

Relazione progetto di Topic Modeling

A cura di Valerio Valentini

1. Introduzione

Questa relazione verte sul progetto di Text Mining da me realizzato, che consiste nell'eseguire topic modeling su una serie di articoli. Questi ultimi sono stati scaricati attraverso metodi di web scraping da me scritti in python, linguaggio utilizzato anche per il resto del progetto. In particolare gli articoli appartengono alla sezione Mathematics della rivista online "MDPI" e costituiscono a tutti gli effetti il dataset utilizzato per questo progetto, di cui è presente una descrizione nella prossima sezione.

The screenshot shows the homepage of the MDPI Mathematics journal. At the top, there is a navigation bar with links for Journals, Topics, Information, Author Services, Initiatives, and About. There are also buttons for Sign In / Sign Up and Submit. Below the navigation bar, there is a search bar with fields for Title / Keyword, Author / Affiliation / Email, Mathematics (selected), All Article Types, a Search button, and an Advanced link. The main content area features the journal logo "Σ mathematics". On the left, there is a "Journal Menu" with links to Mathematics Home, Aims & Scope, Editorial Board, Reviewer Board, Topical Advisory Panel, Instructions for Authors, and Special Issues. Below the menu are social media sharing icons for X, Facebook, LinkedIn, and Share. The central part of the page displays a research article abstract with a mathematical formula: $H^j H'^i(C_\tau(M)) = \frac{\ker(d'_\tau : H''^{j,i}(C_\tau(M)) \rightarrow H''^{j+1,i}(C_\tau(M)))}{d'_\tau(H''^{j-1,i}(C_\tau(M)))}.$ The abstract title is "The Gauge Equation in Statistical Manifolds: An Approach through Spectral Sequences". To the right, there are two circular badges: one yellow for Impact Factor 2.4 and one dark blue for CiteScore 3.5. Below these badges is an "E-Mail Alert" section where users can enter their email address to receive forthcoming issues. Further down, there is a "News" section with a link to the International Conference on Mathematical Analysis and Applications in Science and Engineering (ICMASC'24).

(Pagina iniziale della sezione Mathematics su MDPI)

L'obiettivo del progetto è di analizzare gli argomenti trattati da ognuno degli articoli attraverso algoritmi di topic modeling e assegnare loro un nome sulla base del contenuto di ciascuno. Inoltre, per consolidare e verificare i risultati ottenuti, ho deciso di integrare la parte di topic modeling con algoritmi di text summarization e document clustering. Come già anticipato, il linguaggio di programmazione utilizzato è python, in particolare ho utilizzato un notebook di JupiterLab per scrivere il codice e dividerlo in diverse celle in base agli algoritmi utilizzati.

La relazione è divisa in diverse sezioni: una d'introduzione, dedicata alla descrizione del dataset, una per affrontare gli algoritmi e i metodi adottati durante il processo di topic modeling, una per riportare i risultati dei modelli della sezione precedente, e una finale con un'analisi conclusiva del progetto.

2. Descrizione del dataset

Una descrizione completa ed esaustiva del dataset utilizzato per questo progetto si trova in un documento separato, fornito assieme a questa relazione. Quest'ultimo contiene la Data Card, ovvero un riassunto strutturato in tabelle riguardante numerosi aspetti del dataset e dei dati di cui è costituito.

3. Algoritmi e Metodi

Segue una descrizione delle procedure e degli algoritmi utilizzati in questo progetto.

La prima operazione da me svolta è stata di reperire gli articoli, sui quali eseguire in seguito topic modeling, tramite la tecnica di web scraping. In particolare ho utilizzato due librerie: Selenium e BeautifulSoup. La prima consente l'automazione del browser, e di fatto l'ho utilizzata per raggiungere il sito web di MDPI tramite Chrome, per accettare eventuali cookie e per scorrere la pagina nella sua interezza in modo da non essere limitato dal suo caricamento parziale. Ho quindi utilizzato la seconda libreria per trovare, nella pagina del mese di gennaio, i tag corrispondenti ai link degli articoli, per poi scaricarli e salvarli in formato pdf sulla memoria di archiviazione della mia macchina. Infine ho estratto da ognuno di questi articoli il testo, che sarà l'elemento sul quale eseguiremo tutti gli algoritmi successivi.

Mathematics, Volume 12, Issue 1 (January-1 2024) – 168 articles



Cover Story (view full-size image): The authors introduce a novel option pricing model by adding stochastic interest rates and pure jump Lévy processes to an underlying price process driven by stochastic string shocks. They consider four different jump processes leading to different versions of the model: lognormal and double-exponential jump diffusions, CGMY, and generalized hyperbolic Lévy motion. In each case, they obtain closed or semi-closed form expressions for European call option prices. Moreover, they empirically evaluate the model's performance against S&P 500 call options. The findings indicate that (a) model performance is enhanced with the inclusion of jumps; (b) the model outperforms the alternative models with the same jumps; and (c) the model with CGMY jump offers the best fit across volatility regimes. [View this paper](#)

- Issues are regarded as officially published after their release is announced to the table of contents alert mailing list.
- You may sign up for e-mail alerts to receive table of contents of newly released issues.
- PDF is the official format for papers published in both, html and pdf forms. To view the papers in pdf format, click on the "PDF Full-text" link, and use the free [Adobe Reader](#) to open them.

Order results	Result details	Section
Publication Date	Normal	All Sections

Show export options ▾

[Open Access](#) Article 20 pages, 4598 Kib

Ship Infrared Automatic Target Recognition Based on Bipartite Graph Recommendation: A Model-Matching Method

by Haoxiang Zhang, Chao Liu, Jianguang Ma and Hui Sun
Mathematics 2024, 12(1), 168; <https://doi.org/10.3390/math12010168> - 4 Jan 2024
Viewed by 797

Abstract Deep learning technology has greatly propelled the development of intelligent and information-driven research on ship infrared automatic target recognition (SIATR). In future scenarios, there will be various recognition models with different mechanisms to choose from. However, in complex and dynamic environments, ship infrared [...] [Read more](#).
(This article belongs to the Section Computational and Applied Mathematics)

► Show Figures

(Interfaccia della pagina web contenente gli articoli di gennaio 2024)

La seconda operazione svolta è la tokenizzazione dei testi dei singoli articoli. Questa operazione consiste nella scomposizione del testo in unità fondamentali chiamate appunto “token”. Nelle lingue a ortografia segmentata, come l’italiano, un primo livello di tokenizzazione viene effettuato utilizzando come delimitatori tra i token gli spazi tra le parole. Altri elementi che costituiscono token autonomi sono i segni di punteggiatura, così come le sigle, le abbreviazioni e gli acronimi, le espressioni alfanumeriche e i nomi propri. C’è infine la possibilità di utilizzare un tokenizzatore subwords ovvero in cui i token generati sono costituiti da parti di parole, la cui aggregazione fornisce le parole complete. Nonostante l’utilizzo di questo tipo di tokenizzatori permetta di ridurre le dimensioni del vocabolario, e quindi è preferibile a quello dei tokenizzatori word-based, in cui i token sono costituiti dalle parole nella loro interezza, in questo progetto ho deciso di utilizzare un tokenizzatore word-based per fare in modo di avere delle parole complete e facilmente comprensibili all’interno di ogni topic, come vedremo in seguito, piuttosto che dover applicare successivamente ulteriori algoritmi per ritrasformare le subwords nelle parole originali.

L’operazione successiva è stata la rimozione delle stopwords e dei segni di punteggiatura. Di fatto nell’ambito del topic modeling si utilizza un approccio bag of words, ossia dove non conta né l’ordine delle parole né il loro ruolo grammaticale. Le uniche cose che ci interessano sono le parole e la loro frequenza in ogni documento, e per questo motivo ho proceduto eliminando i token costituiti da segni di punteggiatura e stopwords. Queste ultime sono parole, come articoli, pronomi e particelle, non significative o che comunque non sono utili per eseguire analisi di testi.

Ho anche eseguito la lemmatizzazione dei token, che consiste nel ridurre tutte le forme flesse e derivate a un unico lemma. L’ho fatto per mantenere contenute le dimensioni del vocabolario e per facilitare le analisi successive.

Ho poi creato il vocabolario che sarà quindi costituito dai token rimanenti dopo il pre-processing appena eseguito.

A quel punto ho eseguito un’analisi descrittiva del dataset, in particolare ho calcolato il numero di token per ogni articolo per capire se gli articoli potessero essere considerati lunghi o brevi, e ho realizzato una visualizzazione grafica delle 10 parole più frequenti per ogni documento tramite istogrammi di colore diverso e a barre orizzontali ordinate dal basso verso l’alto in ordine crescente di frequenza.

Una volta eseguito il pre-processing e compiuta l’analisi descrittiva, è il momento di eseguire topic modeling sul nostro dataset.

Prima di analizzare gli algoritmi che ho utilizzato, ci tengo a fare un breve recap sul topic modeling. È una tecnica di text mining che permette di estrarre i topic di cui tratta un documento di testo e di assegnare loro una percentuale (solo con gli algoritmi più avanzati) in modo da ottenere una cosiddetta “miscela di topic”. Un topic non è altro che una lista di parole, ognuna con un peso che rappresenta la sua importanza all’interno del topic. Argomenti differenti possono avere parole in comune così come diversi argomenti possono trattare dello stesso topic. Inoltre, dato che gli algoritmi di topic modeling non si occupano di assegnare un’etichetta, cioè un nome, a ogni argomento, sarà l’utente a dover scegliere per ogni topic il titolo che più gli si addice attraverso l’analisi delle parole di cui è composto.

A questo punto passiamo a esaminare il primo algoritmo di topic modelling che ho applicato, ovvero l’LDA. Si tratta di un algoritmo generativo, il che vuol dire che immagina ci sia un legame probabilistico tra documenti e parole. Inoltre la distribuzione dei topic in un documento è rappresentata da una variabile aleatoria multidimensionale, con tante

dimensioni quanti sono gli argomenti in un documento, e il peso di ognuno di questi topic dipende da una serie di parametri θ . Questi ultimi a loro volta seguono una distribuzione di Dirichlet, da cui l'algoritmo prende il nome, che dipende da altri parametri ancora chiamati a . La procedura di assegnazione dei topic ai documenti funziona in questo modo: dopo la creazione e l'addestramento del modello, si scelgono gli a da cui si genera la distribuzione di dirichlet che a sua volta ci fornisce i θ per ogni documento in base ai quali si sceglie per ogni parola in ogni posizione all'interno di ogni documento il topic più adeguato.

Ho dunque creato e addestrato il modello con un numero di 7 topic (numeri più elevati o più bassi di 7 fornivano risultati peggiori) e ho realizzato una tabella per ogni topic contenente le parole con il loro peso per ognuno di essi. Per rendere i risultati più facilmente visualizzabili ho anche creato 7 istogrammi, uno per ogni topic, con le barre orizzontali ordinate in ordine crescente di peso delle prime 10 parole di ogni argomento. Ho poi riportato la miscela di topic per ognuno dei 10 documenti del mio dataset con la relativa percentuale degli argomenti assegnati a ogni articolo.

Per verificare la coerenza dei topic estratti ho utilizzato un metodo intrinseco, ovvero che usa solamente informazioni contenute dal dataset a disposizione, ossia il metodo UMass, che fornisce per ogni argomento uno score calcolato a partire dalla co-occurenza di ogni coppia di parole nei documenti.

Ho anche creato una raffigurazione visuale interattiva dei topic ottenuti precedentemente che mostra sulla sinistra una rappresentazione nello spazio dei topic e sulla destra un'istogramma a barre verticali con frequenza di ogni termine sia complessiva sia all'interno del topic selezionato.

