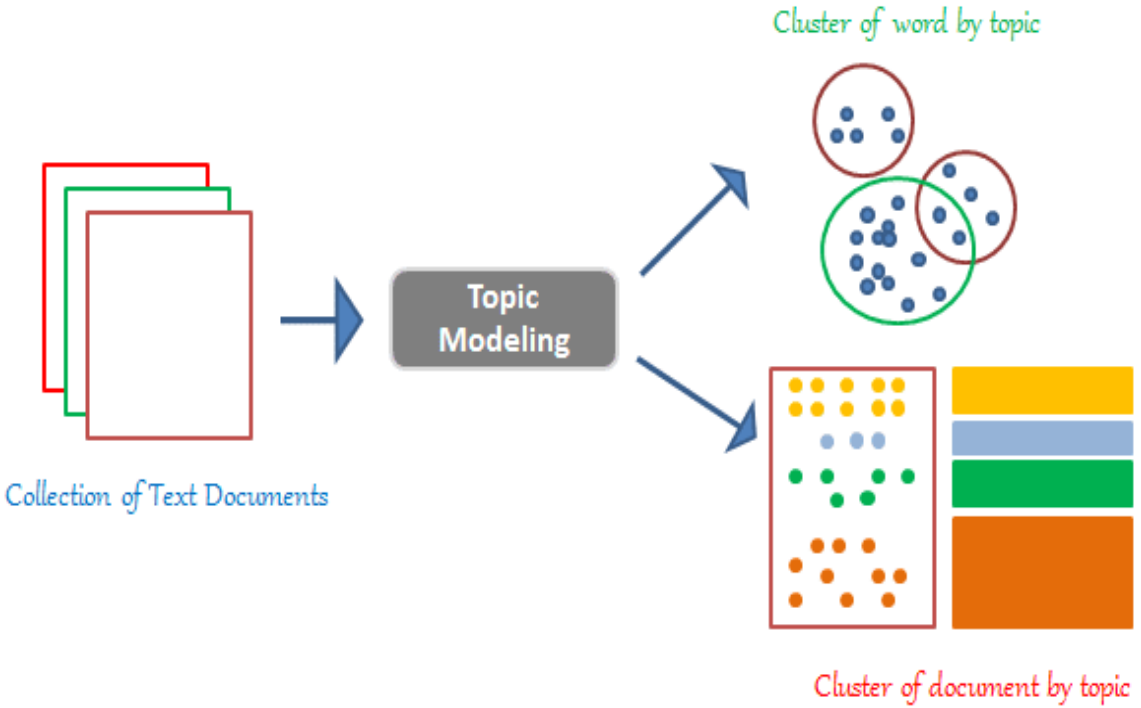
Capstone Project

Topic Modeling on News Articles

Suraj Pandey

Content

- Problem statement
- Data Summary
- Data Preprocessing
- Feature Extraction
- ML Models
- Challenges
- Conclusion

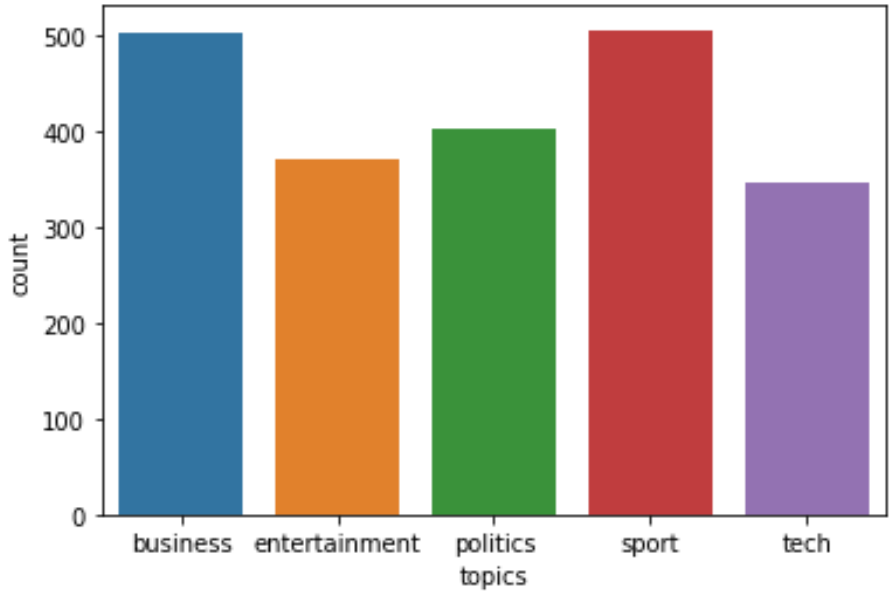


Problem Statement

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

Data Summary

- There are total 5 topics -
  - ➢ Business
  - ➢ Entertainment
  - ➢ Politics
  - ➢ Sports
  - ➢ Technology

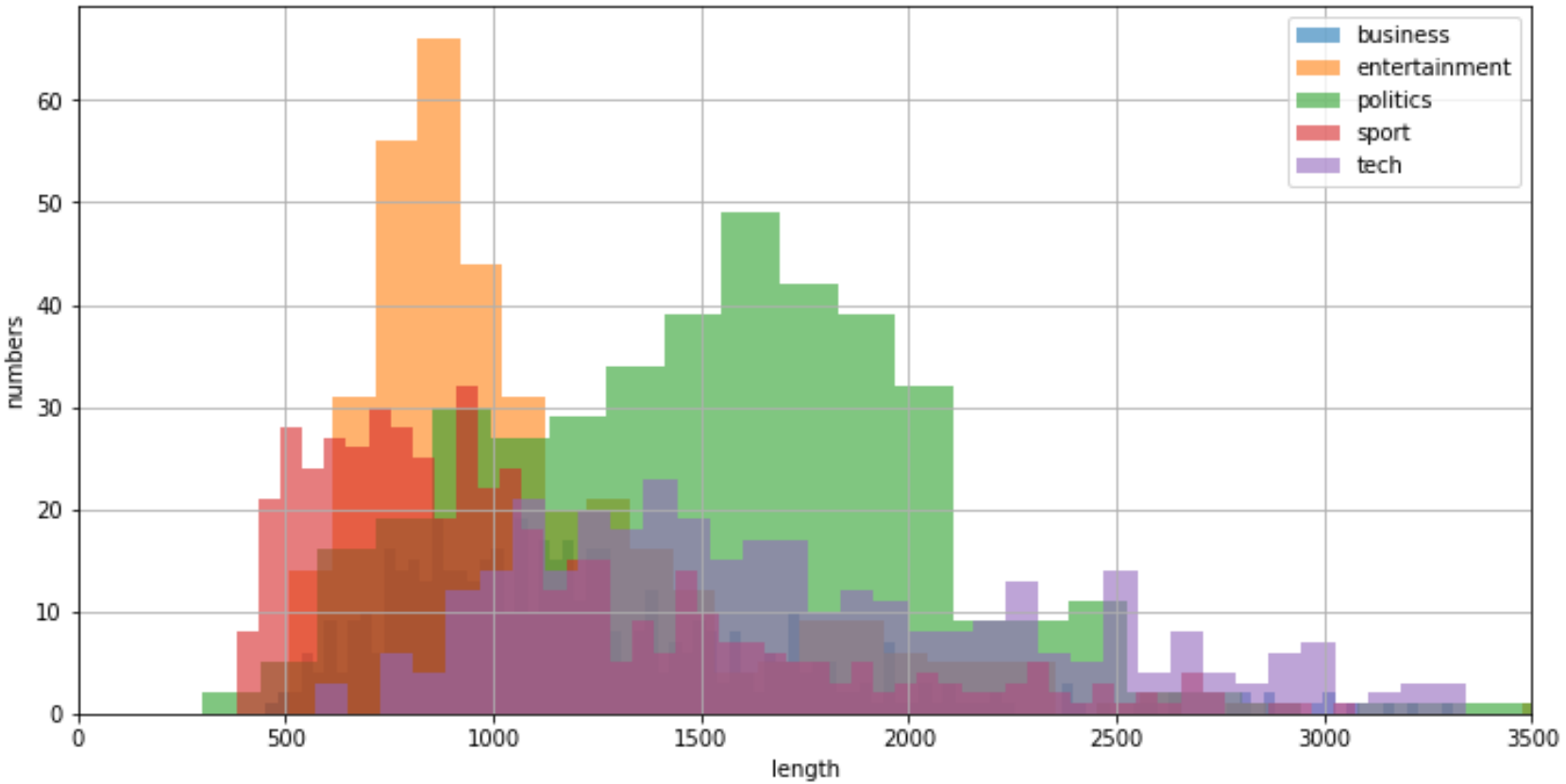


- Dataset consists of total 2225 articles.

