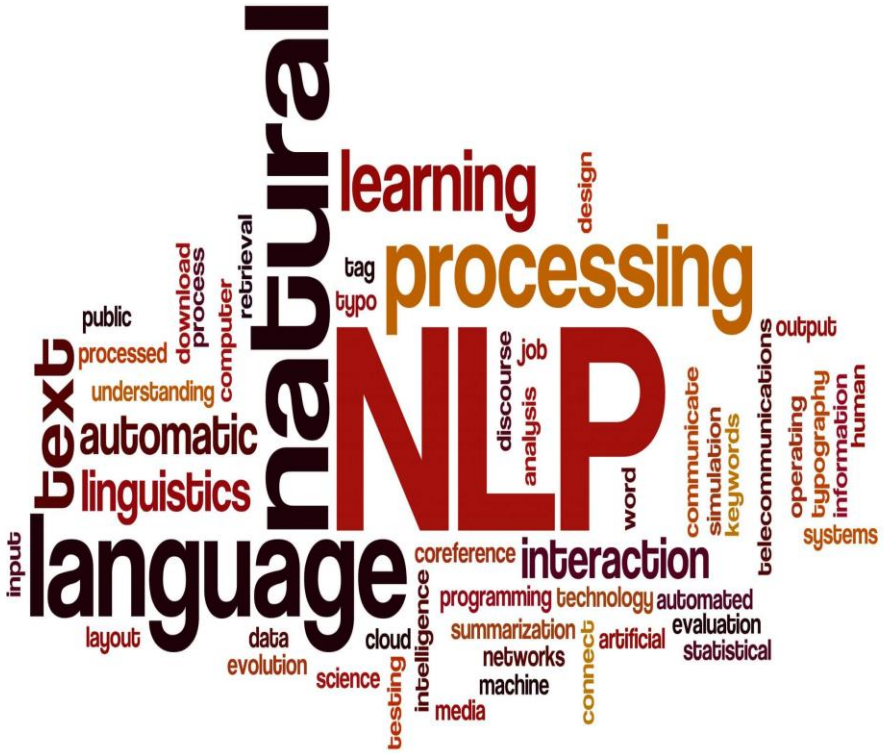
Capstone Project

Topic Modeling on News Articles

Content

- Problem statement
- Data Summary
- Data Processing
- Feature Extraction
- ML Modules
- Challenges
- Conclusion



Problem Statement

- Identify major Topics/themes across a collection of BBC news articles using different topic modeling techniques.

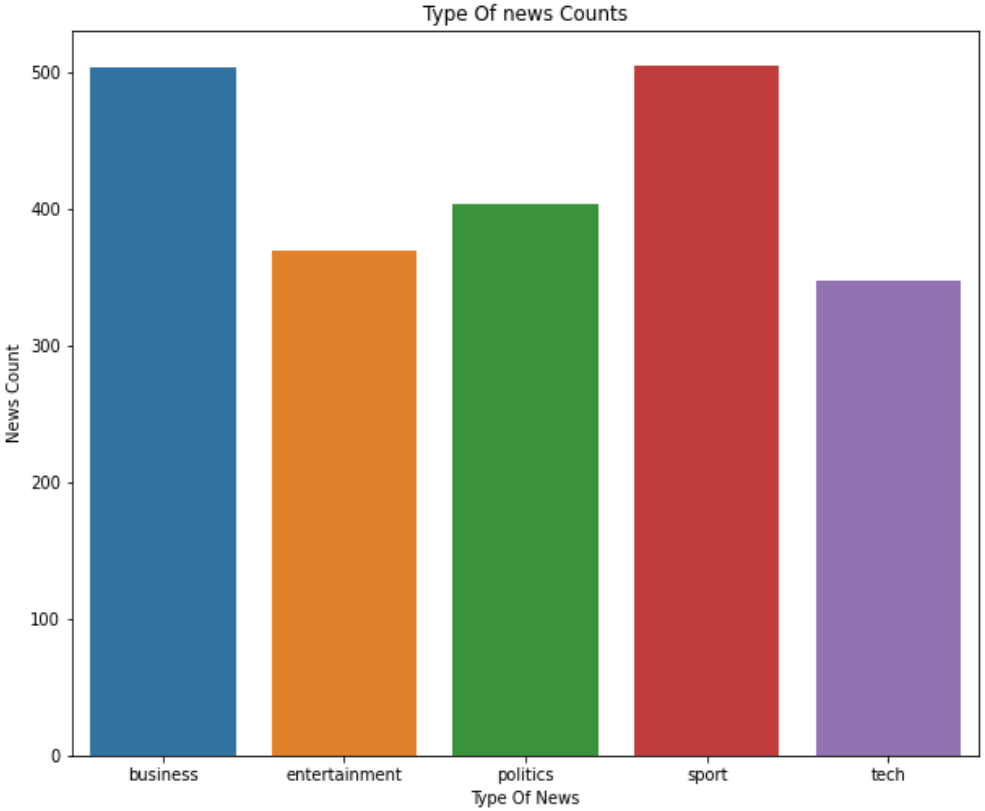
Data Summary

- Dataset has total 2225 rows and 2 columns.
- In that articles are given in news column and article types are given in type column.
- Both columns has object data type.
- There is no null value in dataset.

	news	type
0	b'Yukos unit buyer faces loan claim\n\nThe own...	business
1	b'Ad sales boost Time Warner profit\n\nQuarter...	business
2	b'Dollar gains on Greenspan speech\n\nThe doll...	business
3	b'US trade gap hits record in 2004\n\nThe gap ...	business
4	b'High fuel prices hit BA\'s profits\n\nBritis...	business

EDA

- There are five types of news in our dataset which are business, entertainment, politics, sport and tech.
- In that highest number of news are from sport category having count 505.
- Business category also has higher count with 503 which is very close to sport category.
- Tech category contains lowest count is 347.

