REPORT TECNICO PROGETTO WEB ANALYTICS E ANALISI TESTUALE

Topic Modeling su documenti riguardanti le applicazioni Blockchain nel settore dell' Agri-Food.

Alice Porta 11/82/00142

Francesco Mussetti 11/82/00140

A.a. 2020/2021
Corso di studi: Data Science, Business Analytics e Innovazione

INTRODUZIONE

Il fine di questo progetto è quello di analizzare,sia in un'ottica globale che temporale, quelli che sono e che sono stati i principali argomenti trattati nelle fonti documentali presenti in rete riguardanti la tecnologia Blockchain e, in particolare, in riferimento alle sue applicazioni all'interno della filiera agro-alimentare.
Infatti, per quanto tale tecnologia risulti nota ai più in associazione alle Criptovalute, sempre più frequentemente, grazie alle sue caratteristiche di immutabilità e trasparenza, la sua applicazione è stata resa possibile nei settori più disparati, che l'hanno resa un'ideale tecnologia "base" per lo sviluppo di ulteriori tecnologie, rendendola di fatto una cosiddetta "general-purpose technology" (GPT).
Nel caso trattato, l'applicazione della Blockchain permette la tracciabilità dei prodotti lungo tutti i passaggi della filiera agro-alimentare, diventando così uno strumento fondamentale per garantire la genuinità e qualità dei prodotti forniti ai consumatori e combattere eventuali truffe ai danni dei produttori.
Attraverso l'utilizzo di Python,i documenti sono stati estratti dal DataBase "Scopus", contenente articoli,tesi di ricerca e pubblicazioni scientifiche e sono stati poi elaborati attraverso quattro differenti tecniche di Topic Modeling(LDA, LSA, PLSA e NMF).
Gli algoritmi di Topic Modeling sono dei modelli statistici non supervisionati utilizzati nel Natural Language Processing (NLP) e più in generale nell'apprendimento statistico che consentono di associare un argomento o topic ad un documento presente in una collezione di documenti. Questi algoritmi si basano sul concetto che documenti che trattano lo stesso argomento hanno una probabilità maggiore di contenere termini simili rispetto a documenti nei quali sono trattati argomenti differenti; ovviamente bisogna tenere presente che termini quali articoli, congiunzioni e così via, non possono essere considerati come esplicativi di specifici argomenti in quanto parole troppo comuni presenti in tutti i documenti; per questo motivo, tutti i testi prima dell'applicazione di questi algoritmi di Topic Modeling vengono sottoposti ad una fase di pre-processamento nella quale vengono "ripuliti" da questi termini non decisivi per l'individuazione dei topics. Dunque, l'obiettivo della Topic Modeling è quello di individuare i termini che compongono un particolare Topic e quindi di poter raggruppare in maniera automatica documenti che trattano uno stesso Topic.

PROCEDURA

- Web scraping delle fonti documentali.
- Normalizzazione dei documenti.
- Applicazione degli algoritmi di Topic Modeling implementati in diverse librerie e individuazione del numero ottimale di topics attraverso una GridSearch.
- Confronto risultati ottenuti con i diversi algoritmi.
- Analisi globale e temporale dei topics negli ultimi 5 anni.

Il progetto è stato sviluppato attraverso l'IDE Pycharm ed è stato strutturato in due branches:

1. **topicModeling_with_ScikitLearn**, nel quale sono stati analizzati tutti i punti precedentemente esposti attraverso i quattro differenti algoritmi di Topic Modeling: LDA, NMF, PLSA e LSA, implementando per ciascuno di essi un'analisi grafica;
2. **topicModeling_with_Gensim**, nel quale, oltre ad essere stati posti a confronto i risultati ottenuti con i due algoritmi LSA e LDA, anche secondo il parametro della Coherence, sono state generate WordClouds per entrambi i modelli e infine sono state generate due pagine HTML sulla base del modello LDA che contengono grafici interattivi riguardanti:
   - La visualizzazione dei Clusters delle fonti documentali secondo i diversi Topics ottenuti.
   - La visualizzazione dei Topics in un piano fattoriale che consente di visionare la similarità tra gli stessi e la visualizzazione delle 30 TOP WORD che maggiormente caratterizzano il Corpus documenti, con possibilità della stessa analisi specifica per ciascun topics.

Web Scraping

Il web scraping è un processo automatizzato che permette l'estrazione di contenuti e dati da un sito web che, una volta ottenuti, vengono successivamente sottoposti a differenti forme di analisi al fine di estrarre informazioni utili dai dati grezzi.
Letteralmente “scraping” significa “grattare, raschiare” e infatti nella pratica si estraggono dati e metadati da un sito web attraverso dei software detti “Spider” o “Crawler” che spesso simulano la navigazione umana alla ricerca di informazioni.
Mentre il tool naviga nel sito fa una sorta di copia-incolla e immagazzina dati, testi, immagini che poi vengono archiviati in un database pronto per i successivi passaggi.
Il processo di web scraping da noi eseguito prevedeva l’estrazione di informazioni base dalle fonti documentali quali Titolo,Abstract, Autore e Data di pubblicazione degli articoli disponibili su Google Scholar sotto la ricerca “Blockchain & Agrifood”; tale processo ha visto l’utilizzo di diversi tools e metodologie che di seguito vengono elencate:

A) CREAZIONE DI UN TOOL PERSONALIZZATO IN PYTHON :
Inizialmente, sulla base delle conoscenze ottenute a lezione, si è provato a implementare un Crawler personalizzato che riuscisse ad estrarre le informazioni a noi necessarie dal sito Google Scholar partendo dalla pagina :

https://scholar.google.com/scholar?hl=it&as_sdt=0,5&as_vis=1&q=blockchain+%26+agrifood

A tal fine ci siamo avvalsi dell’utilizzo delle librerie Request e BeautifulSoup.
-Request: dato che gran parte del testo sul Web è sotto forma di documenti HTML, Python permette di automatizzare la ricerca e il salvataggio di una pagina tramite il comando“requests.get()”.
- BeautifulSoup: può essere utilizzato per estrarre dati da HTML e salvarli in formati pronti all’uso, inoltre fornisce differenti funzioni dell’ottica del NLP.

Si è dunque proceduto al tentativo di estrazione dei dati utili all’analisi andando ad esplorare il codice HTML della pagina principale ed individuando nello stesso dei TAG utili al riconoscimento dei link degli articoli e non di quelli “secondari”.
Per rendere possibile tale pratica si è quindi implementato un algoritmo che andava a selezionare tutti i link della pagina principale aventi una profondità pari ad 1, quindi direttamente collegati ad essa, e tra questi, venivano esplorati unicamente quelli aventi un determinato TAG, che li identificava come articoli e pubblicazioni reali.
Una volta dentro ciascun link, si è analizzato ulteriormente il codice HTML, nel tentativo di trovare dei TAG che indicassero i dati d’interesse quali Titolo,Abstract,Autore e Data.

Proprio in tale fase è stata riscontrata una problematica che si è poi presentata in più occasioni: ciascuna pagina risultava avere una struttura differente in quanto pubblicata da differenti fonti, dunque ciascuna di esse risultava avere una struttura HTML peculiare, rendendo molto complicata l'individuazione di ciascun TAG identificatore per Titolo,Abstract,Autore e Data.
Nella speranza che comunque molte fonti potessero condividere i medesimi TAG identificatori, abbiamo proceduto a individuarli singolarmente.

Una volta analizzati i 20 link presenti nella prima pagina principale di Google Scholar, ci siamo scontrati con una seconda problematica: oltre a entrare in ciascun link degli articoli, era necessario automatizzare l'avanzamento nelle pagine principali di Scholar per analizzare nuovi articoli, problematica risolta attraverso l'utilizzo del Tool Selenium.

Selenium è un tool open source per la gestione automatizzata dei browser che in realtà è composto da più tools con differenti scopi, in particolare nel nostro caso ci siamo serviti di Selenium WebDriver, particolarmente utile nell'ottica del web scraping in quanto simula il comportamento di un umano all'interno di un browser e che ci ha permesso di utilizzare le sue funzioni per ciclare all'interno dei link a profondità 1 e di scorrere lungo le pagine principali di Google Scholar per analizzare tutti gli articoli trovati con la nostra query.

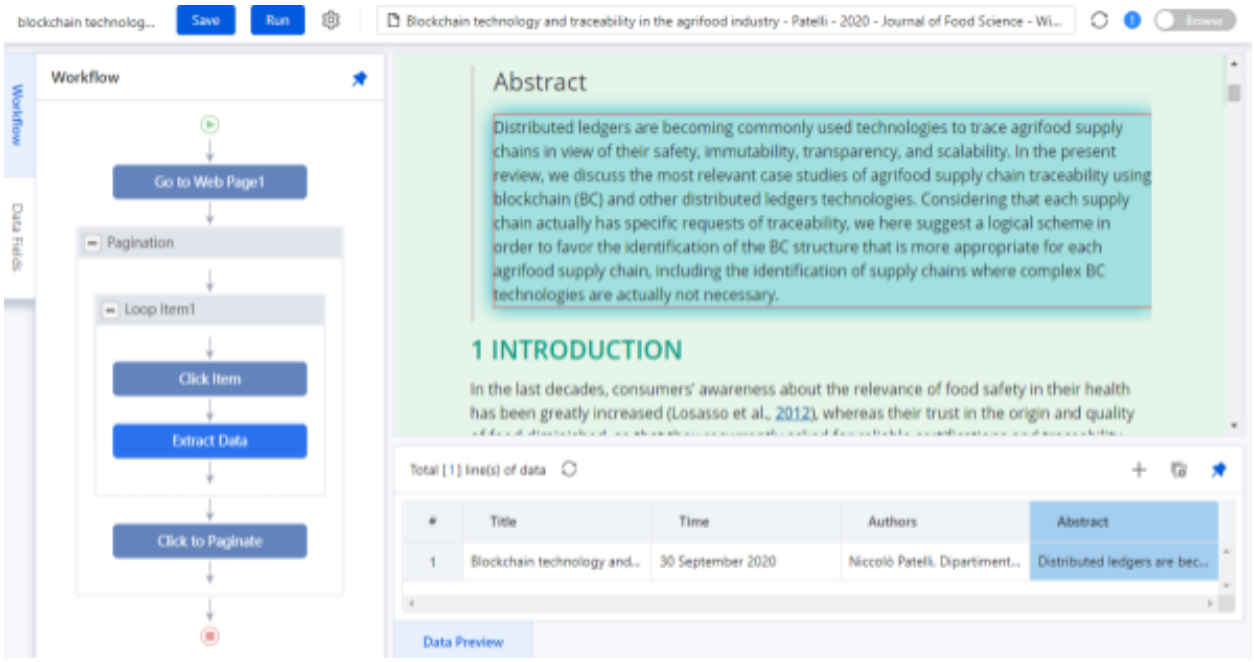
Nonostante l'ausilio del Driver di Selenium, abbiamo comunque ottenuto risultati parziali e difficilmente utilizzabili in quanto:
- Non siamo riusciti a risolvere il problema delle differenti strutture HTML per ciascuna fonte di pubblicazione e conseguentemente dei relativi TAGS d'interesse.
Inoltre abbiamo avuto un'ulteriore conferma della loro diversità, in quanto avendo a disposizione un quantitativo maggiore di articoli analizzati, abbiamo potuto verificare che spesso anche le medesime fonti di pubblicazione potevano presentare strutture HTML differenti e che in rarissimi casi i TAGS coincidevano.
- Il numero di testi effettivamente estratti risultava esiguo rispetto a quello totale, in quanto molto spesso venivano presentati errori di accesso alle pagine, probabilmente collegati a controlli CAPTCHA e alla comparsa di Cookie.

Dati i risultati non ottimali forniti dal Web Scraping effettuato dal Crawler personalizzato, abbiamo deciso di utilizzare vari Tools di Web Scraping Professionali per verificare l'effettiva possibilità di estrarre i dati dal Database di Google Scholar.

B) OCTOPARSE:
Octoparse è un'applicazione per il web scraping che, dunque, permette di estrarre diverse tipologie di dati da sorgenti online. La sua particolarità è che, grazie ad un'interfaccia grafica semplice e abbastanza intuitiva, permette di impostare un'architettura di estrazione senza dover scrivere una riga di codice.

Purtroppo l'utilizzo di questo tool ha prodotto scarsi risultati in quanto per come è strutturato, è molto utile nell'estrarre informazioni di pagine che condividono tutte la stessa struttura HTML (come ad esempio l'estrazione delle principali informazioni sui prodotti su Amazon), ma per il nostro scopo si è rivelato inadatto in quanto la struttura dei singoli articoli in cui lo scraper doveva accedere era diversa pressoché per tutti i link restituiti da Google Scholar.

Di seguito la struttura del task personalizzato con la quale i risultati ottenuti erano minimi.



Struttura task personalizzato Octoparse

Il risultato migliore ricavato con tale strumento è stato ottenuto mediante l'utilizzo dei tasks preimpostati di Octoparse; ovvero utilizzando il task di Google Scholar ciò che viene restituito è: *Autore, Titolo, Anno e solo parte iniziale dell'Abstract*;in quanto questo task opera estraendo le informazioni dalla pagina principale di Scholar in cui i risultati della query vengono restituiti e non entrando in ogni singolo link relativo agli articoli.
Tale fatto ha confermato ulteriormente la teoria secondo la quale risultava difficile, se non impossibile, riuscire ad estrarre i dati d'interesse da ciascun sito di pubblicazione che presentasse una struttura HTML differente e infatti lo stesso Tool professionale forniva risultati buoni solo nel caso in cui si fosse analizzata la sola pagina principale di Google Scholar, dove la struttura HTML è la medesima per tutti i trafiletti degli articoli e pubblicazioni, e dunque anche i TAGS risultano gli stessi.
Purtroppo tali risultati non avrebbero permesso un'analisi adeguata, in quanto avremmo utilizzato esclusivamente la parte iniziale dell'Abstract, perdendo probabilmente il reale senso contenuto in ciascun documento.
Per tali ragioni, si è deciso di provare ulteriori Tools Professionali indicati per il Web Scraping.

C) ENDNOTE,PAPERS,PUBLISH&PERISH:
Il passo successivo è stato dunque quello di provare ad utilizzare i tools più disparati come EndNote,Papers e Publish&Perish, ma, anche in questi caso, con scarsi risultati, in quanto abbiamo avuto conferma che Google Scholar fornisce raramente la possibilità di poter scaricare l'Abstract, infatti sulle tantissime prove fatte con i 3 tools, solo EndNote ha restituito degli Abstract, ma nell'ordine 6 casi rispetto ai circa 3500 analizzati.

