerca di minimizzare la differenza tra i dati di input e i dati ricostruiti dal decoder. Allo stesso tempo, il modello cerca anche di far sì che la rappresentazione latente sia organizzata e continua nello spazio latente.

In sintesi, l'immagine mostra il flusso di informazioni all'interno del VAE, in cui l'encoder riduce la dimensione del vettore di input per ottenere una rappresentazione latente più piccola, e il decoder lavora in senso inverso per ritrasformare la rappresentazione latente nel vettore originale dei dati di input. Questo processo di compressione e decompressione permette di apprendere rappresentazioni latenti significative dei dati.

4.2.1 Preprocessing

Per quanto riguarda ProdLDA, si è scelto di applicare la medesima elaborazione e pulizia dei dati precedentemente effettuata con LDA, con un'unica differenza: non viene eseguito uno studio della legge di Zipf.

Come per LDA, anche per ProdLDA vengono eseguite le medesime operazioni di *preprocessing*, quali: la rimozione della punteggiatura, la conversione in minuscolo, la rimozione delle stopwords, la tokenizzazione e la lemmatizzazione. Tuttavia, a differenza di LDA, ProdLDA non richiede uno specifico studio della distribuzione delle frequenze delle parole secondo la legge di Zipf. Ciò è dovuto al fatto che ProdLDA incorpora una distribuzione di produttività, la quale mira ad evitare il problema che potrebbe verificarsi in LDA se non si applicasse un taglio secondo la legge di Zipf. In LDA, infatti, le parole più comuni potrebbero dominare la rappresentazione dei topic, rendendo difficile la scoperta di argomenti più specifici o meno frequenti. La distribuzione di produttività in ProdLDA aiuta a mitigare questo problema consentendo ai termini più produttivi di influenzare in modo significativo l'assegnazione dei topic, contribuendo a una rappresentazione più equilibrata e accurata degli argomenti nel corpus.

4.2.2 Optimization

Dopo aver ottimizzato sia la *topic coherence* che la *topic diversity*, si sono ottenuti i seguenti risultati:

MISURA	NUMERO DI TOPIC	VALORE DELLA FUNZIONE
Topic coherence	5	-0,11975366
	12	-0,1207879
	15	-0,1570261
Topic diversity	5	0,92
	12	0,7833
	15	0,88

Tabella 12: Risultati ottenuti a seguito delle ottimizzazioni con ProdLDA

1	marsicana	marsi	pescasseroli	habitat	lago
	marie	riserva	tagliacozzo	salto	monte
2	morra	inglobati	cortili	serratura	larino
	cotta	amicizia	recuperati	popolamento	distesa
3	beweb	alunno	eroli	cattolico	culturaitalia
	descrizionel	notealtri	storial	fonica	descrizioneesternola
4	san	chiesa	secolo	parte	centro
	madonna	torre	anno	maria	solo
5	poeti	parteciparono	forconese	certi	valente
	dimesso	teofilo	cina	lex	vagnucci

Tabella 13: I 5 topic estratti mediante l'ottimizzazione di Topic Diversity con ProdLDA

1	città	fino	parte	territorio	centro
	seguito	epoca	comune	roma	zona
2	tredici	guarenna	entusiasmo	grimoaldo	buccia
	justini	casulae	morra	gialloblu	portante
3	avezzano	marsi	marsica	fucino	iccu
	sbn	situato	roveto	geografia	collega
4	guarenna	justini	buccia	grimoaldo	casulae
	gialloblu	maggiorela	cannelle	pallottine	iiii
5	organo	canne	internol	tabernacolo	cripta
	statue	aula	cappelle	rosone	ligneo
6	giardino	sale	collezione	sezione	sala

	ceramica	museale	esposte	oggetti	traiano
7	viaggiatori	fascicolo	fabbricato	fermata	binari
	binario	trenitalia	impiantila	biglietteria	dipartimento
8	castello	abitato	abitanti	comune	società̀evoluzione
	montana	geografia	demograficaabitanti	frazione	parrocchiale
9	riserva	mammiferi	vetta	orfento	cresta
	capriolo	parks	falco	riserve	fauna
10	pinna	guarenna	entusiasmo	alterna	fritti
	maggiorela	dial	buccia	grimoaldo	pasquali
11	chiesa	san	secolo	facciata	madonna
	pietra	portale	maria	monastero	campanile
12	censimento	corciano	stadio	sagra	menotre
	marsciano	popolato	circoscrizione	cattaneo	curva