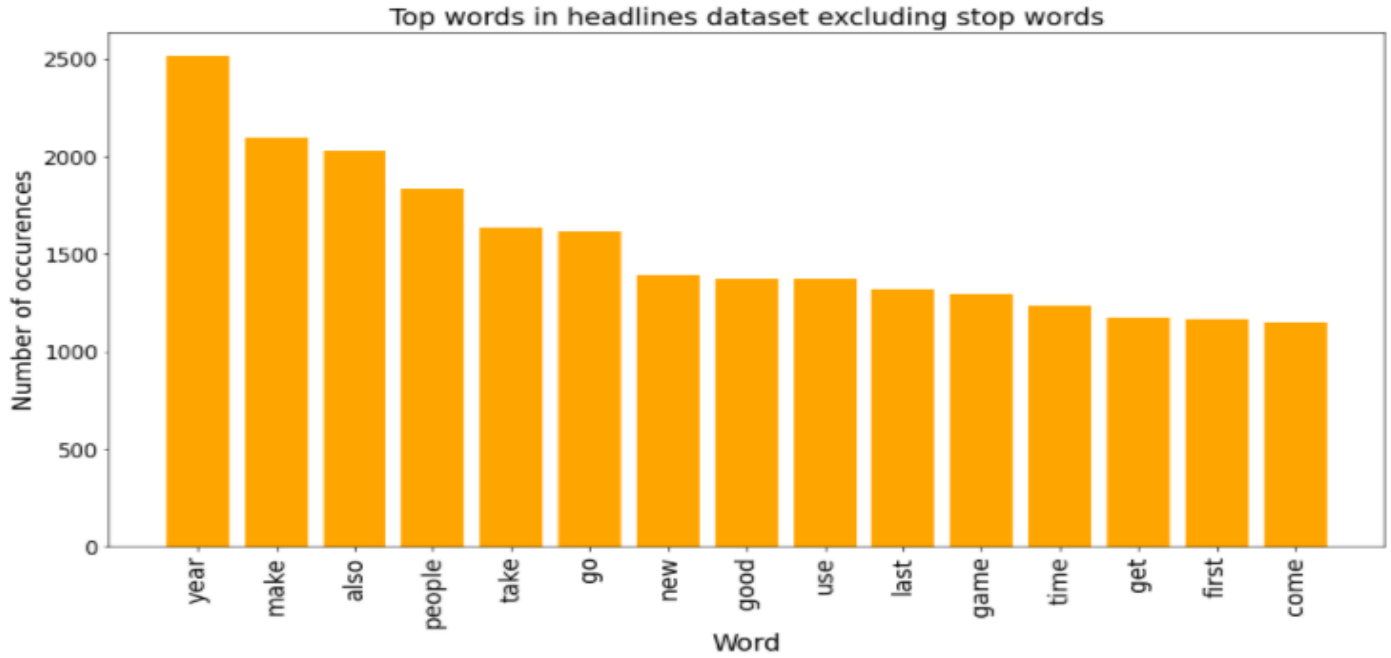


Data Processing

- Removing HTML tags and urls.
- Removing Numbers and Special Characters.
- Removing Stopwords.
- Removing punctuations.
- Applying Lemmatization.

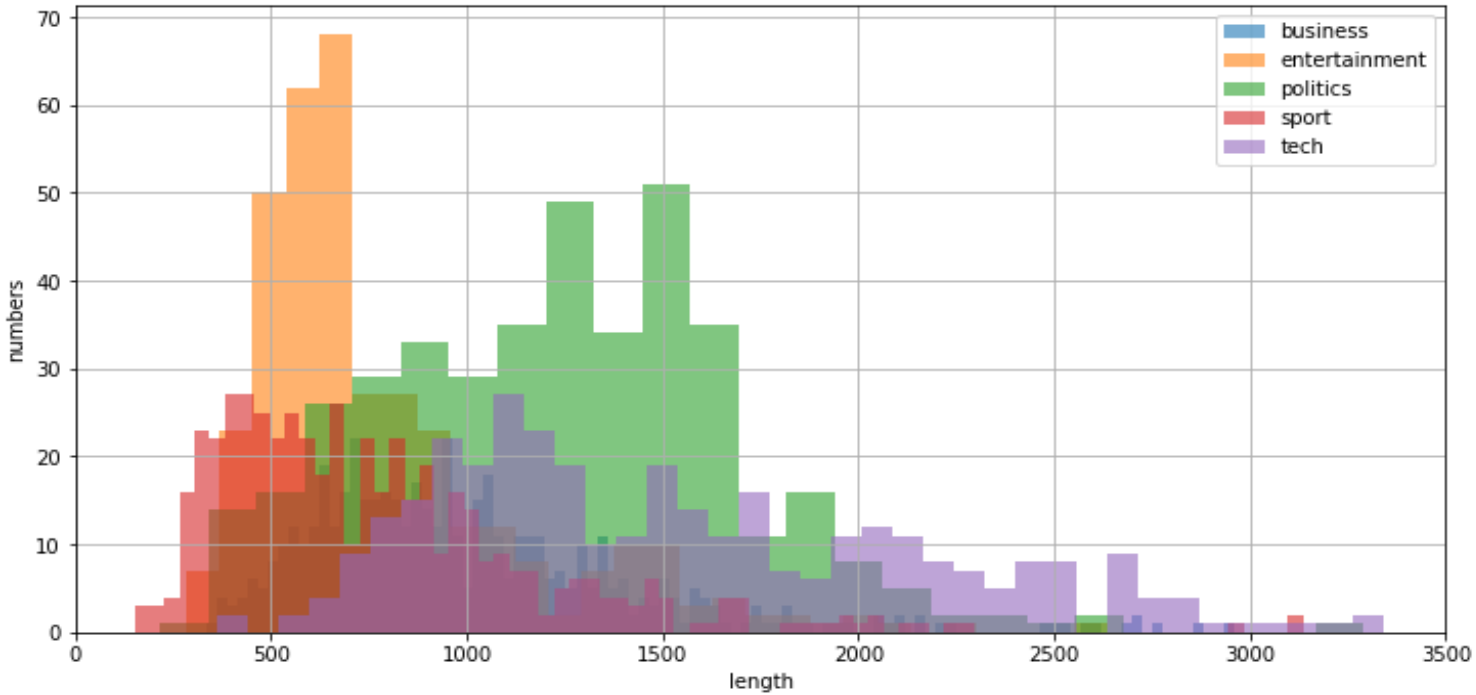
Most Frequent words

Graph shows most frequent word in dataset after removing stop words.

