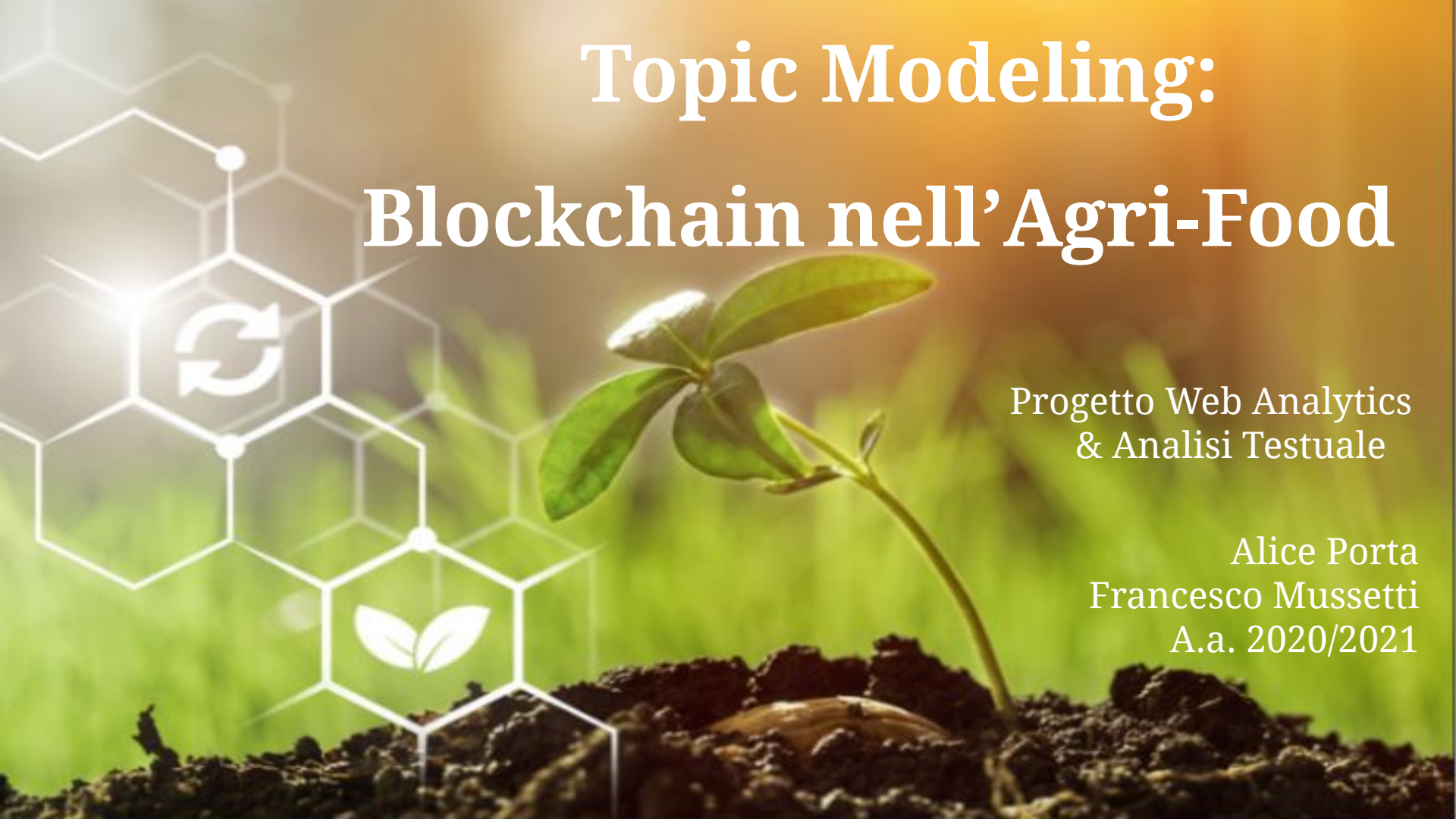


Topic Modeling:

