Master in Data Analytics – A.A. 2022-23

Marta Cutrera

Pedius Education: accessibilità e inclusione mediante l'elaborazione del Linguaggio Naturale (NLP)

Presentazione del lavoro

Il lavoro si origina a partire dalla necessità di proporre una sintesi e una visualizzazione più unitaria agli utilizzatori del prodotto Pedius Education.

Pedius Education offre lezioni universitarie sottotitolate in tempo reale. Il prodotto è rivolto innanzitutto agli studenti con sordità e altre disabilità legate all'apprendimento, come ADHD/ADD o spettro autistico, dislessia, ecc. Ad oggi, questi studenti si trovano in molti casi esclusi dalle lezioni a causa di aule inadeguate e della mancanza di strumenti che facilitino le loro inclinazioni e modalità di apprendimento. Un'istruzione adeguata può aiutare le persone con disabilità a ottenere un maggiore accesso all'occupazione, alla salute e ad altri servizi e a sviluppare una maggiore consapevolezza dei propri diritti.

Solo il 3% della popolazione sorda riesce a conseguire una laurea, mentre la media europea è di circa il 27%, il che dimostra che per uno studente sordo è 10 volte più difficile raggiungere un livello di istruzione elevato. La sordità profonda (soglia uditiva minima superiore a 90dB) è una condizione che colpisce in media una persona su mille.

Obiettivo del presente lavoro è stato quello di elaborare una indicizzazione di un corpus di testo relativo allo stesso corso universitario composto da differenti lezioni al fine di produrre un elenco di parole chiave e di generare una bozza di riassunto per ciascun testo.

Preprocessing dei Documenti con SpaCy per l'Analisi Semantica

E' stata eseguita una analisi del linguaggio naturale (NLP) che utilizza SpaCy per il preprocessing di documenti di testo. La procedura coinvolge l'iterazione attraverso dei documenti testuali posizionati in una specifica cartella, l'elaborazione di ciascun documento con SpaCy per l'estrazione di lemmi, parti del discorso, e l'esclusione di token non rilevanti. Successivamente, viene costruita una lista di lemmi per ogni documento, con conversione in minuscolo per uniformità. Si escludono specifiche parti del discorso e stopwords, mantenendo solo i lemmi rilevanti. Le liste di lemmi per ogni documento vengono infine aggregate in una lista principale denominata "all_lemmas". L'output finale include la stampa delle liste di lemmi e del numero totale di lemmi, svolgendo così una fase preliminare per analisi linguistiche future. L'obiettivo principale è ottenere una rappresentazione semantica coerente e compatta attraverso l'estrazione dei lemmi, contribuendo a semplificare la comprensione rispetto alle forme flessionate delle parole.

POS esclusi

La lista di liste creata contiene tutte i lemmi contenuti nel testo analizzato. Sono stati esclusi dalle liste alcune POS (parti del discorso) L'esclusione di queste etichette aiuta a filtrare il testo per ottenere solo le informazioni rilevanti per l'analisi dei lemmi.

TF-IDF

E' stato poi utilizzato l'algoritmo TF-IDF (Term Frequency-Inverse Document Frequency) per calcolare la rappresentazione di un corpus di lezioni afferenti alla stessa materia. Nella fase di "Build Vocabulary" viene creato un vocabolario contenente tutte le parole uniche nel corpus. La "Document Frequency (DF)" indica quante volte ciascuna parola appare nei documenti, essenziale per calcolare l'IDF. L'"Inverse Document Frequency (IDF)" è il logaritmo del rapporto tra il numero totale di documenti e la DF di ciascuna parola. La "Term Frequency (TF)" rappresenta la frequenza di ogni parola in ogni documento, normalizzata rispetto al numero totale di parole nel documento. Infine, la "TF-IDF Calculation" moltiplica TF e IDF per ottenere la rappresentazione finale di ogni parola in ogni documento. L'output è una matrice in cui ogni riga rappresenta un documento e ogni colonna la rappresentazione TF-IDF di una parola nel vocabolario.
Si è poi proceduto a identificare e stampare parole chiave per ogni documento basandosi sui punteggi TF-IDF calcolati in precedenza.