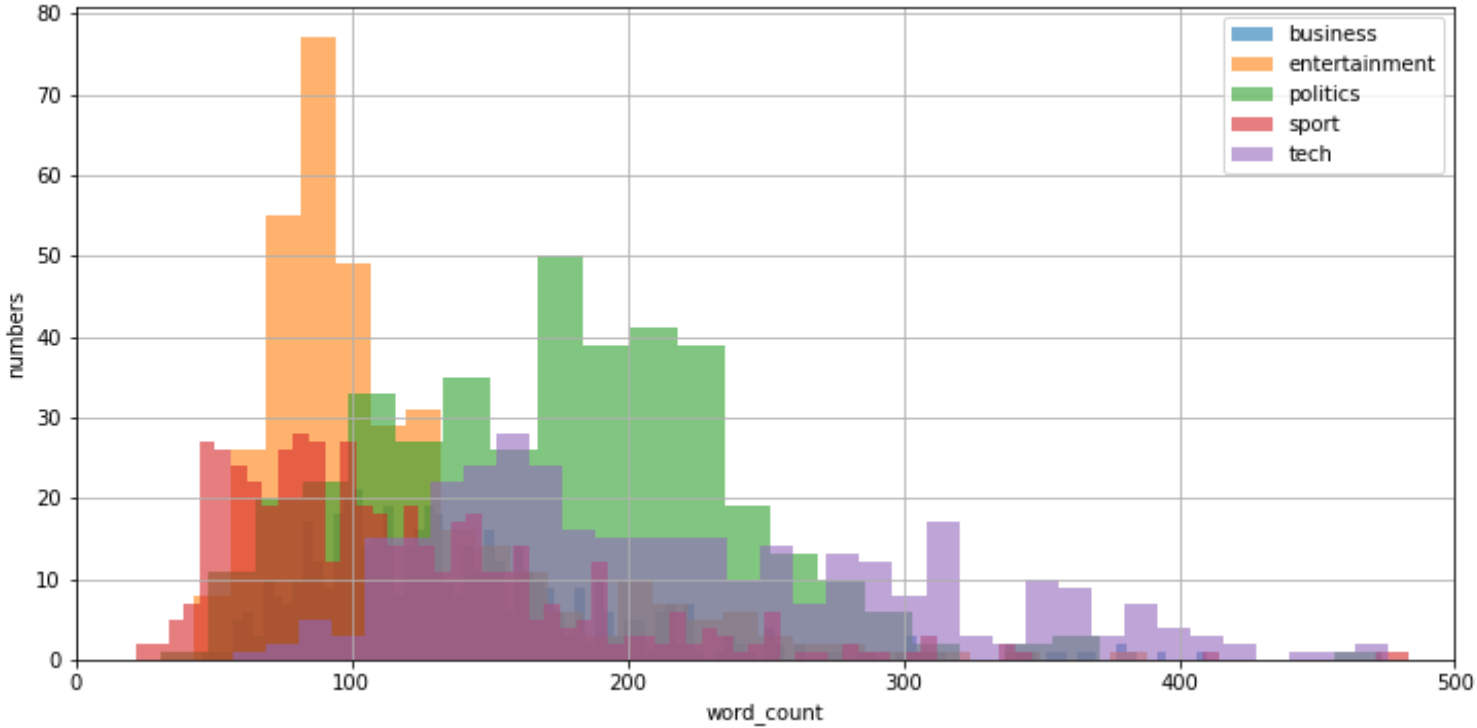


Feature Extraction

Length of Documents -



Number of words in Document -



WordCloud For Business Topic

- In articles of business category we can see that most frequent words are related to business.
- Those words are company, firm, rise, market, make, sales etc.

WordCloud For Tech Topic

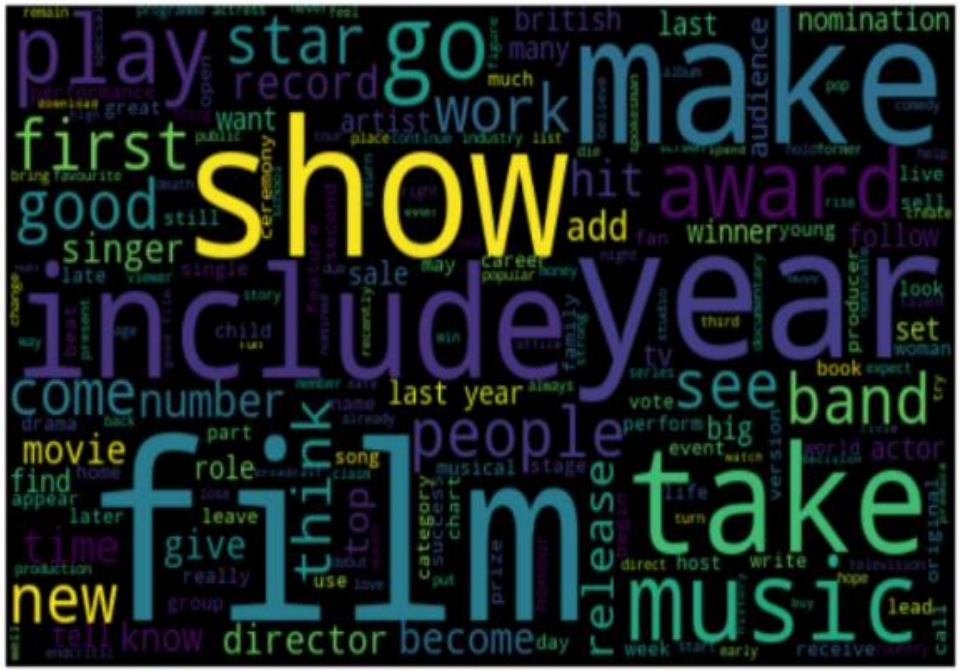
- In articles of Tech category we can see that most frequent words are related to Technology.
- Those words are technology, use, people, game, make, service etc.

WordCloud For sport Topic

- In articles of sport category we can see that most frequent words are related to sport.
- Those words are play, win, player, game, win, go etc.