Infine ho utilizzato un altro metodo per analizzare la coerenza dei topic ottenuti, ovvero attraverso dei grafi. I loro nodi sono costituiti dalle parole di ogni topic e il peso degli archi che collegano le coppie di nodi corrisponde al valore della similarità coseno tra i vettori di embedding delle parole contenute nei nodi. Per ogni grafo si sommano questi valori e si ottiene uno score per quel topic.

Il secondo algoritmo di topic modeling che ho utilizzato è il BTM. A differenza dell'algoritmo precedente, che considerava le occorrenze di parole singole, o unigrammi, come caratteristiche dei topic, il BTM considera le occorrenze dei bigrammi, ovvero coppie di parole adiacenti. Inoltre nell'LDA la distribuzione delle parole per topic può variare da un documento all'altro, il che implica che per addestrare il modello bisogna lavorare sulle occorrenze all'interno di ogni singolo documento, rendendo non adeguati i documenti brevi. Al contrario, nel BTM la distribuzione delle parole per topic vale per l'intero corpus. Questa caratteristica del BTM, assieme all'assegnazione dei topic ai biterm piuttosto che agli unigrammi, rende questo algoritmo più adatto a essere applicato a documenti brevi.

Anche in questo caso ho dunque creato il modello BTM, l'ho addestrato sul corpus a disposizione, ho creato una tabella contenente le parole per ogni topic in ordine decrescente di peso, ho estratto la miscela di topic per ogni documento e l'ho rappresentata anche graficamente attraverso degli istogrammi a barre verticali, la cui altezza rappresenta la percentuale di ogni topic all'interno del documento.

Ho poi calcolato lo score Umass per ogni topic estratto e anche per questo algoritmo ho realizzato un grafo per ogni argomento per calcolare la sua coerenza tramite la similarità coseno tra le coppie di parole per poi ottenere uno score complessivo.

Infine ho riportato anche la rappresentazione grafica dello spazio delle parole dei topic unitamente all'istogramma delle frequenze marginali e condizionate di ogni parola.

Conclusa la parte di topic modeling, per validare i risultati ottenuti, ho deciso di eseguire altre due tecniche di text mining, ovvero text summarization e document clustering.

La prima consiste nell'ottenere il riassunto di un documento facendo in modo che contenga le informazioni più rilevanti del testo del documento di partenza.

Il riassunto deve inoltre essere in grado di coprire tutti i topic presenti in un documento, di evitare ripetizioni, di rispettare le regole grammaticali e sintattiche, e di coprire la lunghezza del documento originale in maniera bilanciata.

L'algoritmo che ho utilizzato è il TextRank, che consiste in un metodo astrattivo di text summarization, ovvero che ha come obiettivo creare un testo semanticamente simile, grammaticalmente e sintatticamente corretto, ma più breve del testo originale. In particolare il TextRank è basato sulla costruzione di un grafo i cui nodi sono le frasi e i cui pesi degli archi consistono nel valore di similarità tra le coppie di frasi. Le frasi più importanti sono quelle con valori di similarità più elevati rispetto alle altre frasi e quindi la loro identificazione si riduce alla ricerca dei nodi più centrali all'interno del grafo. Come misure di centralità dei nodi si usano soprattutto la degree centrality, ovvero il numero di vicini di ogni nodo che è una misura di popolarità, e la closeness centrality, ossia la media della lunghezza dei percorsi più brevi che collegano il nodo in questione con gli altri nodi del grafo che è una misura di raggiungibilità.

Ho poi applicato questo algoritmo ai miei 10 articoli e per ognuno di essi ho fornito il riassunto ottenuto, come pure i topic assegnati a quel documento dalle analisi precedenti per verificare se è possibile riconoscere la loro presenza all'interno dei riassunti.

Ho infine applicato l'ultima tecnica di text mining sopracitata, ovvero il document clustering. Si tratta di una tecnica di analisi dei dati testuali che raggruppa i documenti simili in base al loro contenuto. L'obiettivo è creare dei cluster tali da far sì che i documenti all'interno dello stesso cluster siano più simili tra loro rispetto a quelli di altri cluster. In particolare, nel mio caso ho deciso di eseguire il clustering dei documenti a partire dalle distribuzioni di argomenti generate dal precedente modello di topic modeling di LDA. Per eseguire il vero e proprio clustering ho utilizzato l'algoritmo K-means che funziona assegnando iterativamente ciascun dato al cluster con il centroide più vicino e aggiornando i centroidi dei cluster fino a quando non convergono, ossia non cambiano più significativamente tra iterazioni successive.

Infine ho stampato i documenti divisi in base all'assegnazione a ciascun cluster e ho calcolato l'indice di Silhouette per valutare la qualità del clustering.

4. Risultati

Dopo aver esaminato gli algoritmi utilizzati per l'analisi del dataset, analizziamo ora i risultati ottenuti.

Come risultato del processo di web scraping, ho ottenuto il testo dei dieci articoli, di cui riporto di seguito l'esempio del primo articolo.

Esempio di Articolo (Articolo 1):

1. Introduction
Ship targets are crucial combat units in modern maritime warfare, and it is important to recognize them accurately and credibly to enhance maritime situational awareness and gain an advantage. Infrared (IR) imaging technology is advantageous due to its all-weather capability, long-range perception, and strong concealment. It, along with visible light and synthetic aperture radar (SAR) imaging, forms an important means of acquiring feature information about ship targets and is widely used in ship automatic target recognition (SATR) tasks [1–3]. In the marine environment, different types of ships exhibit varying degrees of thermal characteristics and IR radiation spectra within the IR spectrum. This paper focuses on ship IR automatic target recognition (SIATR) technology, which utilizes sensors to capture IR images of ship targets. By combining image processing, recognition algorithms, and other techniques, this technology automatically and accurately extracts shape, IR radiation, and other feature information from the targets. This extracted information is then compared and matched against a pre-established feature information database to determine the type and identity of the target [4]. In the military domain, this technology provides category-priority information for subsequent tasks like target tracking, threat assessment, and target engagement, thereby delivering reliable target recognition and intelligence support for maritime operations. Additionally, it has significant applications in civilian sectors such as maritime surveillance, safety rescue, and related activities. In recent decades, SATR technology has witnessed rapid development. Traditional SATR research primarily relies on machine learning and pattern recognition algorithms. Specifically, the recognition system preprocesses the acquired images, including steps such as image enhancement and target extraction. Subsequently, through feature extraction and selection, the texture, shape, size, and other information of the targets are transformed into numerical features. Finally, by utilizing a machine learning classifier, the numerical features are analyzed and evaluated, enabling the automated recognition of different targets. For instance, in [5], the original IR images are first segmented to obtain the contour features

Da questo esempio possiamo già osservare che, trattandosi di articoli matematici, contengono numerosi acronimi e cifre, oltre a hyperlinks ad altri articoli. Ciò sarà ancora più evidente durante l'analisi dei topic e delle parole al loro interno.

Dopo la tokenizzazione ho ottenuto 10 liste di token, di cui riporto di seguito un esempio.

Esempio di token del primo documento:

```
['scenarios', ',', 'there', 'will', 'be', 'various', 'recognition', 'models', 'with', 'different']
```

Da questo esempio possiamo constatare la presenza di token costituiti da segni di punteggiatura e parole, come congiunzioni, non significative per il topic modelling che eseguiremo, e che quindi procedo a rimuovere nella fase di pre-processing. Riporto di seguito un esempio del risultato ottenuto dopo il pre-processing.

Esempio di token del primo documento dopo rimozione di stopwords e punteggiatura e di lemmatizzazione:

```
['difference', 'scheme', 'approximate', 'dynamic', 'system', 'considered', 'discrete', 'model', 'phenomenon', 'described']
```

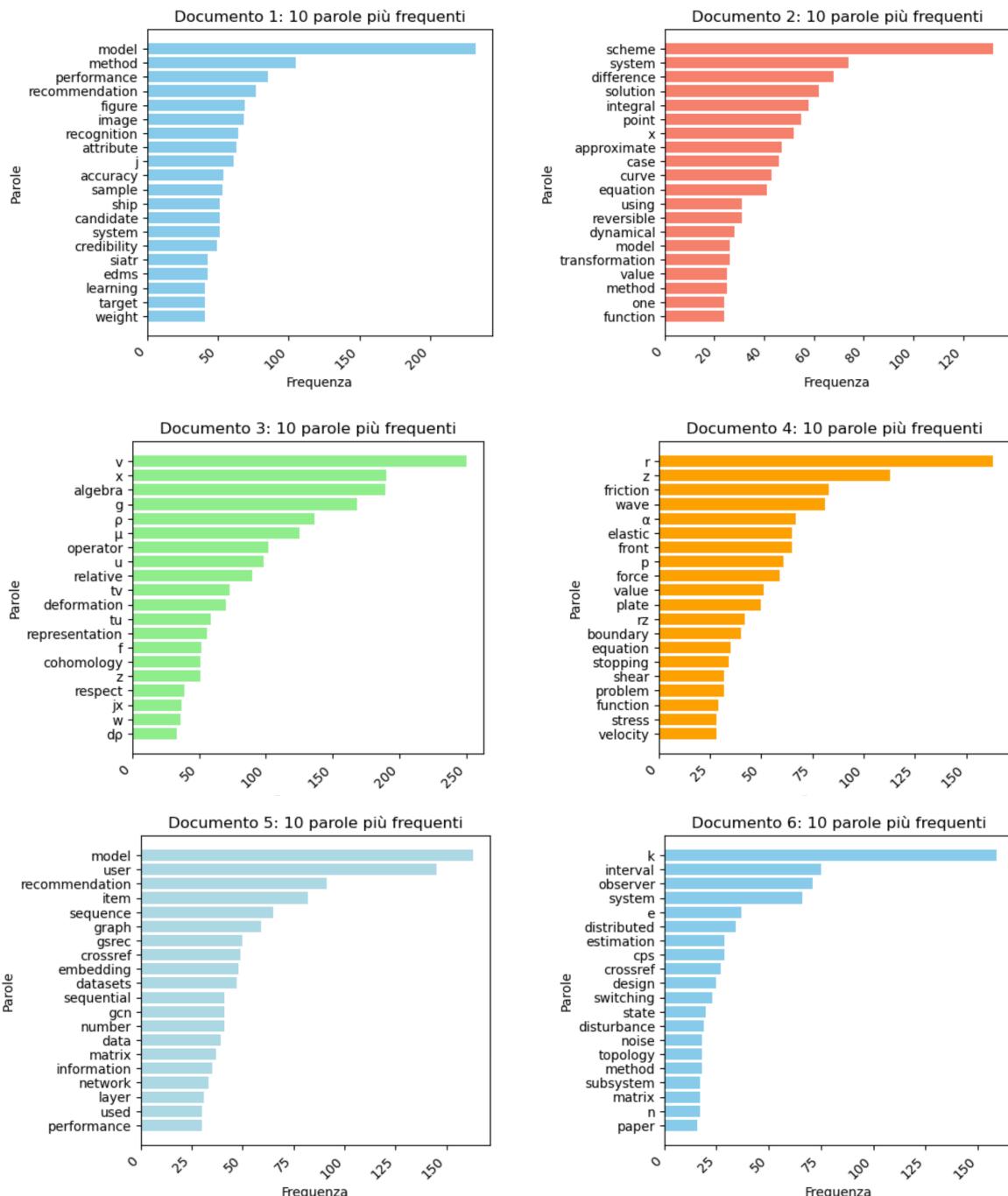
Ho poi proceduto alla creazione del vocabolario e ho calcolato il numero totale di token di cui è costituito (come si osserva sotto).

