Caccese Aurelio
22/05/2024

Information similarity

Introduzione

In questo progetto, mi sono posto l'obiettivo di esaminare gli articoli contenuti nell'ultimo numero della rivista Information, concentrandomi sull'analisi della similarità degli argomenti trattati. Ritengo che comprendere questa similarità sia cruciale per individuare correlazioni, trend emergenti e potenziali aree di ricerca all'interno del campo trattato dalla rivista.
Per conseguire questo obiettivo, ho adottato una metodologia avanzata basata sull'utilizzo dei vettori di embedding delle parole chiave presenti in ciascun articolo. Questi vettori, derivati da modelli di linguaggio preaddestrati, consentono di catturare le sfumature semantiche e il contesto associato alle parole chiave, offrendo così una rappresentazione compatta e significativa del contenuto di ogni articolo.
Ho scelto di utilizzare la similarità coseno come metrica per valutare la relazione tra coppie di articoli. Questa scelta deriva dalla sua ampiezza di utilizzo nell'ambito dell'analisi del testo e dalla sua capacità di fornire una misura robusta della relazione semantica tra documenti.
Attraverso questa analisi dettagliata, mi propongo di individuare pattern e relazioni sottostanti tra gli articoli della rivista Information.

MDPI è un editore accademico che pubblica riviste scientifiche in vari campi, tra cui scienze biologiche, medicina, ingegneria, scienze sociali, e molti altri. "Information" è una delle riviste accademiche pubblicate da MDPI che si concentra sulla ricerca e gli sviluppi nel campo dell'informatica e delle tecnologie dell'informazione. La rivista copre una vasta gamma di argomenti, tra cui intelligenza artificiale, sicurezza informatica, elaborazione dei dati, reti di computer, e molto altro ancora. È una risorsa preziosa per i ricercatori e gli accademici che lavorano nell'ambito dell'informatica e delle scienze dell'informazione.

Caccese Aurelio
22/05/2024

Cos'è la similarità coseno?

La similarità coseno è una misura che valuta quanto due vettori siano simili, utilizzando il coseno dell'angolo tra di essi. Viene utilizzata in vari campi come il trattamento del linguaggio naturale, il clustering di documenti, i sistemi di raccomandazione e la ricerca di informazioni. In pratica, valori più alti indicano una maggiore somiglianza tra i vettori. È una misura efficace e versatile per confrontare la similarità tra vettori in molteplici contesti applicativi.

Per calcolare la similarità coseno tra due vettori, si utilizza la seguente formula:

$$\text{similarity} = A \cdot B \ / \ \|A\| \times \|B\|$$

dove $A \cdot BA \cdot B$ rappresenta il prodotto scalare tra i due vettori e $\|A\|\|A\|$ e $\|B\|\|B\|$ rappresentano le loro rispettive norme Euclidee (misure di lunghezza o grandezza di vettori nello spazio).

Questa misura restituisce un valore compreso tra -1 e 1, dove:

- 1 indica che i vettori sono perfettamente allineati;
- 0 indica che i vettori sono ortogonali (non hanno alcuna somiglianza);
- -1 indica che i vettori sono direttamente opposti.

In pratica, maggiore è il valore della similarità coseno, maggiore è la similarità tra i due vettori.

Dataset con info strutturali

Il dataset utilizzato per questo progetto è costituito da 32 articoli contenuti nell'ultimo numero della rivista Information, disponibili all'indirizzo web: https://www.mdpi.com/journal/information.

Ogni articolo è strutturato secondo il formato standard degli articoli scientifici, comprendente titolo, autori, abstract e testo completo. Inoltre, ciascun articolo presenta un insieme di parole chiave che sintetizzano i concetti chiave trattati nel testo.
Per l'analisi della similarità degli argomenti tra gli articoli, ci concentreremo principalmente sulle parole chiave di ciascun articolo. Queste verranno estratte e rappresentate come vettori di embedding, ottenuti da modelli di linguaggio preaddestrati. I vettori di embedding consentono di catturare informazioni semantiche e contestuali associate alle parole chiave, facilitando così la valutazione della similarità tra gli articoli.