Tabella 14: I 12 topic estratti mediante l'ottimizzazione di Topic Diversity con ProdLDA

1	demograficaabitanti	società̀evoluzione	frazione	poggio	abitanti
	festa	castello	vibrata	categoria	monumenti
2	chiesa	secolo	san	palazzo	presso
	centro	facciata	piazza	storico	maria
3	avezzano	marsica	marsi	iccu	sbn
	situato	fucino	roveto	celano	interessearchitetture
4	interessechiesa	manifestazioniil	terenziano	olivi	vedetta
	polino	calcetto	censimento	portante	artechiesa
5	viaggiatori	rfi	binari	fascicolo	fabbricato
	trenitalia	binario	dipartimento	biglietteria	impiantila
6	ampliò	invito	incappucciati	gladiatori	comandati
	considera	concentrò	abbandonarono	spostandosi	ippolita
7	riferendosi	grandine	incappucciati	petrarca	calura
	concittadini	misurava	prevalgono	coraggio	spostandosi
8	dure	igienico	tabula	ippolita	concittadini
	calura	ripieni	spostandosi	nazi	gregoriano
9	bourbon	cesi	archivi	fabriano	biblioteche
	stanze	sorbello	beniculturali	sassoferrato	borgia
10	riserva	parks	castello	url	stata
	fino	monte	sito	nome	madonna
11	madonna	bambino	opera	cappella	altare
	sinistra	destra	cristo	opere	parete
12	lacu	rise	casoria	abazia	brittoli
	poggiofiorito	omero	nazi	dolii	appartenesse

13	gladiatori	misurava	argenteo	tegoloni	omero
	calura	segnati	cuccagna	fritto	marcata
14	territorio	comune	zona	fino	roma
	provincia	nord	circa	periodo	lungo
15	internol	cattolico	rita	religioso	rosone
	lignea	tabernacolo	campate	beweb	canne

Tabella 15: I 15 topic estratti mediante l'ottimizzazione di Topic Diversity con ProdLDA

Nel quadro delle figure qui sopra presentate, emerge chiaramente che i risultati derivanti dall'applicazione del modello neurale ProdLDA non raggiungono livelli ottimali a causa della debolezza della coerenza interna dei topic identificati.

4.3 CTM

Il terzo ed ultimo modello utilizzato per l'estrazione dei topic model è *Cross-lingual Contextualized Topic Models with Zero-shot Learning* (CTM).

I metodi tradizionali di topic modeling sono specifici della lingua con cui vengono addestrati (si basano su un vocabolario fisso), pertanto soffrono di due limitazioni:

1) non sono in grado di gestire parole sconosciute
2) non possono essere facilmente applicate ad altre lingue.

Quello che invece fa CTM, distinguendosi dagli altri topic model, è un apprendimento "zero-shot".

Il concetto di "zero-shot learning" (apprendimento senza supervisione) è un approccio in cui un modello viene addestrato su un compito specifico, e poi è in grado di generalizzare e applicare le sue conoscenze a compiti correlati per i quali non è stato specificamente addestrato. Nel contesto del topic model CTM "zero-shot learning" si riferisce alla capacità del modello di essere addestrato su una lingua specifica e poi applicare le sue conoscenze su documenti in lingue diverse a cui non ha avuto accesso durante l'addestramento. Si differenzia dagli altri proprio perché è progettato per lavorare in un contesto multilingue.

Ad esempio, se CTM viene addestrato su un insieme di dati in lingua inglese, e successivamente gli viene presentato un testo in lingua francese, il modello può applicare le sue conoscenze sull'inglese per identificare e analizzare i topic nel testo francese. Questa capacità di generalizzazione multilingue rende CTM particolarmente utile per l'analisi di testi provenienti da diverse lingue senza la necessità di addestrare modelli separati per ciascuna lingua [8].

In questo contesto, proporremo un approfondimento sull'architettura del modello in esame, utilizzando la *Figura 14* come punto di riferimento.