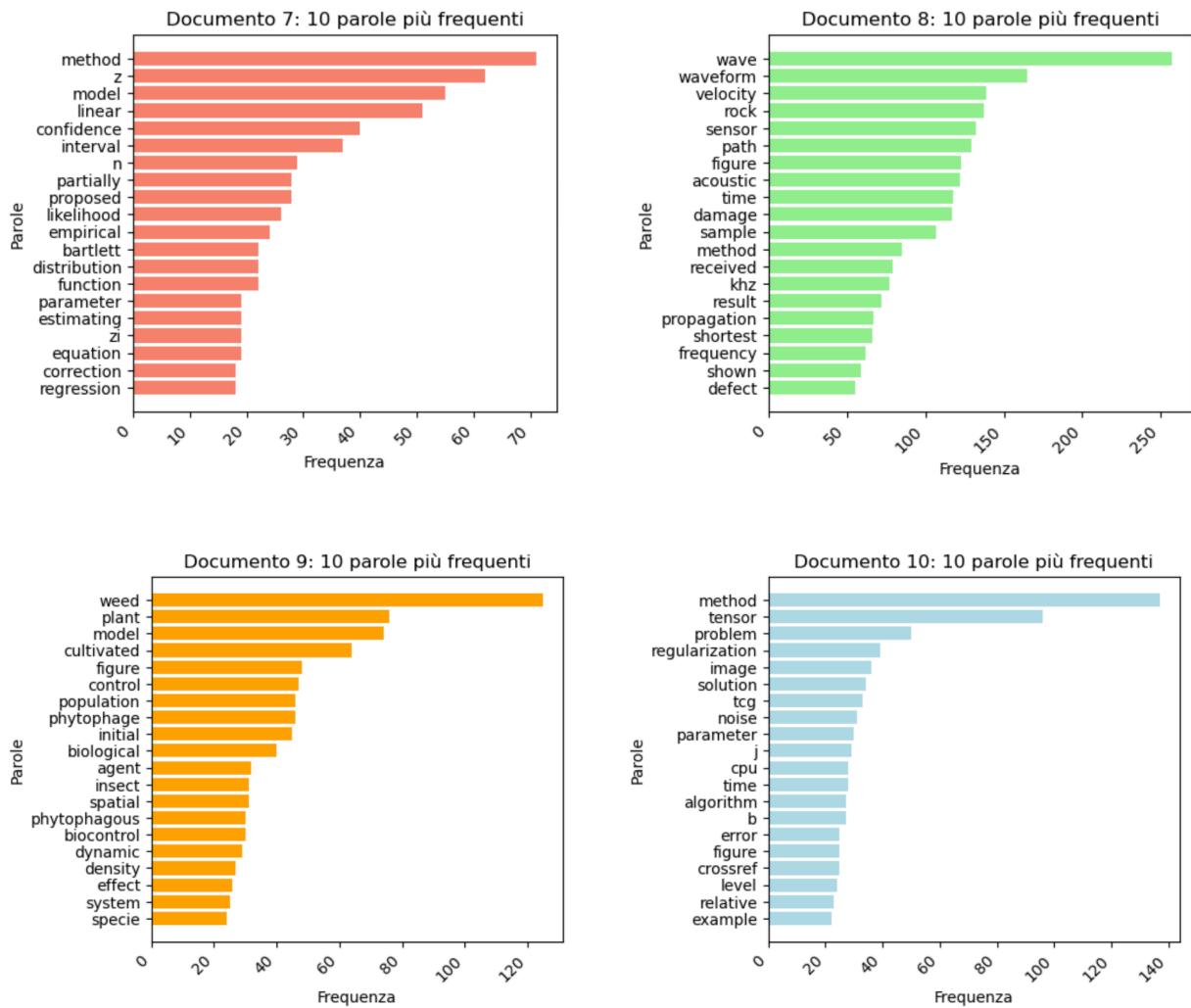
Numero totale di token (dopo pre-processing) nel Vocabolario: 5675

A quel punto ho voluto verificare se avevo a che fare con articoli lunghi o brevi in base al numero di token contenuti in ognuno di essi. Come si può constatare nell'immagine che segue, quasi tutti gli articoli sono composti (dopo il pre-processing) da più di 3000 token, e sono quindi da considerare lunghi.

Numero di token nell'Articolo 1: 6509
 Numero di token nell'Articolo 2: 3901
 Numero di token nell'Articolo 3: 4081
 Numero di token nell'Articolo 4: 3887
 Numero di token nell'Articolo 5: 5064
 Numero di token nell'Articolo 6: 2451
 Numero di token nell'Articolo 7: 2516
 Numero di token nell'Articolo 8: 7316
 Numero di token nell'Articolo 9: 3817
 Numero di token nell'Articolo 10: 3078

Come anticipato nella sezione antecedente, ho realizzato degli histogrammi che riportano le dieci parole più frequenti per ogni documento, come si può vedere di seguito.





In questi grafici possiamo notare che in ogni documento ci sono sempre una o poche parole che hanno una frequenza molto più elevata rispetto alle altre. Per esempio, nel primo documento è la parola "model", che compare più di 200 volte, nel secondo è la parola "scheme", che compare più di 120 volte, etc. La frequenza delle restanti parole varia in un range che va dalle 20 alle 50 occorrenze nei singoli articoli. Questo tipo di analisi è importante soprattutto perché le parole più frequenti all'interno di un documento spesso riflettono i temi principali trattati da quest'ultimo, e quindi ci fanno intuire quali saranno i topic contenuti in ogni documento.

A questo punto procediamo con il primo algoritmo di topic modelling, ossia l'LDA, e visualizziamo di seguito i topic estratti e le 5 parole più rilevanti - con il loro peso - che li compongono.

Tabella con le parole principali del Topic 1 e i loro pesi - LDA

Parole	Peso
0 "method"	0.032
1 "tensor"	0.015
2 "linear"	0.011
3 "z"	0.010
4 "model"	0.009

Tabella con le parole principali del Topic 2 e i loro pesi - LDA

Parole	Peso
0 "v"	0.051
1 "x"	0.039
2 "algebra"	0.039
3 "g"	0.034
4 "p"	0.028

Tabella con le parole principali del Topic 3 e i loro pesi - LDA

Parole	Peso
0 "model"	0.025
1 "k"	0.016
2 "method"	0.013
3 "system"	0.012
4 "performance"	0.009

Tabella con le parole principali del Topic 5 e i loro pesi - LDA

Parole	Peso
0 "model"	0.028
1 "user"	0.025
2 "recommendation"	0.015
3 "item"	0.014
4 "sequence"	0.011

Tabella con le parole principali del Topic 6 e i loro pesi - LDA

Parole	Peso
0 "wave"	0.032
1 "waveform"	0.020
2 "velocity"	0.017
3 "rock"	0.017
4 "sensor"	0.016

Tabella con le parole principali del Topic 7 e i loro pesi - LDA

Parole	Peso
0 "r"	0.019
1 "scheme"	0.015
2 "z"	0.015
3 "solution"	0.010
4 "friction"	0.010

Dato che il topic modelling non si occupa di assegnare un nome a ognuno dei topic estratti, sulla base delle parole in ognuno di essi, ora attribuisco loro un'etichetta.

- Topic 1: Metodi Tensoriali e Modelli Lineari
- Topic 2: Algebra Lineare e Simboli Matematici
- Topic 3: Modelli di Sistema e Performance
- Topic 4: Modelli Agricoli e Piante
- Topic 5: Sistemi di Raccomandazione
- Topic 6: Onde e Sensori
- Topic 7: Schemi Matematici e Soluzioni

Analizziamo più in dettaglio le etichette scelte.

Il primo topic riguarda tecniche matematiche che utilizzano tensori e modelli lineari, e probabilmente è legato al campo dell'analisi dei dati, dove questi metodi vengono utilizzati per elaborare dati multidimensionali.

Il secondo topic s'incentra chiaramente sull'algebra lineare e sui simboli matematici costituiti da lettere del nostro alfabeto e dell'alfabeto greco.

Il terzo topic concerne la modellazione di sistemi - che potrebbero essere tecnologici, ingegneristici o gestionali - e l'analisi e l'ottimizzazione delle loro performance.

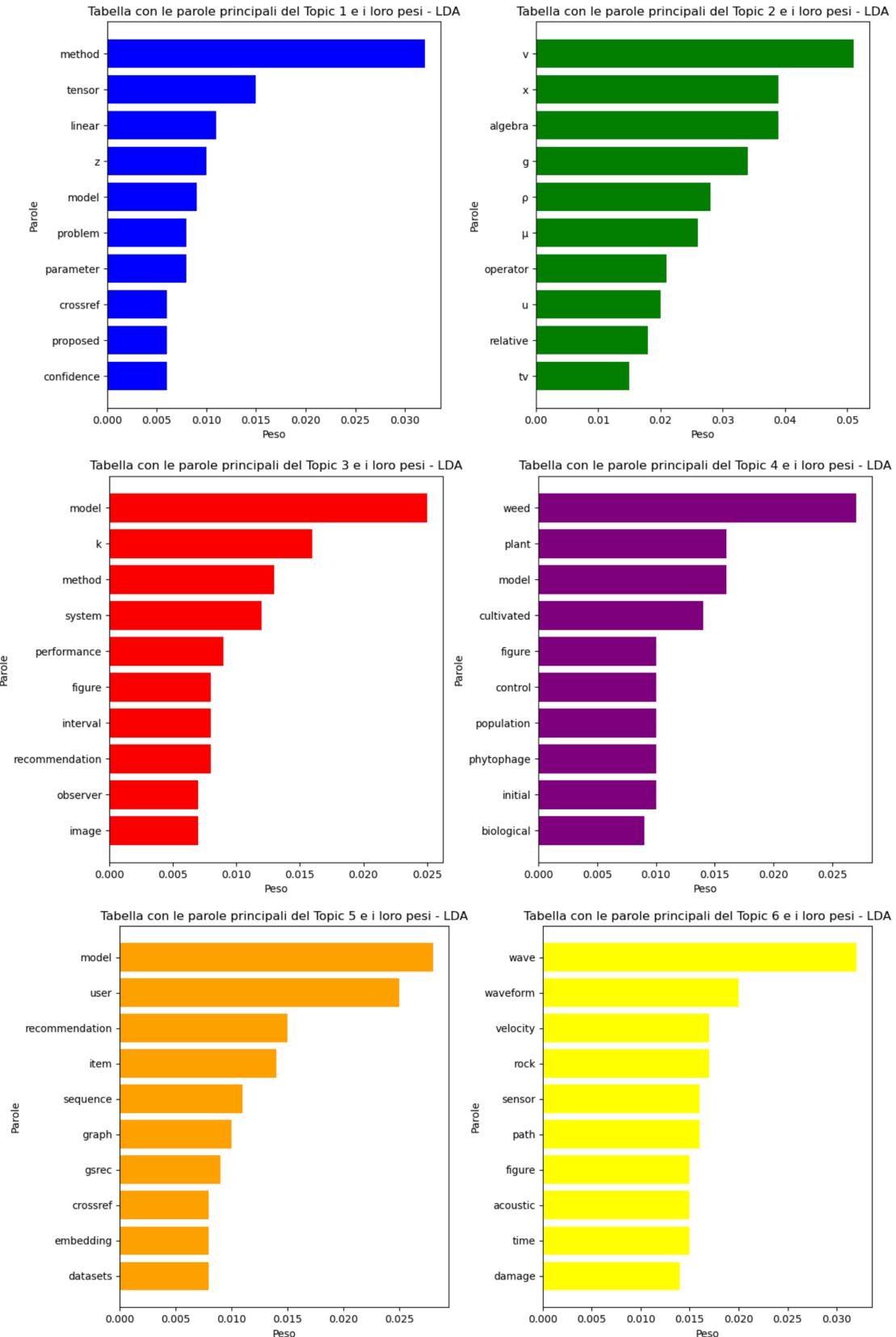
Il quarto topic riguarda l'agricoltura e la botanica, con un'enfasi sui modelli di crescita delle piante e la gestione delle colture.