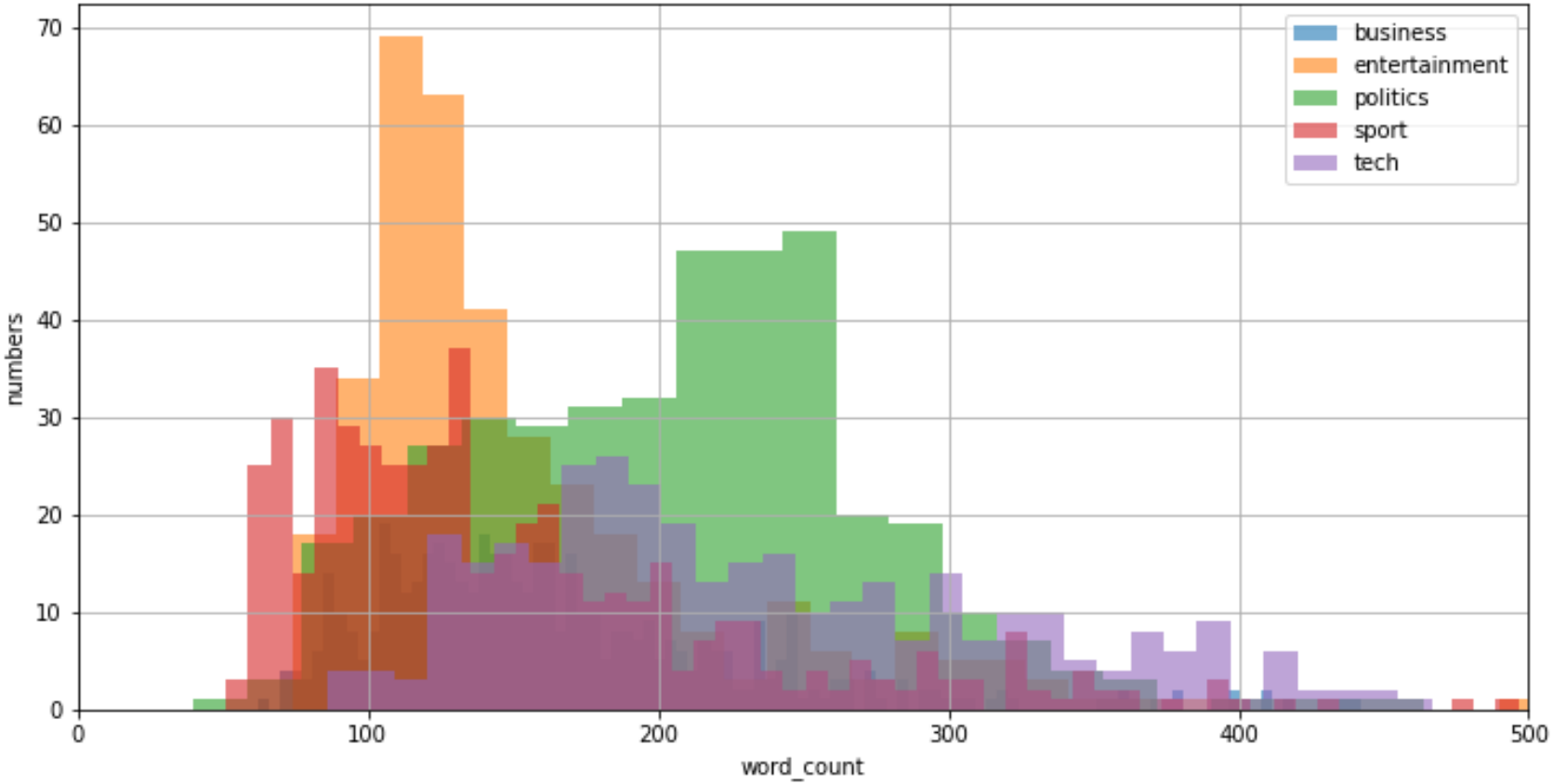
Data Pre-processing

- Remove Html tags and urls
- Convert accented characters to ASCII characters
- Remove punctuations
- Remove numbers
- Split attached words
- Remove small length words
- Remove extra whitespaces
- Spelling corrections
- Lemmatization
- Remove stopwords
- Remove frequent words

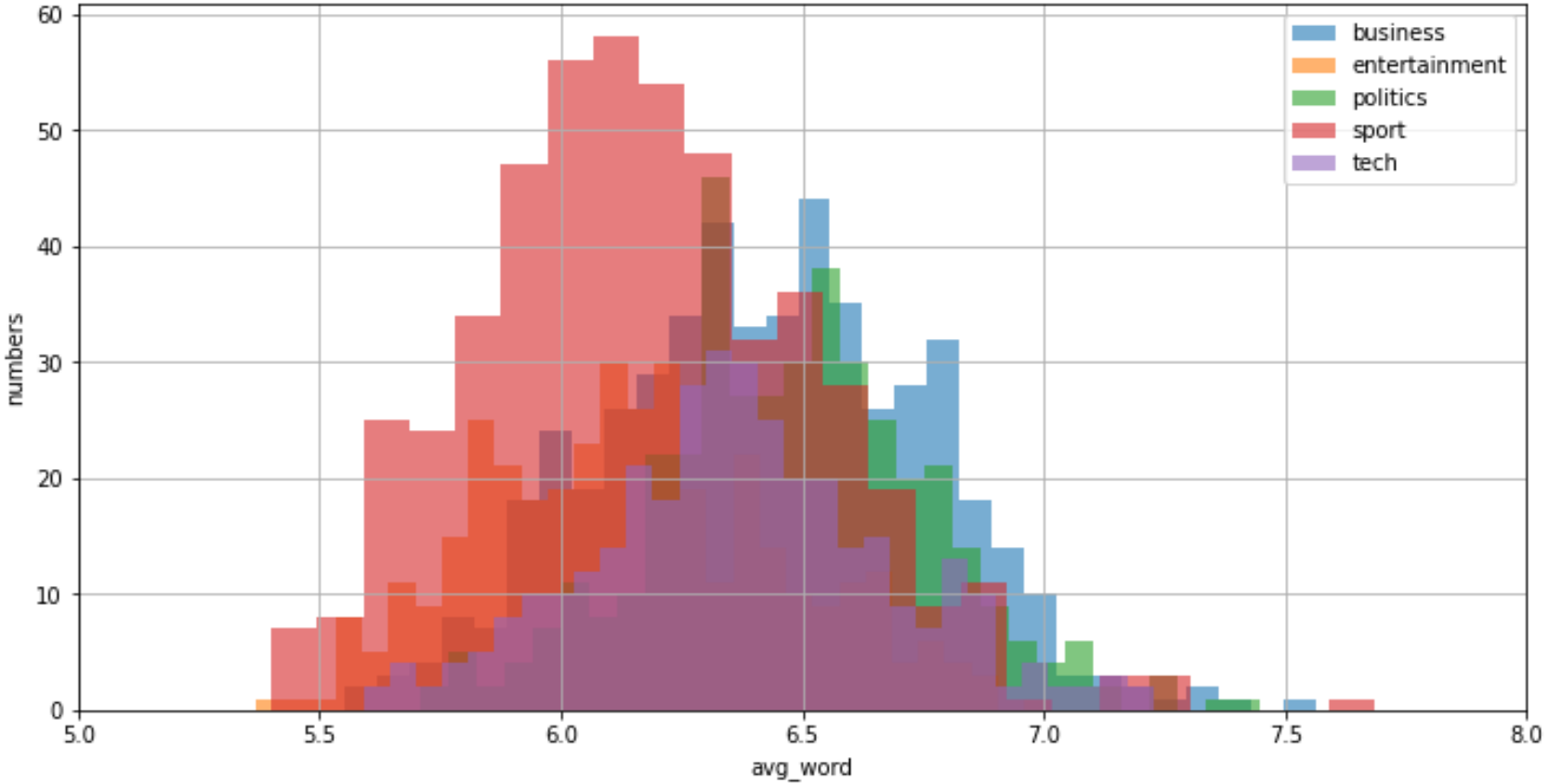
Feature Extraction - Length of documents