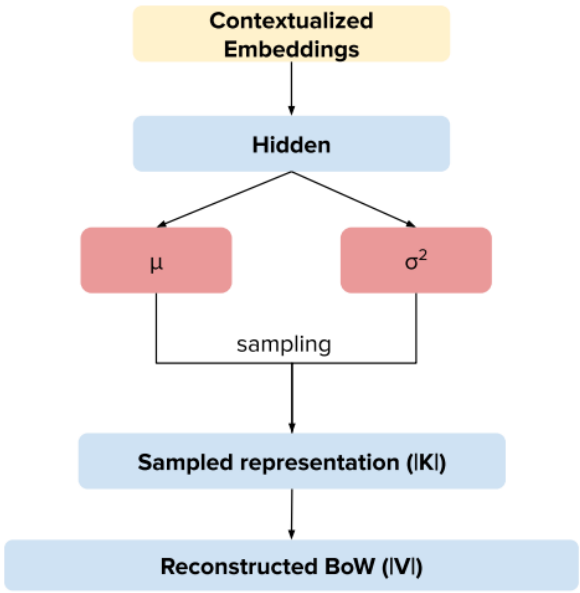

Figura 14: Struttura interna del modello CTM

La struttura del modello CTM è simile a quella di ProdLDA, ma con una differenza fondamentale: invece di ricevere l'intero corpus come input, il modello CTM lavora su un suo sottoinsieme. Questo è dovuto al fatto che CTM utilizza BERT (Bidirectional Encoder Representations from Transformers) come base per generare i topic, il quale ha un limite sulla lunghezza massima del testo che può essere elaborato in un singolo passaggio. Per questo motivo è necessario suddividere i documenti più lunghi in segmenti più piccoli prima di utilizzarli con il modello.

Poiché il CTM utilizza BERT come base per generare i topic, eredita questa limitazione sulla lunghezza massima del testo che può essere elaborato in un

singolo passaggio. In particolare, il modello CTM si arresta a una lunghezza massima di 768 parole, dato che questo rappresenta il limite imposto dalla versione base di BERT.

Tale limite è stato introdotto per gestire le limitazioni di memoria e di elaborazione, in quanto i modelli basati su trasformatori richiedono risorse computazionali significative e il trattamento di testi troppo lunghi potrebbe comportare un aumento eccessivo dei costi computazionali.

Tuttavia, è possibile elaborare testi più lunghi dividendoli in segmenti di 768 parole e applicando il modello CTM a ciascun segmento separatamente. Questa tecnica è nota come "sliding window" (finestra scorrevole) ed è spesso utilizzata per affrontare testi di dimensioni maggiori di quelle gestibili direttamente da BERT o dai suoi modelli derivati. Verrà ripresa nel paragrafo “Preprocessing”.

Anche CTM, come ProdLDA, sfrutta le proprietà dell'encoder-decoder per incorporare il contesto nel processo di addestramento e generare rappresentazioni contestualizzate delle parole. Questo approccio consente al modello di ottenere risultati più precisi e contestualmente rilevanti rispetto ai tradizionali modelli topic-based.

Per *Contextualized Embeddings* (facendo riferimento alla *Figura 13*) si intende una rappresentazione densa di parole o frasi che catturano il significato di un termine in un contesto specifico. A differenza degli embeddings tradizionali, che sono statici e non tengono conto del contesto, gli embeddings contestualizzati sono dinamici e variano in base al contesto in cui la parola appare.

Hidden si riferisce al vettore nascosto (hidden state) generato dall'encoder del modello durante la fase di contestualizzazione. Questo vettore contiene le informazioni semantiche e contestuali della parola o frase in considerazione. *Sampling* è una tecnica utilizzata per estrarre campioni casuali da una distribuzione di probabilità, mentre per Sampled Representation si intende la rappresentazione di parole o frasi che sono state estratte attraverso il campionamento da una distribuzione di probabilità. Infine, la Reconstructed Bag of Words (BoW) indica la rappresentazione finale (ricostruita) dei dati utilizzati nel modello. Bag of Words è una rappresentazione vettoriale dei documenti in cui ogni elemento del vettore

corrisponde a una parola e il valore rappresenta la frequenza della parola nel documento.

4.3.1 Preprocessing

Il topic model CTM richiede molto meno preprocessing dei dati rispetto ad altri modelli di topic modeling, ma non è corretto affermare che non ne richieda affatto. Infatti, anche se la complessità del preprocessing può essere ridotta rispetto ad altri modelli, alcuni passaggi preliminari sono comunque necessari per preparare i dati come, ad esempio, la rimozione della punteggiatura o la conversione del testo in minuscolo.

Inoltre, come anticipato, essendo un modello basato su BERT, si rende necessario suddividere il *corpus*. La maggior parte dei documenti nel corpus, infatti, presenta una lunghezza media di circa 800 parole. Di conseguenza, è stato adottato un approccio di frammentazione dei documenti utilizzando un *sentence recognizer*, implementato tramite la libreria *spacy* di Python. Questa tecnica ha consentito di suddividere ciascun documento in "sotto-documenti", ognuno rappresentante una singola frase.