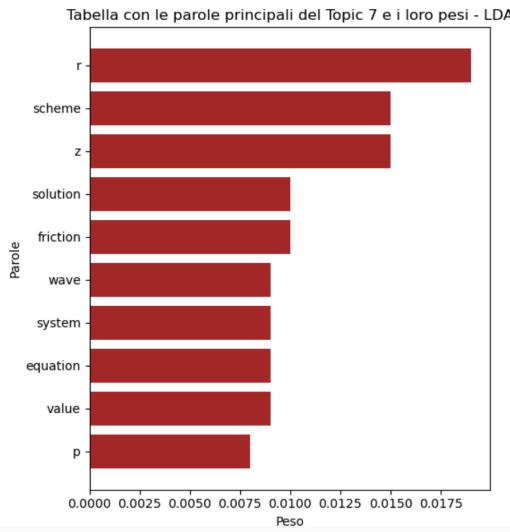
Il quinto topic è legato ai sistemi di raccomandazione che suggeriscono prodotti o contenuti agli utenti basandosi sui loro modelli di comportamento.

Il sesto topic si focalizza sulla fisica delle onde - la loro forma e velocità - e sui sensori utilizzati per misurare tali fenomeni.

Il settimo topic s'incentra su soluzioni a equazioni matematiche e fisiche che potrebbero includere problemi in cui è presente la forza di attrito.

Esaminiamo ora nuovamente le parole di ogni topic, ma stavolta attraverso degli histogrammi, per una visualizzazione più immediata.





A questo punto visualizziamo i topic che sono stati assegnati a ogni documento sempre tramite LDA

Documento 1:
Distribuzione dei topic: Topic 3 con probabilità 0.9999

Documento 2:
Distribuzione dei topic: Topic 7 con probabilità 0.9998

Documento 3:
Distribuzione dei topic: Topic 2 con probabilità 0.9998

Documento 4:
Distribuzione dei topic: Topic 7 con probabilità 0.9998

Documento 5:
Distribuzione dei topic: Topic 5 con probabilità 0.9998

Documento 6:
Distribuzione dei topic: Topic 3 con probabilità 0.9996

Documento 7:
Distribuzione dei topic: Topic 1 con probabilità 0.9997

Documento 8:
Distribuzione dei topic: Topic 6 con probabilità 0.9999

Documento 9:
Distribuzione dei topic: Topic 4 con probabilità 0.9998

Documento 10:
Distribuzione dei topic: Topic 1 con probabilità 0.9997

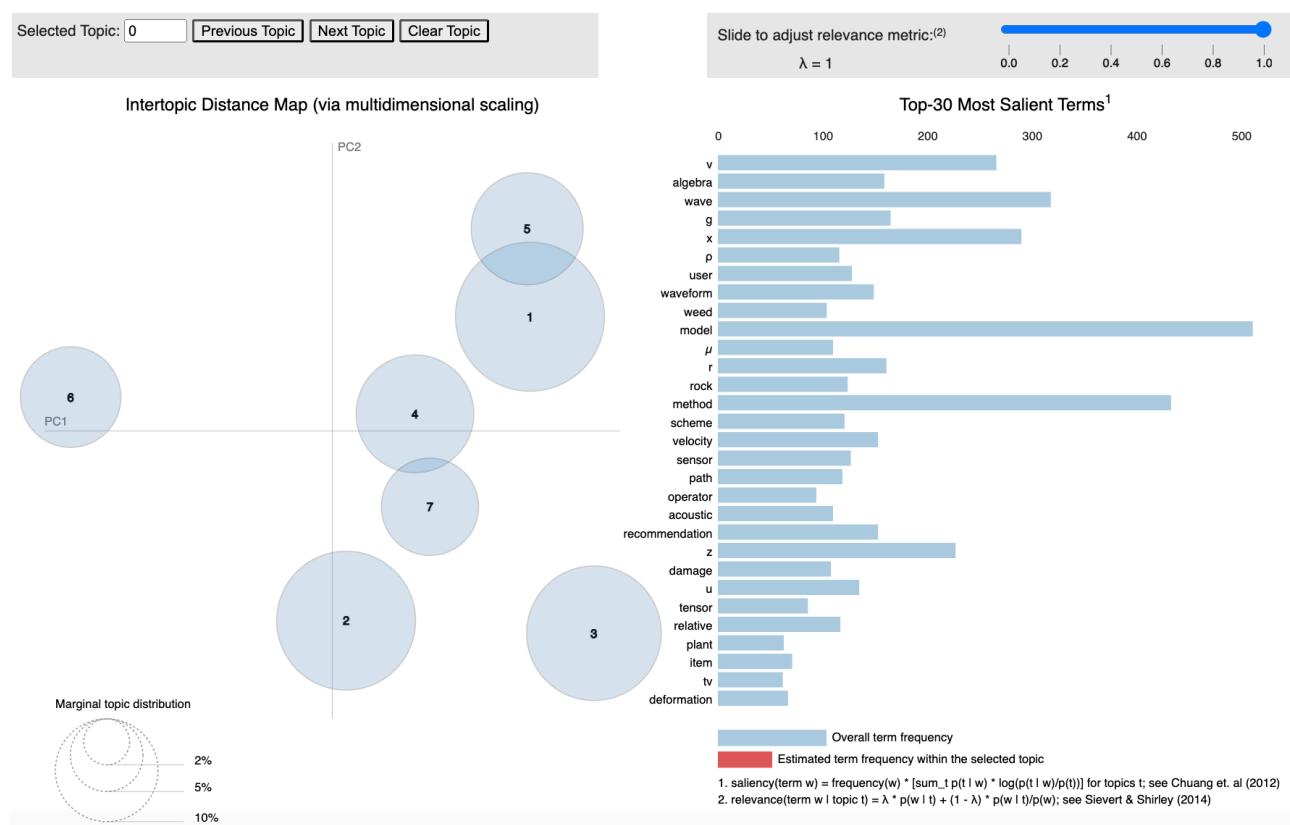
Da questi risultati possiamo notare che a ogni documento viene assegnato un unico topic con una probabilità molto alta. Ciò dimostra che ogni documento è tematicamente ben definito e quindi s'incentra su un unico argomento. Quindi, come si vede nell'immagine sopra, possiamo affermare che i documenti 7 e 10 trattano di modelli tensoriali e lineari, il documento 3 parla di algebra lineare, i documenti 1 e 6 riguardano la modellazione e l'ottimizzazione dei sistemi, il documento 9 s'incentra sui modelli agricoli, il documento 5 si focalizza sui sistemi di raccomandazione, il documento 8 parla di onde e sensori, e infine il documento 4 tratta di equazioni matematiche e fisiche.

Per valutare la coerenza dei topic, calcoliamo lo score UMass per ognuno di essi.

Topic	Coerenza UMass
1	-1.70
2	-0.83
3	-2.13
4	-0.51
5	-0.43
6	-0.55
7	-0.84

Valori fortemente negativi dell'indice UMass denotano una scarsa coerenza all'intento dei topic, mentre valori che si avvicinano allo 0 indicano una sempre maggiore coerenza. Dai nostri risultati possiamo osservare che, in base allo score UMass, quasi tutti i topic sono sufficientemente coerenti.

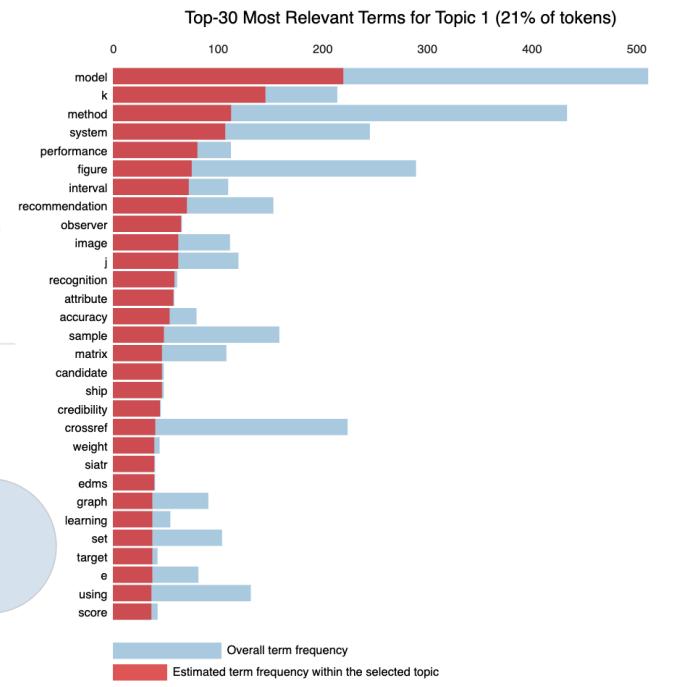
Mostriamo adesso la rappresentazione grafica dello spazio dei topic unitamente all'istogramma che rappresenta la frequenza complessiva di ogni termine nel corpus dei documenti, così come all'interno di ogni topic.



Possiamo subito osservare che i topic 1 e 5 e i topic 4 e 7 si sovrappongono parzialmente, il che indica che contengono delle parole in comune, come si può vedere dalle due figure che seguono (riporto solamente il caso del topic 1 e 5).

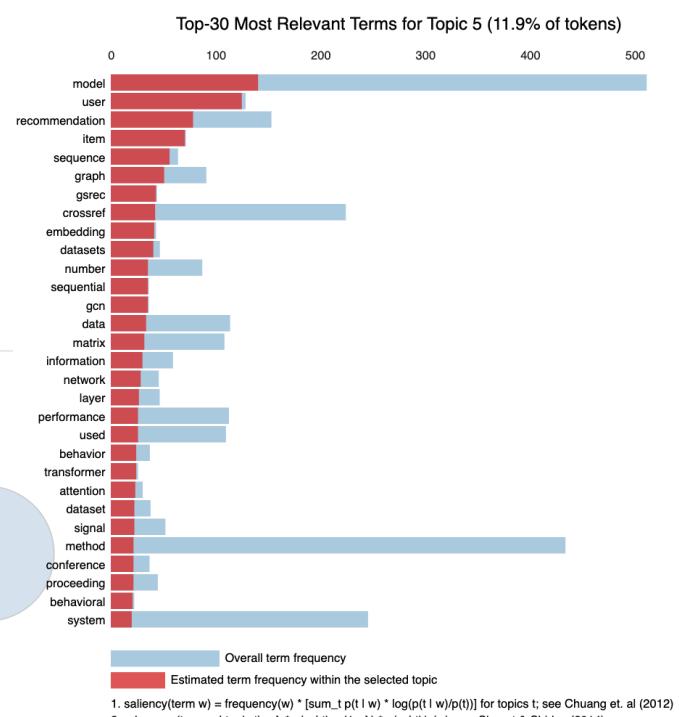
Selected Topic: Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(²)



