Università degli Studi di Perugia
Dipartimento di Matematica e Informatica

Tesi di Laurea Triennale in Informatica

Approccio Deep Learning al Topic Modeling: Analisi di annunci di lavoro tramite BERT

Relatore

Prof. Valentina Poggioni

Laureando

Tommaso Bagiana

Anno Accademico 2024-2025

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Indice

Capitolo 1

Introduzione

Panoramica introduttiva del progetto, motivazione della ricerca e obiettivi principali.

1.1 Cos'è il Topic Modeling

Descrivere in termini generali l'obiettivo del topic modeling, le tecniche classiche e i vantaggi rispetto ad approcci basati su keyword.

1.2 Analisi del paper *The Dynamics of Data Analytics Job Skills: a Longitudinal Analysis*

Riassumere i contributi principali del paper, le metodologie adottate e gli spunti che guidano la replica o l'estensione del lavoro.

1.3 Approccio con BERT

Introdurre il modello BERT, motivare la sua scelta per il topic modeling e spiegare a livello concettuale la pipeline prevista.

Capitolo 2

Data Cleaning

Introdurre le fasi preliminari di pulizia dati e il razionale delle scelte effettuate.

2.1 Strategie di data cleaning usate

Elencare le tecniche applicate (rimozione stopword, normalizzazione, gestione dei duplicati, anonimizzazione, ecc.) indicando per ciascuna motivazioni e strumenti.

2.2 Divisione in paragrafi

Descrivere come i testi degli annunci vengono segmentati in paragrafi o blocchi tematici, includendo eventuali regole euristiche o soglie adottate.

2.3 Classificazione paragrafi

Spiegare il processo di etichettatura o clustering preliminare dei paragrafi, specificando se è supervisionato o meno e come si valida la coerenza delle classi.

Capitolo 3

BERTopic

Descrivere l'architettura di BERTopic e le motivazioni per cui è stata scelta per l'analisi degli annunci.

3.1 Pipeline

Presentare la pipeline end-to-end, evidenziando gli input, le trasformazioni intermedie e l'output finale del modello.

3.1.1 Embedding

Indicare come vengono generati gli embedding con BERT (o varianti), eventuali fine-tuning e impostazioni rilevanti.

3.1.2 Dimensionality Reduction

Illustrare l'algoritmo di riduzione dimensionale adottato (es. UMAP), i parametri principali e l'impatto sulla qualità dei topic.

3.1.3 Clustering

Spiegare il metodo di clustering (es. HDBSCAN), i criteri di scelta dei parametri e come viene determinato il numero di topic.

3.1.4 Tokenizer

Dettagliare il tokenizer utilizzato, le strategie di pre-processing e qualsiasi personalizzazione per il dominio degli annunci di lavoro.

3.1.5 Cosa si può migliorare

Discutere criticità note, possibili ottimizzazioni della pipeline e idee per estensioni future di BERTopic nel progetto.

Capitolo 4

Risultati finali

Descrivere i topic ottenuti, le metriche di valutazione e le principali evidenze emerse dall'analisi degli annunci.

Inserire tabelle, grafici e commenti qualitativi che illustrino chiaramente l'efficacia dell'approccio BERTopic applicato al dataset.

Capitolo 5

Conclusioni

Sintetizzare i risultati principali e delineare i possibili sviluppi futuri.

Bibliografia

Ringraziamenti

Testo dei ringraziamenti da completare.