Blockchain nell'Agri-Food

Progetto Web Analytics
& Analisi Testuale

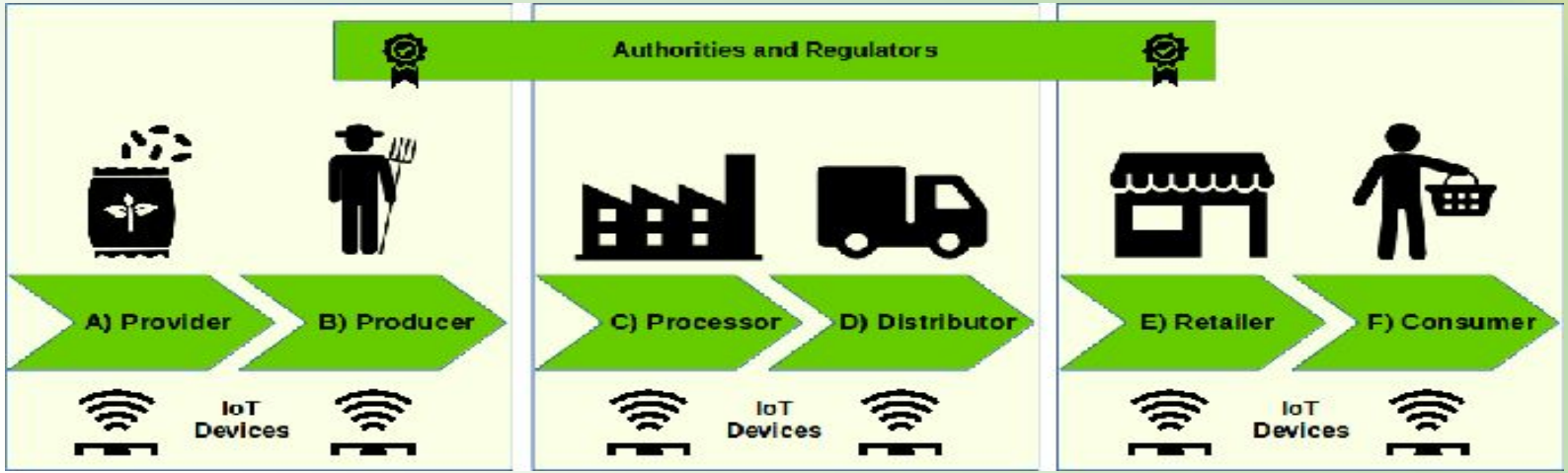
Alice Porta
Francesco Mussetti
A.a. 2020/2021

INDICE

1. Definizione delle applicazioni Blockchain nel settore Agri-Food.
2. Definizione Topic Modeling.
3. Step operativi implementati:
    i) Web scraping
    ii) Normalization
    iii) Applicazione algoritmi di Topic Modeling
    iv) Confronto risultati ottenuti
    v) Analisi temporale
4. Risultati e Conclusioni

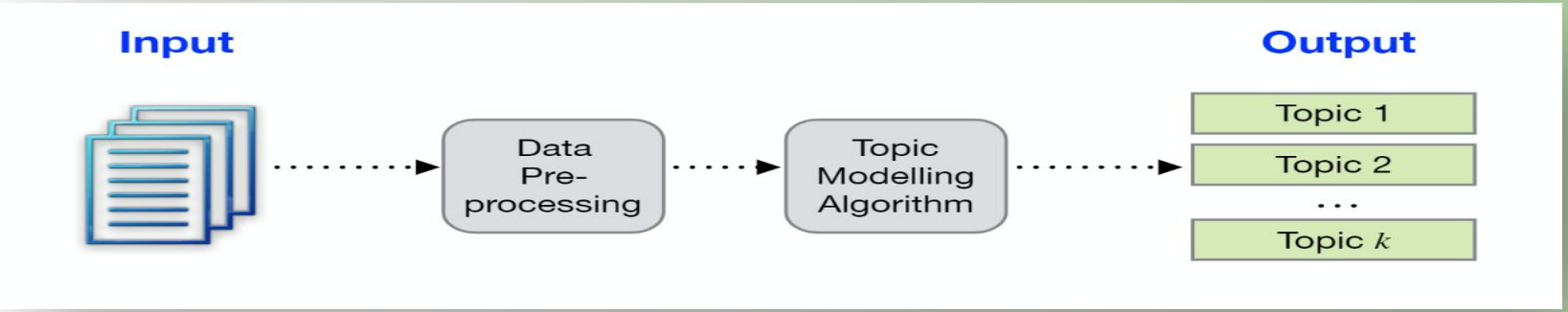
APPLICAZIONI BLOCKCHAIN NEL SETTORE DELL' AGRI-FOOD

- General Purpose Technology
- Tracciabilità
- Garantire genuinità e qualità dei prodotti