Selected Topic: Previous Topic Next Topic Clear Topic

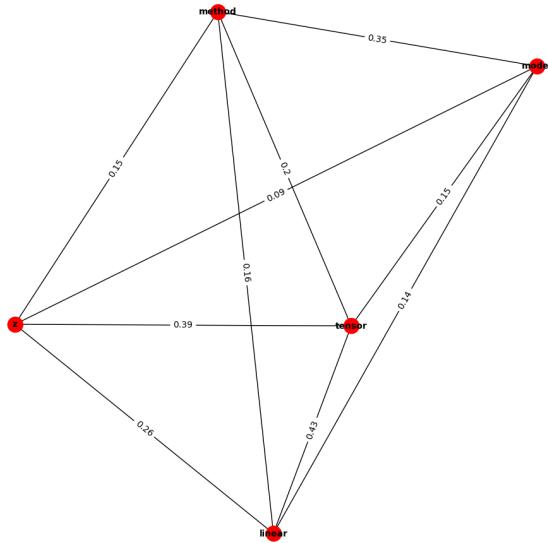
Slide to adjust relevance metric:(²)



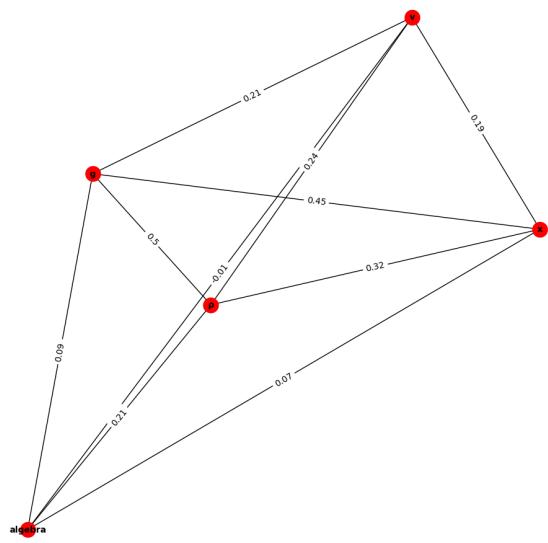
Invece gli altri topic, soprattutto il topic 6, sono tra loro isolati, il che indica che non hanno termini in comune. Di fatto, i topic 1 e 5, parlando rispettivamente di metodi tensoriali e lineari e di sistemi di raccomandazioni, hanno entrambi a che fare con l'elaborazione di dati multidimensionali, mentre per esempio il topic 6, che esamina la fisica delle onde e sulla loro misurazione, è tematicamente più distante dal resto dei topic.

Riporto di seguito il secondo metodo che ho utilizzato per valutare la qualità dell'estrazione delle parole che costituisco i topic, ovvero attraverso grafi pesati.

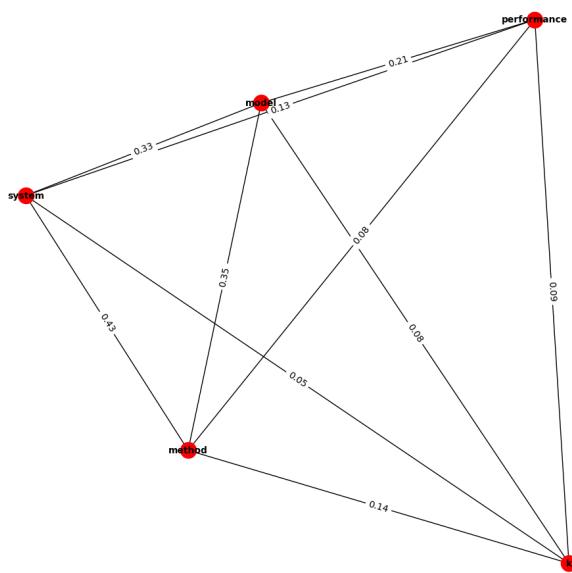
Grafo delle 5 parole più significative per il Topic 1 e della loro similarità coseno - LDA - Score: 2.32



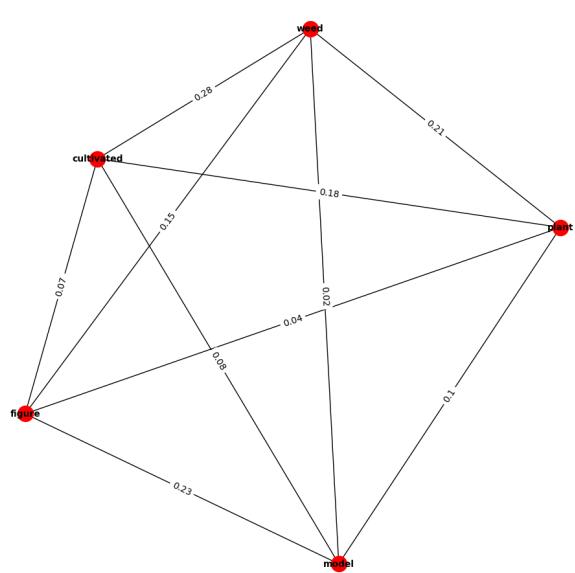
Grafo delle 5 parole più significative per il Topic 2 e della loro similarità coseno - LDA - Score: 2.27



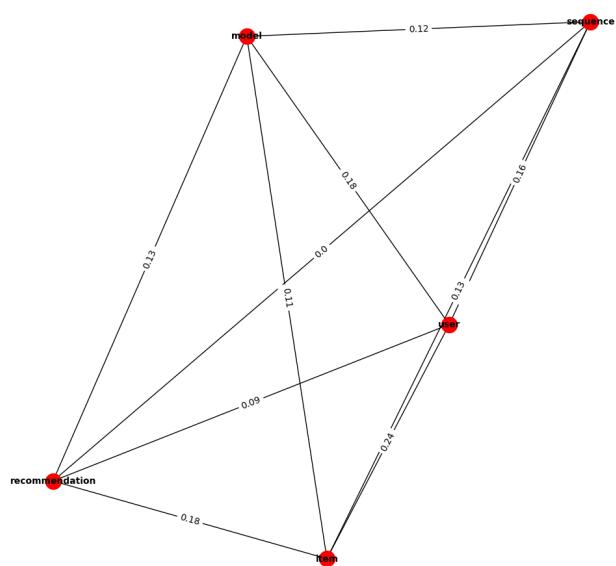
Grafo delle 5 parole più significative per il Topic 3 e della loro similarità coseno - LDA - Score: 1.89



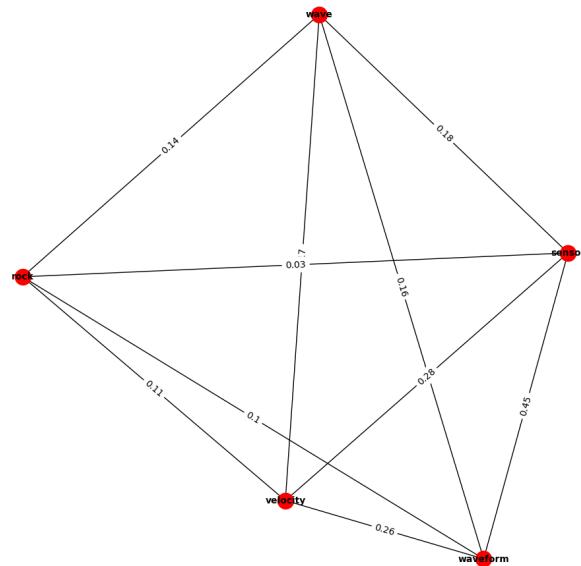
Grafo delle 5 parole più significative per il Topic 4 e della loro similarità coseno - LDA - Score: 1.36

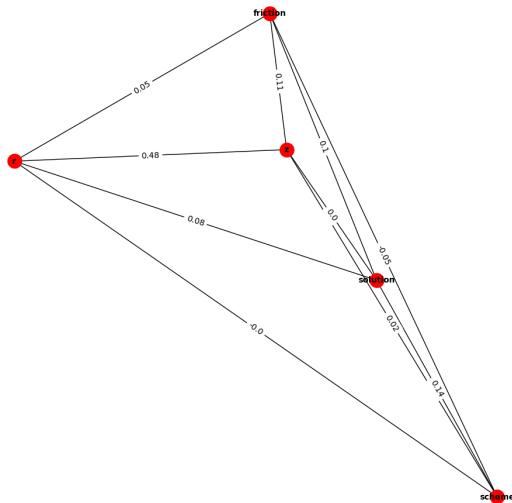


Grafo delle 5 parole più significative per il Topic 5 e della loro similarità coseno - LDA - Score: 1.34



Grafo delle 5 parole più significative per il Topic 6 e della loro similarità coseno - LDA - Score: 1.88





Topic 1: Score = 2.32
 Topic 2: Score = 2.27
 Topic 3: Score = 1.89
 Topic 4: Score = 1.36
 Topic 5: Score = 1.34
 Topic 6: Score = 1.88
 Topic 7: Score = 0.93
 Score Totale: 11.99

Nei grafi riportati sopra, ogni arco è pesato in base al valore della similarità coseno tra le due parole contenute nei nodi che unisce. Possiamo in questo modo calcolare un punteggio per ogni topic e uno complessivo che confronteremo con quelli del prossimo algoritmo.

Analizziamo ora i risultati ottenuti attraverso l'applicazione del secondo algoritmo di topic modelling, ossia il BTM.

Esaminiamo innanzitutto le 5 parole principali per ogni topic riportate in ordine decrescente d'importanza.

Tabella con le parole principali per ogni topic in ordine decrescente in base al loro peso in quel topic – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
0	wave	crossref	algebra	model	crossref	model	method
1	waveform	weed	scheme	method	observer	user	tensor
2	figure	model	operator	performance	interval	item	linear
3	velocity	plant	relative	recommendation	system	matrix	model
4	sensor	cultivated	system	system	rock	sequence	problem
5	time	control	solution	figure	method	graph	parameter

(Nell'immagine che segue non appaiono i pesi delle parole in quanto la libreria che ho utilizzato non li mostra).

Anche in questo caso, assegno dei nomi ai topic estratti.

Topic 1: Dinamiche delle Onde
 Topic 2: Agricoltura e Piante
 Topic 3: Schemi Matematici
 Topic 4: Modelli e Performance
 Topic 5: Sensori e Rilevamento
 Topic 6: Raccomandazioni Utente
 Topic 7: Algebra Lineare e Metodi

Possiamo quindi constatare che i topic estratti tramite BTM sono abbastanza simili a quelli estratti tramite LDA. A variare è però la miscela di topic assegnata a ogni

documento, come mostrato - sia numericamente sia graficamente - dalle immagini che riporto di seguito.