È importante sottolineare che il dataset non include solo il testo degli articoli, ma anche le informazioni strutturali associate, come i metadati degli articoli stessi (autori, titolo, parole chiave.
Queste informazioni seppur importanti non saranno fondamentali per l'analisi della similarità che andrò ad eseguire, per questo verranno eliminate.
Inoltre tutte le altre informazioni più tecniche e dettagliate, relative al dataset, sono presenti nella datacards allegata.

Caccese Aurelio
22/05/2024

Svolgimento

Ho scelto di svolgere questo progetto interamente in Python, su Jupiter Lab, tramite la piattaforma Anaconda Navigator.
Le principali fasi in cui ho diviso il progetto sono 7:

- 1: importazione delle Librerie
- 2: Definizione delle Funzioni per l'Estrazione del Testo e delle Parole Chiave
- 3: Estrazione del Testo e delle Parole Chiave dai PDF
- 4: Calcolo degli Embedding delle Parole Chiave
- 5: Calcolo della Similarità Coseno tra gli Articoli
- 6: Visualizzazione della Matrice di Similarità
- 7: Clustering Gerarchico Agglomerativo

1: importazione delle Librerie

Ho iniziato importando un insieme di librerie necessarie per l'elaborazione dei PDF, l'estrazione delle parole chiave, la creazione di embedding e il calcolo della similarità.
Le principali librerie utilizzate sono state:
"Fitz" di (PyMuPDF) per leggere e manipolare documenti PDF,
"Spacy" libreria open-source di NLP avanzata, progettata per l'elaborazione di grandi volumi di testo. È ottimizzata per velocità e prestazioni.
"Pandas", libreria fondamentale per la manipolazione e l'analisi dei dati in Python. Fornisce strutture dati flessibili come DataFrame, che permettono una facile manipolazione di dati etichettati.
"Scikit-learn" utilizzata per calcolare la similarità coseno tra i vettori di embedding degli articoli. E' una libreria di machine learning che fornisce strumenti semplici ed efficienti per l'analisi dei dati e il data mining.
"Numpy" è stato utilizzato per manipolare i vettori di embedding e calcolare medie e altre operazioni matematiche necessarie per il calcolo della similarità.
"Rake-nltk" RAKE (Rapid Automatic Keyword Extraction) è un algoritmo che è stato utilizzato per estrarre parole chiave rilevanti dal testo degli articoli.
"Seaborn" È una libreria di visualizzazione dei dati basata su matplotlib. È stata utilizzata per creare la heatmap della matrice di similarità coseno, facilitando la visualizzazione delle relazioni tra gli articoli.
"Matplotlib" Utilizzata insieme a seaborn per migliorare la visualizzazione dei dati, è una libreria di plotting per creare grafici statici, animati e interattivi.
"Sentence-transformers" È una libreria per creare e utilizzare modelli di embedding delle frasi, che permette di trasformare frasi e testi in vettori numerici. È stata utilizzata per creare i vettori di embedding delle parole chiave degli articoli, che sono poi stati utilizzati per calcolare la similarità coseno.

2: Definizione delle Funzioni per l'Estrazione del Testo e delle Parole Chiave

In questa fase del progetto, ho definito due funzioni cruciali per l'estrazione del testo dai file PDF e per l'estrazione delle parole chiave dal testo stesso, poiché estraggono informazioni chiave dai PDF in un formato utilizzabile per ulteriori elaborazioni e analisi. Di seguito, descrivo nel dettaglio la definizione e il funzionamento di queste due funzioni.

2.1. Estrazione del Testo dai PDF: è responsabile dell'estrazione del testo dai file pdf, attraverso ogni pagina del file estrae il testo e lo salva in una variabile che viene successivamente utilizzata per contare il numero di parole, e suddividerle in "token" del testo.