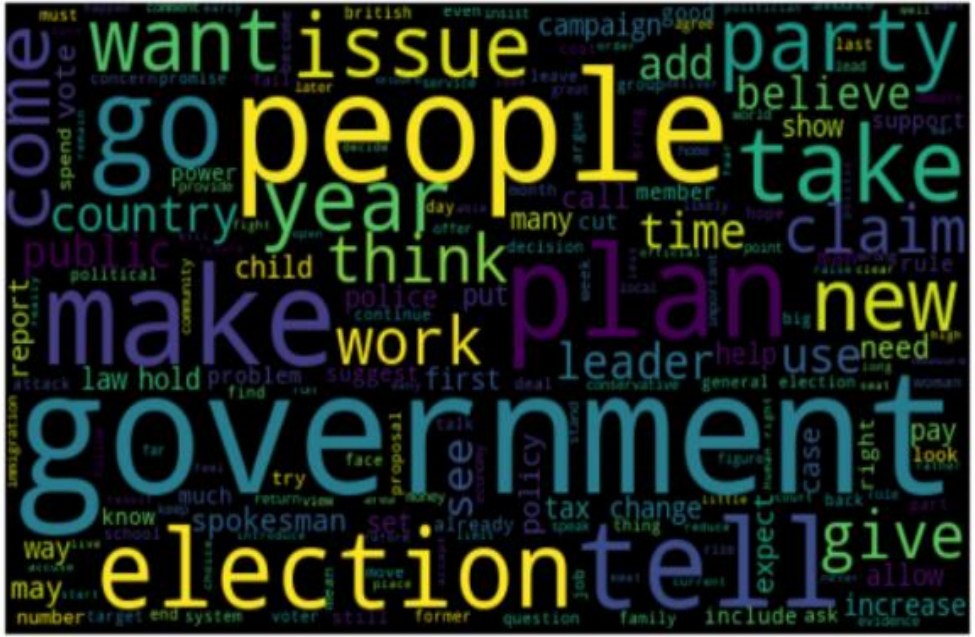
WordCloud For Entertainment Topic

- In articles of Entertainment category we can see that most frequent words are related to Entertainment.
- Those words are film, show, music, award, hit etc.



WordCloud For Politics Topic

- In articles of Politics category we can see that most frequent words are related to Politics.
- Those words are government, people, election, plan, party, issue etc.

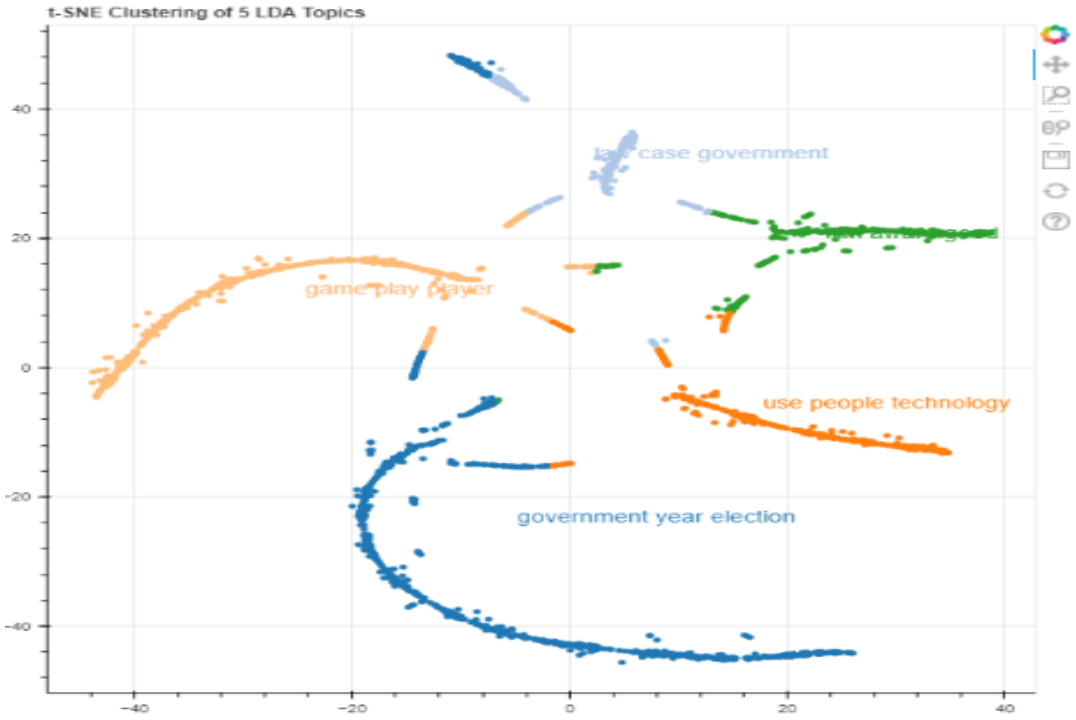


Implementation of ML Modules

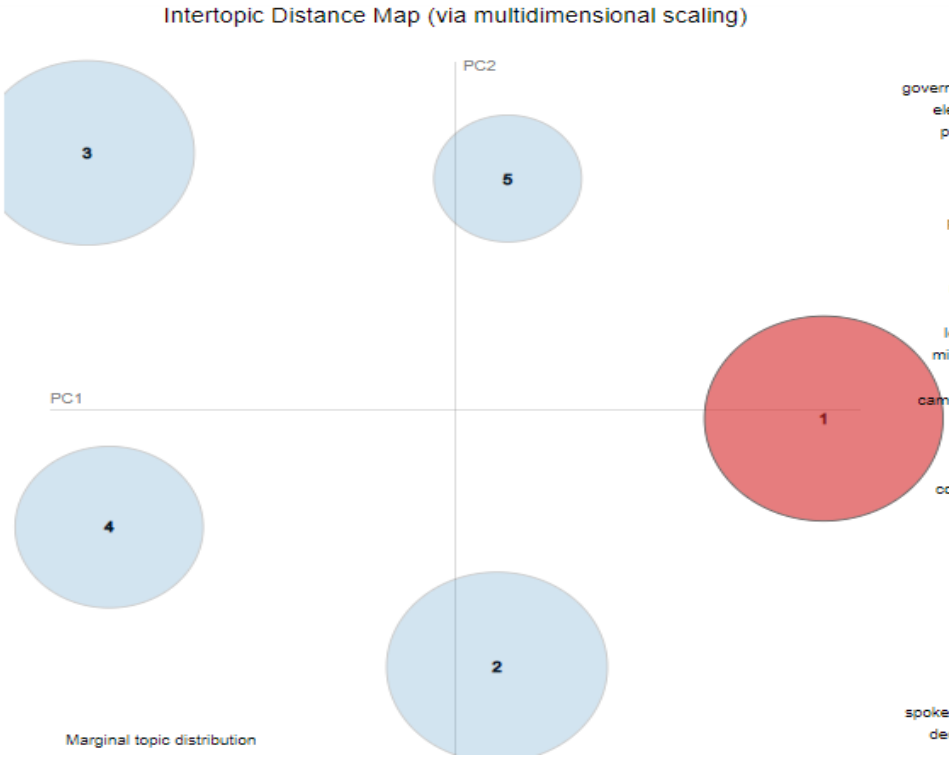
- Latent Dirichlet Allocation (Gensim)
- Latent Dirichlet Allocation (LDA) (Sklearn) with TF-IDF vectorizer
- Latent Dirichlet Allocation (Sklearn) with count-vectorizer and Bi-gram
- Latent Semantic Analysis (LSA)