Author	Year	Abstract	Title	Reference Type
L'Hermitte, C.; Nair, N. C.	2021	This study ...	A blockchain-enabled framework for sharing logistics resour...	Journal Article
Creydt, M.; Fischer, M.	2018	The develo...	Omics approaches for food authentication	Journal Article
Wang, X.; Che, M. Z.; Khali...	2020	Reactive ox...	The role of reactive oxygen species in the virulence of wheat l...	Journal Article
Patelli, N.; Mandrioli, M.	2020	Distributed ...	Blockchain technology and traceability in the agrifood industry	Journal Article
Ferioli, F.; Giambanelli, E.; ...	2017	BACKGROU...	Fennel (Foeniculum vulgare Mill. subsp. piperitum) florets, a t...	Journal Article
Antonucci, F.; Figorilli, S.; ...	2019	BACKGROU...	A review on blockchain applications in the agri-food sector	Journal Article

Tali 6 risultati condividevano effettivamente il medesimo sito di pubblicazione e la stessa struttura HTML, fattore che ci ha portato a confermare la teoria che ciascun articolo avesse una struttura praticamente unica e che anche i tools professionali avessero difficoltà ad estrarre le informazioni per questo motivo.

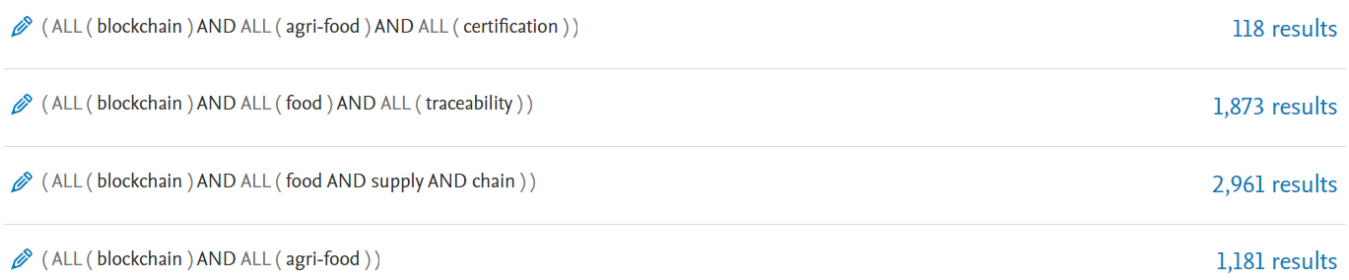
Ottenuti tali ulteriori risultati, seppur non confortanti rispetto al fine dell'analisi da effettuare, abbiamo avuto ulteriori conferme del fatto che per svariati motivi non risulta possibile estrarre gli Abstract dal database di Google Scholar, o perlomeno non è possibile con le nostre attuali conoscenze e strumenti a disposizione.
Per tale ragione si è deciso di utilizzare altri Database online simili a Scholar, ovvero Scopus e Web of Science, contenente articoli, tesi di ricerca e pubblicazioni scientifiche, che fornivano però strumenti integrati per poter scaricare agevolmente le informazioni in formato CSV.

D) SCOPUS

Dopo un'analisi degli articoli presenti in entrambi i Database, si è scelto di utilizzare Scopus in quanto si è riscontrato che restituisse molti più documenti pertinenti all'argomento oggetto di studio oltre a contenere praticamente tutti quelli presenti in Web of Science.

Con lo scopo di ottenere un Dataset più esteso possibile, si sono effettuate più ricerche nel Database Scopus per documenti pubblicati nel periodo dal 2017 al 2021,presentando differenti Query contenenti Keywords pertinenti all'argomento come ad esempio:

- "Blockchain+Agri-Food"
- "Blockchain+Food Supply Chain"
- "Blockchain+Food+Traceability"
-"Blockchain+Agri-Food+Certification"



Dopo aver ottenuto i vari risultati ed averli estratti sotto forma di file CSV, si è proceduto ad una loro ulteriore elaborazione tramite il software MySQL, dove è stato effettuato un JOIN rispetto alle colonne Titolo,Autore,Abstract e Anno di pubblicazione in modo da avere un unico Dataset contenente informazioni coerenti, inoltre si sono mantenuti solo i valori unici in modo da eliminare tutti i doppioni dei documenti originati dalle differenti query.
Come risultato finale si è ottenuto un Dataset di 2960 documenti pertinenti all'argomento oggetto di studio, presenti nel Codice con il nome di "2960_MixQuery_17_21_pulito.csv" sul quale è stato poi possibile eseguire le successive analisi.

Normalizzazione documenti

La normalizzazione dei documenti consiste in tutta una serie di elaborazioni dei testi che vengono fatte per rendere quest'ultimi utilizzabili da un sistema di *information retrieval*. Tra le varie azioni che si intraprendono in questo processo abbiamo lo *stemming* e/o la *lemmatization*, l'eliminazione delle stopwords, della punteggiatura e/o dei numeri. Tutto ciò viene eseguito successivamente alla tokenizzazione del testo, cioè la segmentazione del testo completo in singole clausole o in singole parole.

Nel caso in esame nel branch **topicModeling_with_ScikitLearn** abbiamo tokenizzato i documenti per singole word attraverso l'utilizzo del metodo word_tokenize() di nltk, la quale rappresenta un gruppo di librerie e moduli Python utilizzati nel Machine Learning, nell'Artificial Intelligence e nell'Information Retrieval. Precedentemente a questa operazione abbiamo utilizzato le *regular expression* per una pre-elaborazione del testo, ovvero abbiamo sostituito la punteggiatura e i numeri con uno spazio. Le RegEx o regular expression rappresentano delle nozioni algebriche che descrivono dei pattern di stringhe; infatti, tali funzioni sono utilizzate per filtrare e confrontare stringhe testuali tra loro, come espresso precedentemente.

Mentre nel branch **topicModeling_with_Gensim** la tokenizzazione è stata effettuata sia per singole parole che per bigrammi e trigrammi, in modo tale da riuscire a costruire dei topics maggiormente esplicativi delle argomentazioni trattate nei vari abstracts. Infatti, l'utilizzo dei bigrammi e trigrammi consente di comprendere meglio il senso delle frasi e di conseguenza, il senso dell'intero documento in quanto, per esempio, parole positive possono essere precedute o seguite da parole negative che ne alterano totalmente il significato (per esempio: like e not like).

Per quanto riguarda la lemmatizzazione e lo stemming in entrambi i branch abbiamo preferito optare per la sola lemmatizzazione in quanto quest'ultima considera il contesto e converte la parola nella sua forma base significativa, ovvero nel suo lemma.
Lo stemming, invece, rimuove i suffissi e ciò molto spesso porta a restituire parole prive di significato a differenza della lemmatization che invece restituisce sempre parole che sono presenti nel dizionario preso come riferimento dal metodo.
Di seguito un estratto del testo sottoposto al processo di normalizzazione:

... 'attention', 'pay', 'interaction', 'firm', 'regulator', 'consumer', 'social', 'performance', 'improvement', 'particular', 'innovative', 'business', 'model', 'share', 'economy', 'new', 'disruptive', 'technology', 'blockchain', 'cloud_compute', 'big', 'datum', 'play', 'vital', 'role', 'achieve', 'sustainability', 'informa_uk', 'limited_trade'], ['purity', 'essential', 'property', 'biodiesel', 'purity', 'parameter', 'depend', 'different', 'operating', 'condition', 'direct', 'measurement', 'hard', 'obtain', 'specific', 'range', 'condition', 'therefore', 'work', 'consider', 'least_square', 'machine', 'svms', 'transform', 'operating', 'condition', 'multi', 'dimensional', 'space', 'simulate', 'biodiesel', 'purity', 'wide', 'range', 'operate', 'condition', 'indeed', 'develop', 'reliable', 'ls', 'svm', 'approach', 'model', 'biodiesel', 'purity', 'function', 'catalyst',...

Applicazione degli Algoritmi di Topic Modelling e Ottimizzazione dei parametri

Una volta effettuata la normalizzazione del testo, è stato possibile costruire la cosiddetta Bag of Words (BoW). La Bag of Words è un modello utilizzato nell'Information Retrieval che non tiene conto dell'ordine delle parole nel testo ma considera solo le occorrenze delle parole; più specificatamente, il modello BoW costruisce una sorta di vocabolario in cui il testo viene tokenizzato in numeri, e in cui ogni numero rappresenta un identificatore univoco di ciascuna parola e ad ogni identificatore è associato un secondo numero che invece rappresenta la frequenza di quella parola nel documento codificato. Il risultato che si ottiene è una lista di tuple in cui il primo numero rappresenta appunto l'identificatore della parola e il secondo il numero di occorrenze se si utilizza il CountVectorizer, oppure la sua TF-IDF se si utilizza il TfidfVectorizer.

Nel nostro progetto è stato utilizzato sia il CountVectorizer che il TfidfVectorizer, poichè, il modello LDA utilizza la word count, mentre NMF e LSA utilizzano la TF-IDF. Generalmente l'utilizzo della TF-IDF è da preferire in quanto il conteggio delle parole porta a considerare molto importanti termini come articoli e congiunzioni che, pur essendo molto frequenti, in realtà risultano poco significative nei vettori codificati. L'alternativa è, appunto, quella di calcolare la frequenza inversa delle parole, e il metodo di gran lunga più popolare è chiamato TF-IDF, che sostanzialmente si adegua al fatto che, come precedentemente detto, alcune parole appaiono più frequentemente in generale ma non per questo sono più significative. TF-IDF è un acronimo che significa "Term Frequency - Inverse Document Frequency" le quali rappresentano le metriche risultanti assegnate a ciascun token. In particolare, la *Term Frequency* riassume la frequenza con cui una determinata parola appare all'interno di un documento, quindi sarà tanto più elevata tanto più il termine è frequente; mentre l'*Inverse Document Frequency* misura la frequenza inversa di una parola in tutti i documenti, ciò significa che l'IDF sarà molto alta nei termini specifici per uno specifico documento e molto bassa nei termini molto comuni.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Di seguito un estratto della BoW creata con la libreria Gensim che utilizza la sola TF. Gensim è una libreria utilizzata nel Natural Language Processing per la Topic Modeling e la Document Indexing.

[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 3), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1), (16, 1), (17, 1), …]

Nel branch **topicModeling_with_ScikitLearn**, invece abbiamo utilizzato i Vectorizer offerti dalla libreria Scikit-learn, la quale rappresenta una libreria professionale utilizzata nel Machine Learning, ovvero CountVectorizer e TfidfVectorizer a cui abbiamo passato come parametri:

- il tokenizer da noi costruito;
- un max_df pari a 0,90, il quale permette di ignorare tutti quei termini che sono presenti in più del 90% degli abstracts;
- un min_df pari a 0,07, il quale permette di ignorare tutti quei termini che sono presenti in meno del 7% degli abstracts, poichè una parola con una bassissima frequenza potrebbe indurre solo rumore nel modello;
- infine, max_features è stato impostato a 1000, e ciò indica che nella costruzione del vocabolario verranno considerate solo le prime 1000 parole in ordine di term frequency.

La scelta di settare il parametro precedente max_features, n_component a 4 e learning_decay a 0.5 dei modelli di Topic modeling è dovuta ai risultati ottenuti con una Gridsearch, i cui risultati sono di seguito mostrati:

```
Choosing Optimal Hyperparameter
Best Score Likelyhood: -190323.750
Best parameters set:
    model__learning_decay: 0.5
    model__n_components: 4
    vect__max_features: 1000
    vect__ngram_range: [1, 2]
```

La Gridsearch rappresenta una metodologia utilizzata nel processo di ottimizzazione dei parametri che permette appunto di individuare i coefficienti ottimali relativi a tutti i possibili iperparametri dell'algoritmo di Machine Learning.
Alla luce delle molteplici prove effettuate è da sottolineare che, nel caso della Topic Modeling da noi effettuata, il numero ottimale di Topics da utilizzare nella stessa è risultato essere 4, valore che verrà utilizzato dunque nei successivi modelli.

Nel progetto alla Gridsearch è stata passata una Pipeline contenente come primo step un CountVectorizer e come secondo step il modello di Topic Modeling, ovvero l'oggetto LatentDirichletAllocation.
Una Pipeline può essere vista come una sorta di contenitore di tutti gli step di processamento che permette di ottimizzare e semplificare notevolmente la scrittura del codice in quanto, una volta costruito questo "contenitore", è lui che si occupa di eseguire tutti gli step di processamento sulla base dei dati consegnatogli; senza l'ausilio della Pipeline dovremmo eseguire ogni step separatamente e poi successivamente mettere assieme i risultati ottenuti nei vari step per ottenere il risultato del modello scelto.