Tabella con la miscela dei topic per il Documento 1 – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 1	0.029	0.12	0.025	0.593	0.048	0.146	0.039

Tabella con la miscela dei topic per il Documento 2 – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 2	0.018	0.089	0.752	0.023	0.041	0.03	0.047

Tabella con la miscela dei topic per il Documento 3 – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 3	0.004	0.027	0.843	0.011	0.031	0.056	0.028

Tabella con la miscela dei topic per il Documento 4 – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 4	0.136	0.024	0.664	0.02	0.049	0.072	0.035

Tabella con la miscela dei topic per il Documento 5 – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 5	0.015	0.197	0.019	0.129	0.031	0.577	0.032

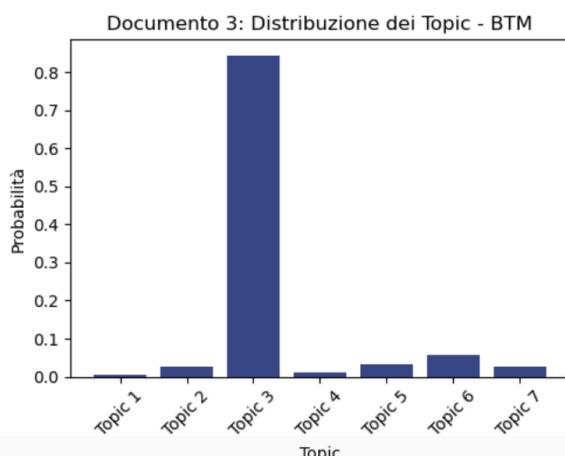
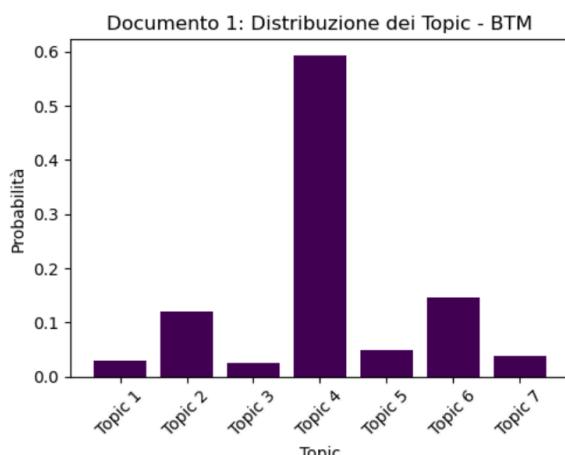


Tabella con la miscela dei topic per il Documento 6 – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 6	0.022	0.045	0.057	0.038	0.592	0.136	0.11

Tabella con la miscela dei topic per il Documento 7 – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 7	0.02	0.041	0.062	0.041	0.057	0.036	0.744

Tabella con la miscela dei topic per il Documento 8 – BTM

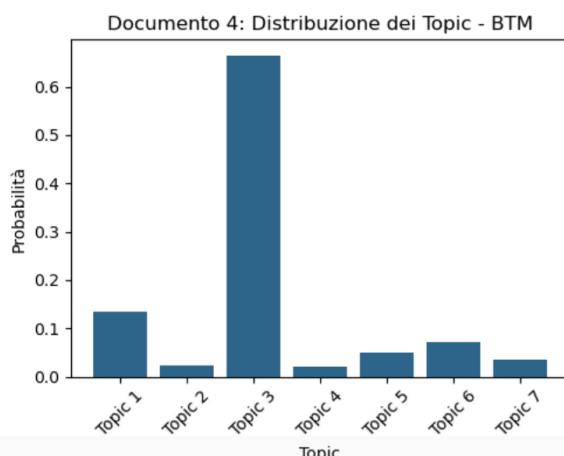
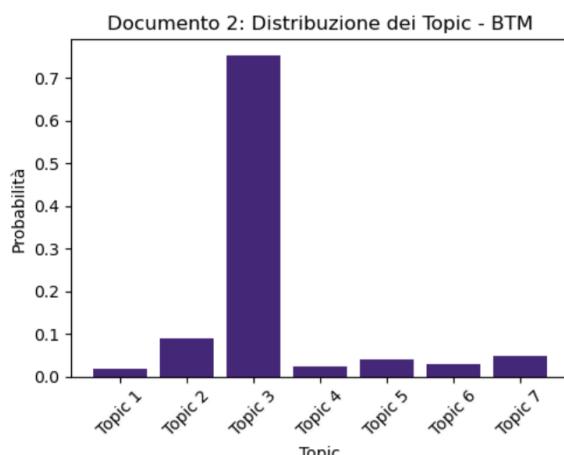
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 8	0.733	0.044	0.024	0.039	0.113	0.014	0.032

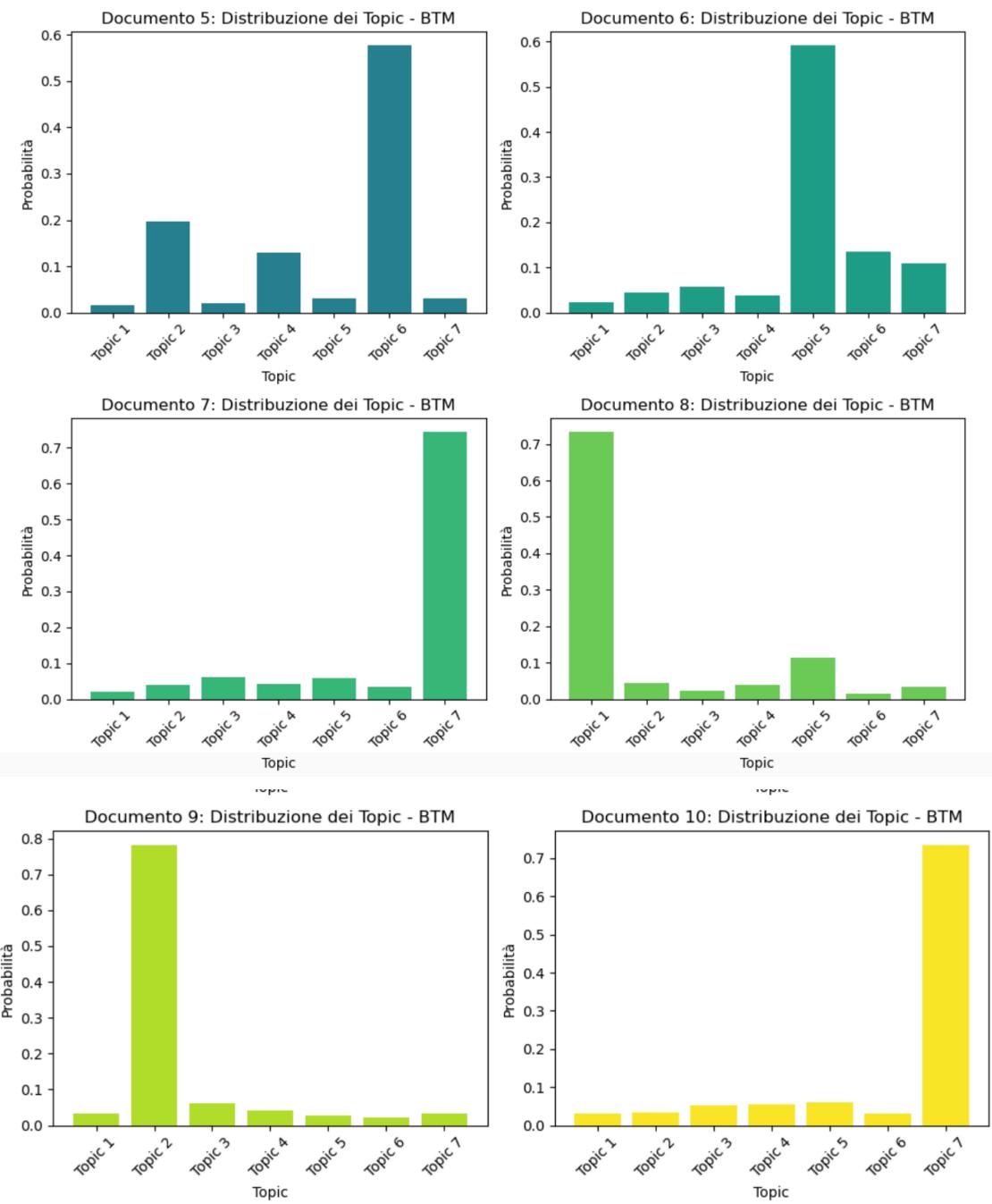
Tabella con la miscela dei topic per il Documento 9 – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 9	0.032	0.782	0.063	0.041	0.027	0.023	0.033

Tabella con la miscela dei topic per il Documento 10 – BTM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Documento 10	0.032	0.034	0.053	0.056	0.06	0.032	0.734





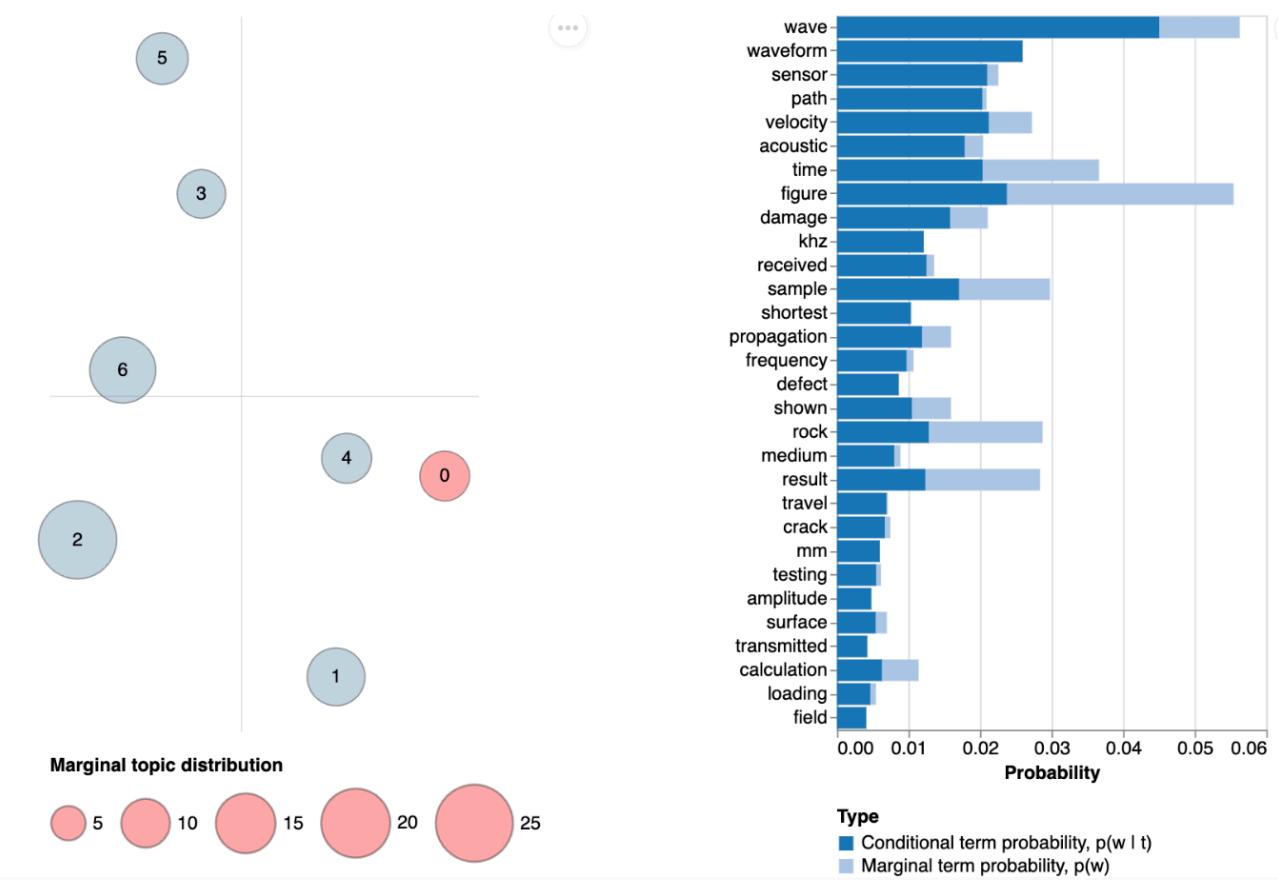
Da queste immagini possiamo constatare che in ogni documento c'è sempre un topic più probabile (dal 60 all'80%), mentre gli altri hanno probabilità molto più basse (non oltre il 20%). Dunque, anche se, come nell'LDA, in ogni documento prevale uno solo topic, nei risultati ottenuti con il BTM in ogni documento viene rilevata la presenza, sebbene in piccola parte, pure di altri topic.

Come anticipato nella sezione precedente, il BTM è un algoritmo efficace soprattutto per documenti brevi e non è pertanto adatto a documenti lunghi come quelli che compongono il mio dataset, come dimostrano i risultati ottenuti calcolando lo score UMass per i topic estratti con il BTM.