2.2. Estrazione delle Parole Chiave dal Testo: questa funzione utilizza l'algoritmo RAKE (Rapid Automatic Keyword Extraction) per estrarre le parole chiave dal testo estratto.
La funzione calcola il numero di parole chiave da estrarre basandosi su una percentuale del numero totale di token nel testo, con un limite massimo di parole chiave. Il numero di parole chiave è calcolato come:
$numero\_parole\_chiave = min(int(num\_token \times percentuale), max\_keywords)$

Dove:

- num_token: Il numero totale di token (parole) nel testo.
- percentuale: La percentuale del numero totale di token che si desidera estrarre come parole chiave. In questo caso ho estratto il 2% delle parole, quindi la percentuale sarà 0.02.
- max_keywords: Pongo un limite massimo per il numero di parole chiave da estrarre, in questo caso 500.

La scelta di questo algoritmo è dovuta a diversi vantaggi pratici, specialmente quando si ha bisogno di rapidità ed efficienza. RAKE è progettato per essere rapido e adatto all'analisi di grandi volumi di testo senza richiedere molta potenza computazionale. Questo lo rende ideale in contesti dove le risorse sono limitate.
Un altro vantaggio significativo è la semplicità di implementazione. RAKE non richiede processi di pre-elaborazione complessi né librerie esterne sofisticate, il che lo rende facilmente integrabile in vari progetti. Inoltre, è un metodo non supervisionato, quindi non necessita di dati etichettati per l'addestramento, semplificando ulteriormente il processo di implementazione.
Inoltre è anche indipendente dalla lingua, funzionando bene con testi in diverse lingue senza necessità di adattamenti significativi. Questo lo rende versatile e applicabile in molti contesti linguistici differenti. Infine, RAKE utilizza un approccio basato su frasi intere per l'estrazione delle parole chiave, catturando meglio il contesto rispetto alle singole parole, offrendo così una rappresentazione più ricca e significativa delle informazioni chiave nel testo

3: Estrazione del Testo e delle Parole Chiave dai PDF

Nella fase di estrazione del testo e delle parole chiave dai PDF, è fondamentale ottenere il contenuto testuale degli articoli scientifici e identificare le parole chiave rilevanti per ciascun articolo.

4: Calcolo degli Embedding delle Parole Chiave

Una volta estratto il testo e le parole chiave da ciascun articolo scientifico, il passo successivo consiste nella creazione dei vettori di embedding per le parole chiave. Questi vettori sono rappresentazioni numeriche delle parole chiave che catturano il loro significato semantico e possono essere utilizzati per calcolare la similarità tra le parole chiave di diversi articoli.

4.1. Utilizzo di Modelli di Embedding delle Frasi:

Ho utilizzato il modello **all-MiniLM-L6-v2** fornito dalla libreria **sentence-transformers** per ottenere i vettori di embedding per le parole chiave.
La libreria Sentence Transformers fornisce modelli pre-addestrati per generare embedding di frasi. Gli embedding di frasi sono rappresentazioni numeriche di frasi o testi che catturano il loro significato semantico in uno spazio vettoriale. Questi modelli utilizzano architetture di deep learning, come ad esempio BERT, RoBERTa, DistilBERT e altri, che sono state pre-addestrate su grandi corpora di testo per imparare a generare rappresentazioni di alta qualità per le frasi.
L'obiettivo è quello di fornire un modo semplice per ottenere embedding di frasi ad alta qualità senza la necessità di addestrare un modello da zero su un corpus specifico. Questi embedding possono essere utilizzati per una varietà di compiti NLP (Natural Language Processing), come ad esempio la classificazione di testo.

La scelta del modello non è stata casuale poiché il modello "all-MiniLM-L6-v2" è stato pre-addestrato su un corpus molto ampio di testi provenienti da una varietà di fonti, inclusi testi generici, testi scientifici, articoli giornalistici e altro ancora. Questo gli consente di catturare una vasta gamma di conoscenze linguistiche e semantiche, rendendolo adatto per una varietà di compiti di NLP. Inoltre è stato progettato per essere leggero e adatto all'uso su sistemi con risorse limitate, mantenendo comunque una buona qualità nell'ottenimento degli embedding delle frasi.