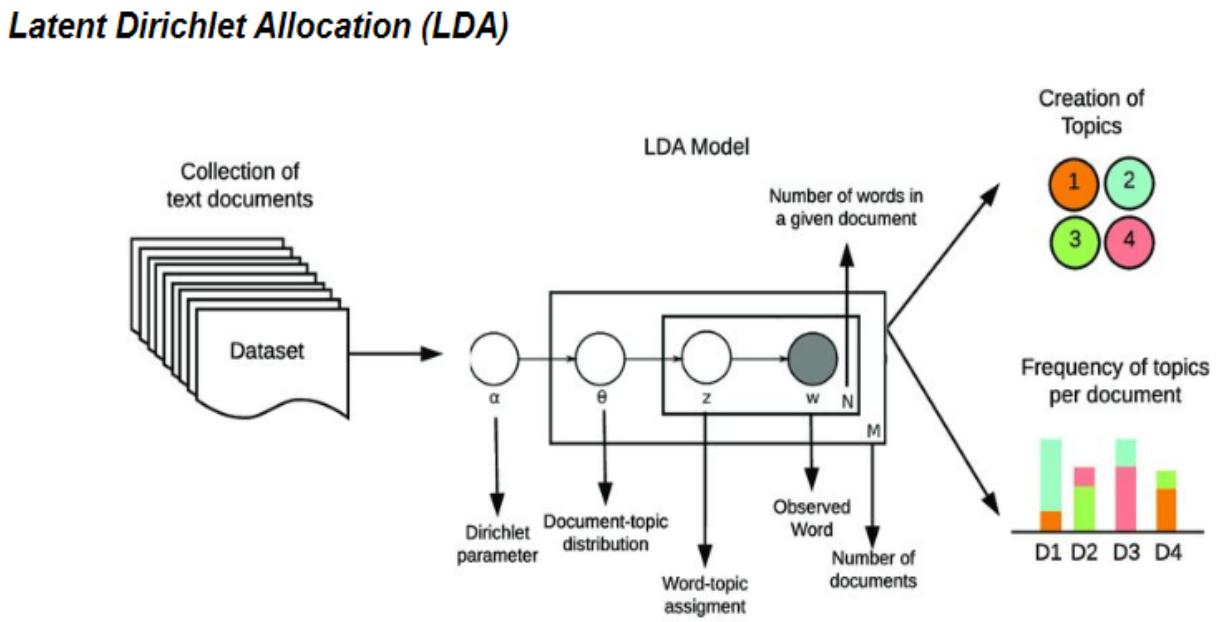
L'impostazione dei parametri,come precedentemente indicato, è stato frutto di molteplici tentativi in cui i parametri min_df e max_df sono stati settati in maniera differente; la scelta è stata fatta andando a cercare un compromesso tra un buon livello dell'indice Likelyhood e un valore dei parametri max_df e min_df che non andasse a scartare troppi termini. Di seguito una piccola tabella riassuntiva di alcuni dei diversi tentativi:

Likelyhood	**Max_df**	**Min_df**	**Commento**
-356610.114	1.0 (default)	1.0 (default)	L'impostazione di default fa si che nessun termine venga scartato, situazione che, non solo non permette di raggiungere un buon livello della Likelyhood, e inoltre, non è da considerare auspicabile a causa della presenza di termini poco significativi.
-4001.290	0.75	0.50	Nonostante la Likelihood sia enormemente migliore, un'impostazione di tali valori non risulta adeguata ad una corretta analisi in quanto i termini scartati sono tantissimi.
-356449.197	0.80	0.25	Nonostante la percentuale dei termini scartati in questo caso possa essere considerata accettabile, il valore assunto dalla Likelyhood rimane abbastanza basso.
-356403.765	0.90	0.20	Situazione pressoché simile alla precedente.
-233751.528	0.90	0.05	Buon compromesso per quanto riguarda termini scartati e livello della Likelyhood.
-212494.162	0.90	0.06	Situazione simile alla precedente.
-190323.750	**0.90**	**0.07**	Settaggio scelto come ottimale in quanto presenta il miglior valore della Likelyhood da noi ottenuto e un buon compromesso con la percentuale dei termini scartati.

Tali parametri sono stati poi utilizzati sia nei Vectorizer che nei modelli di Topic modeling implementati nel progetto.
Come precedentemente indicato, i modelli implementati sono stati la Latent Dirichlet Allocation (LDA), il Non-Negative Matrix Factorization (NMF), la Latent Semantic Analysis o Indexing (LSA o LSI) e, infine, la Probabilistic Latent Semantic Analysis (PLSA).

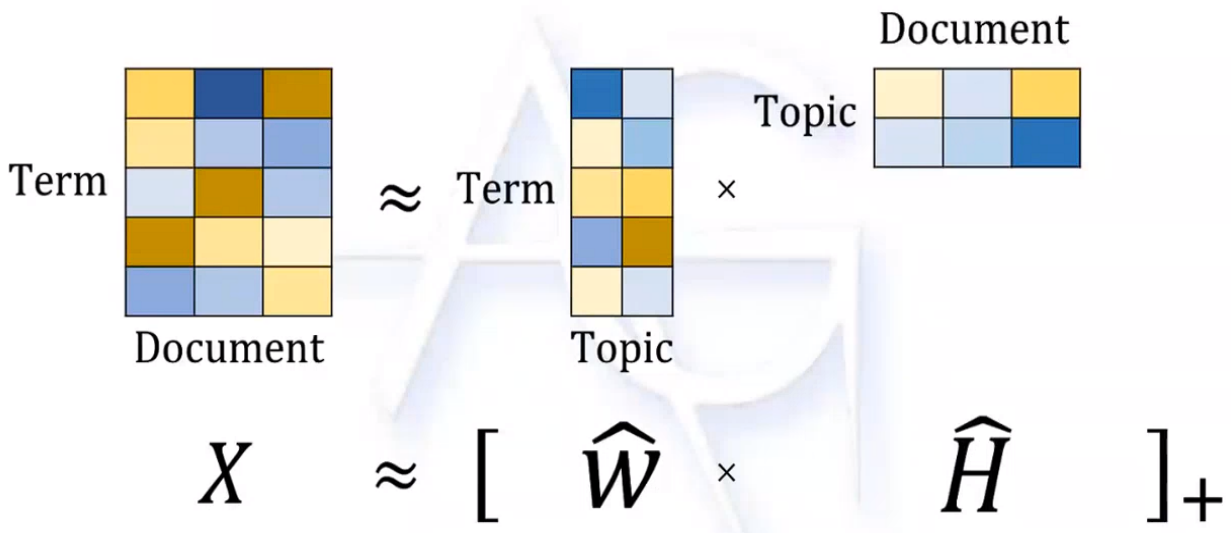
Latent Dirichlet Allocation è un modello probabilistico che consente di estrarre argomenti da un insieme di documenti e si basa sul presupposto che se due termini si trovano spesso in più documenti, probabilmente questi costituiscono il seme di un *topic*. Più precisamente, nel modello LDA ogni documento è considerato come un insieme di parole che, combinate linearmente tra loro, formano uno o più sottoinsiemi di argomenti latenti. In ciascun documento dunque, potranno essere affrontati diversi topics in base alla presenza e alla frequenza delle parole che compongono ciascun topic; ciò significa che, una volta individuati i differenti i topics trattati nel corpus di documenti, è possibile affermare che un particolare documento tratta uno specifico argomento poiché quello è il topic dominante ma non esclusivo nel documento stesso. Quest'ultimo punto è proprio caratteristico dell'LDA, in quanto in questo modello ogni documento viene visto come una mixture di diversi topics, uno dei quali, generalmente, sarà predominante sugli altri.

Di seguito un piccolo grafico che riassume la spiegazione appena esposta.



Il **Non-Negative Matrix Factorization** invece è un modello che si basa sull'algebra lineare e più precisamente sul calcolo matriciale. Più dettagliatamente questo algoritmo prevede la decomposizione della matrice dei dati di partenza, che è una matrice in cui nelle colonne troviamo i termini e nelle righe i vari documenti (nell'immagine sottostante è rappresentata la sua trasposta), in due matrici appunto non negative, che moltiplicate tra di loro restituiscono un'approssimazione della matrice iniziale dei dati.

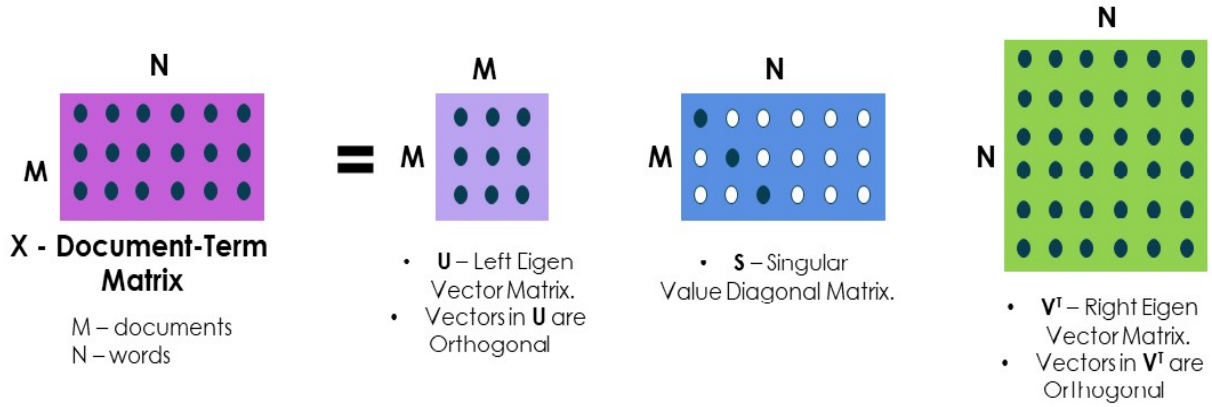
Non-Negative Matrix Factorization (NMF)



Attraverso la scomposizione della matrice di partenza si ottengono quindi delle matrici di dimensione ridotta non negative, una delle quali è una matrice che ci permette di individuare per ogni topics i termini ad esso associati. Successivamente è possibile utilizzare queste informazioni per individuare il topic associato ad ogni singolo documento (seconda matrice).
Il modello NMF rientra tra le tecniche di riduzione della dimensionalità, le quali ci permettono di ottenere le stesse informazioni, in questo caso i topics, ma con un numero inferiore di dati di partenza.

La **Latent Semantic Analysis**, come suggerisce il nome, è una tecnica di analisi semantica che consente di approfondire la conoscenza del contenuto di un documento, oltre ad individuare le relazioni tra i termini che lo compongono.
Anche la LSA parte dal presupposto che parole che hanno un significato simile ricadano in documenti che probabilmente trattano le stesse argomentazioni (ipotesi distributiva).
Nella LSA la ricerca avviene per concetti: ma un concetto non è l'astrazione generalizzazione di un termine (es: golf/vestiario) bensì un insieme di termini correlati (golf, maglia, vestito) detti co-occorrenze o dominio semantico.
I documenti del corpus vengono rappresentati tramite la matrice documenti-termini (le righe rappresentano ciascun documento e le colonne rappresentano parole univoche), successivamente vengono utilizzate tecniche di decomposizione e semplificazione matriciale per ottenere una significativa riduzione delle dimensioni delle matrici di partenza in modo da meglio caratterizzare i documenti contenuti nel corpus.
LSA sfrutta principalmente una tecnica di decomposizione matriciale chiamata Singular Value Decomposition (SVD) che decompone la matrice di partenza in tre matrici; una di queste matrici rappresenta la matrice dei valori singolari del dataset in analisi, dai quali è possibile poi estrarre informazioni circa la similitudine o meno tra i documenti attraverso il calcolo del coseno dell'angolo tra i due vettori (o autovalore) che rappresentano i documenti del corpus.

Latent Semantic Analysis (LSA)

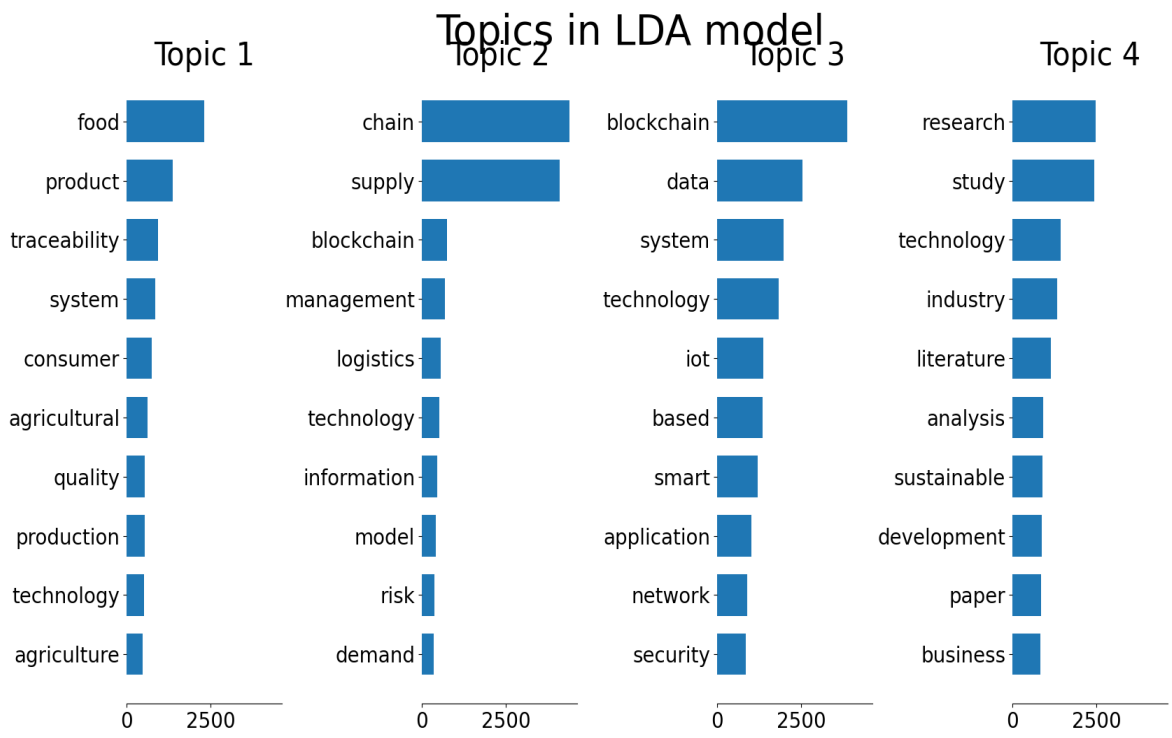


La **Probabilistic Latent Semantic Analysis** deriva da una visione statistica dell'LSA e definisce un proprio modello generativo: invece di utilizzare matrici matematiche e la Singular Value Decomposition per ridurre le dimensioni del problema, utilizza un modello probabilistico secondo il quale è possibile derivare una rappresentazione a dimensionalità ridotta delle variabili osservate in termini di affinità con alcune variabili nascoste considerando sempre le co-occorrenze, proprio come nell'analisi semantica latente.

