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.

2. 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 similarità coseno 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 le 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.

3. 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.

Top 5 risultati per: 'gun'

Top 5 risultati per: 'gun'

Score	Headline	Category
0.649	Where I Stand on Guns	WELLNESS
0.642	Why Didn't the Police Ask About His Guns?	POLITICS
0.583	It's the Guns, Stupid!	CRIME
0.578	Guns, Love and Being Human	WELLNESS
0.568	Almost Nobody Wants To Loosen Regulations On Gun Silencers. Not Even Gun Owners.	POLITICS

Top 5 risultati per: 'supermodel'

Top 5 risultati per: 'supermodel'

Score	Headline	Category
0.541	Women's Capes For Superheroes, Supermodels And Everyone In Between (PHOTOS)	STYLE & BEAUTY
0.425	Supermodel Stephanie Seymour Does Sexy Photo Shoot...With Her Sons (PHOTO)	STYLE & BEAUTY
0.000	If This Isn't The Most Charming Way To Travel, We Don't Know What Is	TRAVEL
0.000	North Korea's Racial Slur of President Obama Is Business as Usual	BLACK VOICES
0.000	6 Tips That Could Help Your Student Loan Repayment	EDUCATION

Top 5 risultati per: 'car'

Top 5 risultati per: 'car'

Score	Headline	Category
0.659	Are You Being Followed on Foot or By Car? What to Do	CRIME
0.455	Hey Little Lady, Come Take A Look At This Toy Car	PARENTING
0.450	The car of 2016 - McLaren 675LT	ENTERTAINMENT
0.423	The Hottest Cars at the Geneva Auto Show	BUSINESS
0.405	Graco Recalls 25,000 Car Seats	PARENTS

Top 5 risultati per: 'bottle'

Top 5 risultati per: 'bottle'

Score	Headline	Category
0.521	Canned vs. Bottled Beer: Can You Really Taste The Difference?	FOOD & DRINK
0.458	Moonshine Marshmallows, The Blue Bottle Way	FOOD & DRINK
0.435	The Art of Blending Wines: From Barrel to Bottle (VIDEO)	FOOD & DRINK
0.403	Weatherman's Hilarious 'Wine Forecast' Tells You How Many Bottles You Need For This Snow	TASTE
0.401	4 Unusual Ways To Reuse Empty Liquor Bottles (PHOTOS)	HOME & LIVING

Top 5 risultati per: 'justice'

Top 5 risultati per: 'justice'

Score	Headline	Category
0.481	There's A Better Way To Get Justice For Sexual Assault Survivors	POLITICS
0.437	Reinvigorating the Faith-Led Movement for Justice	RELIGION
0.427	From Charity to Justice to Love: My Professional Odyssey	IMPACT
0.397	The Department Of Justice Will Still Rely On Private Prisons In A Big Way	POLITICS
0.383	Criminal Justice System Disenfranchises Former Convicts Looking For Work	CRIME

Top 5 risultati per: 'challenge'

Top 5 risultati per: 'challenge'

Score	Headline	Category
0.683	30-Day Challenge to Change the World -- Beginning Right Now	HEALTHY LIVING
0.576	Finding Challenging Work	WELLNESS
0.452	Take Our Fitness Challenge To Move More And Feel Better	HEALTHY LIVING
0.396	IGNITEgood Millennial Impact Challenge (VIDEO)	IMPACT
0.390	Join Me on the 30-Day Blood Sugar Solution Challenge	WELLNESS

Top 5 risultati per: 'crime'

Top 5 risultati per: 'crime'

Score	Headline	Category
0.511	What Does 'Black-On-Black Crime' Have to Do With Ferguson?	POLITICS
0.497	China Sentences 113 On Terror Crimes	WORLDPOST
0.458	3 Skincare Crimes I Know You're Guilty of Committing	STYLE & BEAUTY
0.414	Hate Crimes In Libraries See Post-Election Spike	ARTS & CULTURE
0.379	Congressional Democrats Tell DOJ To Do More To Address COVID-19 Hate Crimes	POLITICS

Top 5 risultati per: 'paper'

Top 5 risultati per: 'paper'

Score	Headline	Category
0.558	The Paper Cuts That Don't Heal	PARENTING
0.480	Paper Straws In The Best Patterns (PHOTOS)	FOOD & DRINK
0.468	Rock, Paper, Scissors: A Dialogue on Who We Are	WELLNESS
0.460	Holiday Gift Bags vs. Wrapping Paper	HOME & LIVING
0.430	About A 50/50 Chance A Computer Threatens To Steal Your Job: Paper	BUSINESS

Top 5 risultati per: 'meat'

Top 5 risultati per: 'meat'

Score	Headline	Category
0.467	HuffPost Tastemakers: Artisan Meat Share	FOOD & DRINK
0.464	Red Meat and Eggs on Trial Again, But Jury Is Still Out	WELLNESS
0.454	The Fascinating Case For Eating Lab-Grown Meat	TASTE
0.451	You Won't Even Miss The Meat With These Delicious Vegetarian Sandwiches	TASTE
0.422	'Good Wife' Recap: All of Alicia's Men in 'Red Meat'	ENTERTAINMENT

Top 5 risultati per: 'bone'

Top 5 risultati per: 'bone'

Score	Headline	Category
0.497	Bone Health And Your Diet: The Worst Foods For Your Bones	WELLNESS
0.476	Earliest Human Cancer Found in 1.7-Million-Year-Old Bone	HEALTHY LIVING
0.399	Marti Noxon Poured Her Own Life Into 'To The Bone,' A Movie About Anorexia	ENTERTAINMENT
0.000	6 Tips That Could Help Your Student Loan Repayment	EDUCATION
0.000	NC Official Changes Stance On Pepper Spray In Trans-Friendly School Bathrooms (UPDATE)	QUEER VOICES

4. 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.