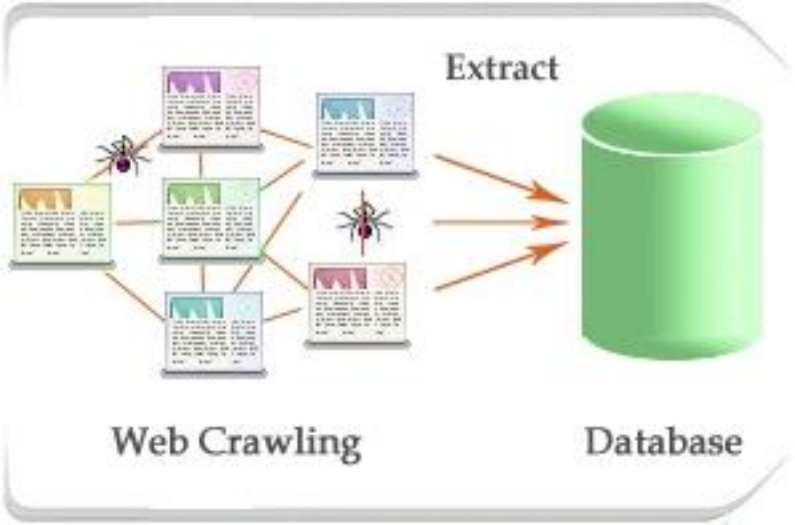
TOPIC MODELING

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 un corpus di documenti.



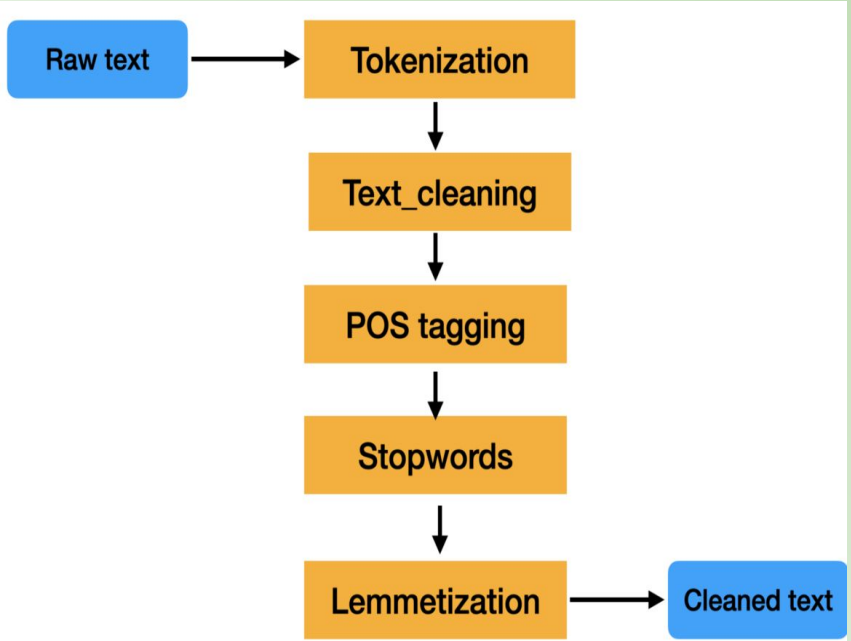
WEB SCRAPING

Il **web scraping** è un processo **automatizzato** che permette l'estrazione di contenuti e **dati** da un sito web.