4.2. Calcolo dei Vettori di Embedding:
Per ogni articolo, ho calcolato il vettore di embedding medio per le sue parole chiave estratte. Questo è stato fatto calcolando il vettore di embedding per ciascuna parola chiave e quindi ottenendo la media di questi vettori per rappresentare l'intero insieme di parole chiave dell'articolo.
Infine ho memorizziamo questi vettori di embedding in un dizionario, utilizzando il nome del file PDF come chiave.

Questa fase è essenziale per preparare i dati per il calcolo della similarità tra gli articoli, poiché fornisce rappresentazioni numeriche delle parole chiave che possono essere utilizzate per confrontare il contenuto semantico degli articoli.

5: Calcolo della Similarità Coseno tra gli Articoli

5.1. Implementazione del Calcolo della Similarità Coseno: Per calcolare la similarità coseno tra gli articoli, ho utilizzato la libreria scikit-learn, che fornisce una semplice implementazione di questa misura tramite la funzione cosine_similarity.

5.2. Interpretazione dei Risultati: Una volta calcolata la similarità coseno per tutte le coppie di articoli, possiamo interpretare i risultati per comprendere quanto siano simili gli articoli tra loro. Un valore di similarità vicino a 1 indica che gli articoli sono molto simili, mentre un valore vicino a 0 indica una bassa similarità ed un valore negativo indica una dissimilarità.

La matrice di similarità rappresentata in seguito è stata creata utilizzando una scala di riferimento normalizzata, utilizzando come estremi i valori minimi e massimi calcolati tramite la funzione "cosine_similarity".
In questo modo viene semplificata l'interpretazione visiva dei dati. Potendo facilmente individuare le aree della heatmap che rappresentano i livelli più alti e più bassi di similarità, evidenziando le relazioni più rilevanti tra gli articoli.