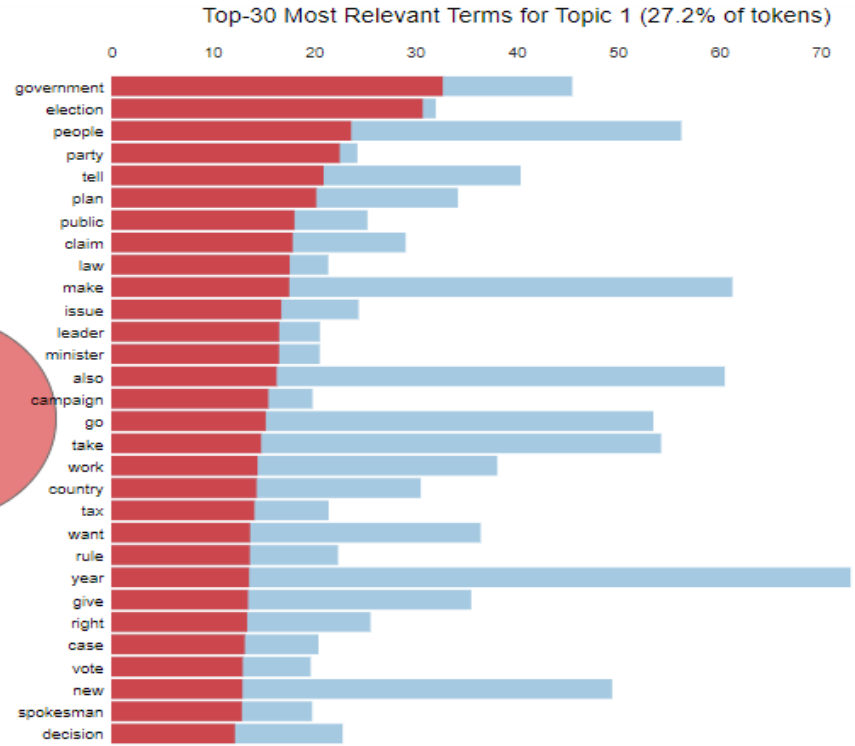
Latent Dirichlet Allocation (LDA)