Scopus

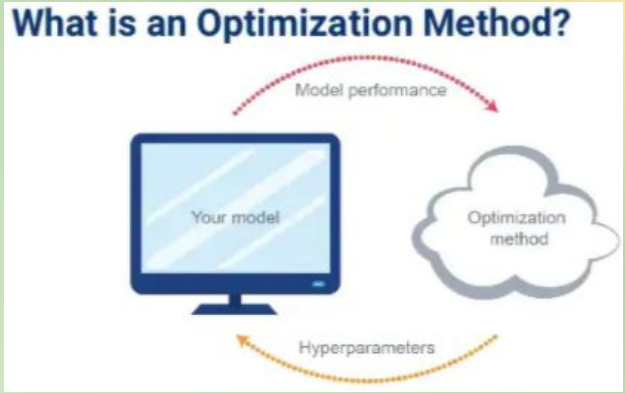
NORMALIZZAZIONE DEI DATI TESTUALI



RISULTATO 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',..

TOPIC MODELING: BoW e OTTIMIZZAZIONE



Likelyhood	Max_df	Min_df
-356610.114	1.0 (default)	1.0 (default)
-4001.290	0.75	0.50
-356449.197	0.80	0.25
-356403.765	0.90	0.20
-233751.528	0.90	0.05
-212494.162	0.90	0.06
-190323.750	**0.90**	**0.07**

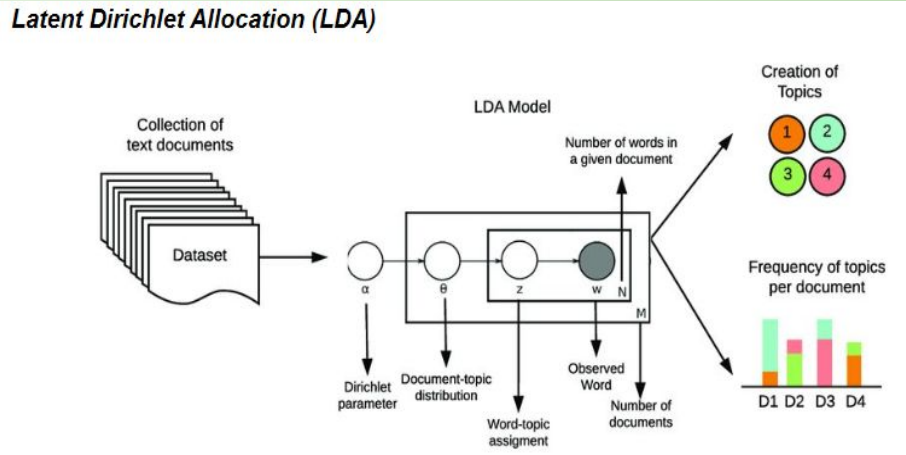
RISULTATO OTTIMO:

```
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]
```

RISULTATO BoW:

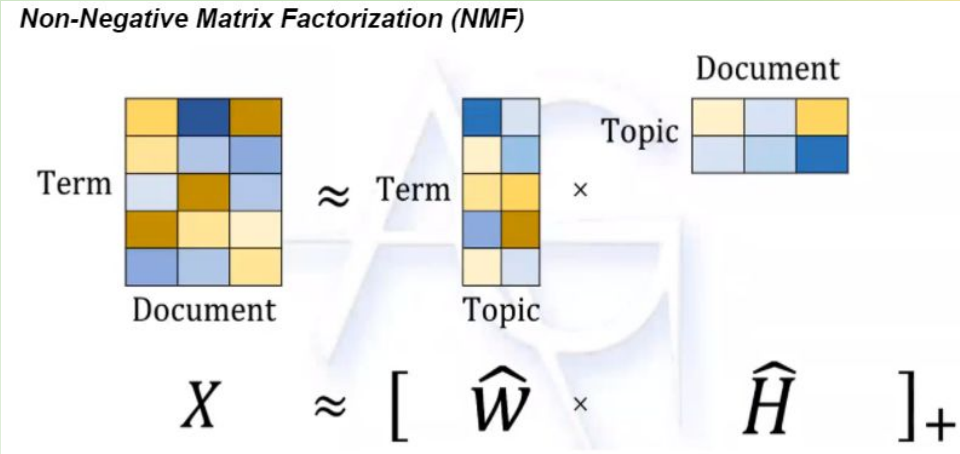
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 3), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1), (16, 1), (17, 1), …]

