Relazione Esercizio 4

Pipeline di Clustering e Topic Modeling

Dataset

Per questa esercitazione sono stati utilizzati due dataset:

1. Books Dataset, contenente svariate informazioni di circa 103,063 libri.
2. News Dataset, contenente circa 210,000 titoli di notizie raccolte dal 2012 al 2022 su HuffPost.

La scelta di **utilizzare una coppia di dataset** deriva dal fatto che il primo, quello dei libri, non contiene una categorizzazione prestabilita degli elementi al suo interno. Le notizie contenute nel secondo dataset invece, hanno già una classificazione con la categoria di tratta la notizia. Questo ha permesso di effettuare ulteriori analisi e confronti sui risultati del clustering ed in generale dell'esercizio.

Entrambi i dataset sono stati sottoposti ad un processo di riduzione della dimensionalità, e per quanto riguarda il secondo è stato sono state anche effettuate operazioni di bilanciamento delle categorie.

Embedding Generation

Per generare gli embedding sono state utilizzate le "short description" dei libri nel primo dataset e le descrizioni delle notizie nel secondo dataset, entrambe sottoposte ad un processo di Text Cleaning comprendente: tokenization, stopwords filtering e lemmatization.

A partire da questi dati sono stati generati gli embeddings tramite il modello General Text Embeddings (GTE), la cui dimensionalità è stata poi ridotta da 384 a 6.

Clustering

Gli embeddings ottenuti sono stati clusterizzati tramite il modello HDBSCAN.

I risultati della clusterizzazione di entrambi i cluster sono stati poi salvati all'interno di un dataframe pandas per una migliore futura gestione e visualizzazione.

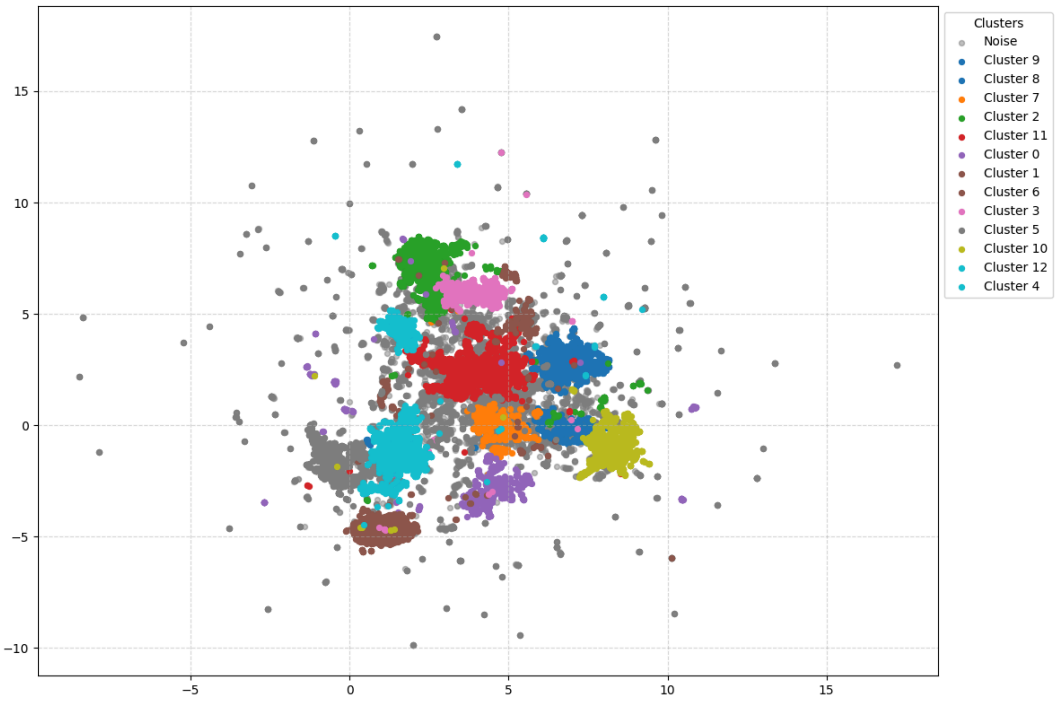
CLUSTER DATASET 1
- CLUSTER 7: 1655
- CLUSTER 5: 1547
- CLUSTER 12: 1121
- CLUSTER 14: 1082
- CLUSTER 13: 927
- CLUSTER 10: 854
- CLUSTER 3: 846
- CLUSTER 1: 740
- CLUSTER 11: 676