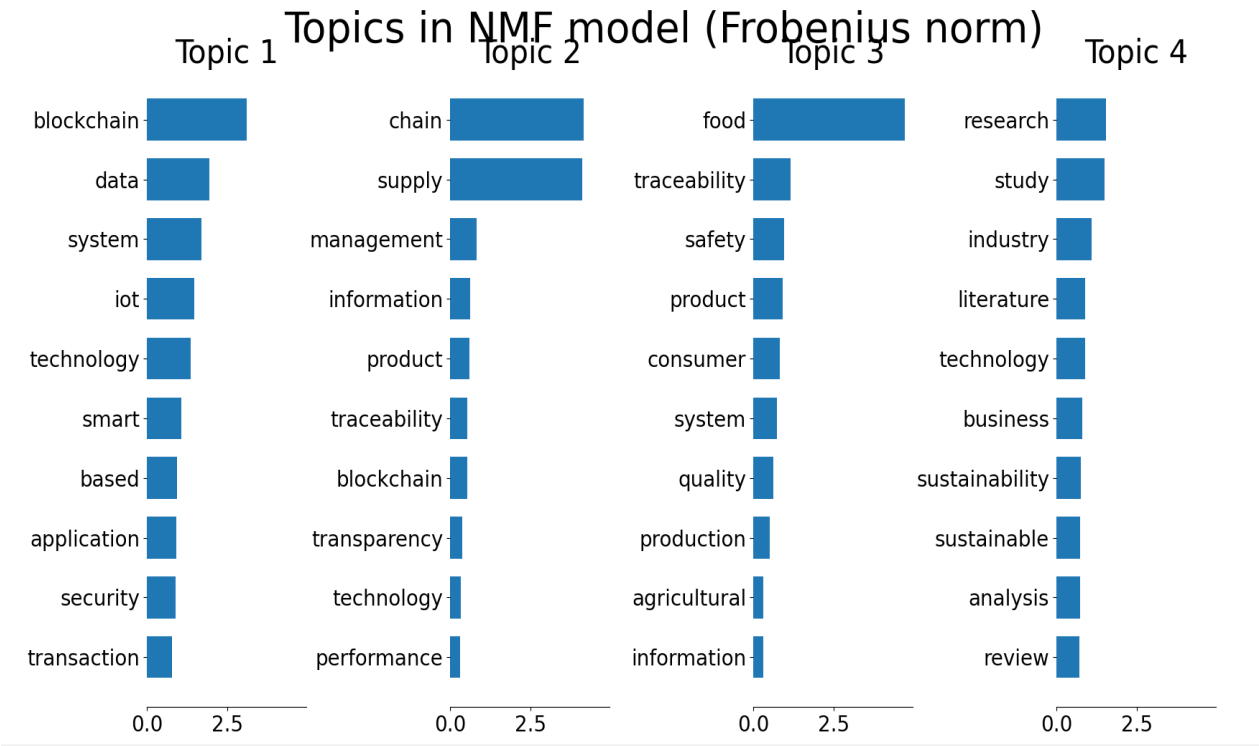
Dal punto di vista computazionale la PLSA è abbastanza pesante, per questo motivo nel progetto da noi proposto è stata utilizzata una sua approssimazione con il modello NMF che utilizza la divergenza di Kullback-Leibler; infatti, è stato dimostrato che i risultati ottenuti con questa variante del modello NMF convergono ai risultati ottenuti con la PLSA.

Di seguito si mostrano i risultati ottenuti con il modello LDA, LSA, NMF, NMF con divergenza di Kullback-Leibler che, come precedentemente esposto, equivale alla PLSA. Tutti e quattro i modelli sono stati eseguiti impostando un numero di topics pari a 4, numero ottimale indicato precedentemente dalla GridSearch.

L'output è un grafico dei Topics, ciascuno rappresentato come grafico a barre dove sono indicate le 10 parole più caratteristiche del Topic stesso, in ordine di importanza in base ai pesi utilizzati (NMF utilizza la TF-IDF mentre LSA e LDA utilizzano solo la TF).
Il codice dei seguenti risultati si trova nel branch **topicModeling_with_ScikitLearn.**



Nel grafico sovrastante possiamo osservare i 4 Topics ottenuti mediante il modello LDA,che sembrano offrire un buon livello di interpretabilità e distinguibilità tra Topics.
-TOPIC 1: Potrebbe far riferimento alla produzione di prodotti agro-alimentari e all'implementazione di un sistema di tracciabilità basato sulle nuove tecnologie con il fine di fornire prodotti di qualità ai consumatori.
- TOPIC 2: Potrebbe far riferimento al miglioramento della Gestione della Logistica della Catena di Distribuzione attraverso l'utilizzo della tecnologia Blockchain.
- TOPIC 3: Potrebbe far riferimento alle nuove applicazioni intelligenti della Tecnologia Blockchain basate sulla raccolta di dati quali la IoT in un'ottica di miglioramento della sicurezza.
- TOPIC 4: Potrebbe far riferimento ai nuovi studi e ricerche sulle Tecnologie applicabili nell'economia industriale, in un'ottica di sviluppo sostenibile.



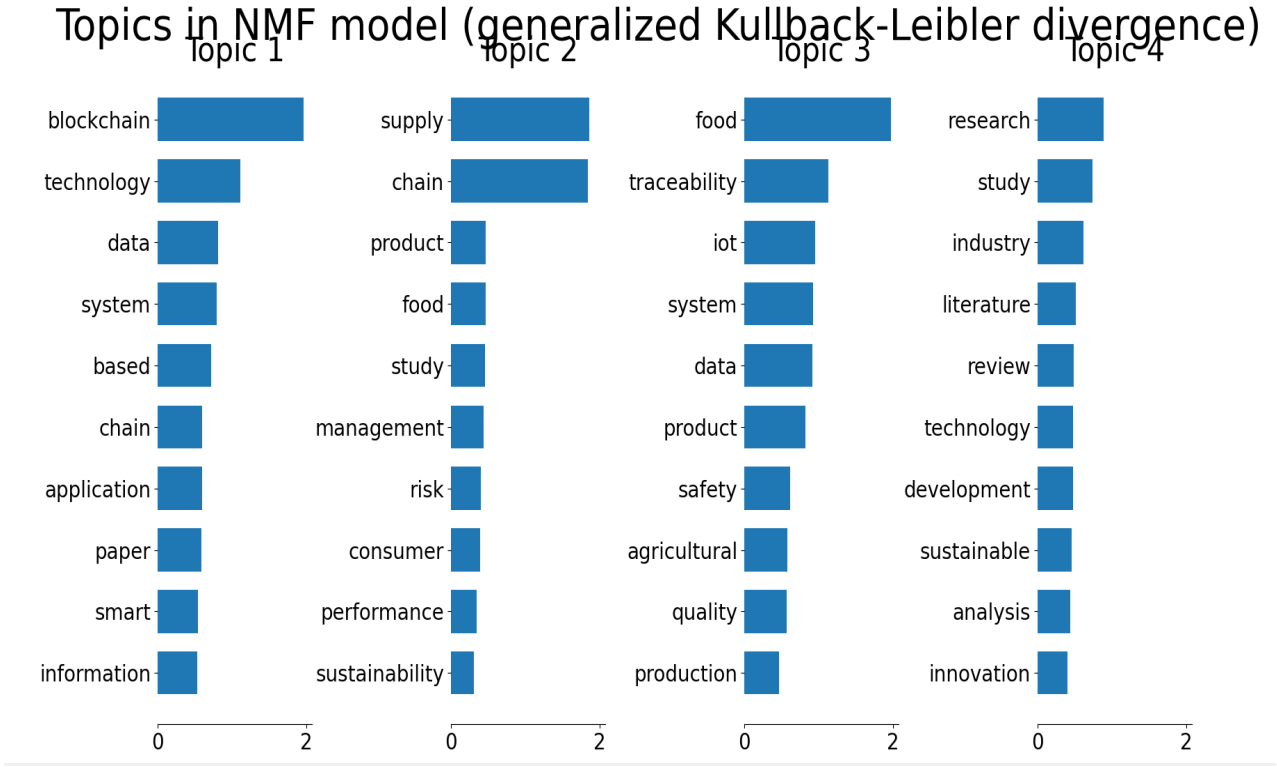
Nel grafico sovrastante possiamo osservare i 4 Topics ottenuti mediante il modello NMF classico ,che sembrano offrire risultati molto simili semanticamente a quelli della LDA anche se con un differente ordine di Topics e frequenze dei termini:

-TOPIC 1: Potrebbe far riferimento alle nuove applicazioni intelligenti della Tecnologia Blockchain basate sulla raccolta di dati quali la IoT in un'ottica di miglioramento della sicurezza delle transazioni.

- TOPIC 2: Potrebbe far riferimento al miglioramento della Gestione della Catena di Distribuzione attraverso l'utilizzo della tracciabilità offerta tecnologia Blockchain al fine della migliore informazione e trasparenza.

- TOPIC 3: Potrebbe far riferimento ad un sistema di tracciabilità dei prodotti agro-alimentari in un'ottica di miglioramento della sicurezza, dell'informazione e della qualità per i clienti.

- TOPIC 4: Potrebbe far riferimento ai nuovi studi e ricerche sulle Tecnologie applicabili nell'economia industriale, in un'ottica di sviluppo sostenibile.



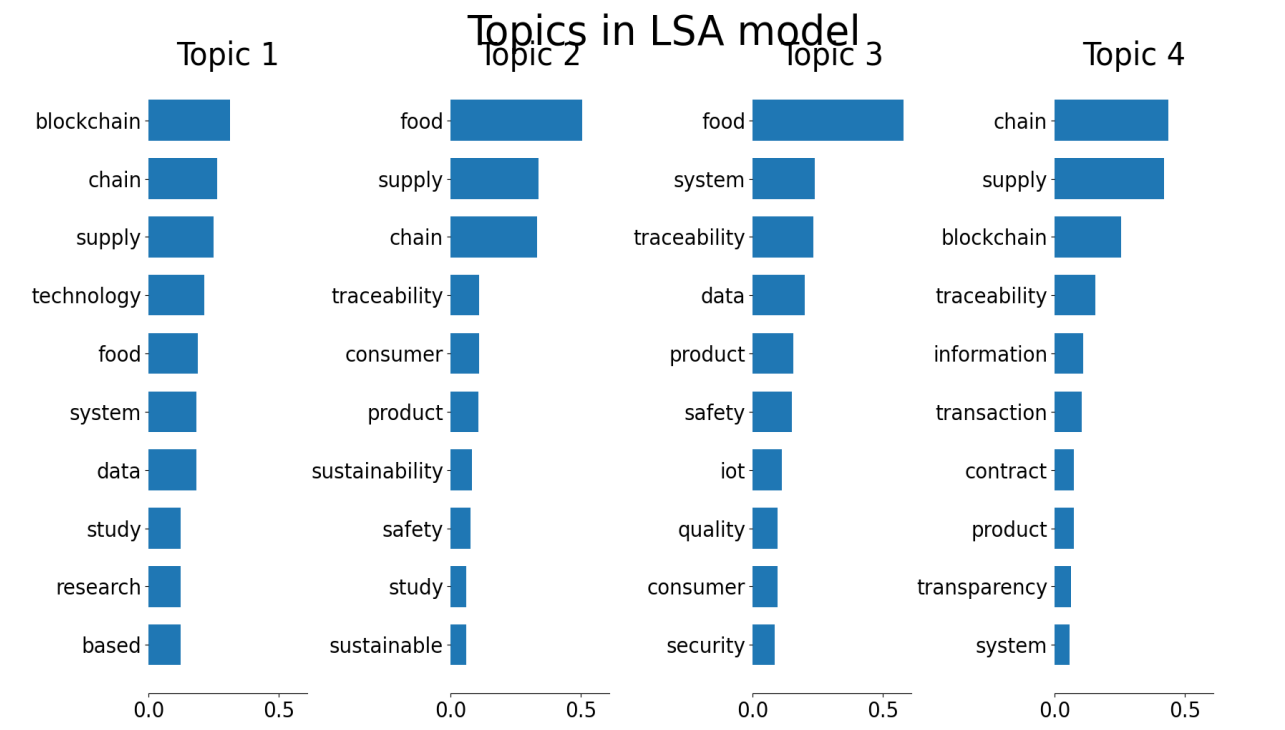
Nel grafico sovrastante possiamo osservare i 4 Topics ottenuti mediante il modello PLSA ottenuto come esposto precedentemente ,che sembrano offrire risultati simili a quelli della NMF classica a livello semantico anche se con minime differenze nelle frequenze e ripetizioni di termini che rendono meno distinguibili i Topics:

-TOPIC 1: Potrebbe far riferimento alle nuove applicazioni intelligenti della Tecnologia Blockchain basate sulla raccolta di dati in un'ottica di miglioramento dell'informazione

- TOPIC 2: Potrebbe far riferimento al miglioramento della Gestione della Catena di Distribuzione dei prodotti alimentari in un'ottica di sostenibilità.

- TOPIC 3: Potrebbe far riferimento ad un sistema di tracciabilità dei prodotti alimentari in un'ottica di miglioramento della sicurezza e della qualità.

- TOPIC 4: Potrebbe far riferimento ai nuovi studi e ricerche sulle Tecnologie applicabili nell'economia industriale, in un'ottica di sviluppo sostenibile.