LDA clustering of 5 Topics

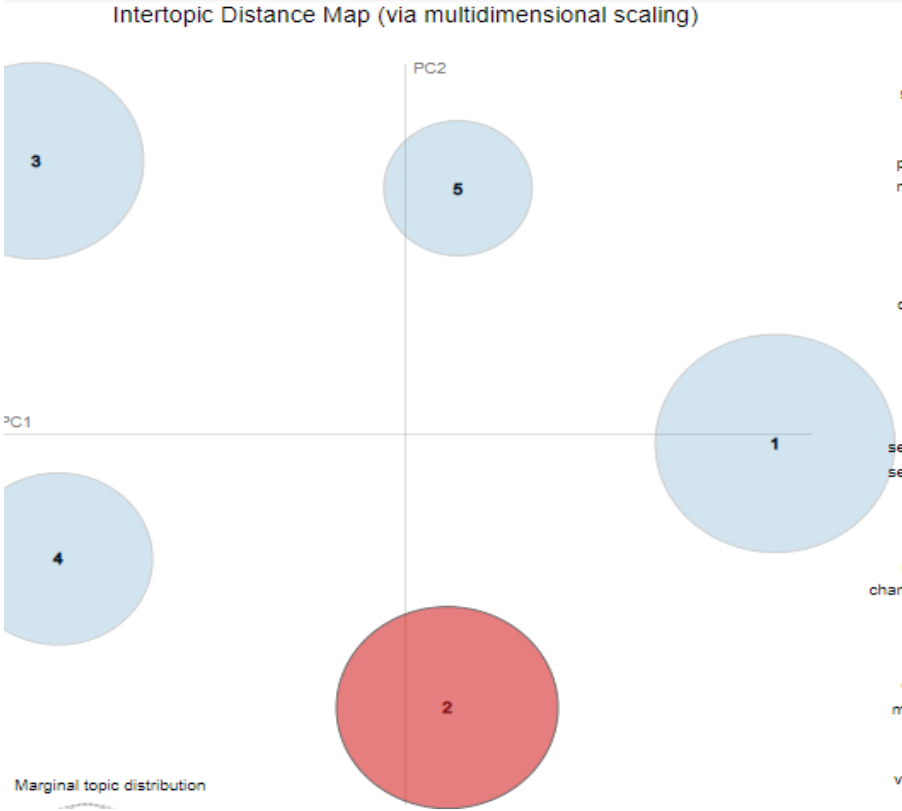


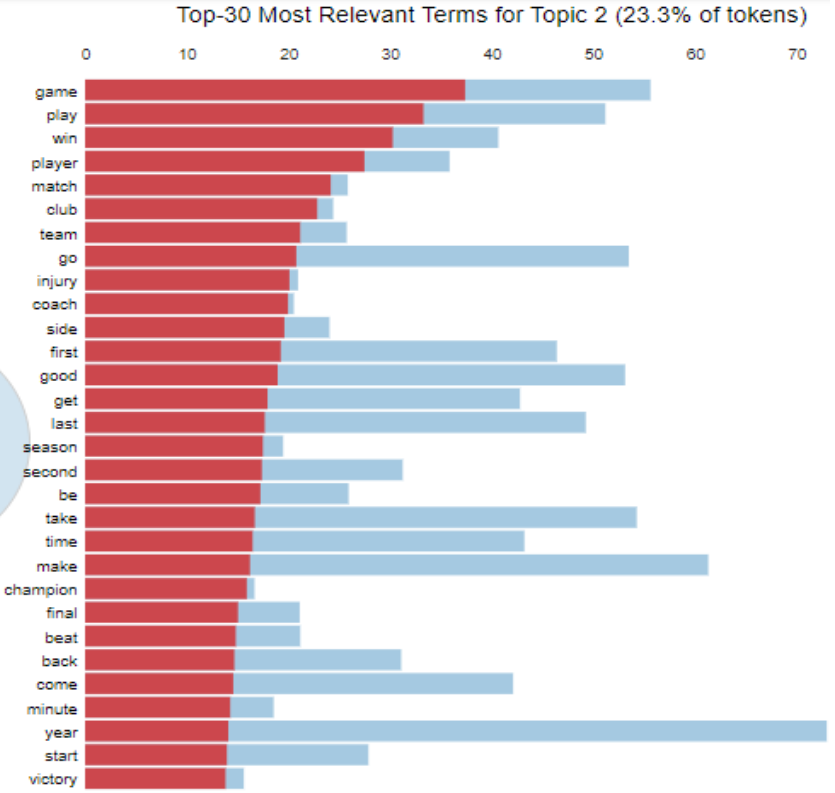
LDA - Cluster 1 : Politics



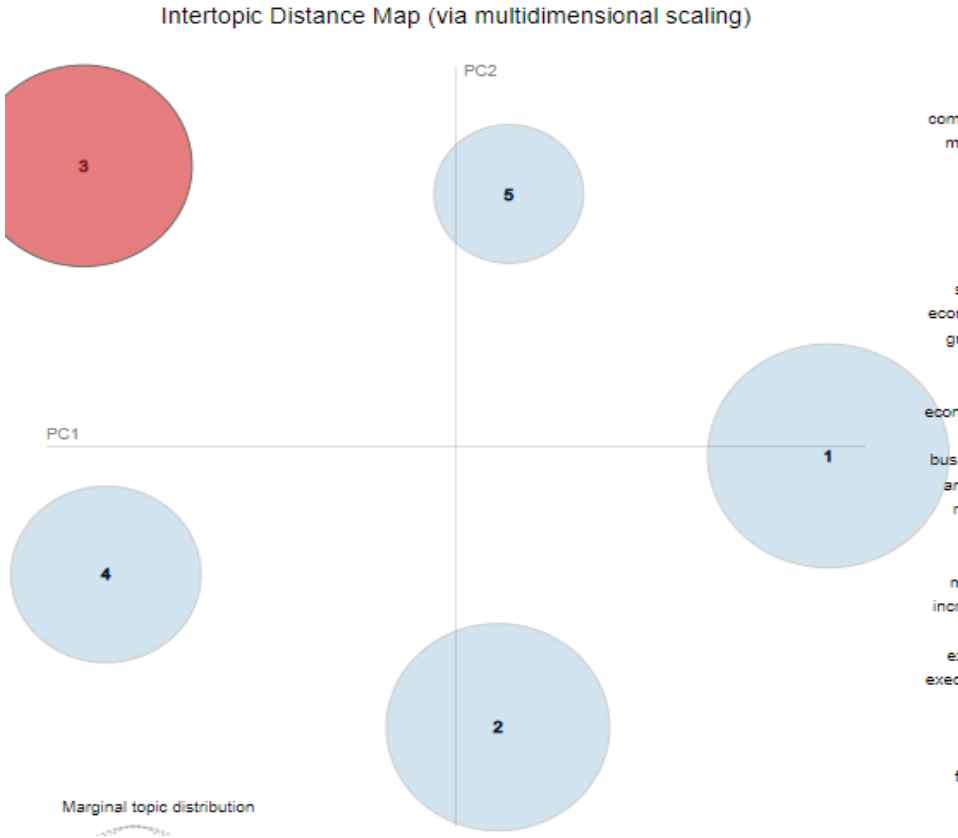


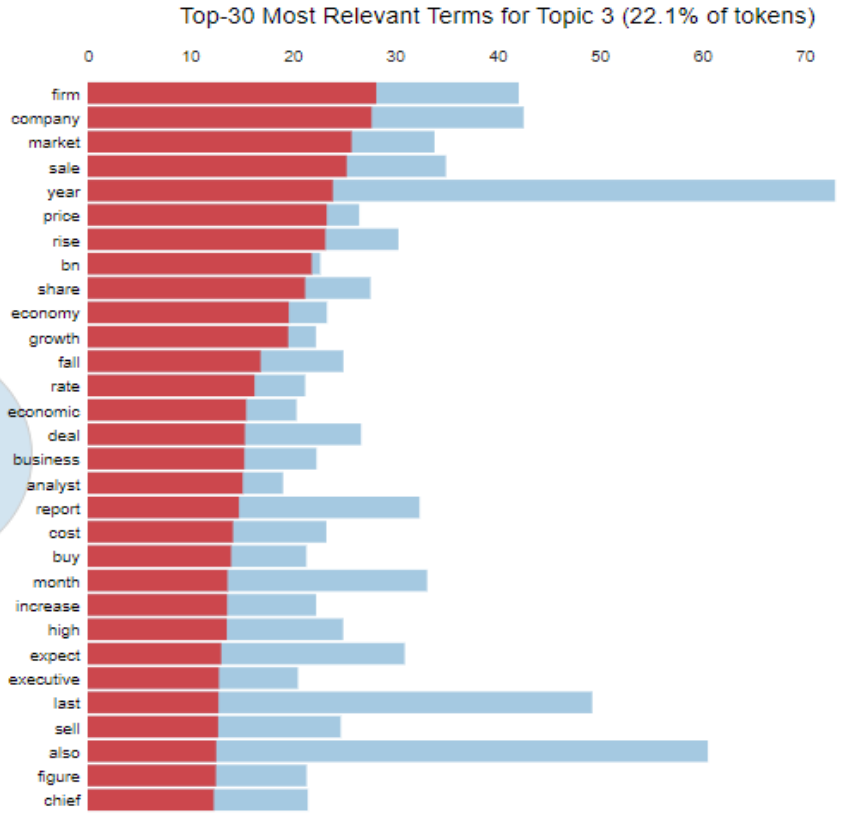
LDA - Cluster 2 : Sport



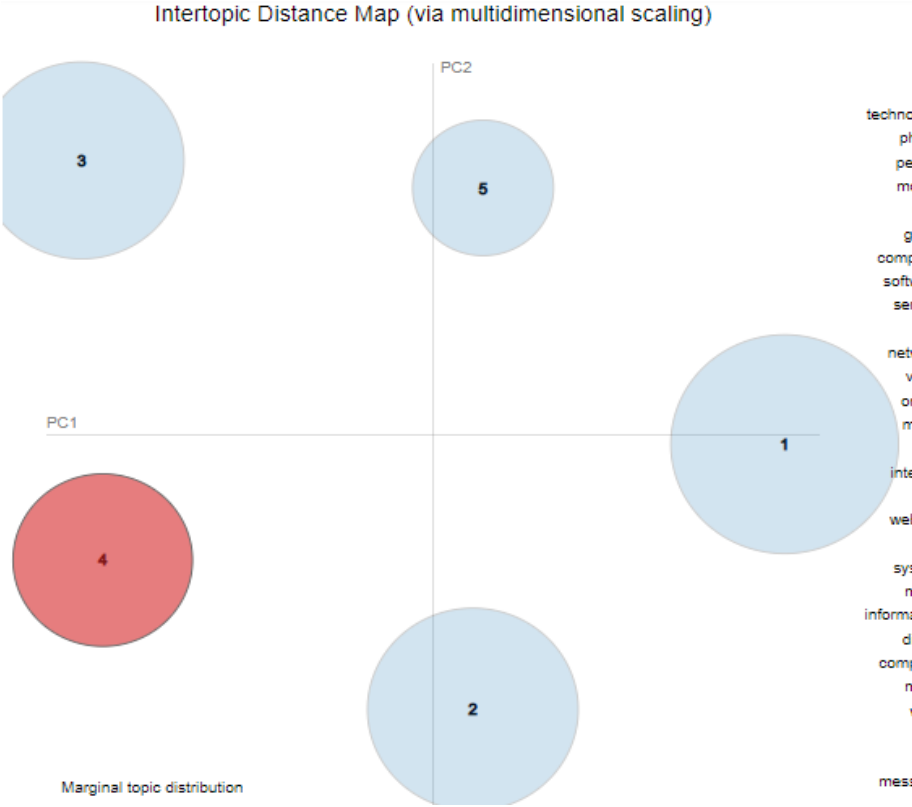


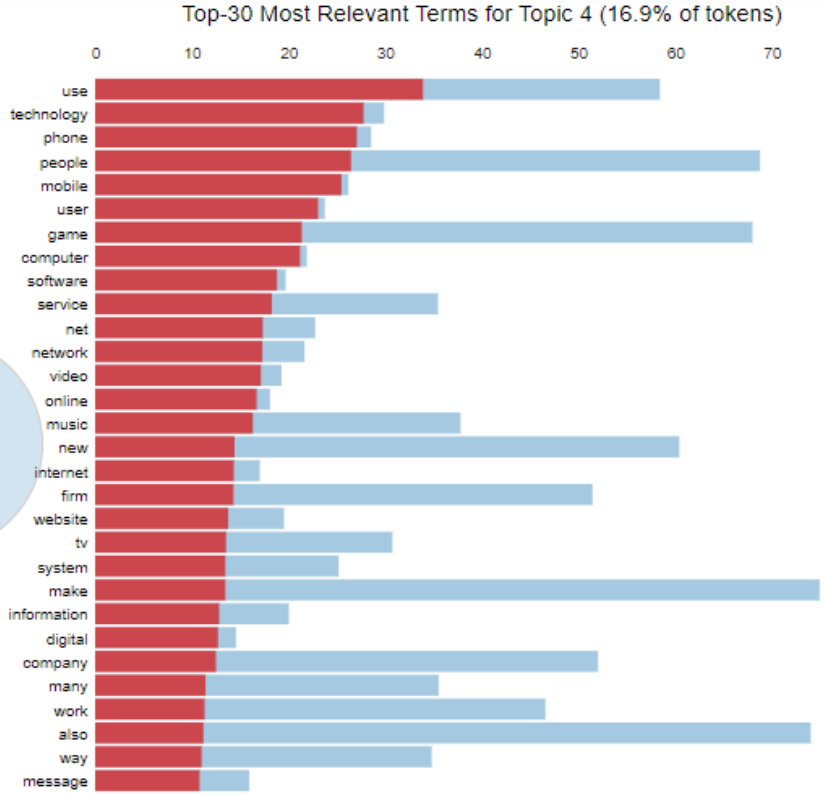
LDA - Cluster 3 : Business



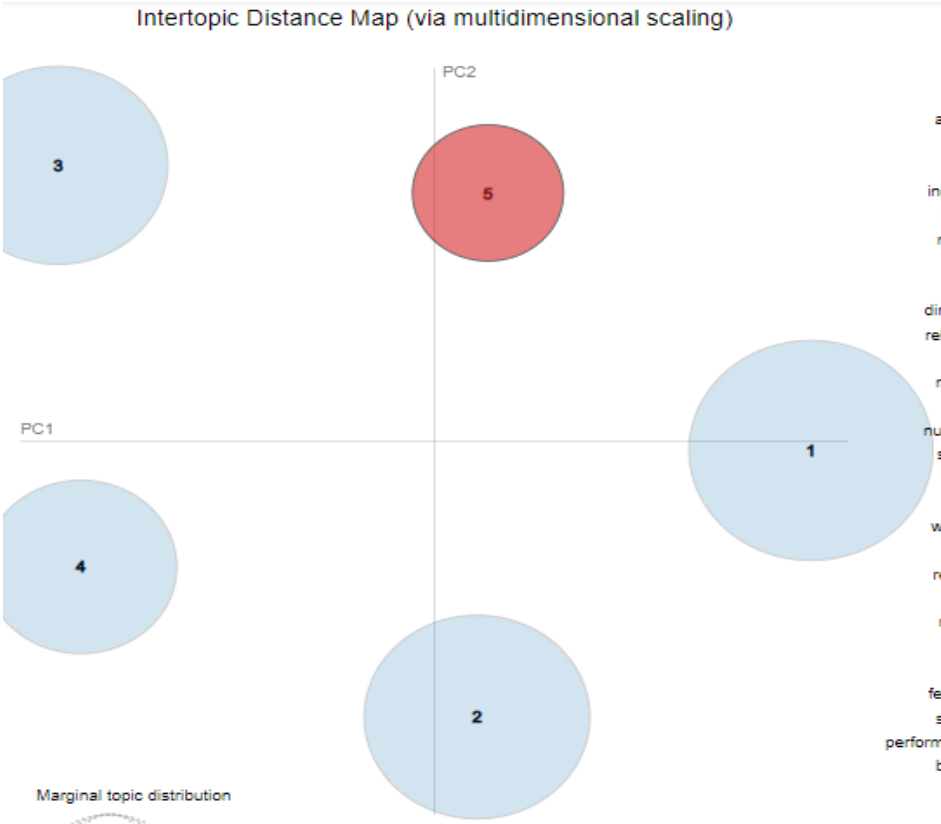


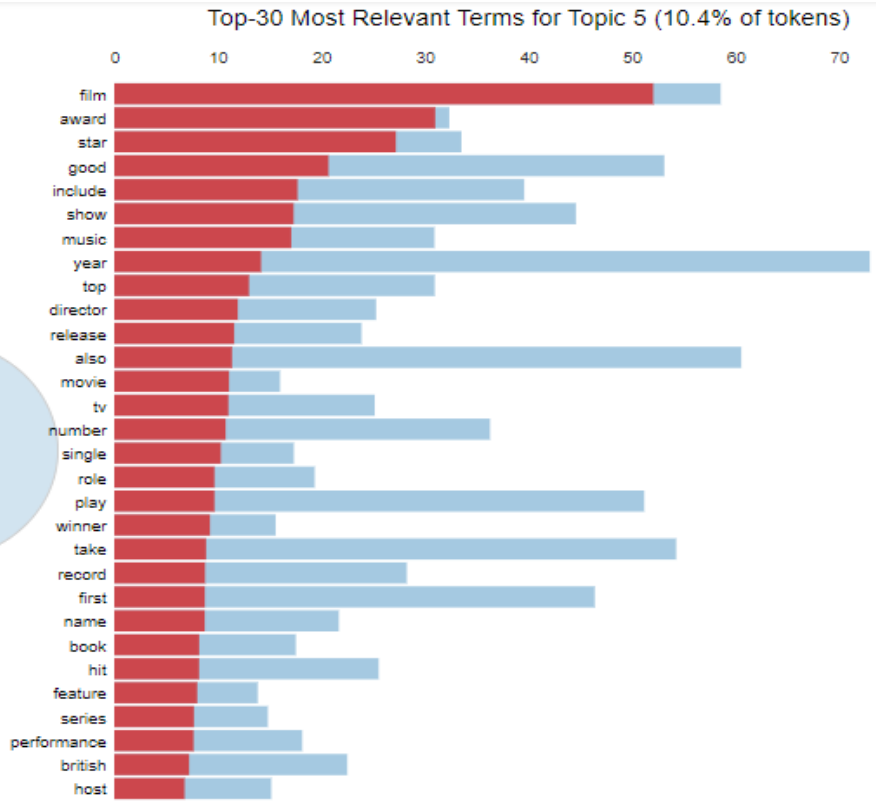
LDA - Cluster 4 : Technology