6: Visualizzazione della Matrice di Similarità

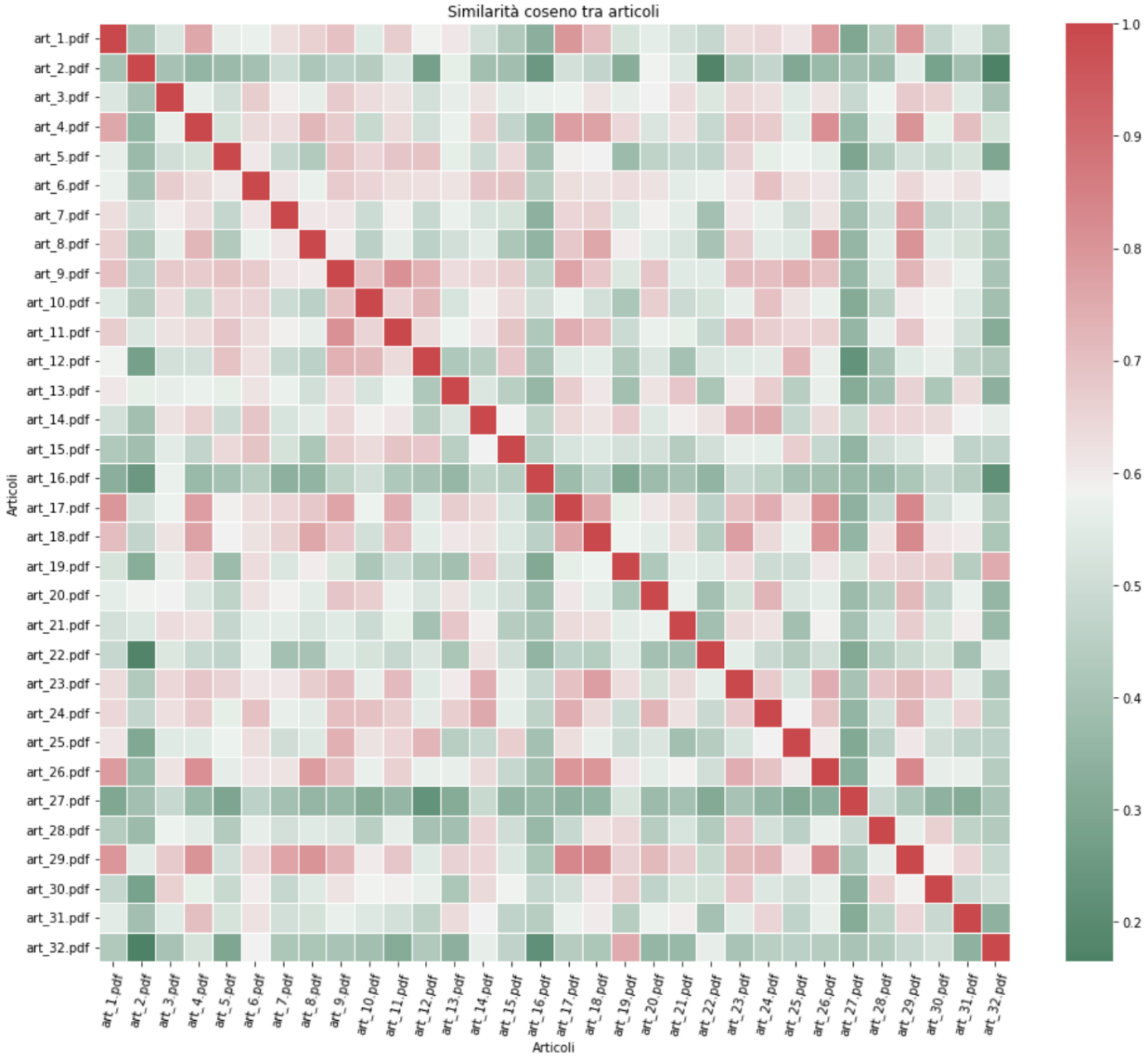
La similarità coseno è rappresentata attraverso una mappa di calore (heatmap) in cui il colore varia dal verde al rosso, con verde che indica bassa similarità e rosso che indica alta similarità.
La diagonale principale (dall'angolo in alto a sinistra all'angolo in basso a destra) è rossa e ha un valore di 1. Questo è atteso poiché ogni articolo è perfettamente simile a se stesso.

Gli altri confronti fra gli articoli mostrano la similarità in un intervallo dal valore più basso calcolato, ovvero 0.2 al valore più alto 1.0 che rappresenta la similarità perfetta.
Escludendo la diagonale principale, possiamo osservare che i valori più alti sono praticamente assenti e solo alcune celle sono colorate di un rosso più intenso come, ad esempio, l'intersezione fra gli articoli 9 ed 11, 17 e 18, 29 e 26, 29 e 18, 29 e 17.
Soffermandoci sui pattern emersi dalla matrice possiamo osservare che gli articoli 16, 19, 22, 27, 28, 31 e 32 hanno un coefficiente di similarità rispetto agli altri articoli molto basso, per questo si evidenziano delle righe quasi interamente verdi.
Al contrario gli articoli che mostrano un'elevata similarità costante con gli altri articoli sono praticamente assenti. Solamente le righe degli articoli 9, 17, 18, 23, 24 e 29 sono caratterizzate dalla presenza in maggioranza di celle rosse, che comunque non essendo colorate di un rosso intenso, non rappresentano elevati valori di similarità.
Le uniche eccezioni in cui possiamo rilevare un'elevata similarità sono presenti sulle colonne degli articoli 26 e 29.