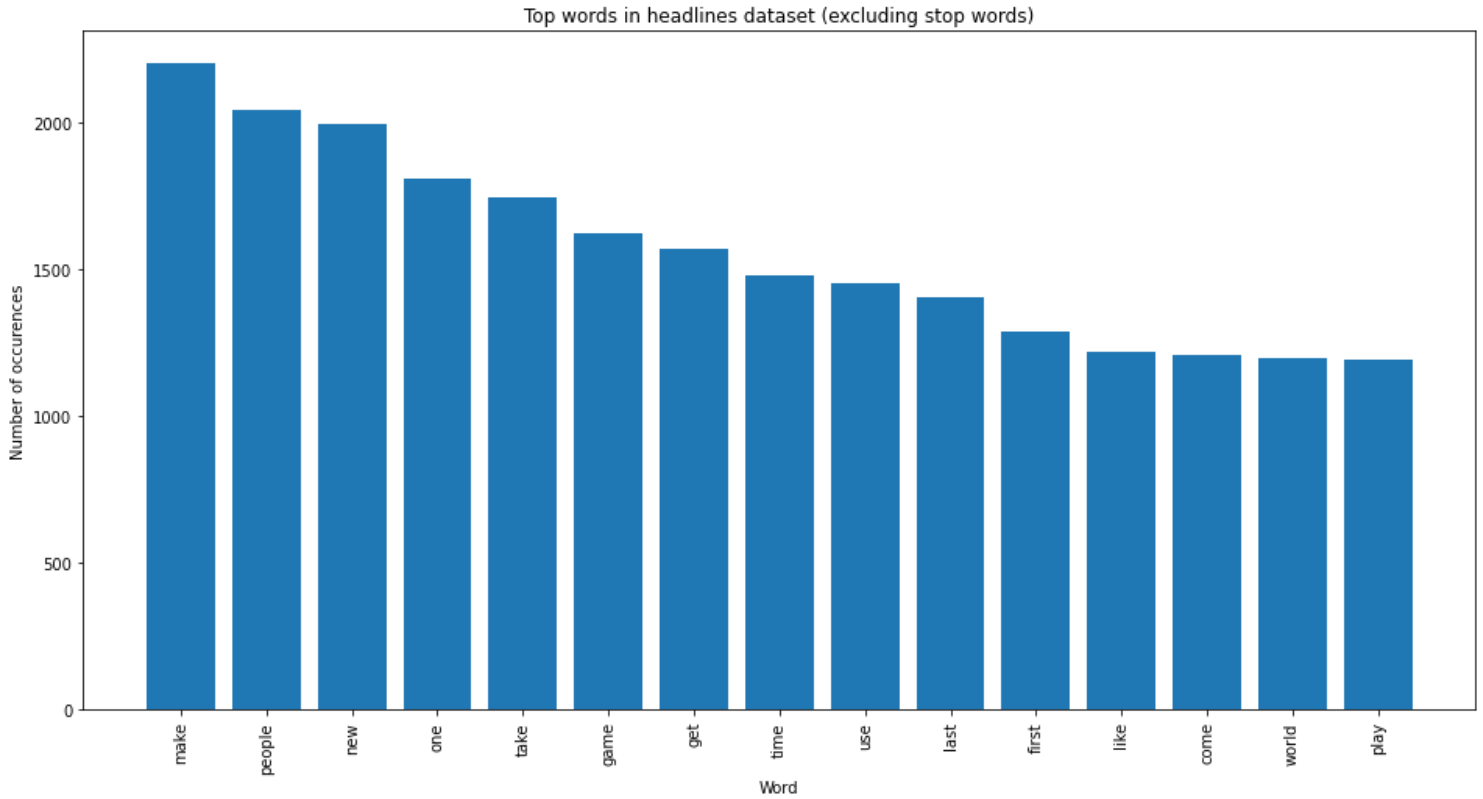
Number of words in documents



Average number of words in documents

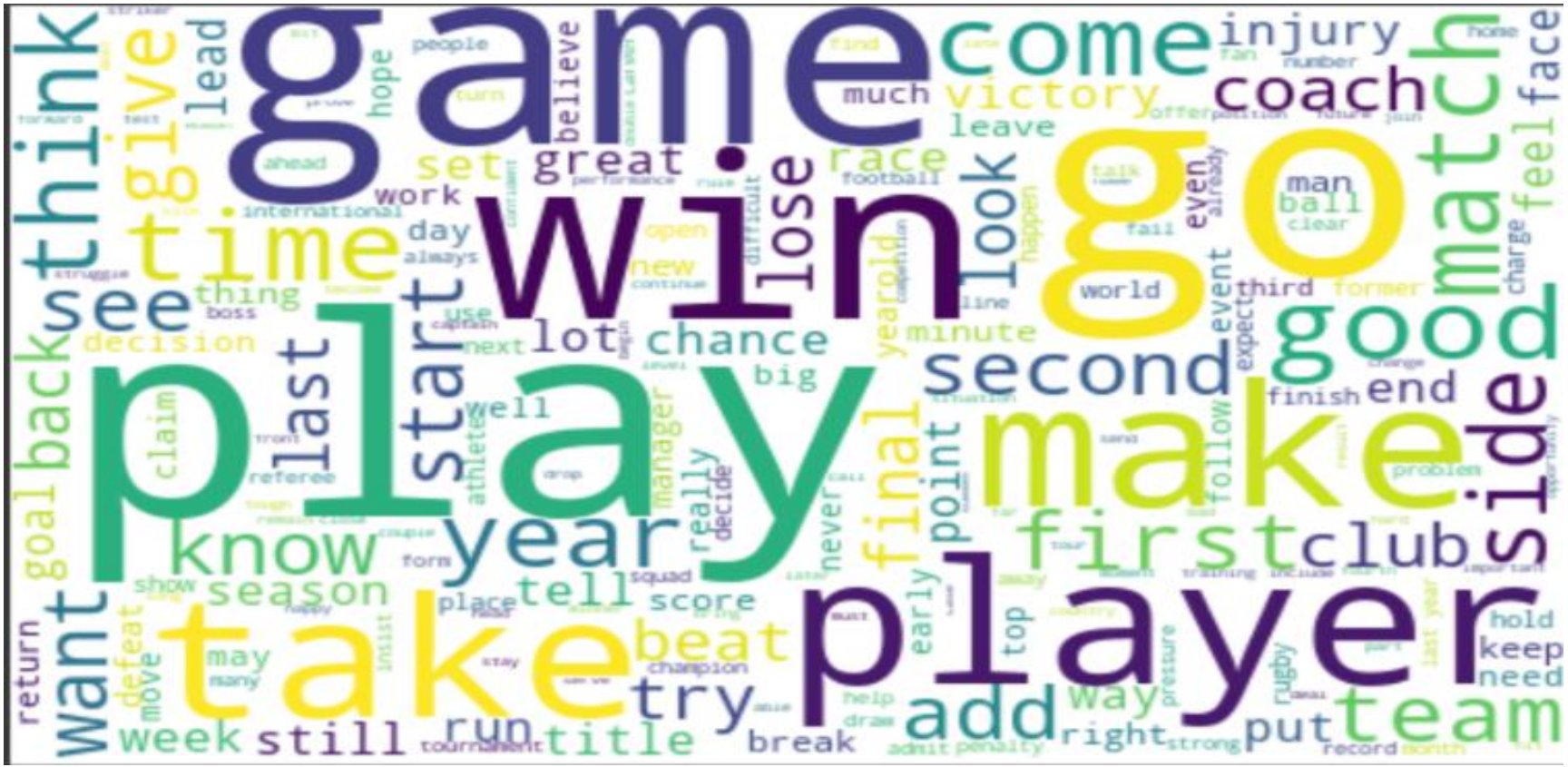


Frequent words in all documents