CLUSTER DATASET 2
- CLUSTER 11: 2567
- CLUSTER 4: 2028
- CLUSTER 2: 2008
- CLUSTER 8: 1691
- CLUSTER 5: 1520
- CLUSTER 10: 1499
- CLUSTER 1: 1276
- CLUSTER 0: 1202
- CLUSTER 3: 860

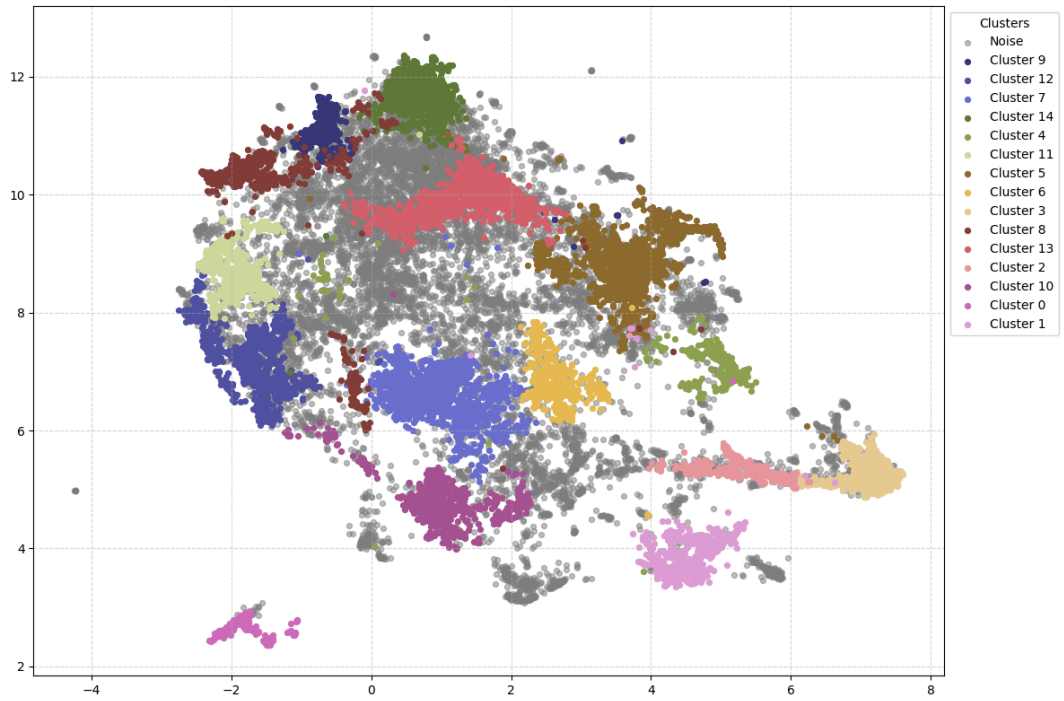
– CLUSTER 8: 625
– CLUSTER 6: 523
– CLUSTER -1: (noise): 11528

– CLUSTER 7: 560
– CLUSTER 9: 552
– CLUSTER 12: 468
– CLUSTER 6: 415
– CLUSTER -1: (noise): 7354

Notiamo come in entrambi i casi si riscontra una quantità molto alta di elementi rumorosi. Questo è dovuto al fatto che (come si evince dai successivi plot) gli elementi sono molto vicini tra loro, e solo in zone con una densità estremamente alta l'algoritmo riesce ad individuare un cluster.