Metodo BTM – Coerenza del Topic 1 secondo il metodo UMass: -2.54
 Metodo BTM – Coerenza del Topic 2 secondo il metodo UMass: -14.39
 Metodo BTM – Coerenza del Topic 3 secondo il metodo UMass: -38.81
 Metodo BTM – Coerenza del Topic 4 secondo il metodo UMass: -34.42
 Metodo BTM – Coerenza del Topic 5 secondo il metodo UMass: -24.63
 Metodo BTM – Coerenza del Topic 6 secondo il metodo UMass: -3.41
 Metodo BTM – Coerenza del Topic 7 secondo il metodo UMass: -67.53

Possiamo notare che la maggior parte dei topic ottiene uno score UMass fortemente negativo, che indica una scarsa coerenza al loro interno. Tali risultati divergono naturalmente da quelli ottenuti calcolando lo score UMass sul modello di LDA, i quali erano migliori in quanto si avvicinavano molto di più a 0.

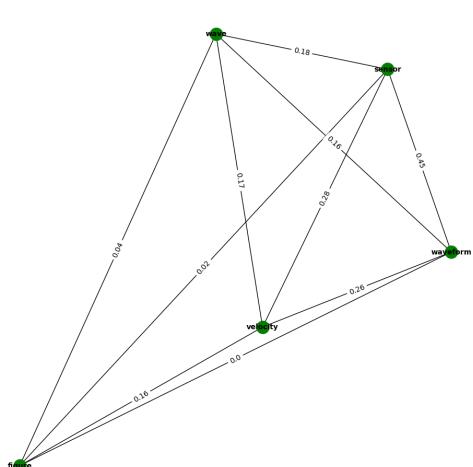
Anche per questo algoritmo rappresentiamo i topic nello loro spazio, unitamente all'istogramma contenente le probabilità marginali e condizionali di ogni parola, ovvero la sua probabilità di occorrenza rispettivamente nell'intero corpus e in ogni topic.



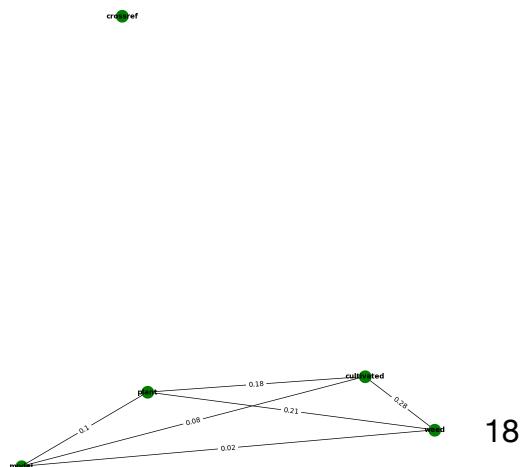
Per fare un esempio, questa immagine mostra la probabilità marginale e condizionale delle parole del primo topic (chiamato topic 0 in quanto la numerazione dei topic in questo grafico parte da 0).

Anche per questo algoritmo calcoliamo la coerenza dei topic con un altro metodo, cioè mediante dei grafi, come già fatto per l'LDA.

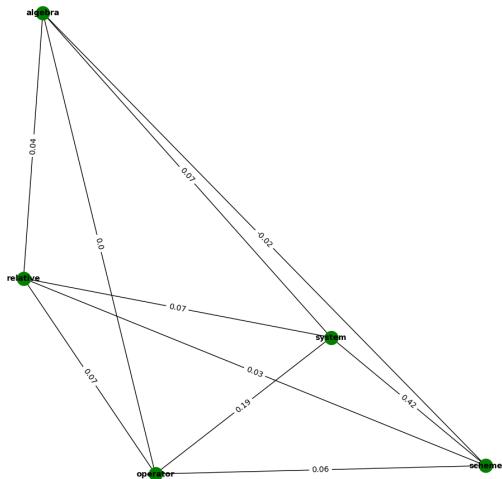
Grafo delle 5 parole più significative per il Topic 1 e della loro similarità coseno - BTM - Score: 1.72



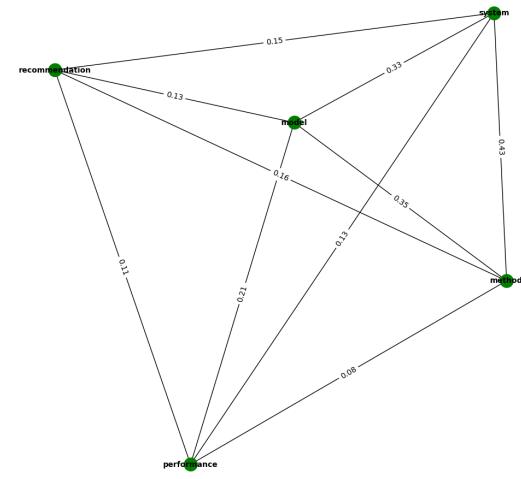
Grafo delle 5 parole più significative per il Topic 2 e della loro similarità coseno - BTM - Score: 0.87



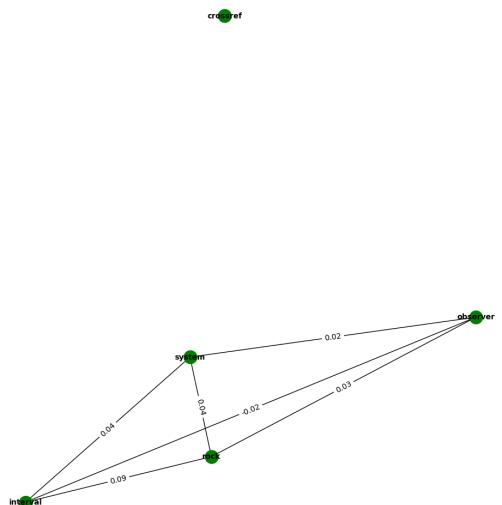
Grafo delle 5 parole più significative per il Topic 3 e della loro similarità coseno - BTM - Score: 0.93



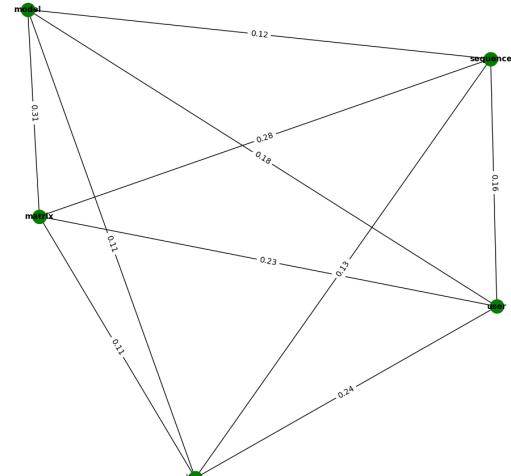
Grafo delle 5 parole più significative per il Topic 4 e della loro similarità coseno - BTM - Score: 2.08



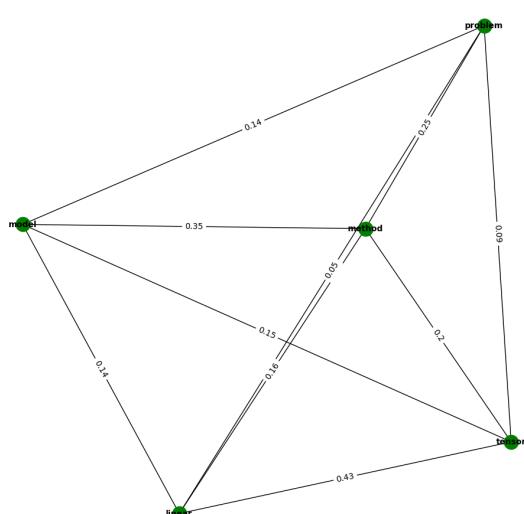
Grafo delle 5 parole più significative per il Topic 5 e della loro similarità coseno - BTM - Score: 0.20



Grafo delle 5 parole più significative per il Topic 6 e della loro similarità coseno - BTM - Score: 1.87



Grafo delle 5 parole più significative per il Topic 7 e della loro similarità coseno - BTM - Score: 1.96



Topic 1: Score = 1.72
Topic 2: Score = 0.87
Topic 3: Score = 0.93
Topic 4: Score = 2.08
Topic 5: Score = 0.20
Topic 6: Score = 1.87
Topic 7: Score = 1.96
Score Totale: 9.63

Anche attraverso questo metodo, in particolare confrontando lo score totale, possiamo notare che la coerenza dei topic ottenuti con BTM è minore rispetto a quella ottenuta con LDA, con un punteggio di 9,63 per il BTM contro 11.99 per l'LDA. Nei grafici riportati sopra

possiamo anche osservare che alcune parole non presentano archi incidenti e sono quindi nodi isolati. Questo indica che la solo similarità coseno con le altre parole del loro topic è nulla, un altro segnale di scarsa coerenza.

Conclusa la parte di topic modelling, ho proseguito nell'analisi del dataset applicando un'altra tecnica di topic modelling, ossia la text summarization, in particolare tramite l'algoritmo del TextRank. Per ogni articolo ho quindi generato un riassunto, sulla base delle frasi più rilevanti e ho riportato il topic precedentemente assegnato a quell'articolo mediante l'LDA, dato che si è rilevato essere l'algoritmo di topic modelling migliore per i nostri documenti.

Di seguito riporto i risultati ottenuti con il TextRank.

Riassunto dell'articolo 1:

This is because the training of the SIATR model mainly involves learning the feature space distribution of the samples, including the shape of the target, the brightness of t he imaging, and the characteristics of the background region. The calculation process of the weight matrix UP; three candidate model examples and the acquisition method of the weights Uj, iand Pj, ifor a single sample with a certain labe l set to .Mathematics , of * For the i-th input image, the mask image is obtained using the method in Figure , and the attribute partition interval of the image is obtaine d using the following formula: Ai=h Attribute 1ti, .

Numero del topic assegnato a questo articolo da LDA: 3 con probabilità: 0.9999 %

Riassunto dell'articolo 2:

So, for example, there was active discussion on the Internet about wing nut somersaults in zero gravity, which were observed by Janibekov, and which, nevertheless, are perfectly described by this model .Mathematics , of Following , let us denote the coordinates of the angular velocity vector relative to the principal axes of inertia as p,q,r, and the principal moments of inertia as A,B,C; then, the angular velocity evolution is described by a dynamic system $\dot{p}=aqr, \dot{q}=bpr, \dot{r}=cpq$, the coefficients of which are expressed in terms of the principal moments of inertia $a=-C-B, b=-A-C, c=-B-A$. In analytical theory , when studying the Riccati equation, they switched from the affine line to the projective line in the second half of the 19th century. Mathematics , of Thus, we consider xas a point in the projective space Pn, and the discrete model of the dynamical system as a transition from layer to layer: "x=Cx, described by the Cremona transformation Cdepending on At.

Numero del topic assegnato a questo articolo da LDA: 7 con probabilità: 0.9999 %

Riassunto dell'articolo 3:

The key role played in this step is to construct a representation of this Lie-Yamaguti algebra T, J, \dots, K (ong viewed as the representation space), that is, we shall present the explicit formulas of linear maps $g: V \rightarrow g$, $g: V \rightarrow V$ -gland Dg, m , which are linked with the representation and the relative Rota-Baxter operator T, such that the triple becomes a representation of Lie-Yamaguti algebra V. Consequently, we obtain the corresponding Yamaguti cohomology of T, J, \dots, K with coefficients in the representation . Let $T: V \rightarrow V$ be a relative Rota-Baxter operator on a Lie-Yamaguti algebra J, \dots, K with respect to a representation : Two linear deformations $T_1: t=T+tT_1$ and $T_2: t=T+tT_2$ are said to be equivalent if there exists an element $X \in \Lambda^2 g$ such that $T_2 = T + tT_1$. A linear deformation $T: t=T+tT$ of a relative Rota-Baxter operator T is said to be trivial if there exists an element $X \in \Lambda^2 g$ such that $T = T + tX$ is a homomorphism from T_2 to T_1 . Then, $Id + tX$ is a Lie-Yamaguti algebra homomorphism of g, i.e., the following equalities hold: for all $x, y, z \in g$, $[x, [y, z]] = [[x, y], z] + [y, [x, z]]$, $[x, [y, z]] = [[x, y], z] + [y, [x, z]]$.