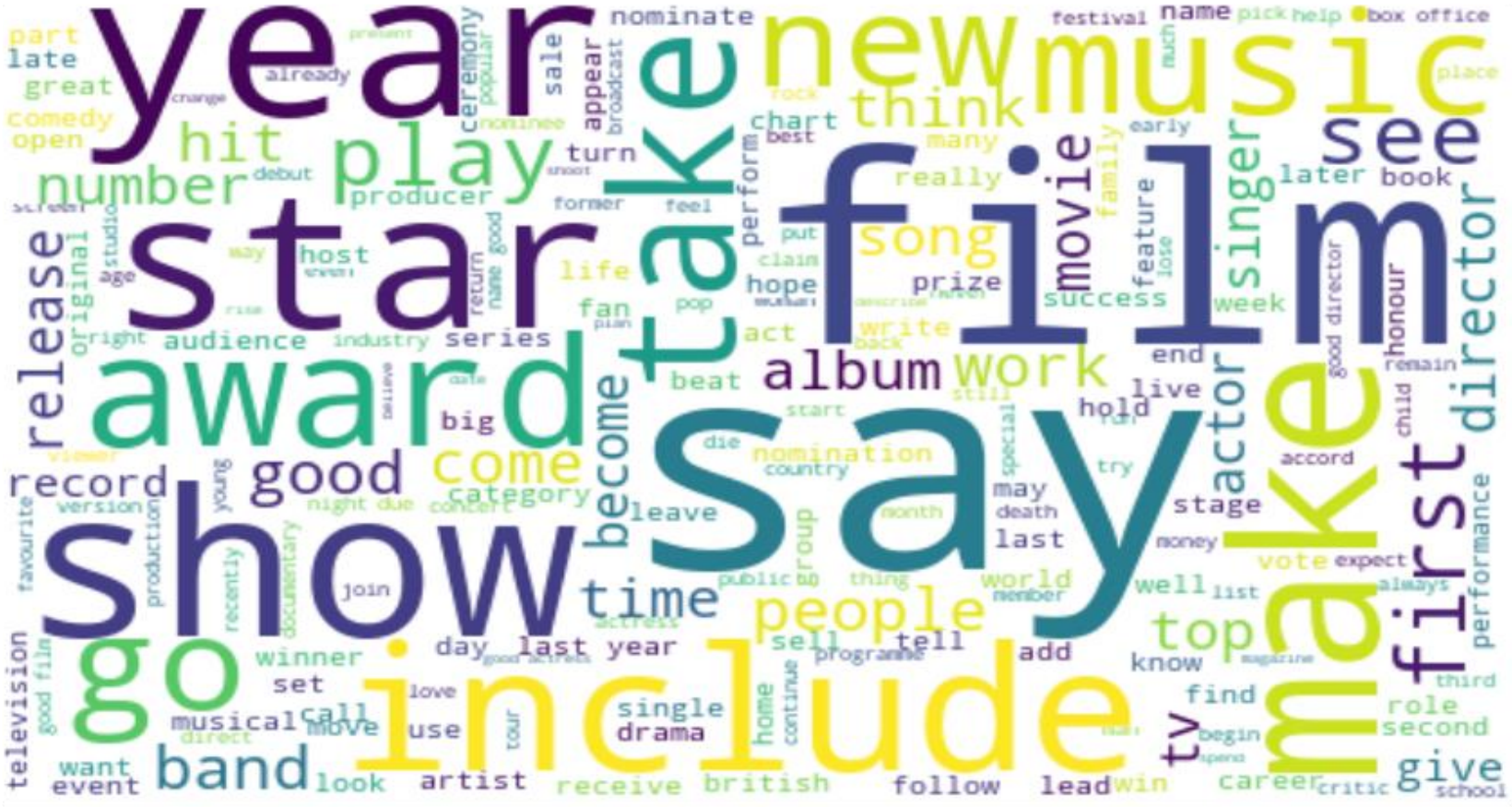
WordCloud - Business

WordCloud - Sport



WordCloud - Tech

WordCloud - Entertainment

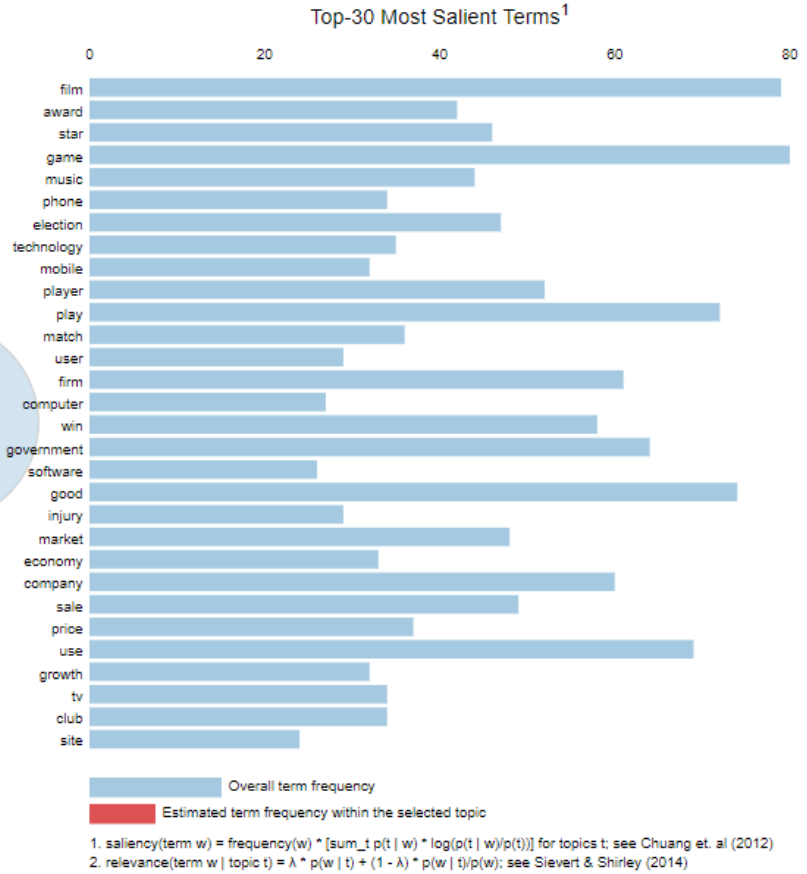


WordCloud - Politics