Un *sentence recogniser* è un sistema di NLP che ha il compito di identificare le frasi all'interno di un testo più ampio, suddividendo il testo in singole frasi con precisione e coerenza.

Per garantire la coerenza e la riconducibilità dei sotto-documenti al documento originale, è stata mantenuta una chiave univoca, ossia l'idpages della pagina Wikipedia corrispondente al POI. Questa chiave ha permesso di associare e collegare i sotto-documenti al documento di riferimento, semplificando eventuali operazioni di ricostruzione successiva del POI in questione. [9]

4.3.2 Optimization

Dopo aver ottimizzato sia la *topic coherence* che la *topic diversity,* si sono ottenuti i seguenti risultati:

MISURA	NUMERO DI TOPIC	VALORE DELLA FUNZIONE
Topic coherence	5	-0,16135623
	12	-0,21760201
	15	-0,12840205
Topic diversity	5	0,96
	12	1
	15	0,98

Tabella 16: Risultati ottenuti a seguito delle ottimizzazioni con CTM

1	campanile	laterali	arco	decorata	navata	portale	sesto	pianta	timpano	cornice
2	trovandosi	produzioni	sottoposta	alture	andando	portava	contesa	avviata	site	formò
3	territorio	nome	monte	valle	del	via	nel	dal	pressi	paese
4	catalogo	arcangelo	patronale	meteorologica	milita	turistiche	istituto	ufficio	fermata	guida
5	mezza	descritta	svolgeva	livio	posizionata	salendo	paradiso	voli	progressiva	costiera

Tabella 17: I 5 topic estratti mediante l'ottimizzazione di Topic Diversity con CTM

1	del	della	alla	piazza	palazzo
	prima	convento	costruzione	fine	delle
2	galleria	movimento	luco	franco	clima
	cammino	cinema	bruno	drappo	guida
3	infrastrutture	sportivi	milano	trasporti	tradizioni
	club	impianti	touring	censiti	folclore
4	attività	gli	questi	specie	stati
	numerosi	state	presenti	non	molti
5	monti	monte	valle	statale	comune
	frazione	nord	collega	confine	val
6	amministrazione	descrizione	associazioni	architettura	classificazione
	economia	urbanistica	cultura	manifestazioni	religiose
7	iii	feudo	ducato	passò	figlio
	militare	dominio	successivamente	divenne	contea
8	sport	note	esattamente	cittadella	proteggere
	gradualmente	recuperato	distrutte	chiudere	dava
9	santi	parrocchiale	nicola	madonna	dedicata
	maria	santuario	michele	santa	bambino
10	protette	musei	geografiche	istat	siti
	coordinate	fauna	flora	società	monumenti
11	originale	catalogo	rete	istituto	turistiche

	fermata	civico	febbraio	dipartimento	agosto
12	campanile	pianta	cornice	arco	navate
	facciata	archi	copertura	unica	portale

Tabella 18: I 12 topic estratti mediante l'ottimizzazione di Topic Diversity con CTM

1	movimento	vinta	architettura	muzii	realizzando
	dialetti	erroneamente	agricoltura	affermare	cinema
2	numerosi	presenti	state	reperti	diversi
	questi	specie	stati	molti	diverse
3	classificazione	manifestazioni	cultura	economia	strade
	storia	flora	fauna	bibliografia	istruzione
4	feudo	iii	passò	conte	dominio
	conti	contea	ducato	divenne	successivamente
5	pianta	campanile	cornice	arco	sesto
	facciata	rettangolare	portale	unica	navate
6	sport	note	cittadella	medesimo	compito
	creata	voci	operato	pensa	permise
7	guida	istituto	club	sbn	squadra
	originale	cura	edizioni	agosto	settembre
8	civili	religiose	geografiche	protette	monumenti
	demografica	coordinate	luoghi	architetture	evoluzione
9	piazza	palazzo	convento	porta	metà
	medievale	cattedrale	chiese	casa	presso
10	monte	monte	valle	sasso	val
	situato	statale	ovest	torrente	piana
11	per	che	delle	dello	della
	alla	del	dei	guerra	seconda
12	fermata	italiano	stazione	ferroviaria	rete
	ferrovia	contiene	servizio	servita	dialetto
13	rocco	michele	sebastiano	nicola	lorenzo
	arcangelo	caterina	battista	santi	parrocchiale
14	amministrazione	simboli	seguenti	associazioni	infrastrutture
	descrizione	drappo	urbana	trasporti	impianti
15	movimento	dialetti	cucina	galleria	architettura
	tesoro	turismo	origini	bar	clima

