

UNIVERSITA’ DEGLI STUDI DI TORINO

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

Anno Accademico 2024/2025

**Tecnologie del Linguaggio Naturale-Parte II Prof. D. Radicioni**

Autore: Annalisa Sabatelli Matr. 866879

ESERCITAZIONE 2 (2 punti)

Document Similarity and Retrieval using the VSM:

A. With Sparse Embeddings

1. **Introduzione**

La seguente esercitazione ha come finalità l’applicazione del Vector Space Model (VSM) per il recupero e la categorizzazione automatica di testi, utilizzando Sparse Embeddings basati su TF-IDF. L’attività è stata svolta impiegando il *News Category Dataset*, un corpus contenente circa 200.000 titoli di notizie provenienti da HuffPost, ciascuno associato a una specifica categoria tematica come politica, tecnologia, sport ed economia. A partire da questo dataset, è stato selezionato un sottoinsieme casuale di 10.000 titoli, sui quali è stata applicata una pipeline di pre-elaborazione che ha incluso la tokenizzazione, la conversione in minuscolo, la rimozione di stopwords e punteggiatura, e la lemmatizzazione. I documenti preprocessati sono stati quindi trasformati in vettori TF-IDF tramite la libreria fornita da scikit-learn. Successivamente, è stato implementato un sistema di recupero documentale basato sulla cosine similarity tra i vettori TF-IDF che sono stati precedentemente creati e query testuali composta da uno o più parole inserite dall’utente. Ciascuna query viene preprocessata e confrontata con i documenti estratti per identificare e restituire i 5 titoli più rilevanti.

1. **Struttura del codice**

Il codice è suddiviso in due file:

• *utils.py*: contiene metodi di utility tra cui:

* *def extraction\_lemmi\_from\_sentence(sentence):* metodo che prende in input una frase e la elabora restituendo una stringa costituita dai lemmi delle parole che la costituiscono. In particolare, l’elaborazione si articola nei seguenti passi:
  + tokenizzazione;
  + rappresentazione di tutte le parole in lower case;
  + rimozione delle stop-words;
  + rimozione della punteggiatura;
  + rimozione degli spazi vuoti;
  + creazione di una lista di lemmi;
  + trasformazione della lista di lemmi in una stringa dove ogni lemma è separato dall’altro da un solo spazio.
* *def pipeline\_vectorize\_training (sentences, vectorizer):* il metodo prende in input una lista di frasi e un oggetto vectorizer già addestrato. Per ogni frase applica la funzione di preprocessing che estrae i lemmi. Successivamente, trasforma le frasi preprocessate in una rappresentazione numerica usando fit\_transform, che adatta il vettorizzatore ai dati e genera la matrice TF-IDF.
* *def pipeline\_retrieval (queries, vectorizer):* il metodo prende in input una lista di query testuali fornite dall’utente e un oggetto vectorizer già addestrato. Ogni query viene preprocessata per estrarne i lemmi. Le query preprocessate vengono poi trasformate in vettori TF-IDF usando transform (senza rifare il fit).  
  In questo modo, le query sono rappresentate nello stesso spazio vettoriale dei documenti di training. La funzione restituisce la matrice vettoriale delle query, utile per confronti e calcolo della similarità.
* *def search\_and\_display\_queries (query\_vector, queries, X\_tfidf, df\_sampled, TOP\_N):* il metdo si occupa di cercare e visualizzare i documenti più rilevanti per ciascuna query fornita. Prende in input i vettori delle query (query\_vector), le query testuali originali, la matrice TF-IDF dei documenti (X\_tfidf), un DataFrame con i titoli e le categorie (df\_sampled), e il numero di risultati da mostrare (TOP\_N). Per ogni query, calcola la cosine similarity tra il vettore della query e tutti i documenti. Ordina i documenti per similarità decrescente e seleziona i TOP\_N risultati più rilevanti. Crea poi una tabella formattata con score di similarità, titolo della notizia e categoria. Infine, stampa la tabella su console per ciascuna query fornendo, quindi, un output leggibile che mostra i documenti più pertinenti rispetto a ciascuna query.