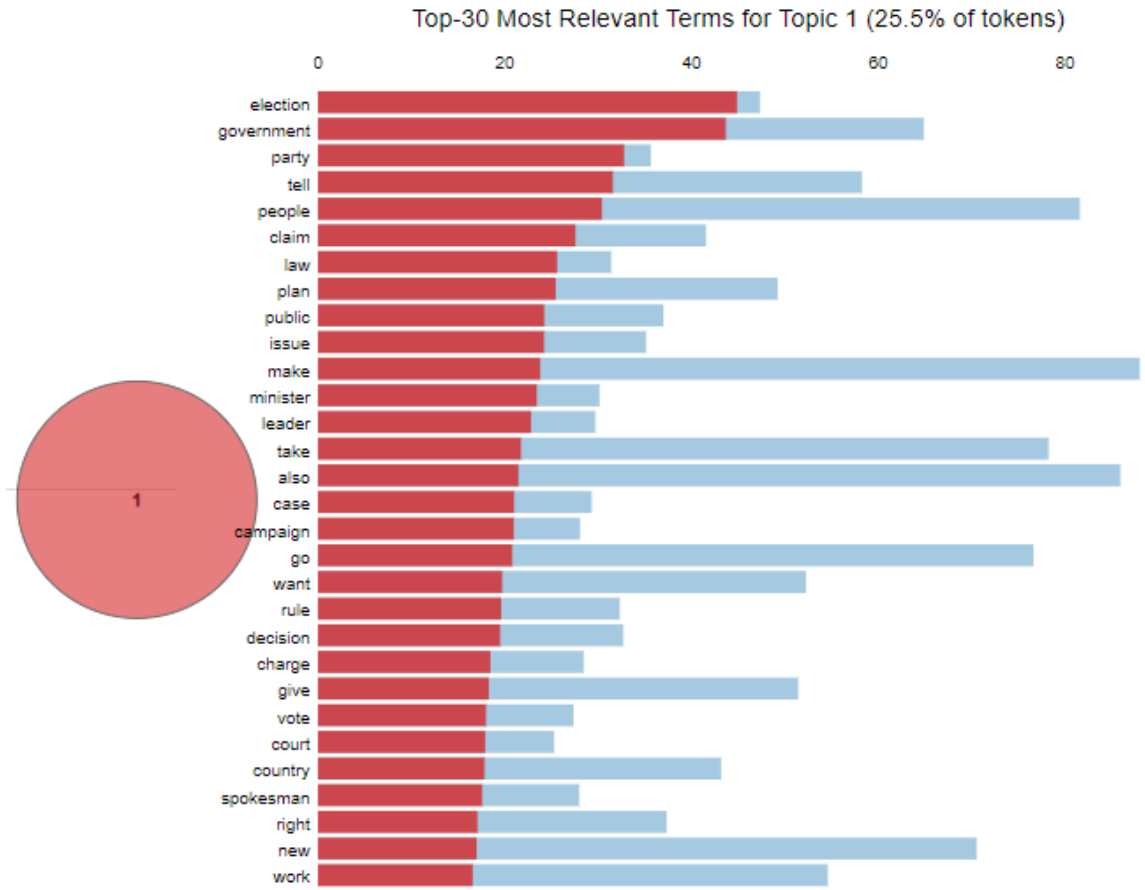
Implementation of ML Models

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

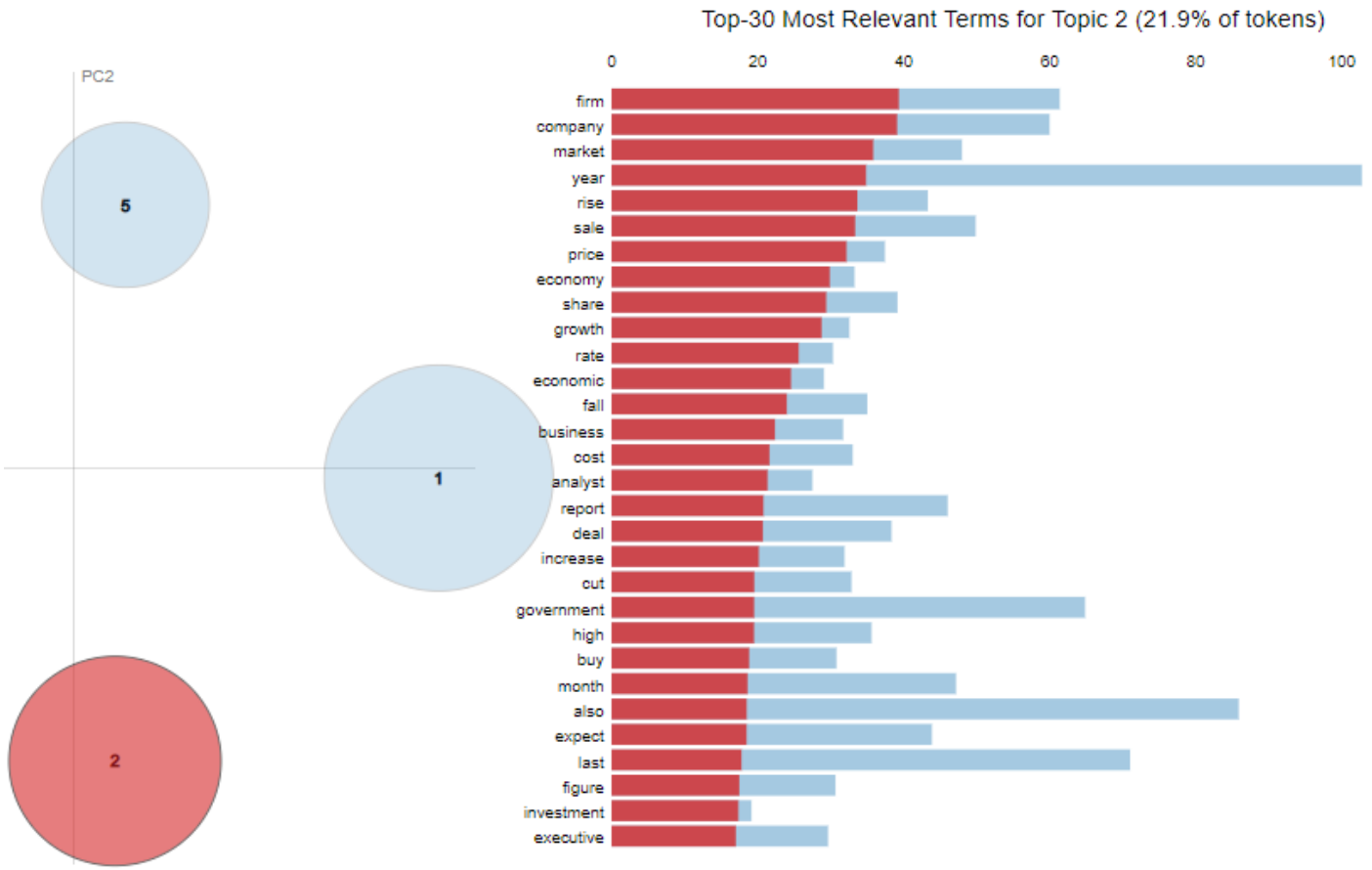
Latent Dirichlet Allocation (LDA)