News Dataset Clustering



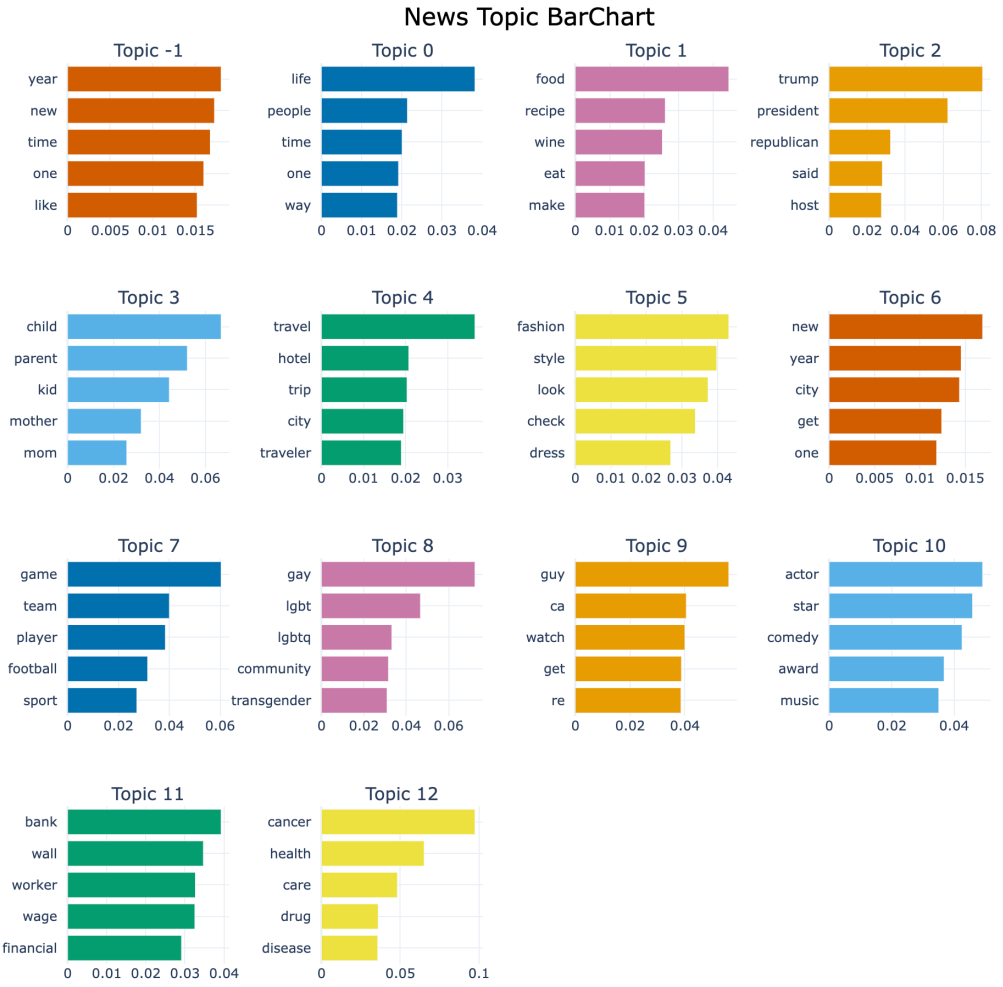
Books Dataset Clustering

Topic Modeling

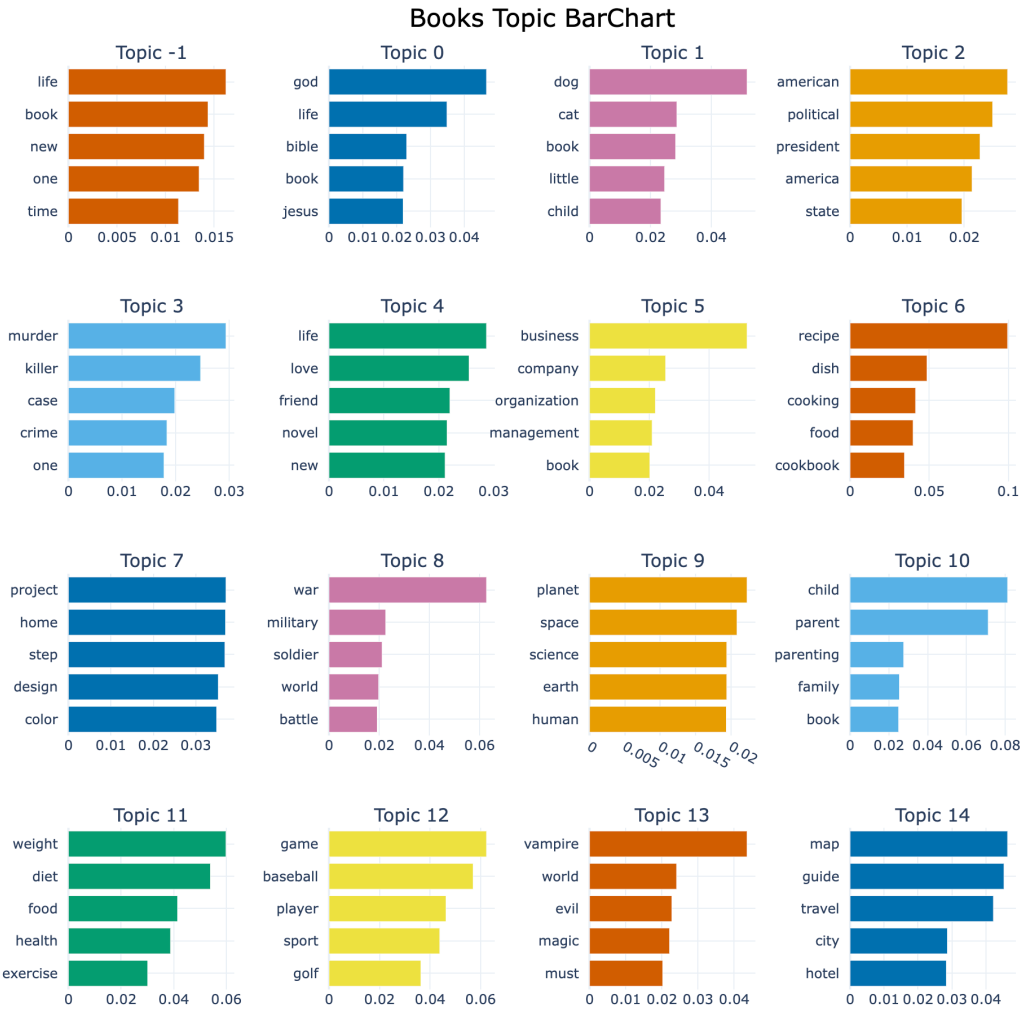
I cluster ottenuti sono stati infine sottoposti all'ultimo step della pipeline, quello di Topic Modeling, in cui si è cercato di ottenere le distribuzioni tf-idf dei termini all'interno dei cluster stessi.

Per quanto riguarda il dataset contente le news, come intuibile, i topic ottenuti rispecchiano semanticamente quelle che sono le categorie iniziali tramite le quali sono state classificate le notizie all'interno del dataset. Ad esempio:

- Il Topic 1 con "FOOD & DRINK"
- Il Topic 4 con "TRAVEL"