Caccese Aurelio
22/05/2024

Per l'articolo 26 in corrispondenza degli incroci con gli articoli 4, 8, 17, 18. tutti gli articoli menzionano l'uso di modelli di deep learning per affrontare problemi complessi. Che si tratti di rilevamento di cellule della malaria, hashing per il recupero delle immagini, segmentazione di nuvole di punti, rilevamento di uccelli acquatici o rilevamento di immagini generate dal computer. Il deep learning è la tecnologia centrale, con l'applicazione delle reti neurali, e l'utilizzo di un approccio multidimensionale combinano dati da diverse sorgenti per ottenere migliori risultati.
Mentre per l'articolo 29 in corrispondenza dell'incrocio con gli articoli 8, 17, 18 e 26 gli articoli condividono diverse caratteristiche comuni, nell'applicazione di tecniche di visione artificiale e deep learning a vari problemi di rilevamento e classificazione.

Inoltre gli articoli 17: "Modelli di deep learning per il rilevamento degli uccelli acquatici e Classificazione nelle immagini aeree" e 18: "Rileva con stile: un quadro di apprendimento contrastivo per Rilevamento di immagini generate dal computer", trattando entrambi temi che si basano sul riconoscimento di immagini, hanno una forte similarità fra di loro, mostrando di conseguenza lo stesso risultato quando vengono confrontati con gli altri articoli.