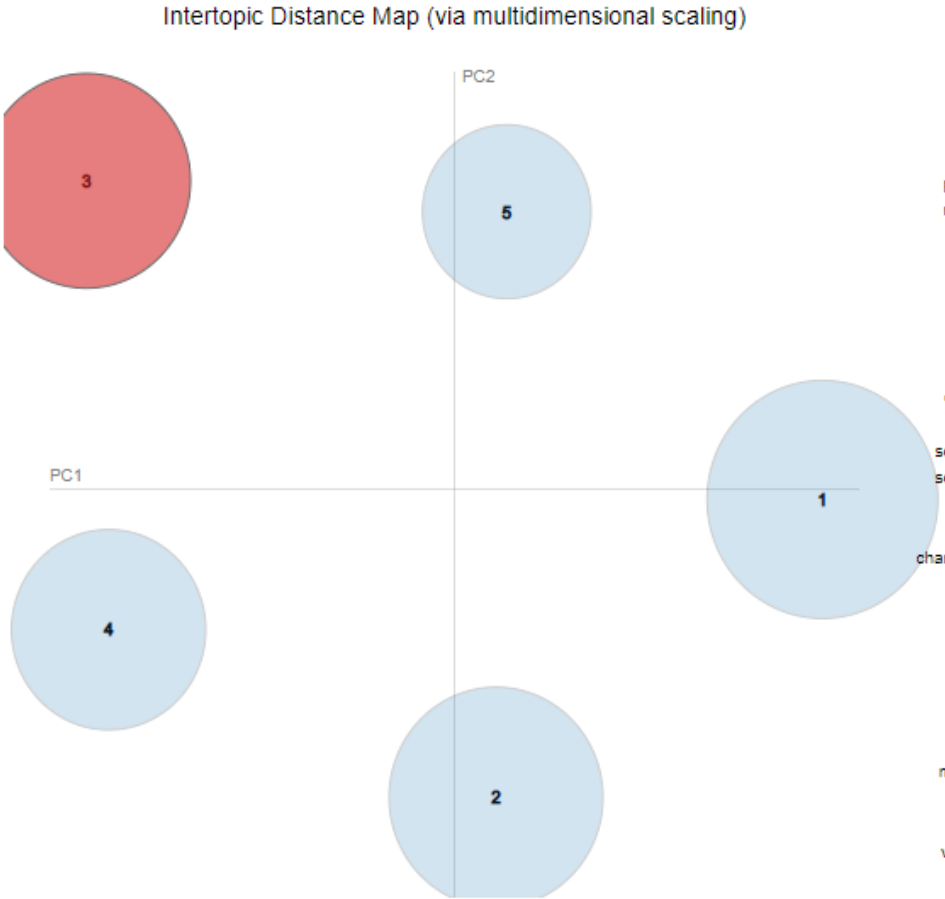
LDA - Cluster 1 : Politics

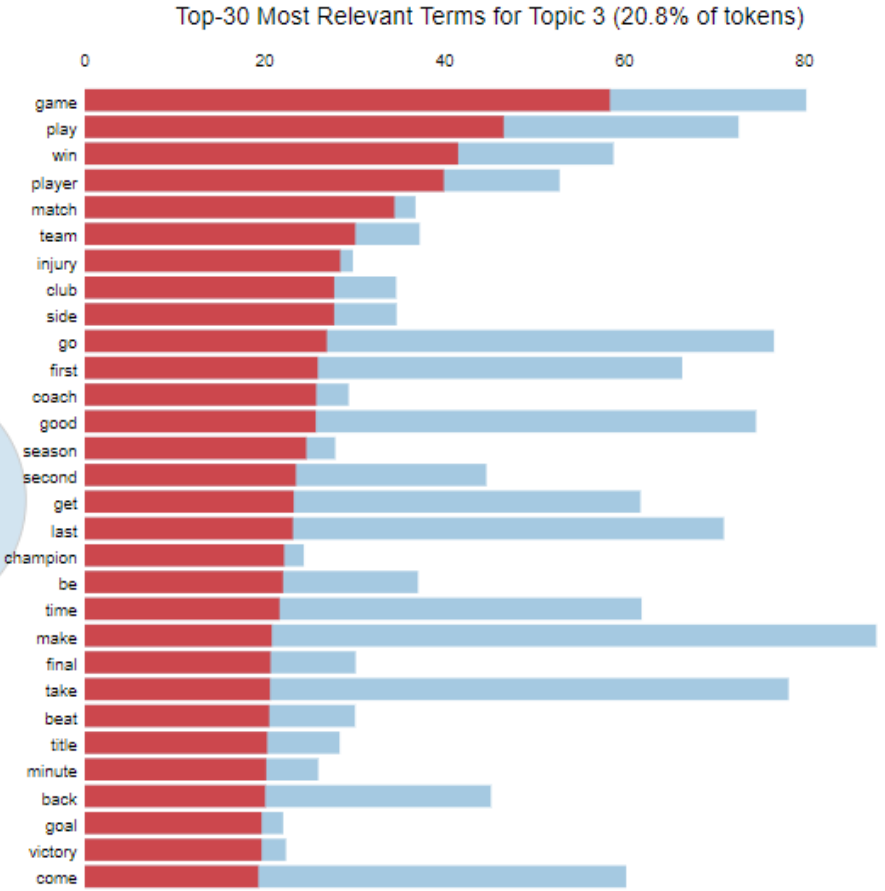


LDA - Cluster 2 : Business

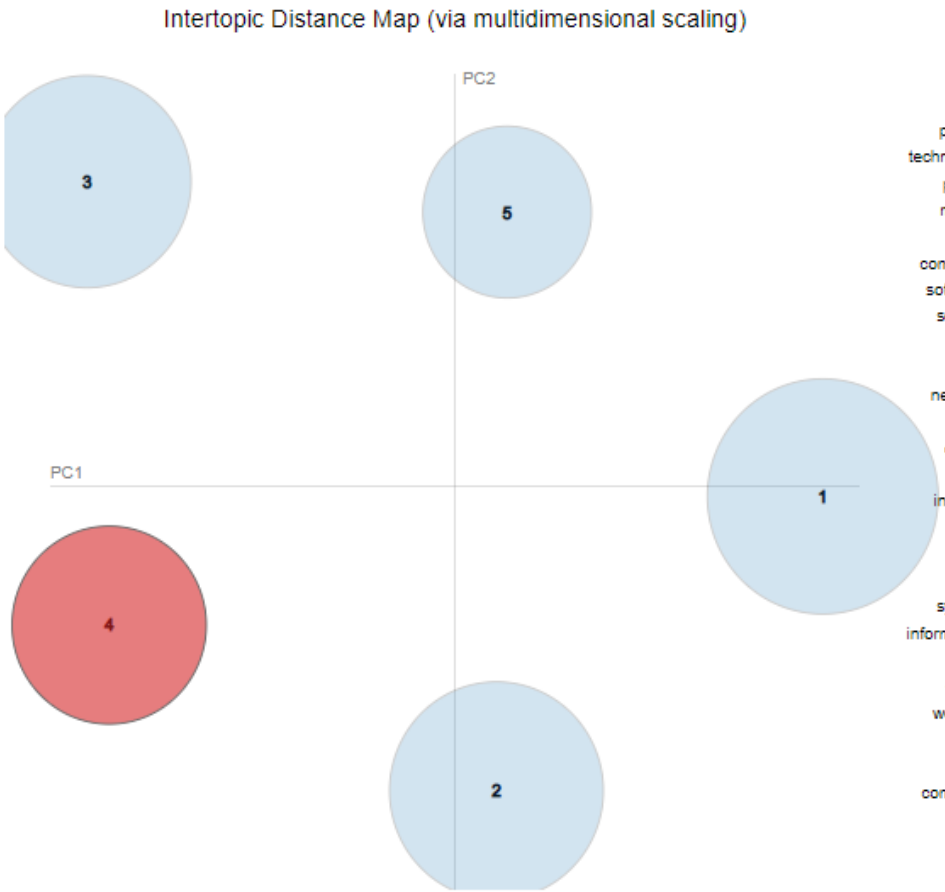


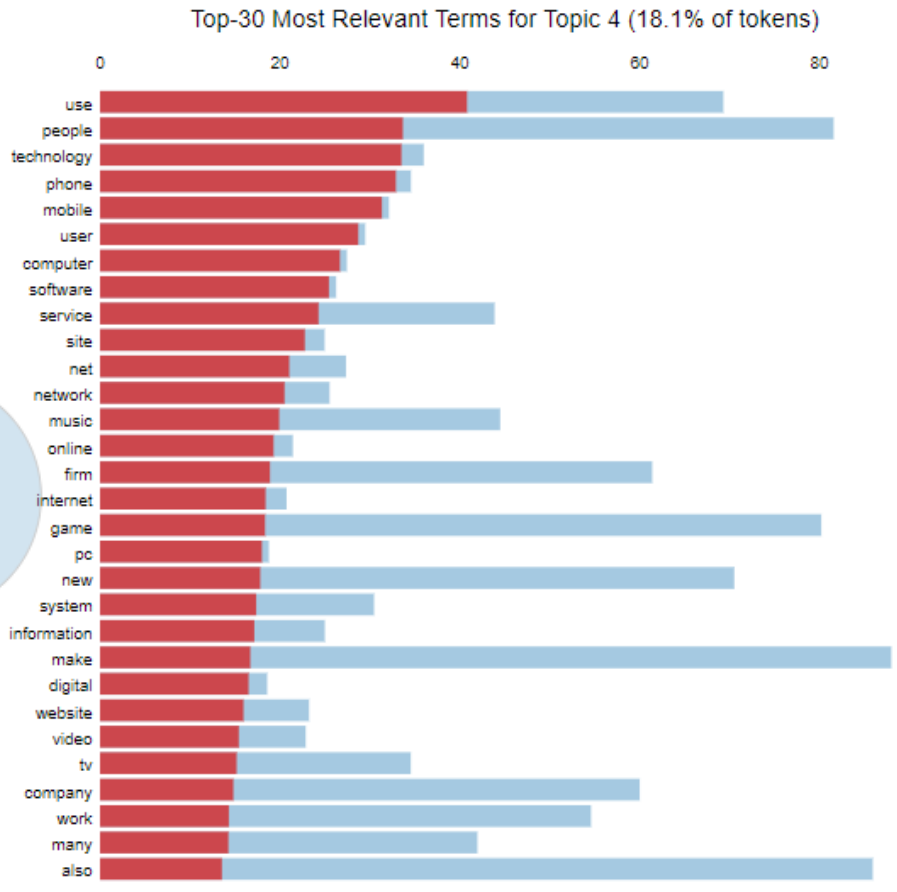