Tabella 19: I 15 topic estratti mediante l'ottimizzazione di Topic Diversity con CTM

4.4 Risultati a confronto

Mettendo a confronto i vari risultati ottenuti dalle diverse ottimizzazioni, emerge chiaramente come l'utilizzo di reti neurali, in particolare con il modello CTM, porti ad un notevole miglioramento della topic diversity.
Nella *Figura 15* vengono comparati i risultati dell'ottimizzazione della topic diversity per diverse quantità di topic desiderati (5, 12 e 15).

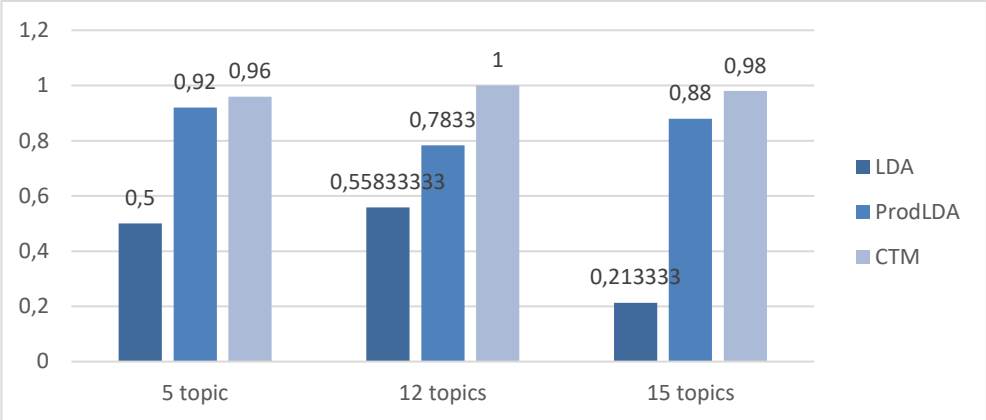

Figura 15: Topic Diversity a confronto

Per quanto riguarda la coherence, invece, i risultati non sono ottimali (si veda la *Figura 16*). Questo può essere dovuto al fatto che il dataset considerato presenta una grande varietà di argomenti, per cui non è in grado di produrre risposte coese, poiché cerca di adattarsi a una vasta gamma di temi. Se il modello viene addestrato su un insieme di dati molto eterogenei, può apprendere associazioni deboli o casuali tra parole o concetti, producendo quindi risposte non correlate e incoerenti.

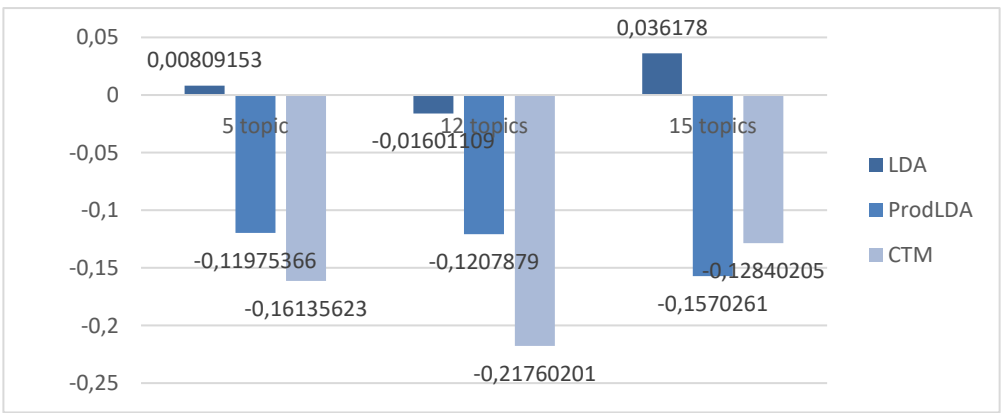

Figura 16: Coherence a confronto

Capitolo 5
Conclusione e sviluppi futuri

In questa tesi si è affrontato il problema dell'identificazione di argomenti da un insieme di documenti mediante l'utilizzo di modelli statistici e neurali di *topic modeling*.