Numero del topic assegnato a questo articolo da LDA: 2 con probabilità: 0.9998 %

Riassunto dell'articolo 4:

In the case when his small compared to the characteristic propagation length of the elastic pulse, the stress change over the thickness of the plate can be neglected and the friction force can be considered as a volumetric one with density $\eta = N/h$. Mathematics , of Let us choose the beginning of the polar coordinate system in the center of the cutout . Then, due to the symmetry of the problem, the tangential stresses τ and the transversal velocity v satisfy the following system of partial differential equations: $\partial_t \partial_r v + \partial_r^2 v = 0$, $\partial_t \tau + \partial_r \tau = 0$, $\partial_t v - \partial_r v = 0$. Dimensionless variables and quantities are used here: the stress is related to the shear modulus g of the plate material, the velocity is related to the velocity of transverse elastic waves $c = \sqrt{\mu/g}$, the radial coordinate r is related to the radius of the hole r_0 , time t by the time the shear wave travels a distance equal to the radius of the hole r_0 , and the friction parameter $\eta = r_0/N = r_0/v$ is introduced.

Numero del topic assegnato a questo articolo da LDA: 7 con probabilità: 0.9998 %

Riassunto dell'articolo 5:

Key Description nu the number of users ni the number of items U the set of users, $U=\{u_1, u_2, \dots, u_n\}$ I the set of items, $I=\{i_1, i_2, \dots, i_m\}$ I the identity matrix, $I \in \mathbb{R}^{L \times L}$ the length of behavioral sequence Su the historical behavioral sequence of the user $u \in U$ s the length of training sequence A adjacency matrix that has interactive information between users and items D degree matrix, $D \in \mathbb{R}^{L \times L}$ the embedding size d an embedding of user $u, e_u \in E$ an embedding of item $i, e_i \in E$ an embedding matrix, $E \in \mathbb{R}^{d \times L}$ position embedding g, $PERs \in \mathbb{R}^{d \times d}$ incentive factor in adjacency matrix, $\theta = \text{avglength}$ index the distance between an item and the last item the user clicked or purchased GSRec is devised to predict top-N ranked items with which the user will likely interact by exploiting existing user-item interaction information. The Adam optimizer is an adaptive optimizer that combines the RMSProp optimizer and the Momentum optimizer, which can adjust the learning rate based on historical gradient information: $\Delta w = \alpha \nabla V + \beta_1 w - \beta_2 \Delta V + \gamma g$ where α is the learning rate, β_1 is the momentum of the current step, V is the variance of the current step, β_2 is a coefficient that increases the stability of the denominator, β_1 is the historical momentum retention rate, β_2 is the historical variance retention rate, and g is the gradient.

Numero del topic assegnato a questo articolo da LDA: 5 con probabilità: 0.9998 %

Riassunto dell'articolo 6:

The following is the ith subsystem with disturbances and noise: $x_i = Ax_i + Bu_i + pi, y_i = Cx_i + q_i$, where $x_i \in \mathbb{R}^n$ is the state, $u_i \in \mathbb{R}^m$ is the control input, $y_i \in \mathbb{R}^m$ is the output, $p_i \in \mathbb{R}^n$ is the disturbance, $q_i \in \mathbb{R}^m$ is the noise. In the future, we will study the necessary conditions for the design of interval observers, and in conjunction with this paper, we will give the necessary and sufficient conditions for the design of interval observers.

Numero del topic assegnato a questo articolo da LDA: 3 con probabilità: 0.9999 %

Riassunto dell'articolo 7:

The bootstrap procedure of estimating the Bartlett correction factor in the new linear model follows the procedure shown below: .Generate bootstrap resamples of size n by sampling with replacement from the sample $\{Y-E\}_n$ and $\{X1-E\}_n$, respectively, after the projection; then, calculate $-2\log(R_k)$ based on the resamples, where $\hat{\beta}$ is the global max imum empirical likelihood estimator of β based on the original sample $\{Y-E\}_n$ and $\{X1-E\}_n$. In that case, the length of the confidence interval for the Bartlett correction is larger than that of the normal approximation, the gam method,Mathematics , of and the em pirical likelihood without Bartlett correction, but the coverage probability is the closest to the nominal level %.

Numerico del topic assegno a questo articolo da LDA: 1 con probabilità: 0.9997 %

Riassunto dell'articolo 8:

According to Radon changes, the following is true: $t_i = Z \cdot L_{i1} \cdot v_{jdl} = Z \cdot L_{ifjd},$ where v_j is the wave speed of the j th small unit, and f_j is the reciprocal of the j th small unit's wa ve speed.Mathematics , of When the wave velocity grid of the small element assumed for wave velocity inversion is small enough, the distance can be considered as a constant, and Equation can be changed as follows: $t_{im} = \sum_j a_{11}x_{j1} + a_{12}x_{j2} + \dots + a_{1m}x_m$ $t_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2m}x_m \dots t_n = a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nm}x_m,$ where t is the mo nitored arrival time; a_i is the propagation length obtained by the Bellman-Ford algorithm; x_i is the slowness of wave velocity. In Figure , is the spectrum diagram of the waveform of the straight path passing through the defect, and is the spectrum diagram of the waveform of the straight path not pa ssing through the defect.

Numerico del topic assegno a questo articolo da LDA: 6 con probabilità: 0.9999 %

Riassunto dell'articolo 9:

It should be noted that the weed always outnumbers the cultivated plant at the initial phase of the transient dynamics (as in the cases presented above in Figure 1b,c).Mathem atics , of .. Case of 1D Spatial Domain Numerical experiments reveal that, due to accounting for spatial effects, even the 1D version of the model demonstrates dependence of the simulated weed control mea- sures on both parameter values and initial distribution of species densities, manifesting the capability of the model to capture conditions determining successful suppression of the weed foci, explaining the synergistic effect of the combined application of classical and phytocenotic methods. Moreover, according to our observations, successful suppression of weeds in the spatial model -can be achieved with much lower initial abundance of the phytophage population than in the point model , e.g., with $Z=..$. Next, with the same parameters of Set presented in Table , we analyzed the mech- anism providing synergistic effect of complex app lication of classical and phytocenotic methods of weed control, specifying the initial distributions of weed, cultivated plant, and phytophagous biocontrol agent in order to consider separation of weed-infested and cultivated areas, and the concentrated release of phytophagous insects in the zone of maxi- mum weed density.

Numerico del topic assegno a questo articolo da LDA: 4 con probabilità: 0.9998 %

Riassunto dell'articolo 10:

Example : The original image Lena, the blurred and noised image and reconstructed images by the tCG method, the A-tCG-FFT method, the A-CGLS-FFT method, the auto-tCG me thod, the auto-ttCG and the auto-ttpCG method according to the noise level $v=-3$ in Table . Example : The second frame image of the original video, the blurred and noisy image and recovered images by the tCG method, the A-tCG-FFT method, the A-CGLS-FFT method, the auto-tCG method, the auto-ttCG and the auto-ttpCG method according to the noise level $v=-3$ in Table .

Numerico del topic assegno a questo articolo da LDA: 1 con probabilità: 0.9997 %

Dai risultati possiamo notare che la maggior parte di questi riassunti contiene molti simboli, oltre a lettere e operatori matematici. Inoltre, i testi generati non sempre sono sintatticamente e grammaticalmente corretti. Ciò è dovuto alla natura e alla struttura degli articoli originali, che contengono numerose espressioni matematiche o fisiche con lettere e caratteri che solitamente non compaiono in altri tipi di testi, come ad esempio quelli narrativi, che si prestano meglio a essere riassunti tramite questo algoritmo.

In alcuni riassunti è comunque possibile riconoscere il topic principale, che coincide con quello individuato dall'LDA. Ad esempio nel riassunto dell'articolo 9 dove, già dalla prima frase “*It should be noted that the weed always outnumbers the cultivated plant at the initial phase of the transient dynamics*”, possiamo individuare l'argomento principale, ossia i modelli agricoli e le piante. Lo stesso avviene nel riassunto dell'articolo 8 dove la frase “*the spectrum diagram of the waveform of the straight path passing through the defect, and is the spectrum diagram of the waveform of the straight path not passing through the defect*” rivela già il topic dell'articolo, ovvero onde sonore e sensori per la loro rilevazione.

L'ultima tecnica di text mining che ho applicato, e i cui risultati riporto di seguito, è il document clustering, tecnica eseguita a partire dalle distribuzioni dei topic nei documenti ottenute tramite l'LDA. Nello specifico, il clustering avviene tramite l'algoritmo del K-means e un numero di cluster pari a 4.

Cluster 1:

- Documento 6
- Documento 7

Cluster 2:

- Documento 1
- Documento 5
- Documento 8

Cluster 3:

- Documento 2
- Documento 3
- Documento 4
- Documento 9

Cluster 4:

- Documento 10

Silhouette Score medio: 0.66

Dai risultati possiamo osservare che il quarto cluster contiene un unico documento, mentre gli altri racchiudono più documenti, fino a un massimo di 4. Per valutare la coerenza interna dei cluster, ho calcolato anche l'indice di silhouette, il cui valore è di 0.66. Sulla base di tale valore, ho deciso qual era il numero di cluster più corretto per far sì che il valore si avvicinasse il più possibile a 1. Questo valore dell'indice di silhouette indica che gli elementi all'interno dello stesso cluster sono tra loro vicini e distanti da quelli degli altri cluster. Al contrario, se il valore si avvicina a 0, vuol dire che gli elementi si trovano lungo i confini tra coppie di cluster. Nel nostro caso il valore dell'indice di silhouette ci suggerisce che i cluster sono abbastanza coerenti al loro interno, ma non del tutto. Di fatto, documenti come il 2 e il 4, che trattano lo stesso topic - schemi matematici e le loro soluzioni -, appartengono allo stesso cluster. Allo stesso tempo, documenti come l'1 e il 6, che trattano anch'essi lo stesso topic - modelli di sistemi e le loro performance -, vengono invece attribuiti a cluster diversi.

5. Conclusioni

Riassumendo i risultati della parte di topic modelling, possiamo affermare quanto segue.

1. Mettendo a confronto i risultati ottenuti tramite l'LDA con quelli ottenuti mediante il BTM, abbiamo notato che, vista la lunghezza dei documenti, l'algoritmo più adatto all'estrazione dei loro topic è l'LDA. Di fatto, i topic ottenuti tramite l'LDA sono più coerenti rispetto a quelli del BTM. Inoltre, l'assegnazione dei topic ai documenti è più immediata.
2. Una porzione delle parole che fanno parte dei topic è costituita da lettere, dato il carattere scientifico degli articoli analizzati.
3. Nello spazio dei topic, alcuni si sovrappongono, e ciò indica che hanno molte parole in comune, mentre altri sono tra loro isolati, avendo poche parole in comune.

Passando alla parte di text summarization, possiamo dire che, nonostante i riassunti ottenuti contengano grandi quantità di lettere e sequenze di caratteri non sempre comprensibili, è comunque possibile riconoscere la presenza del topic assegnato a ognuno dei documenti originali.

Infine, attraverso il document clustering, abbiamo notato che, anche se alcuni articoli che condividono lo stesso topic appartengono allo stesso cluster, ce ne sono altri che, pur condividendo lo stesso topic, vengono assegnati a cluster diversi.