LDA - Cluster 3 : Sport



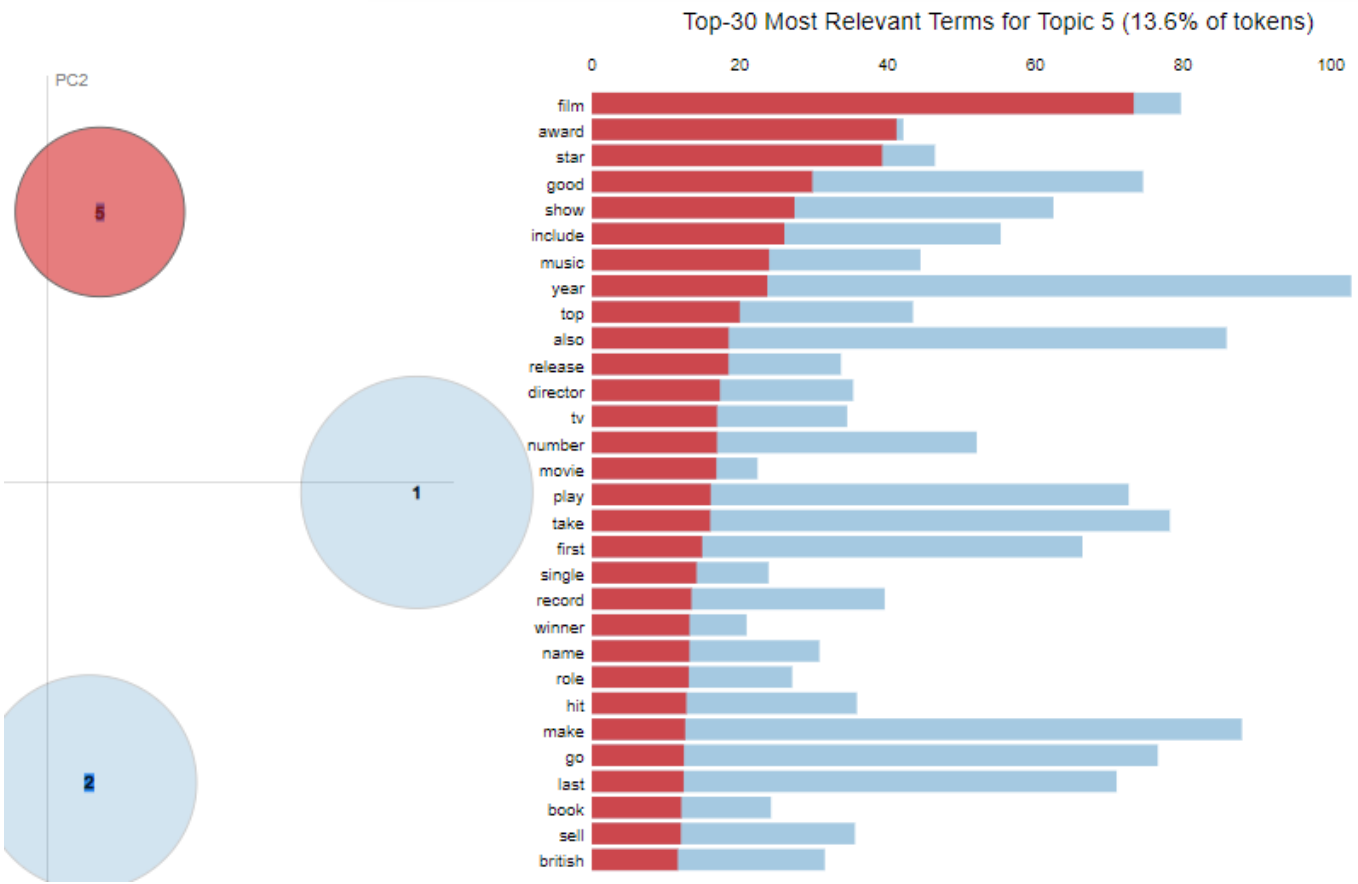


LDA - Cluster 4 : Tech

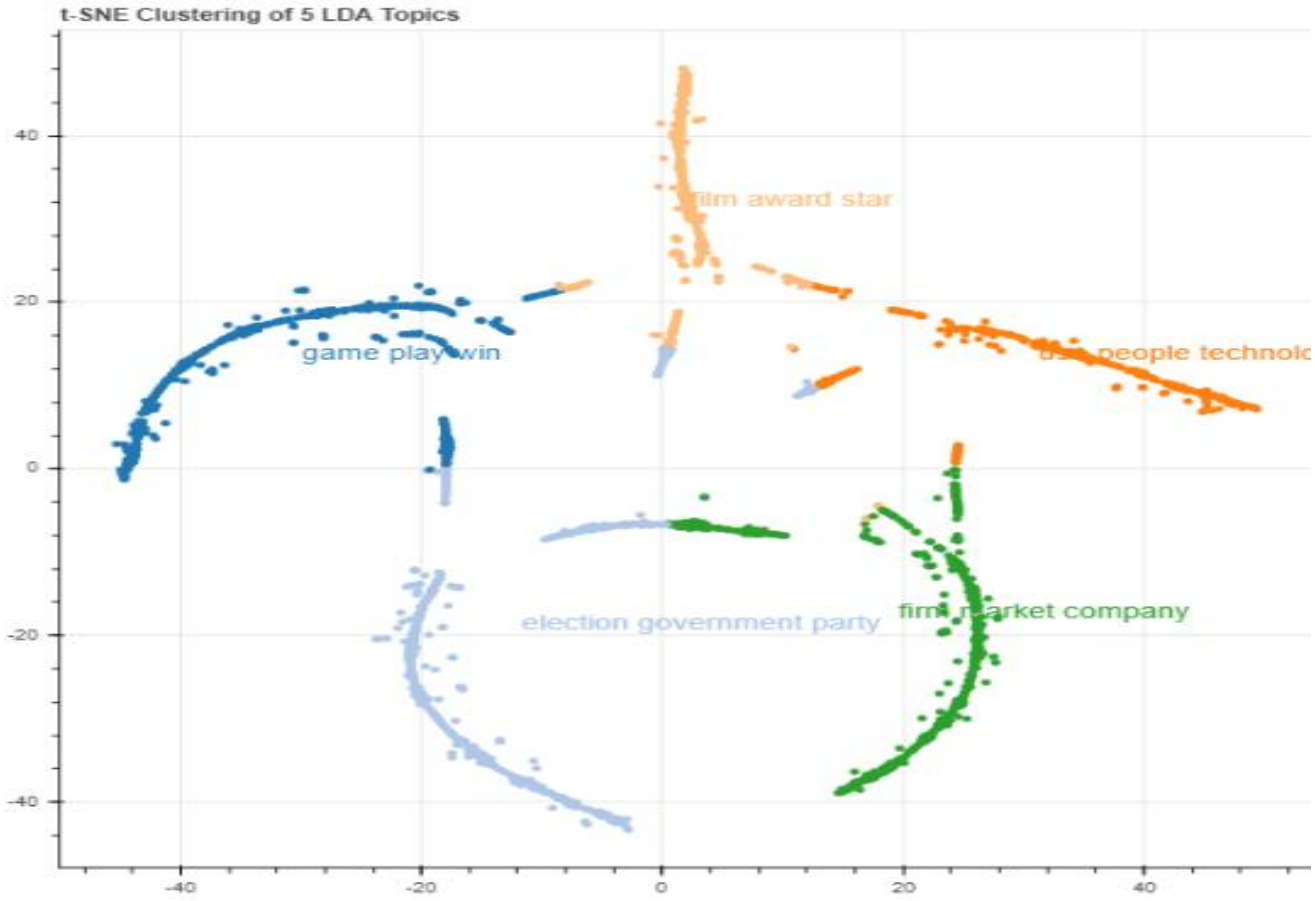




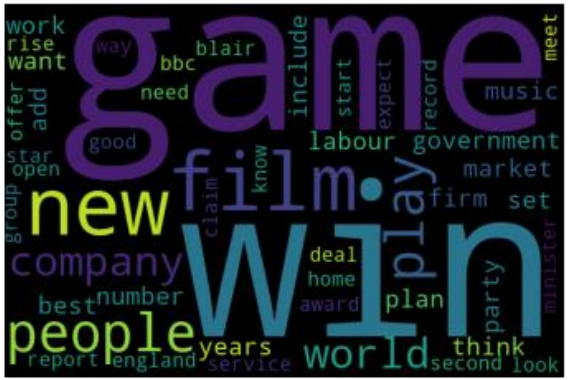