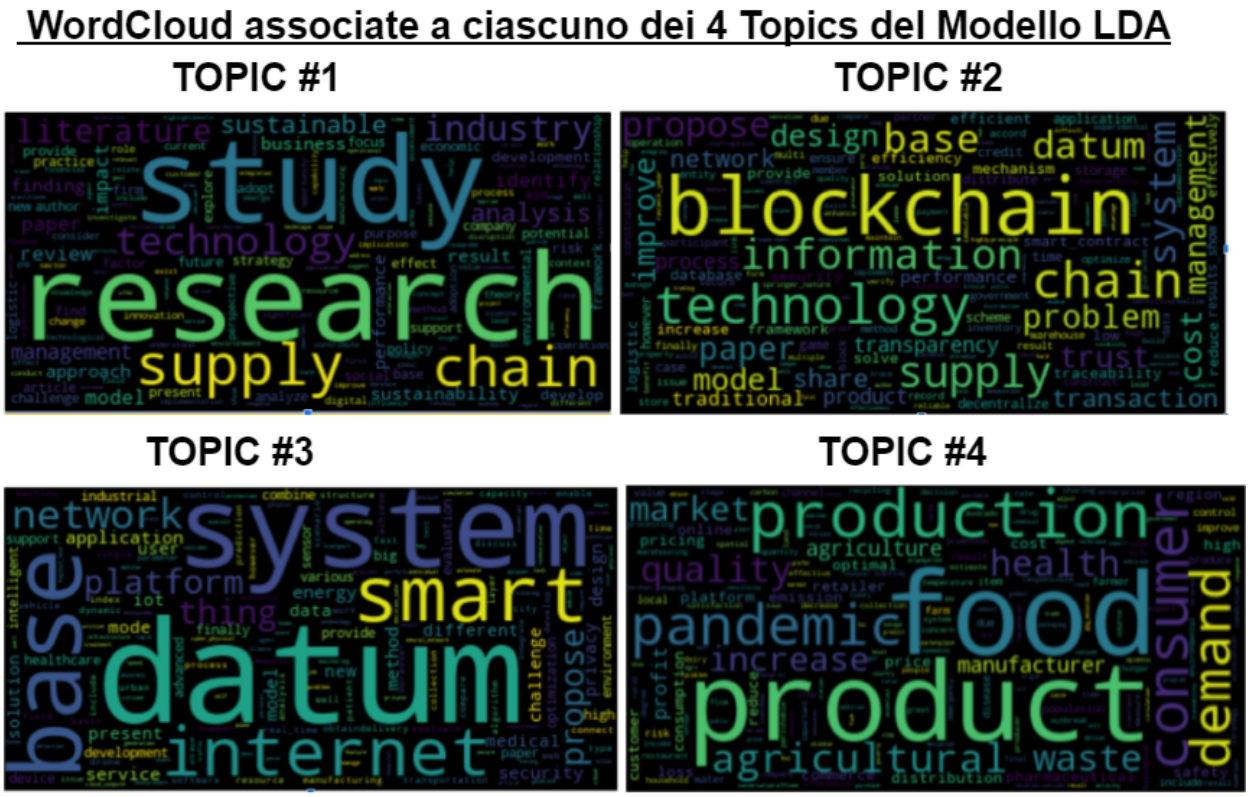


Nel grafico sovrastante possiamo osservare i 4 Topics ottenuti mediante il modello LSA, che per quanto offrano un buon livello di interpretabilità per ciascun Topics, risultano abbastanza scarsi nell'ottica della distinguibilità tra Topics.
Infatti come si può osservare i termini più caratterizzanti sono gli stessi in praticamente tutti i Topic, rendendo molto difficile l'individuazione di differenze tra gli stessi.
-TOPIC 1: Potrebbe far riferimento ai nuovi studi e ricerche basate sulla Tecnologia Blockchain applicabili alla Catena di Distribuzione Alimentare.
- TOPIC 2: Potrebbe far riferimento alla Tracciabilità della Catena di Distribuzione dei prodotti alimentari al fine di migliorare la sicurezza per i consumatori e la sostenibilità ambientale.
- TOPIC 3: Potrebbe far riferimento ad un Sistema di Tracciabilità dei prodotti alimentari al fine del miglioramento della qualità,della sicurezza e protezione forniti ai consumatori.
- TOPIC 4: Potrebbe far riferimento ad un Sistema di Tracciabilità dei prodotti basato sulla Blockchain al fine dell'aumento di informazione e trasparenza rispetto a contratti e transazioni .

Riassumendo le informazioni contenute nei 4 Modelli, possiamo dire che tutti forniscono risultati abbastanza simili, coerenti sotto l'aspetto semantico con l'argomento oggetto di studio e caratterizzati da lievi distinzioni date dall'ordine dei Topics e relative frequenze dei termini. In conclusione a tale analisi possiamo dire che la LDA sembra performare meglio rispetto alle altre in quanto presenta alta interpretabilità per ciascun Topic e una buona distinguibilità tra gli stessi.

Per quanto riguarda invece il branch **topicModeling_with_Gensim**, i modelli proposti sono due, LDA e LSA, entrambi con un numero di topics, anche in questo caso, pari a 4. In questo caso, per visualizzare il contenuto di ciascun topic, abbiamo optato per l'utilizzo della tecnica delle WordCloud, le quali rappresentano i termini secondo un font di grandezza differente in base al peso associato a ciascun termine, peso che in questo caso è rappresentato dalla Term Frequency.

Di seguito vengono mostrate le WordCloud associate ai topics del modello LDA e LSA.



-TOPIC #1: Potrebbe far riferimento ai nuovi studi e ricerche sulle Tecnologie applicabili alla Catena di Distribuzione nell'economia industriale, in un'ottica di sostenibilità.
-TOPIC #2: Potrebbe far riferimento alle nuove proposte di utilizzo della Tecnologia Blockchain rispetto ai problemi di gestione della Catena di Distribuzione, che attraverso un modello caratterizzato dalla trasparenza delle transazioni e condivisione delle informazioni possa migliorare la fiducia dei clienti e ridurre i costi.
-TOPIC #3: Potrebbe far riferimento ai nuovi "Sistemi Intelligenti" dell'IoT basati sull'analisi dei Dati, che attraverso la loro estrazione dalle varie piattaforme internet possono essere applicati ai settori più disparati.
-TOPIC #4: Potrebbe far riferimento alle modalità di produzione dei prodotti agro-alimentari, in un'ottica di aumento dell'attenzione dei consumatori rispetto alla domanda di prodotti di qualità, sani e che riducano gli sprechi.

WordCloud associate a ciascuno dei 4 Topics del Modello LSA

TOPIC #1

TOPIC #2

TOPIC #3

TOPIC #4

-TOPIC #1: Potrebbe far riferimento alla Tecnologia Blockchain per migliorare la Gestione della catena di Distribuzione dei cibi sulla base delle informazioni fornite dai dati.
-TOPIC #2: Potrebbe far riferimento alla Tecnologia Blockchain basata sui dati per migliorare il Sistema di Sicurezza delle Transazioni.
-TOPIC #3: Potrebbe far riferimento alla produzione di prodotti agro-alimentari, in un'ottica di implementazione del Sistema di Tracciabilità dei prodotti che ne garantisca la qualità e la sicurezza per i consumatori.
-TOPIC #4: Potrebbe far riferimento ai nuovi studi e ricerche sulle Tecnologie Innovative applicabili al mercato in un'ottica di sostenibilità ambientale.

Riassumendo le informazioni contenute nelle WordCloud dei 2 modelli, possiamo dire che anche in questo caso entrambe forniscono risultati molto simili, coerenti sotto l’aspetto semantico con l’argomento oggetto di studio e caratterizzati da leggere distinzioni. In conclusione, anche in tale analisi possiamo dire che la LDA sembra performare leggermente meglio rispetto alla LSA in quanto, non presentando ripetizioni di termini chiave nei differenti Topics, mantiene una migliore distinguibilità tra gli stessi.

Volendo effettuare un’analisi più approfondita e un'ulteriore confronto rispetto ai modelli LDA e LSA, si è deciso di implementare due differenti algoritmi per entrambi i modelli che si basano sulla metrica di valutazione detta “Coherence”.
Infatti, nonostante le metriche di Likelihood e Perplexity siano ampiamente utilizzate per la valutazione dei modelli linguistici, studi recenti hanno dimostrato che la tali metriche di valutazione e il giudizio umano spesso non siano correlati, e talvolta anche leggermente anti-correlati. Dunque l'ottimizzazione per la Likelihood(o alternativamente Perplexity) potrebbe non produrre argomenti interpretabili dall'uomo.

Sulla base di tali limiti delle metriche “tradizionali” e nel tentativo di modellare il giudizio umano, la Coherence dei Topics si pone come nuova metrica alternativa.
Le misure di coerenza dei Topics valutano un singolo argomento misurando il grado di somiglianza semantica tra le parole con punteggio elevato all'interno dello stesso. Queste misurazioni aiutano a distinguere tra argomenti che sono argomenti interpretabili semanticamente e argomenti che sono artefatti di inferenza statistica.

Esistono diverse misure di coerenza che si distinguono in base al modo in cui vengono calcolate, ma in tale lavoro utilizzeremo solo quella “C_v” che si basa su una finestra mobile, una segmentazione “one-set” delle parole più caratteristiche e una misura di conferma indiretta che utilizza l'informazione mutua puntuale normalizzata (NPMI) e la “Cosine Similarity”.

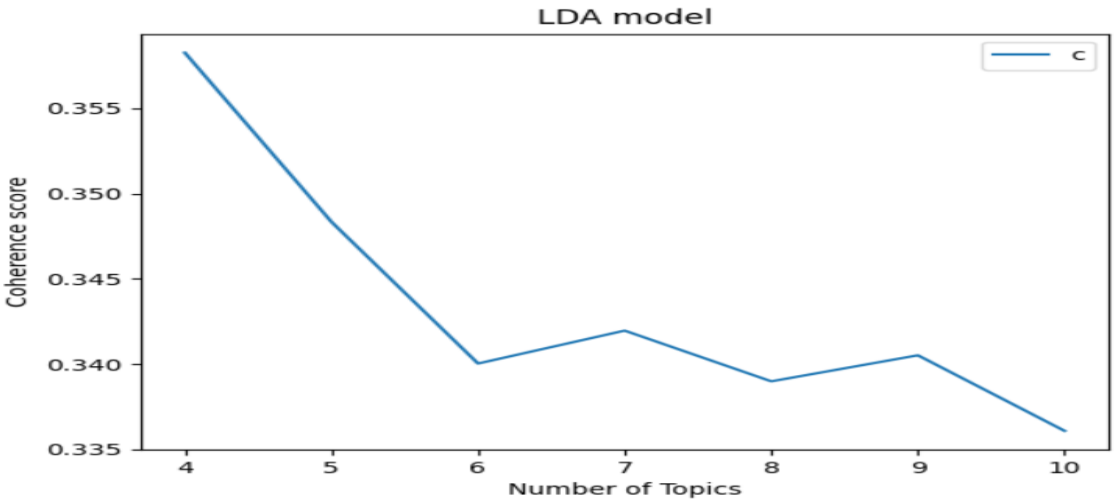
Entrando nel merito del lavoro eseguito, si sono implementati due differenti algoritmi che calcolano la Coherence per entrambi i modelli LDA e LSA.
-Un primo algoritmo fornisce un'unica Coherence in base ai parametri e al numero di Topics prefissati in maniera discrezionale.

```
LDA Coherence Score for 4 Topics: 0.4206258747523106
```

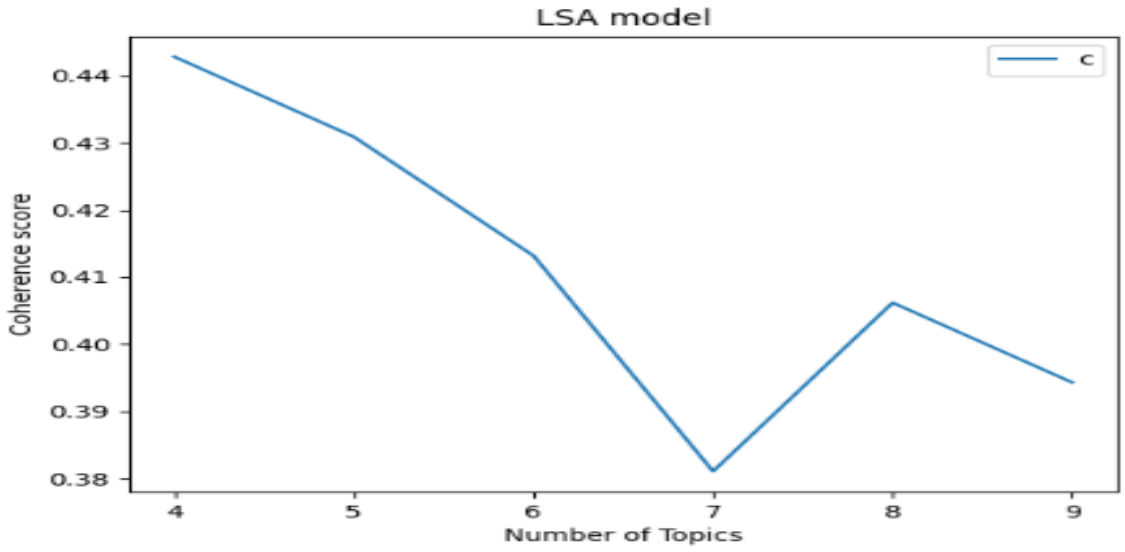
```
Coherence LSA Score for 4 Topics: 0.442853803306906 93
```

Secondo un’analisi generale, essendo entrambi i valori tra il 40% e il 45% di Coherence, possiamo affermare che non siano ottimi. In un’ottica comparativa, differentemente dai modelli visti in precedenza, anche se solo del 2%, il modello LSA sembra performare leggermente meglio rispetto al modello LDA .

-Un secondo algoritmo fornisce più Coherence associate a ciascun numero di Topics dato da a un range di valori da noi indicato, permettendo così una valutazione ulteriore del miglior numero di Topics,ma stavolta in base alla Coherence.
Inoltre lo stesso algoritmo fornisce una visualizzazione grafica che fornisce la stessa informazione in modo molto intuitivo.




```
^^^^ LDA Coherence values for each number of topics:  ^^^^
Num Topics = 4  has Coherence Value of 0.3583
Num Topics = 5  has Coherence Value of 0.3483
Num Topics = 6  has Coherence Value of 0.34
Num Topics = 7  has Coherence Value of 0.342
Num Topics = 8  has Coherence Value of 0.339
Num Topics = 9  has Coherence Value of 0.3405
Num Topics = 10  has Coherence Value of 0.3361
```




```
^^^^ LSA Coherence values for each number of topics:  ^^^^
Num Topics = 4  has Coherence Value of 0.4429
Num Topics = 5  has Coherence Value of 0.4309
Num Topics = 6  has Coherence Value of 0.4132
Num Topics = 7  has Coherence Value of 0.381
Num Topics = 8  has Coherence Value of 0.4062
Num Topics = 9  has Coherence Value of 0.3943
```

Analizzando i risultati forniti dal secondo algoritmo possiamo affermare che:
- Anche la scelta ottimale di Topics attraverso la metrica di valutazione della Coherence, ci conferma i risultati forniti in precedenza, ovvero che il numero ottimale di Topics risulta essere 4, sia per la LDA che per la LSA.
- Anche nel caso della MultiCoherence, i valori della LSA risultano mediamente migliori rispetto a quelli della LDA, con uno scarto ancora maggiore rispetto ai valori forniti dal primo algoritmo.