TOPIC MODELING: LDA & NMF



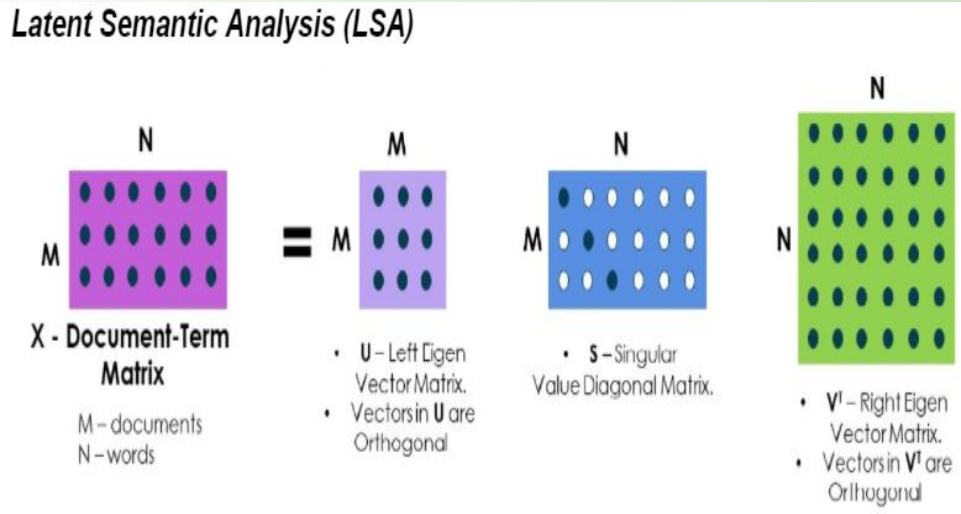
La **LDA** è un modello probabilistico che consente di estrarre argomenti da un insieme di documenti e si basa sul presupposto che se due termini si trovano spesso in più documenti, probabilmente questi costituiscono il seme di un *topic*.

La **NMF** invece è un modello che si basa sull'algebra lineare e più precisamente sul calcolo matriciale.
Attraverso la scomposizione della matrice di partenza si ottengono delle matrici di dimensione ridotta che permettono di individuare per ogni topics i termini ad esso associati e di individuare poi il topic associato ad ogni singolo documento.