- Il Topic 7 con "SPORTS"
- e così via…



Mentre per quanto riguarda il primo dataset, notiamo come i topic ottenuti vadano a definire i vari possibili generi a cui appartengono i libri.

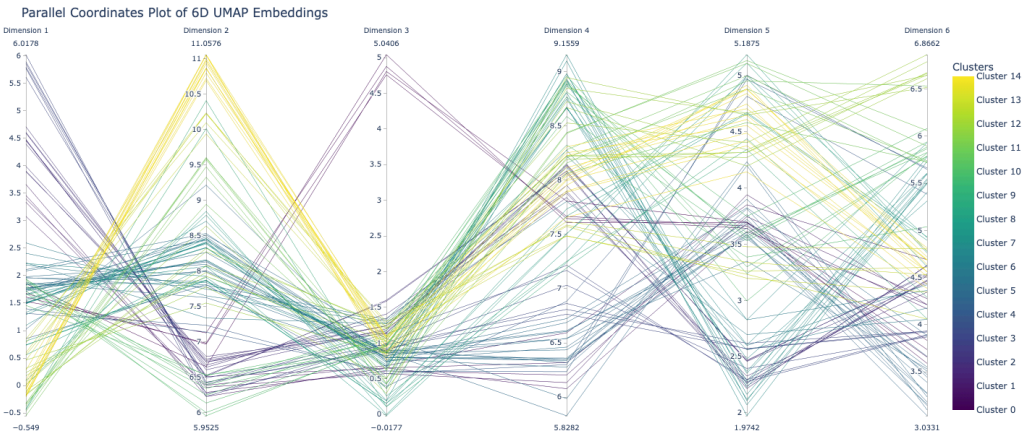


Visualizzazione dei Topic

Per visualizzare i topic identificati è stata utilizzata la tecnica delle Parallel Coordinates, evidenziando come i valori degli embedding 6-dimensionali si distribuiscono nei vari insiemi.

I grafici ottenuti mostrano chiaramente i pattern tra i valori su ciascuna dimensione, formando così i topic estratti.

Books Topics Parallel Coordinates



News Topics Parallel Coordinates