Analisi delle Percentuali di Sovrapposizione tra Parole Chiave: Confronto tra FreqDist e TF-IDF

Documento	Parole Chiave Comuni (%)	Parole Chiave Diverse TF-IDF (%)
Documento 1	62.50%	37.50%
Documento 2	62.50%	37.50%
Documento 3	62.50%	37.50%
Documento 4	62.50%	37.50%
Documento 5	62.50%	37.50%
Documento 6	75.00%	25.00%
Documento 7	75.00%	25.00%
Documento 8	87.50%	12.50%
Documento 9	100.00%	0.00%
Documento 10	62.50%	37.50%
Documento 11	75.00%	25.00%
Documento 12	50.00%	50.00%
Documento 13	87.50%	12.50%
Documento 14	75.00%	25.00%
Documento 15	62.50%	37.50%
Documento 16	62.50%	37.50%
Documento 17	75.00%	25.00%
Documento 18	50.00%	50.00%
Documento 19	62.50%	37.50%
Documento 20	75.00%	25.00%
Documento 21	50.00%	50.00%
Documento 22	50.00%	50.00%

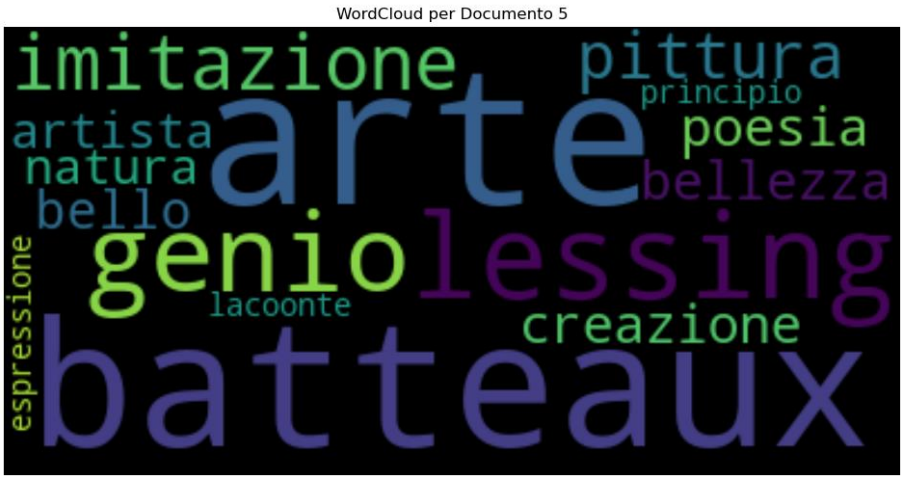
Il codice confronta le parole chiave estratte da due approcci, FreqDist e TF-IDF, per ogni documento di un corpus. I risultati sono organizzati in un DataFrame pandas che mostra chiaramente le parole chiave comuni e diverse, consentendo un'analisi dettagliata delle differenze tra gli approcci. La tabella include le percentuali di parole chiave comuni e diverse rispetto all'approccio TF-IDF, offrendo una panoramica delle similitudini e delle differenze nelle parole chiave identificate.

Osservazioni generali:

La maggior parte dei documenti mostra una significativa sovrapposizione nelle parole chiave tra l'approccio basato sul conteggio delle frequenze e quello basato su TF-IDF. Tuttavia, alcuni documenti presentano una percentuale più bassa di parole chiave comuni, indicando differenze nella rilevanza delle parole chiave tra i due approcci.

Generazione della Wordcloud

La WordCloud è una rappresentazione visuale di un insieme di parole, dove la grandezza di ogni parola è proporzionale alla sua frequenza o importanza nel testo di origine. In una WordCloud, le parole sono disposte in modo casuale e spesso vengono collocate in modo da formare una forma o una struttura visuale interessante.



Riassunti con la libreria Gensim

Sono stati infine generati dei riassunti automatici delle lezioni, consentendo agli studenti di avere una visione concisa e riassuntiva del contenuto dei documenti. Il riassunto è un ausilio allo studio importante soprattutto per persone con bisogni educativi speciali (BES). Questa attività, svolta attualmente prevalentemente a mano, risulta praticamente indispensabile allo studente in quanto:

• gli permette di concentrare la memoria su una selezione di informazioni (poiché è molto complesso, oltre che poco utile, ricordarsi tutto il contenuto);

• gli permette di costruire nella sua memoria degli schemi logici personali e quindi più funzionali (apprendimento significativo)