Come primo passo sono stati illustrati i passaggi iniziali riguardanti il dataset utilizzato e la sua procedura di creazione.

In seguito, è stata fornita una breve panoramica sul tema centrale di questa ricerca, vale a dire il Natural Language Processing, focalizzandosi specificatamente su OCTIS, il framework adottato per l'estrazione degli argomenti di interesse.

Infine, sono stati presentati e discussi tre modelli di topic modeling, illustrando i vari risultati conseguiti nell'ambito di questa analisi.

5.1 Sviluppi futuri

L'approccio proposto si dimostra un utile supporto per la generazione dei topic desiderati. Tuttavia, per raffinare ulteriormente i risultati ottenuti, sarà necessario apportare alcune correzioni alle fasi del processo di individuazione dei topic. In futuro, il lavoro dovrà concentrarsi su diversi aspetti, tra cui sicuramente:

- Sviluppare un parser personalizzato per processare i dati estratti da Wikipedia, in quanto il database utilizzato è stato acquisito da questa fonte e presenta alcune imprecisioni che necessitano di essere corrette.
- Migliorare il preprocessing dei dati, ad esempio inserendo ulteriori parole che ricorrono nel corpus ma che non sono utili al nostro fine tra le stopword.
- Per arricchire il database utilizzato durante l'addestramento dei modelli, integrare ulteriori risorse al fine di ampliare la quantità e la diversità delle informazioni disponibili.
- Valutare l'impatto di differenti metriche di valutazione sulla qualità dei topic prodotti.
- Effettuare operazioni di sentence splitting sul corpus per poter permettere ai modelli, soprattuto neurali, di coprire l'intera lunghezza della descrizione.
- Affrontare eventuali limiti o lacune dei modelli attuali per garantire una migliore copertura di argomenti e tematiche.

Queste azioni mirano a perfezionare l'approccio esistente e a potenziare la capacità di generare topic coerenti e di alta qualità per soddisfare le esigenze degli utenti e migliorare l'efficacia del sistema complessivo.

Bibliografia

[1] M. a. P. W. Haklay, "Openstreetmap: User-generated street maps." IEEE Pervasive computing 7.4," 2008.

[2] M. Montebello, "Elaborazione delle richieste in Linguaggio Naturale (NLP) nell'ambito dell'enterprise search e degli agenti conversazionali.," *Natural Language Processing in enterprise search and conversational agents.*, 2018.

[3] D. M. Blei, "Probabilistic topic models," *IEEE Signal Processing Magazine*, 2010.

[4] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano and A. Candelieri., "{OCTIS}: Comparing and Optimizing Topic models is Simple!," *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2021.

[5] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of machine Learning research*, Gennaio 2003.

[6] S. Urban, "Uno studio di metodi di Topic Modeling," 2021.

[7] A. Srivastava and C. Sutton, "AUTOENCODING VARIATIONAL INFERENCE FOR TOPIC MODELS," 2017.

[8] F. Bianchi, S. Terragni, D. Hovy, D. Nozza and E. Fersini, "Cross-lingual Contextualized Topic Models with Zero-shot Learning," *The 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

[9] J. D. a. M.-W. C. a. K. L. a. K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

[10] H. Schnoering, "Short Text Topic Modeling: Application to tweets about Bitcoin," 17 Marzo 2022.

Appendice A

Chiavi e relativi valori ottenuti dopo la prima scrematura

(etichette selezionate da https://wiki.openstreetmap.org/wiki/IT:Map_Features)

amenity	• library • research_institute • training • music_school • school • university • arts_centre • cinema • community_centre • fountain • music_venue • planetarium • public_bookcase • theatre • grave_yard • monastery • place_of_worship • city_wall • national_park • protected_area • college • language_school • studio • courthouse • townhall • ditch
landuse	• education • cemetery • religious
man maid	• bridge • obelisk • observatory • tower • watermill • windmill

building	• cathedral • chapel • church • kingdom_hall • monastery • mosque • presbytery • religious • shrine • synagogue • temple • civic • college • university • castle • bridge • government • school • pavilion • sports_hall • stadium • hangar
leisure	• nature_reserve • stadium
tourism	• artwork • attraction • gallery • museum • viewpoint • zoo

historic	• aircraft • aqueduct • archaeological_site • battlefield • bomb_crater • building • cannon • castel • castell_wall • charcoal_pile • church • city_gate • citywalls • fort • gallows • highwater_mark • locomotive • manor • memorial • mine • milestone • monastery • monument • pillory