• *main.py*: contiene il metodo main. Dopo aver scaricato il dataset di riferimento, lo carica in formato JSON direttamente in memoria, selezionando casualmente 10.000 titoli di notizie. Per ogni notizia selezionata, estrae le headline e la preprocessa in lemmi usando la funzione *extraction\_lemmi\_from\_sentence*. Le frasi preprocessate vengono convertite in una matrice TF-IDF tramite TfidfVectorizer. Tramite standard input, l’utente inserisce 10 query testuali che a loro volta vengono preprocessate e trasformate in vettori TF-IDF nello stesso spazio vettoriale dei documenti. Infine, per ogni query, viene calcolata la cosine similarity rispetto a tutti i documenti, e i 5 più simili vengono visualizzati in una tabella con punteggio, titolo e categoria.

Lo script consente di valutare l’efficacia del recupero basato su VSM per interrogazioni libere.

1. **Risultati ottenuti**

Di seguito si riportano, a titolo esemplificativo, i risultati ottenuti a seguito dell’esecuzione del codice con le 10 query mostrate in figura. Le query sono state selezionate con l’intento di coprire un’ampia gamma di ambiti semantici, includendo sia concetti astratti sia concreti, provenienti da diversi contesti tematici.

Immagine che contiene testo, schermata, Carattere, software

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, schermata, Carattere, software

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, schermata, software, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Immagine che contiene testo, schermata, software, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

1. **Conclusioni**

I risultati appaiono in generale coerenti con le query inserite, anche se in alcuni casi si notano anomalie nei punteggi o nella pertinenza. Si analizzano di seguito separatamente le singole query.

1. **Gun:** i titoli selezionati sono chiaramente incentrati sul tema delle armi, con riferimenti espliciti a regolamentazioni e contesti politici, confermando una buona rilevanza semantica.
2. **Supermodel:** i primi due articoli sono adeguati, ma i successivi sembrano completamente irrilevanti, probabilmente a causa di una bassa soglia di similitudine o mancanza di sinonimia semantica nel modello.
3. **Car:** i risultati spaziano da sicurezza a modelli automobilistici, con coerenza tematica, indicando una buona copertura lessicale nel corpus per questa parola.
4. **Bottle:** i titoli trovati sono molto centrati su bevande, riuso e cucina. Probabilmente il termine viene trattato prevalentemente in un contesto alimentare influenzando il recupero dei documenti.
5. **Justice:** i risultati mostrano una varietà di angolazioni (politica, religione, impatto sociale), che suggerisce un buon raggio semantico del termine, anche se alcune sfumature possono risultare vaghe.
6. **Challenge:** i titoli sono molto coerenti con l'idea di “sfida personale” o “iniziativa”, riflettendo un contesto motivazionale o di crescita, probabilmente favorito da frasi ricorrenti nei dataset.
7. **Crime:** alcuni risultati sono solo figurativamente legati al termine, suggerendo un limite nella capacità del sistema di distinguere usi metaforici da quelli letterali.
8. **Paper**: i risultati sono misti ma tutti tematicamente validi. Si passa dalla carta fisica alla pubblicazione accademica indicando una certa ricchezza polisemica del termine nel corpus.
9. **Meat:** i titoli estratti sono tutti coerenti con il cibo e anche con implicazioni etiche e di salute dimostrando una rilevanza contestuale ben conservata.
10. **Bone:** i primi tre articoli sono rilevanti, mentre gli ultimi due sembrano completamente fuori tema.

In sintesi, l’approccio mostra buone capacità di recupero per termini con alta frequenza e contesto ben definito, ma presenta debolezze nei casi di polisemia, ambiguità semantica o bassa soglia di similarità.