L'algoritmo basato sulla MultiCoherence fornisce anche la possibilità di ottenere degli output che ci indicano il contributo in percentuale di ciascun termine all'interno di ciascun Topic, secondo il numero di Topics ottimale fornito dall'algoritmo.
Per questioni di visualizzazione vengono esposti solo i primi 10 termini più importanti sia per il modello LDA che per il modello LSA.

```
^^^^  Optimal LDA model's topics: ^^^^
[(0,
  '0.017*"blockchain" + 0.017*"technology" + 0.015*"chain" + 0.014*"supply" + '
  '0.011*"system" + 0.009*"paper" + 0.008*"study" + 0.008*"research" + '
  '0.008*"food" + 0.007*"datum"'),
 (1,
  '0.023*"blockchain" + 0.017*"chain" + 0.016*"supply" + 0.012*"technology" + '
  '0.010*"system" + 0.009*"study" + 0.009*"food" + 0.008*"information" + '
  '0.008*"base" + 0.008*"datum"'),
 (2,
  '0.016*"blockchain" + 0.011*"chain" + 0.010*"food" + 0.010*"supply" + '
  '0.009*"technology" + 0.009*"system" + 0.008*"research" + 0.008*"study" + '
  '0.008*"datum" + 0.006*"base"'),
 (3,
  '0.016*"chain" + 0.015*"technology" + 0.013*"supply" + 0.012*"datum" + '
  '0.012*"system" + 0.012*"blockchain" + 0.008*"food" + 0.008*"base" + '
  '0.007*"model" + 0.007*"propose"')]
```

```
^^^^  Optimal LSA model's topics: ^^^^
[(0,
  '0.387*"chain" + 0.380*"blockchain" + 0.347*"supply" + 0.290*"technology" + '
  '0.198*"system" + 0.177*"food" + 0.167*"datum" + 0.151*"study" + '
  '0.139*"research" + 0.137*"base"'),
 (1,
  '-0.543*"chain" + -0.531*"supply" + 0.433*"blockchain" + 0.207*"technology" '
  '+ 0.191*"system" + 0.191*"datum" + 0.104*"application" + 0.087*"base" + '
  '0.084*"security" + -0.072*"food"'),
 (2,
  '0.706*"food" + -0.430*"blockchain" + 0.267*"system" + 0.172*"traceability" '
  '+ -0.161*"supply" + 0.158*"datum" + -0.150*"technology" + 0.144*"product" + '
  '-0.127*"chain" + 0.108*"safety"'),
 (3,
  '0.358*"study" + 0.357*"research" + -0.307*"blockchain" + -0.224*"system" + '
  '-0.203*"chain" + 0.199*"industry" + -0.195*"traceability" + '
  '0.145*"literature" + -0.137*"supply" + -0.134*"datum"')]
```

Nonostante LSA presenti valori di Coherence maggiori e anche contributi percentuali più alti, si può notare come all'interno di alcuni Topic, siano presenti valori di contributi percentuali negativi, difficilmente interpretabili nel contesto dell'analisi linguistica.

Riassumendo le informazioni forniteci dagli algoritmi basati sulla Coherence ed i relativi grafici, possiamo dire che,innanzitutto abbiamo un' ulteriore conferma sul numero ottimale di Topics, ovvero 4. Inoltre, se osserviamo esclusivamente i livelli di Coherence, a differenza dei modelli precedenti , in questo caso la LSA sembra performare meglio della LDA. Detto ciò, se si va ad osservare nel dettaglio la composizione dei Topics, come detto precedentemente, i valori negativi presenti nella LSA non permettono un'adeguata interpretazione e unendo tale informazioni ai risultati mostrati innanzi, non siamo convinti nell' affermare che LSA sia globalmente migliore della LDA, che invece mantiene un alto livello interpretativo, parametro che riteniamo più importante in questo contesto di analisi.

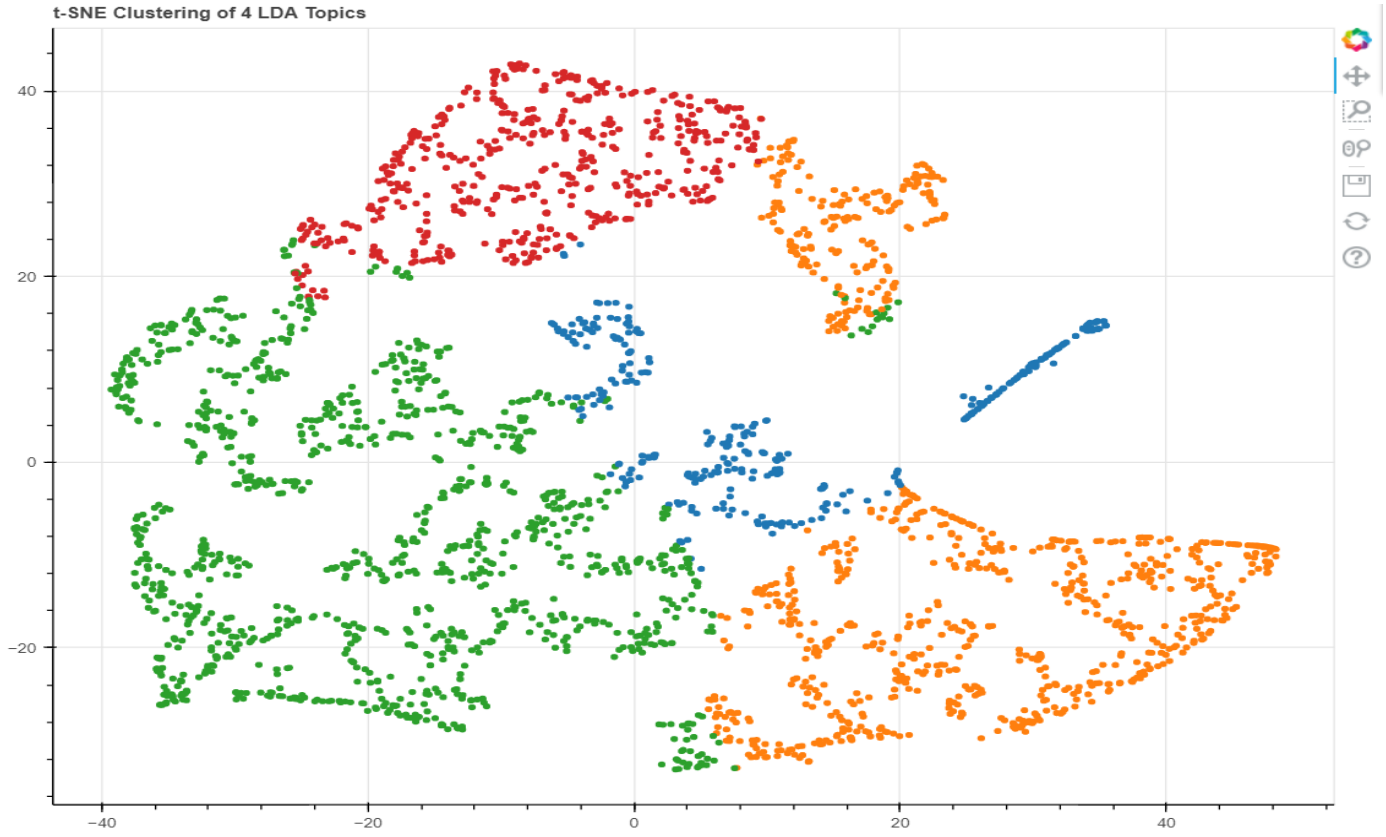
Dunque, sulla base dei risultati globali ottenuti, che indicano come modello ottimale al nostro lavoro quello LDA con 4 Topics, abbiamo voluto approfondire l'analisi di tale modello con alcuni strumenti molto interessanti visti a lezione che permettono la generazione di grafici interattivi sotto forma di pagine HTML, permettendo così un'analisi altamente dinamica e dal contenuto informativo imparagonabile rispetto ai grafici tradizionali. Nel nostro caso specifico utilizzeremo le librerie TSNE e pyLDAvis.

- **TSNE** : Tale sigla sta ad indicare il t-Distributed Stochastic Neighbor Embedding (t-SNE) che è una tecnica in grado di visualizzare i dati ad alta dimensione su uno spazio bi o tridimensionale.
t-SNE è un algoritmo non lineare per la riduzione della dimensionalità in grado di rilevare la presenza di Cluster di diverse dimensioni tramite una funzione che consente di rappresentare punti di dati simili vicini tra loro e, allo stesso tempo, dati diversi lontani tra loro.
Tale risultato è ottenibile convertendo le distanze euclidee tra i punti dati in probabilità condizionali che rappresentano le somiglianze.
Questa tecnica si articola in due fasi principali:
- Una prima fase in cui viene costruita una distribuzione di probabilità che ad ogni coppia di punti nello spazio originale ad alta dimensionalità associa un valore di probabilità elevato se i due punti sono simili, basso se sono dissimili.
-Una seconda fase in cui viene definita una seconda distribuzione di probabilità, analoga alla prima, nello spazio a dimensione ridotta.
L'algoritmo quindi minimizza la divergenza di Kullback-Leibler delle due distribuzioni tramite il gradiente discendente, riorganizzando i punti nello spazio a dimensione ridotta.

Di seguito sono mostrati i risultati che mostrano le caratteristiche con le quali è stato generato il grafico t-SNE nel nostro caso specifico.

```
^^^^ Informazioni relative alla generazione dei Cluster: ^^^^
[t-SNE] Computing 91 nearest neighbors...
[t-SNE] Indexed 2835 samples in 0.000s...
[t-SNE] Computed neighbors for 2835 samples in 0.063s...
[t-SNE] Computed conditional probabilities for sample 1000 / 2835
[t-SNE] Computed conditional probabilities for sample 2000 / 2835
[t-SNE] Computed conditional probabilities for sample 2835 / 2835
[t-SNE] Mean sigma: 0.045727
[t-SNE] KL divergence after 250 iterations with early exaggeration: 65.439018
[t-SNE] KL divergence after 1000 iterations: 0.773206
```

Successivamente è mostrata la visualizzazione del dataset con t-SNE vera e propria, dove è possibile vedere in maniera intuitiva che i punti rappresentanti i documenti sono raggruppati in 4 diversi cluster in base alle loro caratteristiche semantiche.
I diversi colori dei cluster indicano l'assegnazione del documento ad un argomento specifico. Data la quasi ottima distinzione tra Clusters, si può affermare che tale rappresentazione sia buona, e rispetto a quei documenti che si posizionano in un cluster di colore diverso,lo fanno perché probabilmente non hanno un topic dominante molto forte.

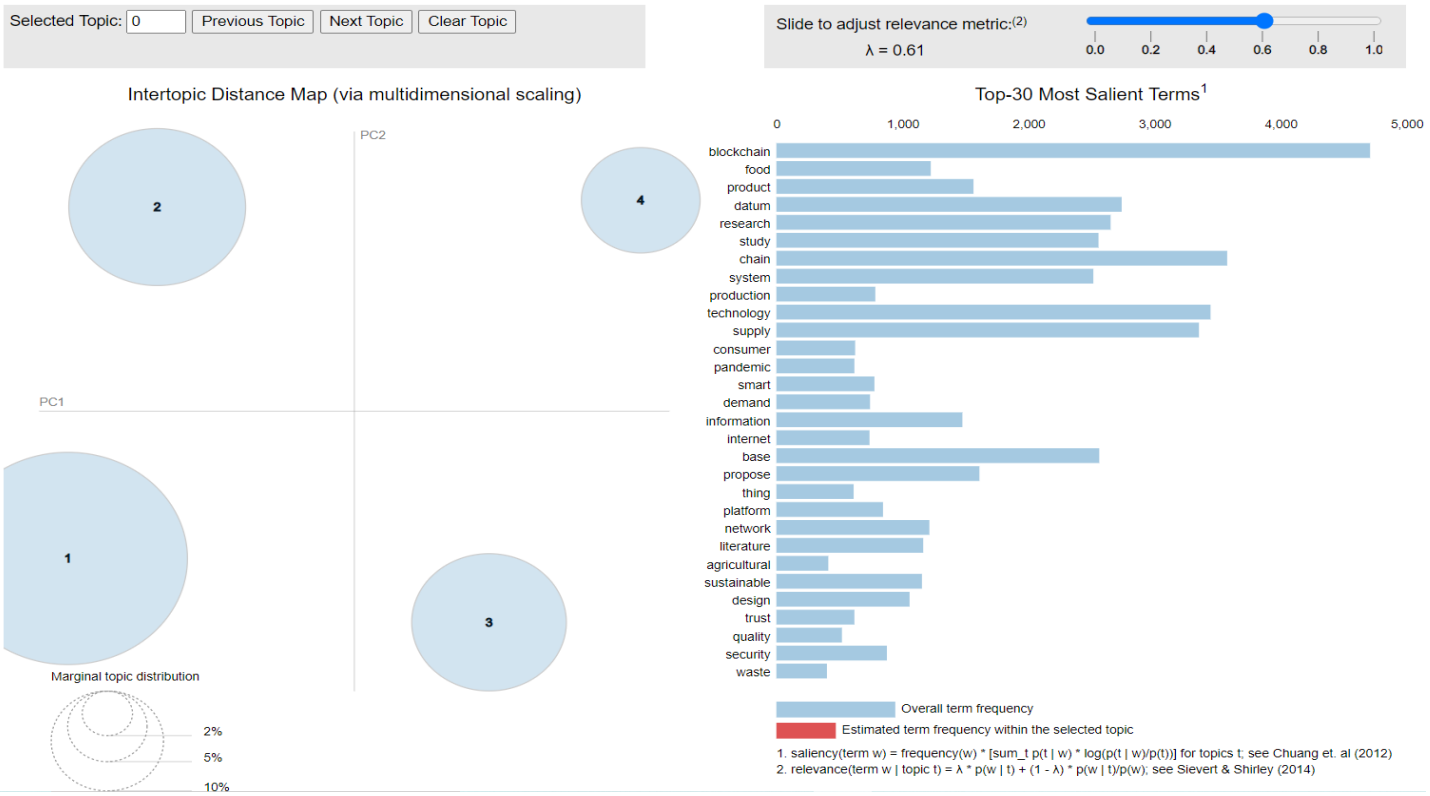


- **pyLDAvis**: grazie a questa libreria viene generato un grafico interattivo che fornisce sia una visione globale degli argomenti e di come differiscono l'uno dall'altro, sia una visione locale consentendo allo stesso tempo un'analisi approfondita dei termini più strettamente associati a ciascun singolo argomento.
La visualizzazione ha due parti fondamentali:

1) Il pannello di sinistra,chiamato Intertopic Distance Map, visualizza gli argomenti come cerchi nel piano bidimensionale e più il cerchio è grande, più la sua frequenza relativa rispetto al Corpus è alta. Gli indici all'interno del cerchio indicano la popolarità di ciascun topics e, in particolare con il numero 1 viene indicato l'argomento più popolare e con il numero 4 l'argomento meno popolare. La poca distanza tra due “bolle” rappresenta la maggior somiglianza tra argomenti e viceversa.
   Tuttavia, questa è solo un'approssimazione della matrice di similarità dell'argomento originale poiché la distribuzione spaziale di tutti e 4 gli argomenti viene riadattata alle due dimensioni in quanto il grafico utilizzato è un grafico bidimensionale.
   Un singolo argomento può essere selezionato per un esame più attento facendo clic sul suo cerchio o inserendo il suo numero nella casella "argomento selezionato" in alto a sinistra.