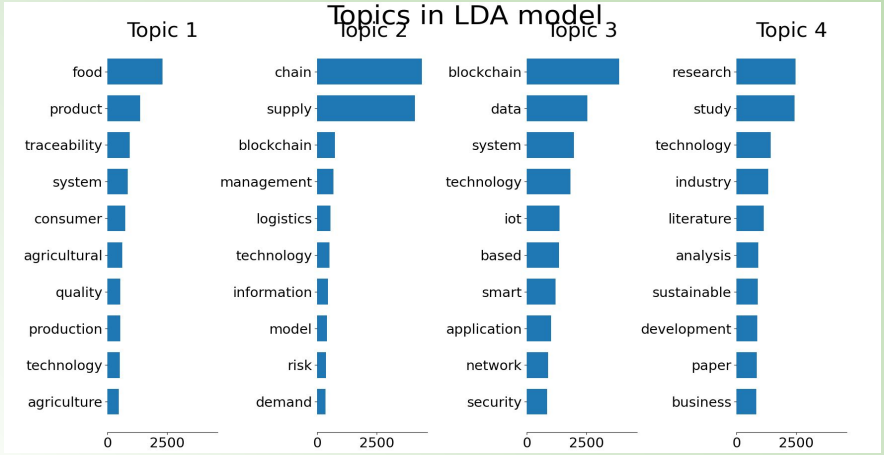
TOPIC MODELING: PLSA & LSA

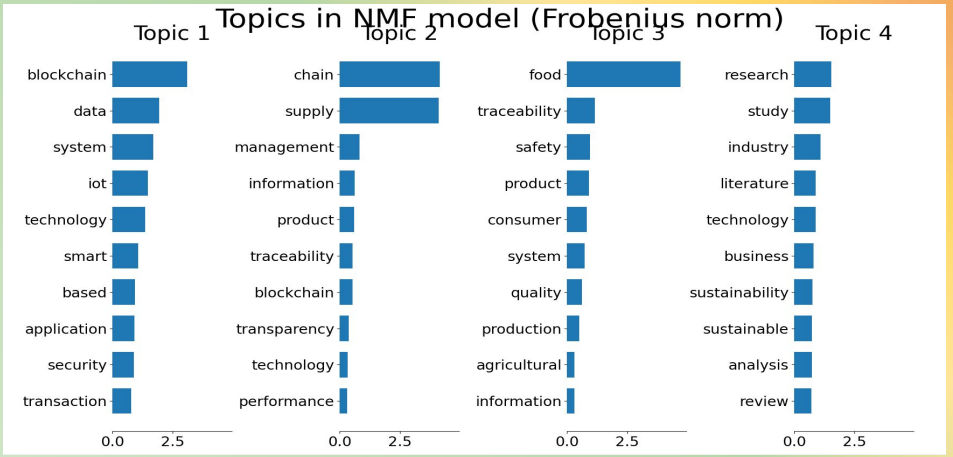
La **PLSA** deriva da una visione statistica dell'LSA e definisce un proprio modello generativo: invece di utilizzare matrici matematiche e la SVD per ridurre le dimensioni del problema, utilizza un modello probabilistico secondo il quale è possibile derivare una rappresentazione a dimensionalità ridotta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

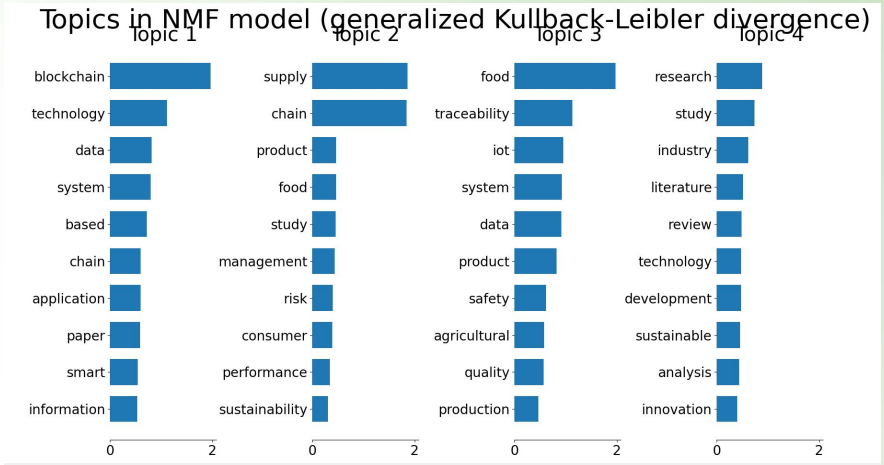


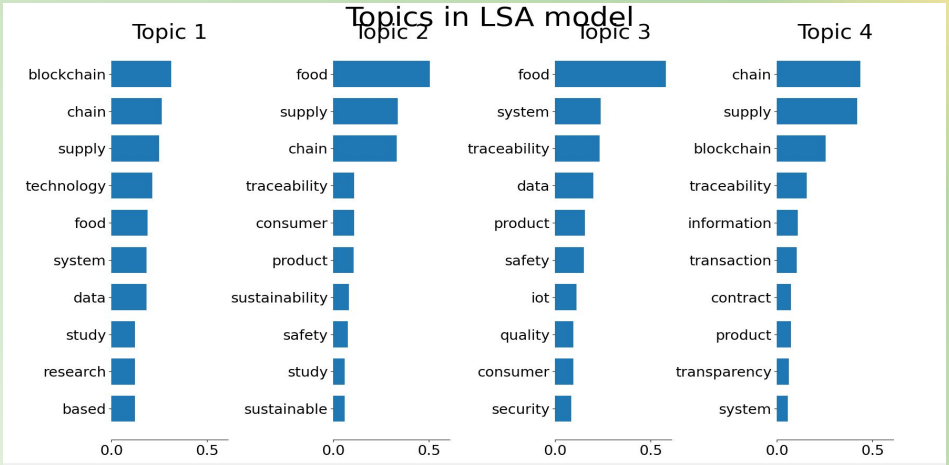
La **LSA** sfrutta principalmente una tecnica di decomposizione matriciale chiamata Singular Value Decomposition (SVD) che decompone la matrice di partenza in tre matrici; una di queste matrici rappresenta la matrice dei valori singolari del dataset in analisi, dai quali è possibile poi estrarre informazioni 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.