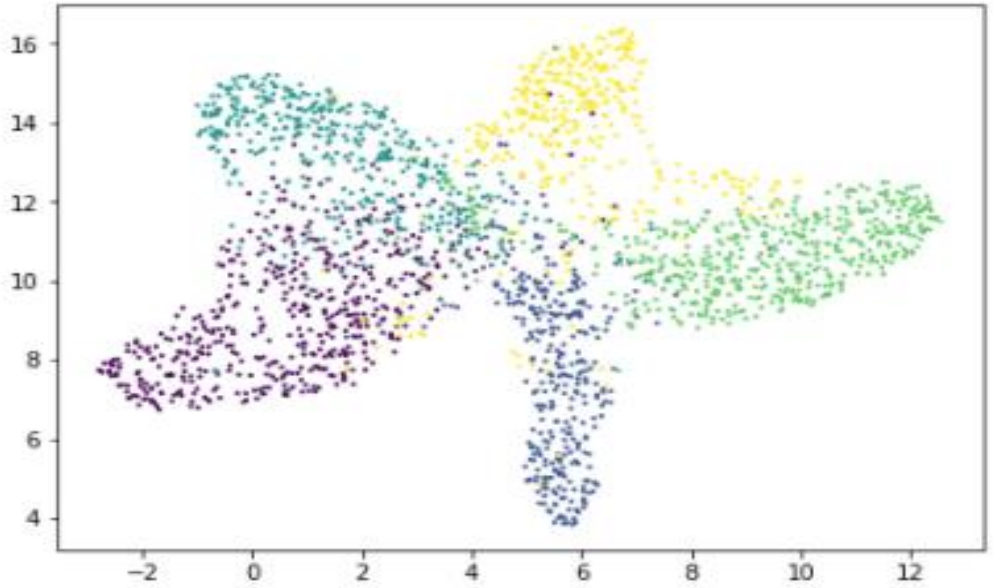
LDA - Cluster 5 : Entertainment





Latent Semantic Analysis (LSA)

- LSA is a technique in NLP, used in analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.
- LSA assumes that words that are close in meaning will occur in similar pieces of text



Challenges

- Some text pre-processing technique took too much time to execute.
- Limited visualization techniques to identify model performance
- Less availability of information of different algorithms implementation technique in python.

Conclusion

- LDA (Sklearn) with TF-IDF vectorizer along with NMF were best to identify the 5 given clusters.
- Scope of implementing neural network in future.
- As a future work, using one of the topic modeling algorithms, we can implement various applications for recommending research articles, analyzing news articles etc, which can be used for segregation of documents from topic.

Q & A