2) Il pannello di destra raffigura un grafico a barre orizzontali le cui barre rappresentano i 30 termini più utili per interpretare l'argomento attualmente selezionato a sinistra; i termini vengono mostrati seguendo un ordine decrescente di rilevanza. Quando nessun argomento è selezionato nel grafico a sinistra, il grafico a barre mostra i primi 30 termini più "salienti" nel corpus.
   La salienza di un termine è una misura sia di quanto sia frequente il termine nel corpus e sia di quanto sia utile nel distinguere tra diversi argomenti.
   La barra blu rappresenta la frequenza complessiva del termine e la barra rossa indica la frequenza stimata del termine all'interno dell'argomento selezionato.
   Quindi, se una barra appare sia rossa che blu, significa che il termine è presente in più di un topics.

Il cursore λ consente di classificare le parole in base alla pertinenza del termine; si tratta di una metrica regolabile che bilancia la frequenza di un termine in un particolare argomento con la frequenza del termine nell'intero corpus di documenti.
Per impostazione predefinita, i termini di un argomento sono classificati in ordine decrescente in base alla loro probabilità specifica per argomento ($\lambda = 1$).
Quando aggiustiamo la pertinenza con un lambda più basso significa che i termini che sono frequenti in tutti gli argomenti vengono penalizzati, quindi riducendo λ le parole che sono molto frequenti nell’intero corpus assumeranno una rilevanza minore.
Il valore "ottimale" suggerito è di 0,6.

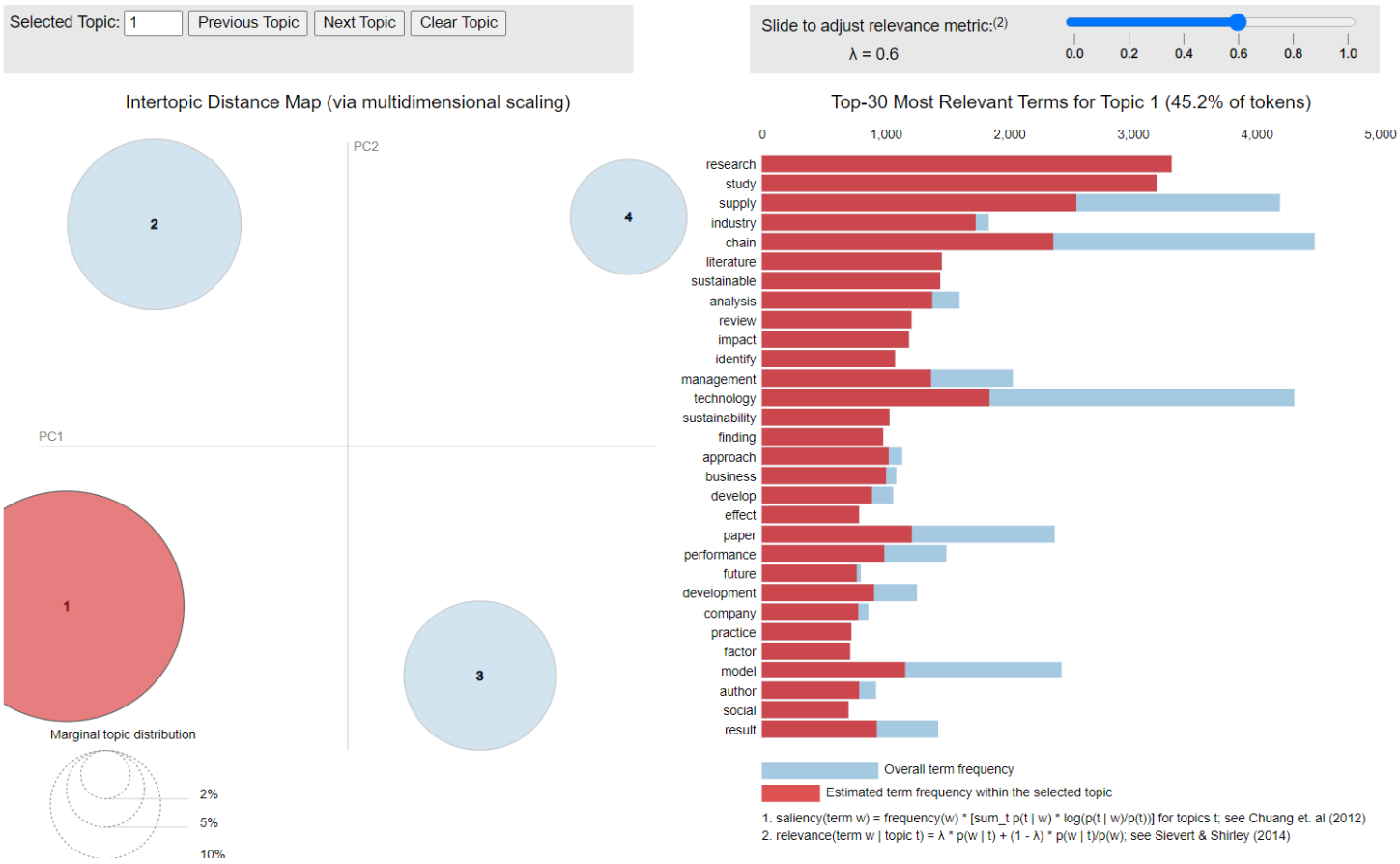


ANALISI GLOBALE:

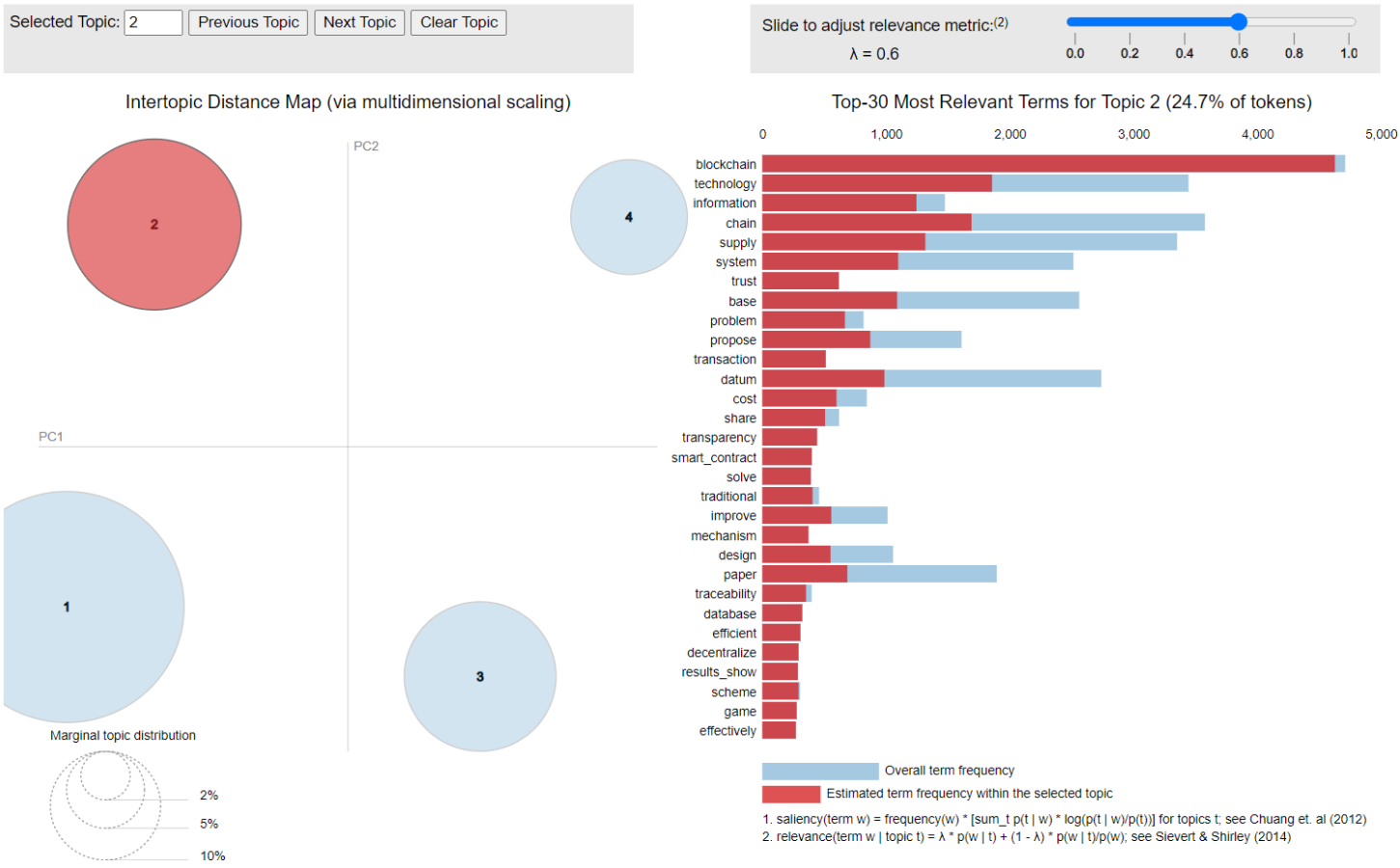
Secondo un'analisi generale rispetto al Corpus delle fonti documentali analizzate, i 30 Termini più rilevanti sembrano indicarci che l'argomento trattato potrebbe essere:

"Studi e Ricerche sull' Utilizzo della Tecnologia Blockchain per migliorare il Sistema della catena di distribuzione dei prodotti alimentari nell'ottica della sicurezza e qualità dei prodotti, lotta agli sprechi, sostenibilità ambientale e aumento di fiducia e livello di informazione dei clienti , attraverso un uso intelligente dei dati, della Tecnologia e dei Network."

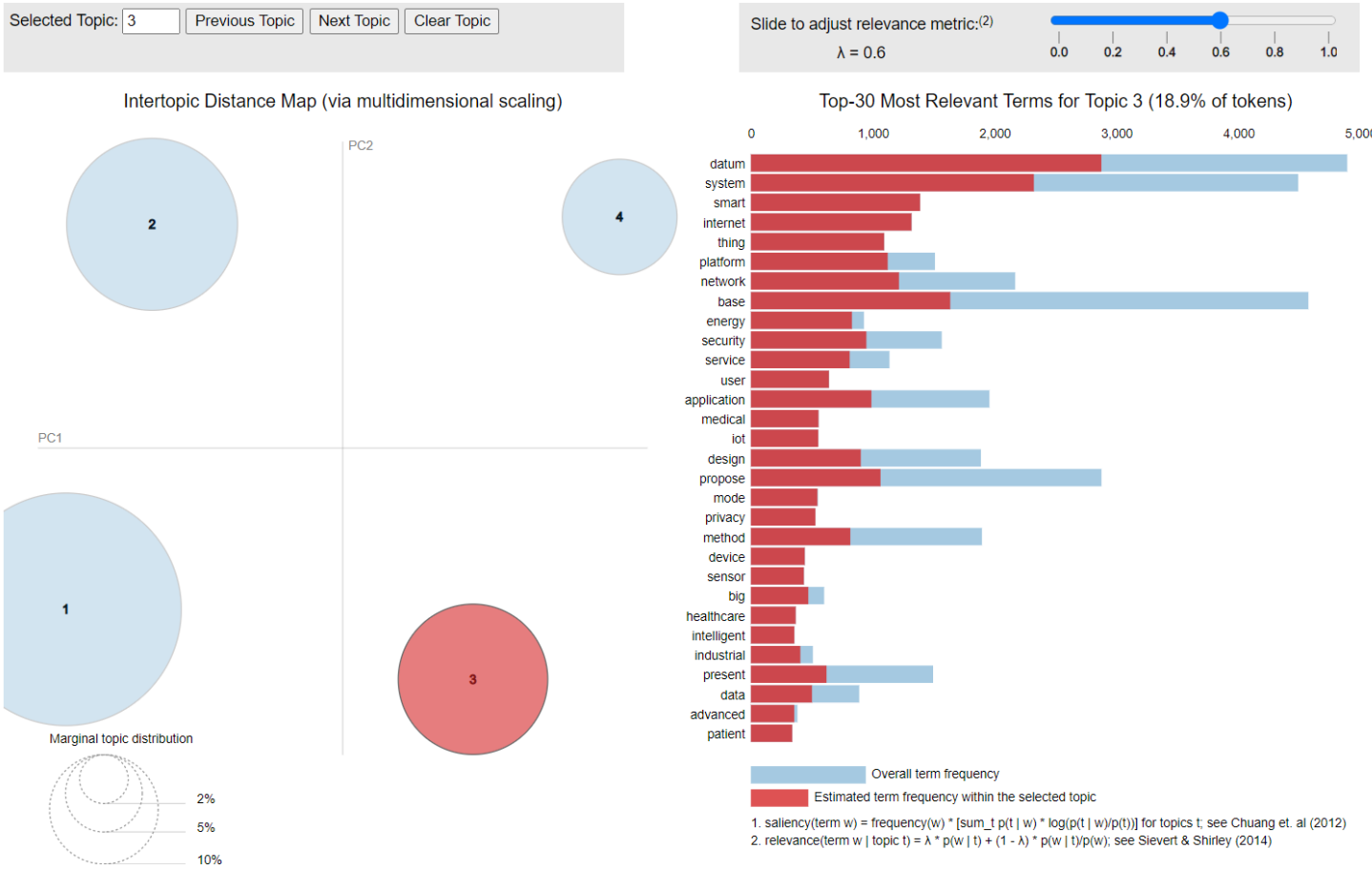
Tale possibile interpretazione sembra essere perfettamente in linea con la base di argomenti oggetto di studio.