ANALISI COMPARATIVA: LDA, NMF, PLSA, LSA









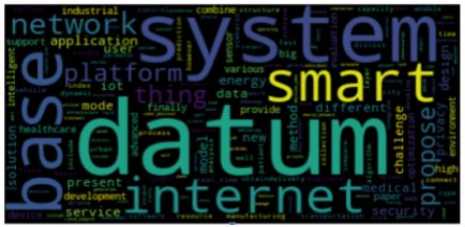
ANALISI COMPARATIVA LDA Vs LSA: WordCloud

WordCloud associate a ciascuno dei 4 Topics del Modello LDA

TOPIC #1

TOPIC #2

TOPIC #3



TOPIC #4

WordCloud associate a ciascuno dei 4 Topics del Modello LSA

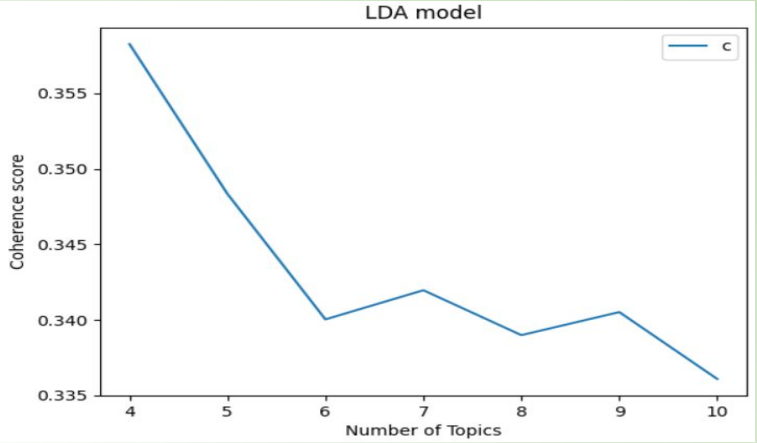
TOPIC #1

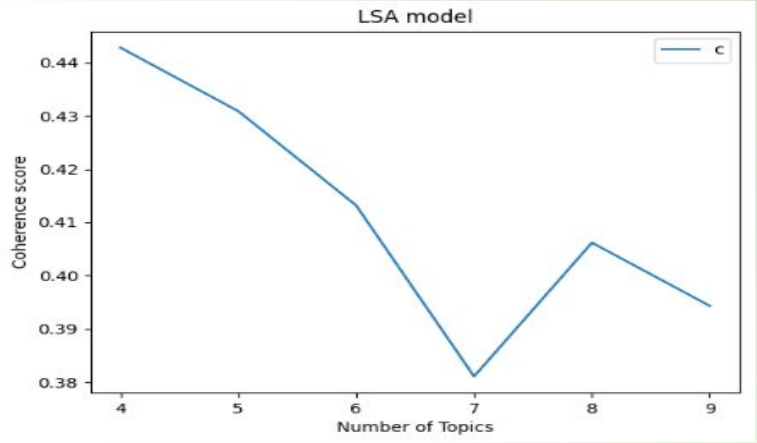
TOPIC #2

TOPIC #3

TOPIC #4

ANALISI COMPARATIVA LDA Vs LSA: COHERENCE



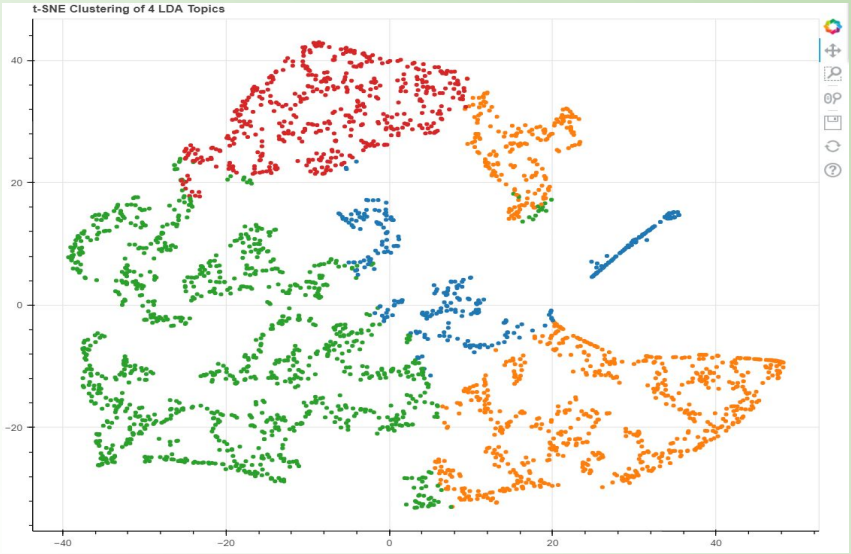


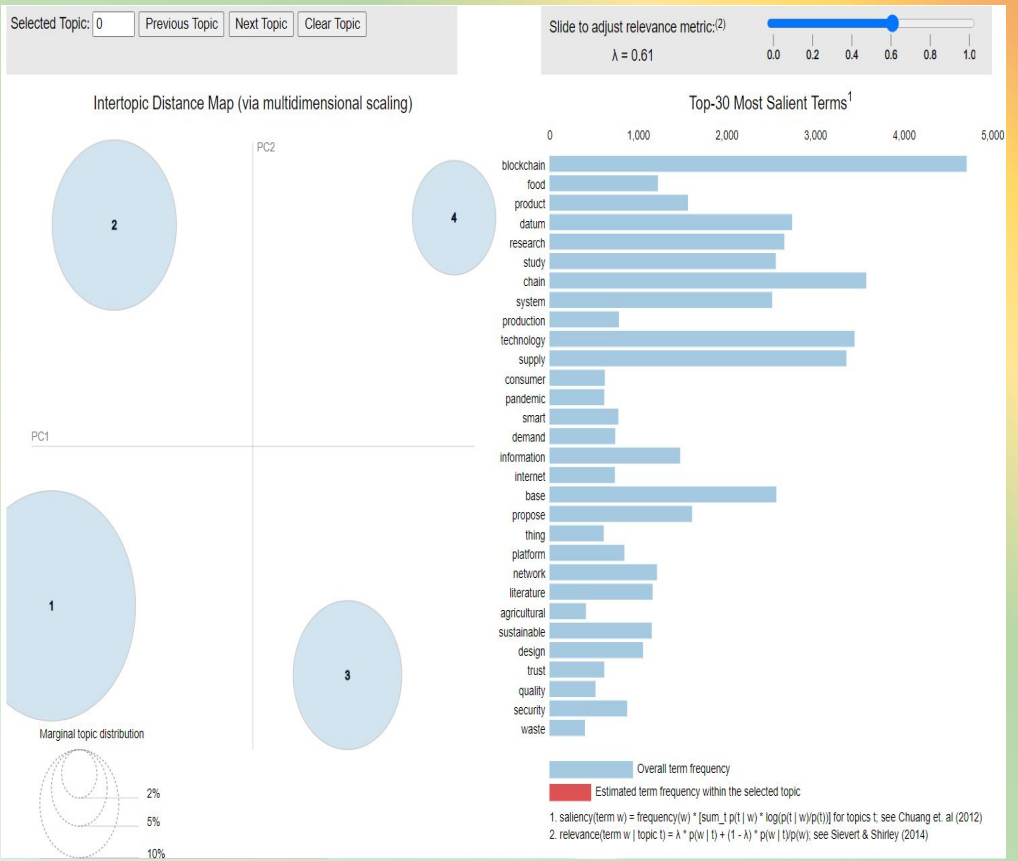
Risultati:

- Numero ottimale di Topics 4;
- Coherence maggiore nel modello LSA

Nonostante il modello LSA raggiunga livelli della Coherence più elevati, il modello LDA risulta comunque preferibile poichè presenta un grado di interpretabilità maggiore.

GRAFICI INTERATTIVI LDA





- Librerie TSNE e pyLDAvis;
- Divisione in Cluster dei documenti in base alla predominanza di un Topic;
- Analisi generale e specifica per ogni Topic.

Analisi temporale

```
######## 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
```

Risultati finali

- Nonostante le difficoltà affrontate nella fase iniziale di Web Scraping, 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.
- Attraverso l'analisi Temporale è stato possibile fornire anche un interpretazione dei trend dei Topics che si sono susseguiti anno per anno nell'arco temporale 2017-2021, 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.