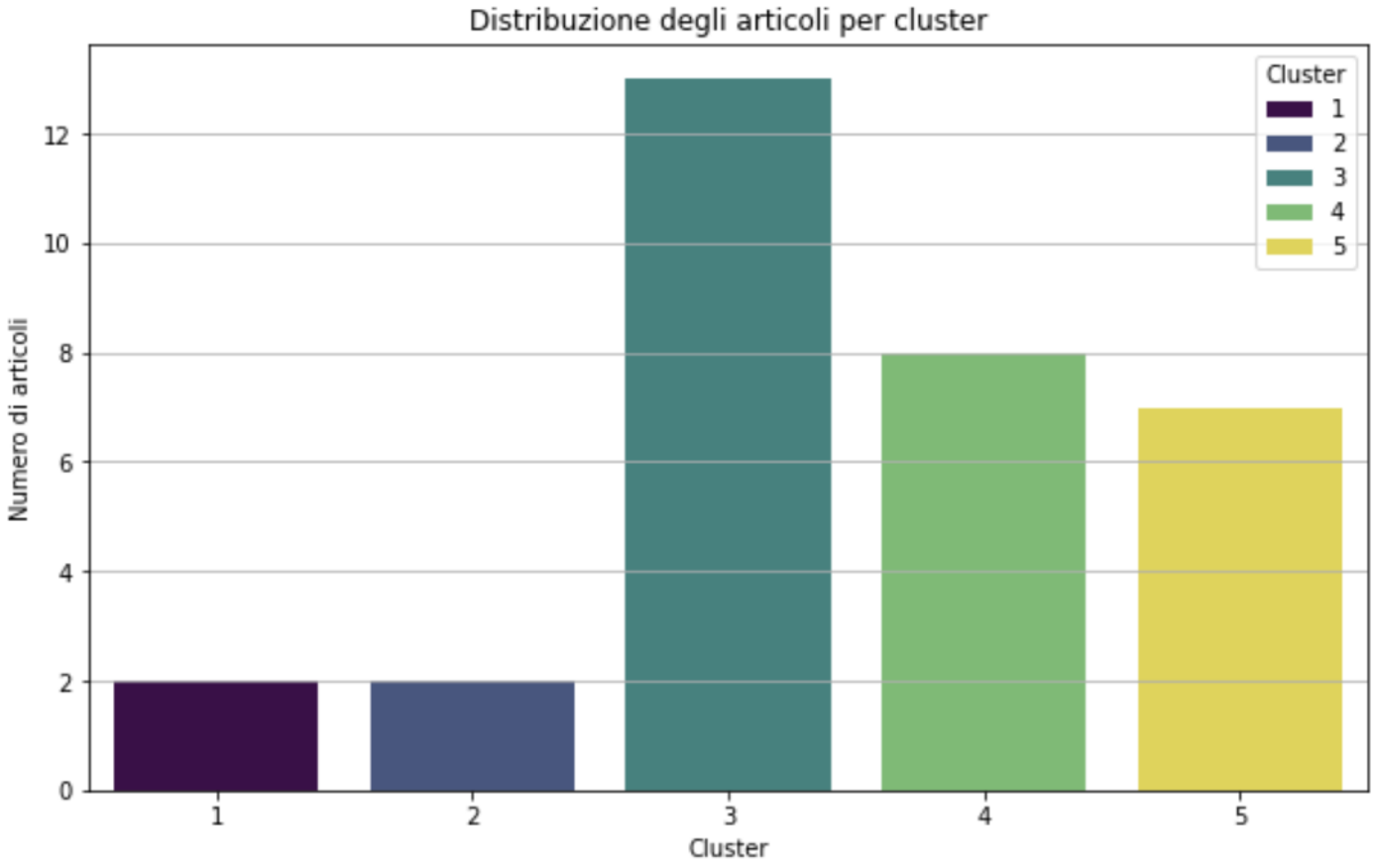
Caccese Aurelio
22/05/2024

7: Clustering Gerarchico Agglomerativo

Ho scelto di raggruppare gli articoli tramite il metodo di clustering gerarchico in modo da formare gruppi che contenessero articoli simili, sulla base della similarità coseno calcolata in precedenza.

Per fare ciò ho definito il numero di cluster desiderati e inizializzato il modello di clustering agglomerativo gerarchico. In questo caso, dopo vari tentativi e controlli ho scelto di creare 5 cluster diversi per rappresentare al meglio la divisione degli argomenti trattati negli articoli.
Ho applicato il modello di clustering ai vettori di embedding delle parole chiave estratte dagli articoli, assegnando un'etichetta di cluster a ciascun articolo. In seguito ho aggiunto le etichette dei cluster al DataFrame contenente la matrice di similarità degli articoli.

Infine, salvo i risultati dei cluster in un file CSV, creando una tabella, in modo da poterla consultare e analizzare ulteriormente in futuro.



Come possiamo notare dal grafico a barre, la distribuzione degli articoli nei cluster che sono stati generati non è uniforme. I primi due cluster hanno entrambi 2 articoli, il terzo cluster identifica come simili 13 articoli, il quarto cluster 8 articoli ed il quinto cluster 7 articoli.
Soffermandoci sui titoli degli articoli possiamo capire il perché di questa suddivisione.

Gli articoli nel Cluster 1 sembrano trattare argomenti piuttosto distinti a prima vista. Il primo articolo si concentra sullo sviluppo di una simulazione in realtà virtuale per i sistemi ferroviari, mentre il secondo si occupa di algoritmi di dimensionamento dei lotti per minimizzare i costi di inventario. Tuttavia, la loro presenza nello stesso cluster è dovuta alle tecniche algoritmiche citate negli articoli, evidenziando una connessione non immediatamente evidente nei loro titoli.

Il Cluster 2 è più coerente tematicamente. Entrambi gli articoli trattano di immagini, anche se con obiettivi differenti: uno si focalizza sul riconoscimento dei veicoli in immagini SAR (Synthetic Aperture Radar) utilizzando meccanismi di attenzione mista, mentre l'altro riguarda un algoritmo di cifratura delle immagini.

Anche questa connessione tematica suggerisce che entrambi gli articoli utilizzano tecniche avanzate di elaborazione delle immagini e crittografia, indicando una stretta relazione nelle loro metodologie.

Il Cluster 3 raggruppa una vasta gamma di argomenti, dalla disaggregazione dei profili dei clienti nelle reti di distribuzione a bassa tensione, all'analisi del comportamento degli utenti, fino all'uso di tecniche AI in applicazioni IoT, alla diagnosi medica e al rilevamento di intrusioni di rete. Questa varietà di temi indica che gli articoli condividono tecniche di intelligenza artificiale e machine learning applicate a diversi domini. La presenza di articoli con tematiche simili ma con applicazioni diverse sottolinea l'importanza di effettuare delle analisi approfondite, senza fermarsi ai primi risultati ottenuti.