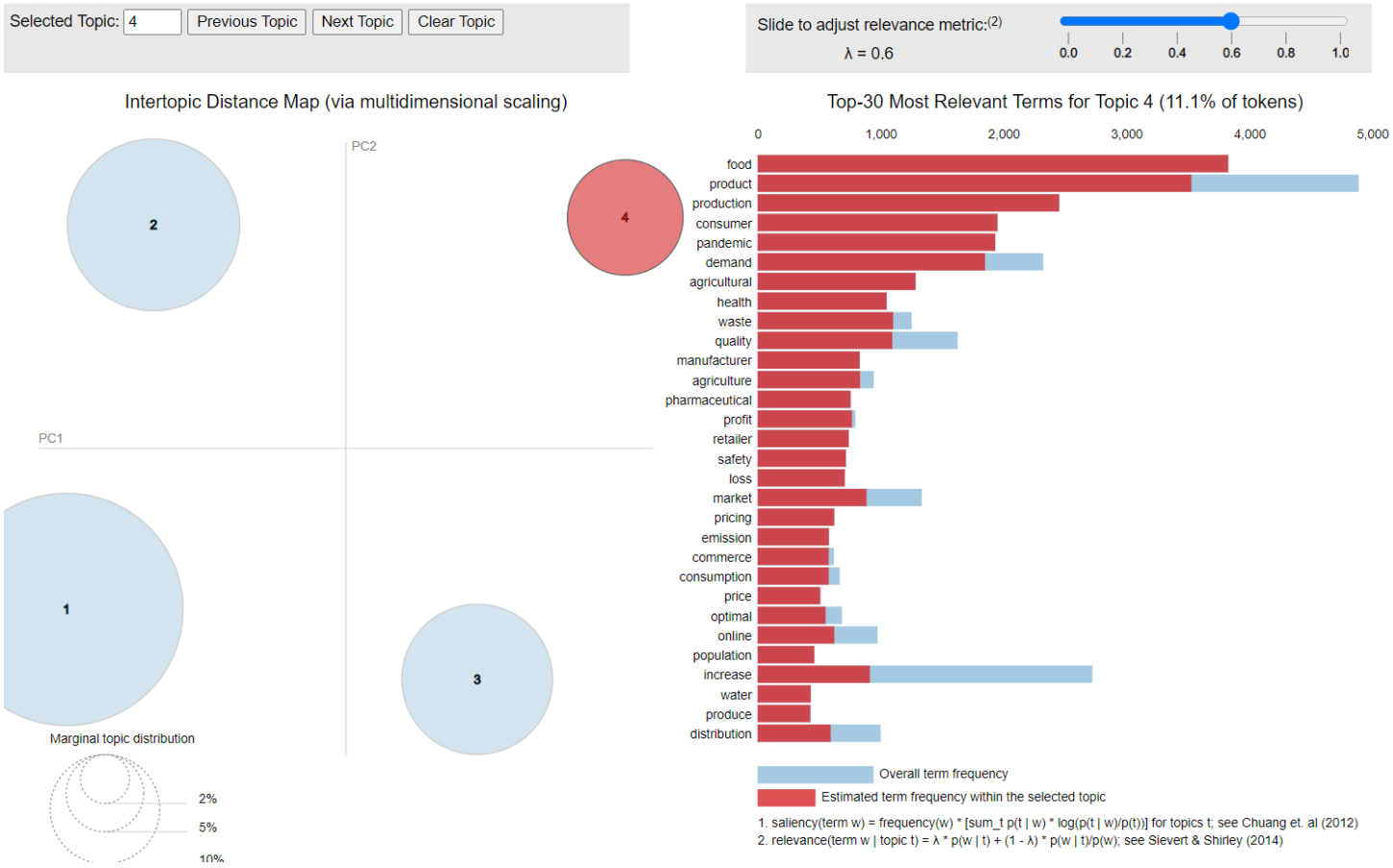
TOPIC #1: Potrebbe far riferimento ai nuovi studi e ricerche sulle Tecnologie applicabili alla Gestione della Catena di Distribuzione nell'economia industriale, in un'ottica di sostenibilità.
Le imprese cercano nuovi modelli di sviluppo e performance che impattino il meno possibile, in grado di garantire quindi una sostenibilità futura.
Si può notare una buonissima corrispondenza con la WordCloud del Topic#1 della LDA.



TOPIC#2: Potrebbe far riferimento alle nuove proposte di utilizzo della Tecnologia Blockchain rispetto ai problemi di Gestione della Catena di Distribuzione, che attraverso un modello caratterizzato dalla trasparenza e tracciabilità delle transazioni e condivisione delle informazioni possa migliorare la fiducia dei clienti e ridurre i costi, portando così ad un sistema più efficiente rispetto a quello tradizionale.
Si può notare una buonissima corrispondenza con la WordCloud del Topic#2 della LDA.



TOPIC#3: Potrebbe far riferimento ai nuovi "Sistemi Intelligenti" dell' IoT basati sull'analisi dei Dati, che attraverso la loro estrazione dalle varie piattaforme internet possono essere applicati ai settori più disparati, specialmente in ambito industriale, fornendo risparmi di energia e in quello sanitario, migliorando notevolmente la salute delle persone.
Si può notare una buonissima corrispondenza con la WordCloud del Topic#3 della LDA.



TOPIC#4: Potrebbe far riferimento alla produzione di prodotti agro-alimentari, in un'ottica di aumento dell'attenzione dei consumatori rispetto alla domanda di prodotti di qualità, sani e che riducano gli sprechi.
Anche nella prospettiva di un'aumento della popolazione mondiale, si cercano nuovi metodi di produzione per il mercato di massa che riescano comunque a limitare i problemi legati ai consumi (vedi acqua ed emissione gas).
Si può notare una buonissima corrispondenza con la WordCloud del Topic#4 della LDA.

Riassumendo le informazioni fornite dai Grafici interattivi t-SNE e da quelli pyLDAvis, possiamo dire che gli stessi confermano come migliore il modello la LDA.
Infatti grazie all'alto contenuto informativo fornito dai grafici interattivi, è stato possibile non solo trovare delle corrispondenze con i contenuti semantici riscontrati nelle precedenti analisi LDA basate su differenti algoritmi, ma bensì è stato possibile ampliare il significato fornito da tali analisi, mantenendo un alto livello di coerenza semantica con le interpretazioni precedenti. Tale risultato si è ottenuto sia a livello globale in riferimento all'intero Corpus di Documenti, sia a livello locale in riferimento ai singoli Topics.

Analisi temporale dei Topics negli ultimi 5 anni.

L'obiettivo dell'analisi storica sotto riportata è stato quello di mettere in evidenza eventuali trend nelle argomentazioni trattate nei 2960 articoli raccolti nel periodo compreso tra il 2017 e il 2021. I topics di seguito elencati sono il risultato dell' applicazione del modello Latent Dirichlet Allocation presente nella libreria Scikit-learn.

```
 ######## HISTORICAL TREND 2017-2021 ######

######    TOPICS LDA  2017    ######
Topic 1:
chain technology blockchain supply system management data research information quality
Topic 2:
chain product supply food agricultural information traceability system based technology
Topic 3:
technology blockchain iot chain application system business supply service research
Topic 4:
data blockchain storage product logistics system ledger store chain based
```

```
######    TOPICS LDA  2018    ######
Topic 1:
food chain supply data system blockchain technology product based traceability
Topic 2:
blockchain technology business smart industry service application based paper iot
Topic 3:
blockchain system chain supply capability data distribution management service transaction
Topic 4:
blockchain research data technology new paper management information area emerging
```

```
######    TOPICS LDA  2019    ######
Topic 1:
food system production data information traceability product agricultural method based
Topic 2:
data iot security internet thing blockchain device based application sensor
Topic 3:
blockchain chain supply system technology product data based transaction information
Topic 4:
technology research blockchain chain supply study paper industry business food
```

```
######    TOPICS LDA  2020    ######
Topic 1:
food chain supply product traceability consumer quality system safety risk
Topic 2:
data system iot based smart blockchain information proposed internet chain
Topic 3:
chain study supply research industry technology paper business literature model
Topic 4:
blockchain technology application paper management system transaction data security trust
```

```
######    TOPICS LDA  2021    ######
Topic 1:
research study industry analysis management author model performance manufacturing literature
Topic 2:
chain supply blockchain technology management information study logistics based paper
Topic 3:
blockchain data system technology based iot smart food application security
Topic 4:
study food research covid sustainability approach finding social pandemic literature
```

Sulla base dei risultati forniti dalla Topic Modeling per anno, possiamo affermare che nel 2017 i Documenti analizzati trattavano il tema della Tecnologia Blockchain rispetto alla Catena di Distribuzione e alle sue caratteristiche informative, senza però una decisiva prevalenza di tale tema legato alla Filiera Agro-Alimentare e approcci Innovativi quali l'IoT.
Probabilmente perché, seppur si intravedevano potenzialità in tale tecnologia applicata al Agri-Food, l'argomento era ancora "prematuro" perché vi fosse una preminenza nei documenti pubblicati.

A favore di tale interpretazione possiamo vedere che dal 2018 al 2020, il tema della Tracciabilità dei Prodotti Alimentari nella Catena di Distribuzione attraverso l'utilizzo della Tecnologia Blockchain, basata su un approccio come l'IoT,nell'ottica di garantire la qualità dei prodotti,la sicurezza dei consumatori e una produzione sostenibile prende sempre più piede e diventa argomento centrale dei Documenti pubblicati.
Tale trend è interpretabile probabilmente con la maggior consapevolezza globale acquisita negli anni rispetto all'effettiva efficacia di tali applicazioni e dei conseguenti miglioramenti in termini di sviluppo,con conseguente aumento degli studi e pubblicazioni su tale tematica.

Interessante l'analisi del 2021, in quanto il precedente Trend sembra arrestarsi, infatti l'argomento della Blockchain in relazione alla tracciabilità dei prodotti alimentari sembra perdere di centralità a favore di una sua più generale applicazione a livello di studi e ricerche per il miglioramento delle performance nel settore industriale e manifatturiero.

Riguardo a tale punto, un'interpretazione possibile potrebbe essere quella secondo cui dopo un grande fermento e generazione di applicazioni degli anni precedenti, si stiano cercando nuovi modelli applicativi per la Blockchain sulla base di nuovi studi e ricerche.

Particolarmente interessante risulta essere come in tale anno, a seguito della Pandemia mondiale da Sars-Cov19 scoppiata nel 2020,nel Topic 4 siano presenti termini come "pandemia" e "covid".
Una possibile chiave di lettura potrebbe essere connessa all'inversione di tendenza nel comportamento dei consumatori causata dalla pandemia. Più precisamente, diversi studi hanno confermato la tendenza nata durante i diversi Lockdown, nei quali il consumatore, spinto da una ragione di forza maggiore ad acquistare prodotti alimentari dai piccoli rivenditori sottocasa, abbia favorito indubbiamente,in maniera consapevole o meno,ad alimentare quel sistema di produzione sostenibile e distribuzione di prodotti di qualità che si vuole valorizzare attraverso l'utilizzo della Blockchain applicata al Settore Agro-Alimentare.
Una seconda interpretazione potrebbe essere legata al fatto,che,visto l'aumento dell'attenzione rispetto a temi quali salute e ambiente provocato da una catastrofe di tali dimensioni,molte persone siano state spinte a rivalutare le proprie abitudini,tra cui quelle alimentari, in termini di miglioramento dei prodotti consumati e di una produzione sostenibile degli stessi.

Detto ciò, una precisazione risulta doverosa: l’individuazione dei trend sopra elencati è stato possibile considerando il 1° Topic come maggiormente diffuso all’interno del corpus documenti, in quanto se si dovessero considerare con egual peso tutti i 4 Topics anno per anno l’individuazione di eventuali trend risulta essere più difficoltosa.

La spiegazione di questo fatto potrebbe, a nostro avviso, essere ricondotta al fatto che l’analisi temporale è stata fatta su un corpus di documenti già precedentemente filtrato attraverso le Query iniziali, che tratta quindi un argomento molto specifico.
Inoltre l'analisi su un arco temporale considerabile abbastanza breve, soprattutto in relazione ad un tema così innovativo come l’applicazione della tecnologia Blockchain nel settore dell’agri-food, non può che portare ad una decisamente circoscritta vastità di differenziazione nel contenuto degli articoli scientifici.

RISULTATI

Giunti alla fine della nostra analisi, riportiamo brevemente i risultati ottenuti dalla nostra ricerca riguardante la Topic Modeling effettuata su fonti documentali riguardanti le applicazioni della Tecnologia Blockchain nel Settore dell'Agri-Food.

- Nonostante le difficoltà affrontate nella fase iniziale di Web Scraping, probabilmente date dalle caratteristiche della fonte utilizzata, siamo riusciti a trovare un'alternativa che ci permettesse di ottenere un Set di Dati coerente con l'oggetto di studio e adatto all'implementazione delle analisi successive che ci eravamo prefissati.
- Siamo riusciti a implementare adeguatamente gli algoritmi di pre-processing del contenuto testuale del Dataset sulla base delle conoscenze ottenute.
- Abbiamo implementato più algoritmi Topic Modeling sul Dataset basati sui principali approcci LDA,LSA,PLSA e NMF ed effettuato un'analisi comparativa tra gli stessi tramite l'uso delle principali metriche di valutazione conosciute e l'uso di strumenti di visualizzazione grafica.

In tal senso, basandosi sui risultati ottenuti, la LDA è risultata il modello più performante in termini di livello interpretabilità del significato espresso dai Topics individuati e in termini di distinguibilità tra i Topics stessi.

- Siamo riusciti a identificare gli argomenti trattati nelle fonti documentali pubblicate su Scopus dal 2017 al 2021 sia sotto l'aspetto Globale sia sotto l'aspetto Temporale, ottenendo in entrambi i casi un buon grado di coerenza e interpretabilità.

Attraverso l'analisi Temporale è stato possibile fornire anche un interpretazione dei trend dei Topics che si sono susseguiti anno per anno, fornendo così un idea di come si sono evoluti gli argomenti trattati nelle pubblicazioni riguardanti le applicazioni della Tecnologia Blockchain nel Settore dell'Agri-Food.

CONCLUSIONE

In conclusione,grazie ai risultati ottenuti, possiamo affermare che un’analisi basata sulla Topic Modeling come quella da noi effettuata, possa avere un carattere informativo enormemente più ampio e soddisfacente se applicata a un DataSet più vasto che tratti un argomento non troppo specifico, giacché scopo della Topic Modeling stessa risulta essere quello di trovare argomenti che possano racchiudere grandi quantità di dati variegati in insiemi ridotti.
Come espresso in precedenza infatti, crediamo che il nostro DataSet, generato già a partire da un filtraggio di query e riguardante quindi un argomento molto specifico oltre che nuovo, risulti penalizzato nell'individuazione di Topics distinguibili globalmente e anche nell'individuazione di Trend temporali.
Nonostante ciò,possiamo affermare di aver raggiunto tutti gli obiettivi prefissati e che la nostra Analisi riesca a fornire comunque un significativo contributo informativo sulla base dell'argomento oggetto di studio.