Gli articoli del Cluster 4 sono incentrati su tecniche di visione artificiale e deep learning applicate a vari scenari, come la segmentazione delle immagini, il riconoscimento di strutture, la guida autonoma, il rilevamento di malattie e la rilevazione di usura in attraversamenti pedonali.
La coerenza tematica di questo cluster evidenzia l'importanza delle tecniche di visione artificiale in molteplici settori, dimostrando come queste metodologie possano essere adattate per risolvere problemi complessi in vari contesti applicativi.

Il Cluster 5 contiene articoli che spaziano dalla classificazione di tweet legati a disastri, alla modellazione di epidemie, alla percezione sociale dei disturbi neurologici, alla preservazione degli ecosistemi e ai giochi digitali. Nonostante la diversità degli argomenti, tutti questi articoli condividono l'uso di tecniche di machine learning e analisi dei dati su tematiche ambientali e sociali. Questo suggerisce che, sebbene i contesti applicativi possano variare ampiamente, le metodologie di base per l'analisi dei dati e il machine learning sono trasferibili e adattabili a molteplici campi di studio.

Conclusioni

La matrice di similarità coseno tra gli articoli mostra un panorama complesso e variegato di relazioni tematiche. Gli articoli analizzati, nonostante le differenze nei contesti applicativi specifici, condividono l'uso di tecniche avanzate di deep learning e visione artificiale per affrontare problemi di rilevamento e classificazione. I valori di similarità coseno indicano che alcuni articoli, come quelli dedicati alla segmentazione semantica e al rilevamento di immagini, presentano forti correlazioni tra loro, riflettendo approcci metodologici simili e tematiche sovrapposte. Altri articoli, invece, mostrano minor similarità, suggerendo una diversità nelle applicazioni e nei dati utilizzati. Questa analisi evidenzia l'interconnessione e la diversità all'interno del campo della visione artificiale e del deep learning, sottolineando come tecniche simili possano essere adattate per risolvere una vasta gamma di problemi pratici.

L'analisi dei cluster ottenuti attraverso l'algoritmo di clustering gerarchico ha rivelato diverse intuizioni interessanti sui contenuti degli articoli e sulle loro relazioni tematiche e metodologiche. La suddivisione degli articoli in cinque cluster distinti ha permesso di identificare sia similarità evidenti che connessioni più sottili tra i vari documenti.

L'analisi della similarità coseno tra coppie di articoli condotta per questo progetto fornisce una panoramica iniziale delle relazioni tematiche e metodologiche tra i vari documenti. Tuttavia, è importante notare che i risultati ottenuti sono soggetti a limiti computazionali della macchina che ho utilizzato per lo svolgimento del progetto.
Questi limiti possono influenzare la precisione e la profondità dell'analisi.

La macchina utilizzata ha una capacità di calcolo un po' limitata, il che ha comportato tempi di elaborazione prolungati e la necessità di ridurre la complessità dei modelli utilizzati per l'analisi. Inoltre, la gestione di grandi volumi di dati può essere problematica, portando a compromessi nella qualità della rappresentazione delle caratteristiche degli articoli.

Caccese Aurelio
22/05/2024

Per migliorare significativamente la qualità dell'analisi, l’utilizzo di macchine con maggiore potenza di calcolo e memoria può permettere l'uso di modelli più complessi e la gestione di dataset più grandi. Inoltre, l'adozione di algoritmi ottimizzati per l'efficienza computazionale può ridurre i tempi di elaborazione senza compromettere la precisione dei risultati. Un'analisi più approfondita che include tecniche di pre-elaborazione dei dati, tuning dei parametri dei modelli e valutazione continua dei risultati può portare a una maggiore accuratezza nella misurazione della similarità tra gli articoli e nella divisione in cluster.