LDA - Cluster 5 : Entertainment

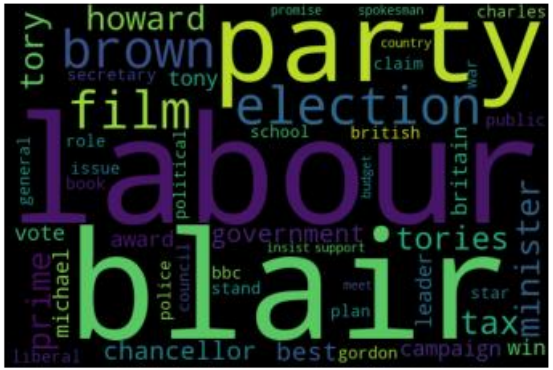


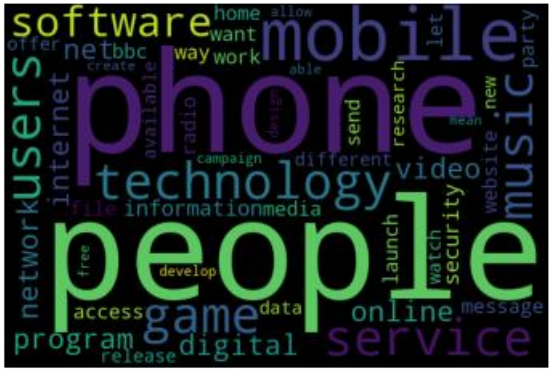
LDA - t-SNE Clustering



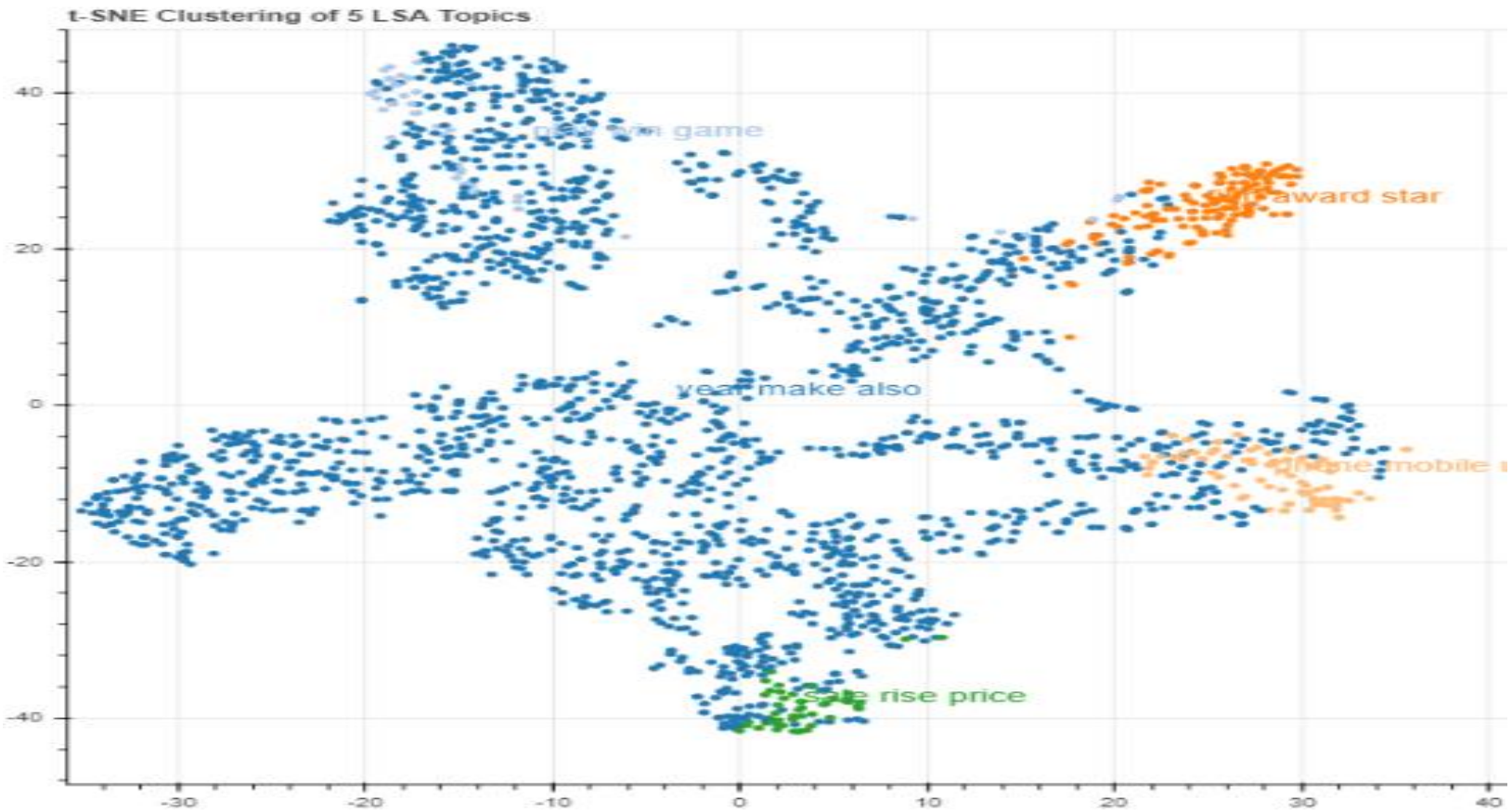
Latent Semantic Analysis (LSA)



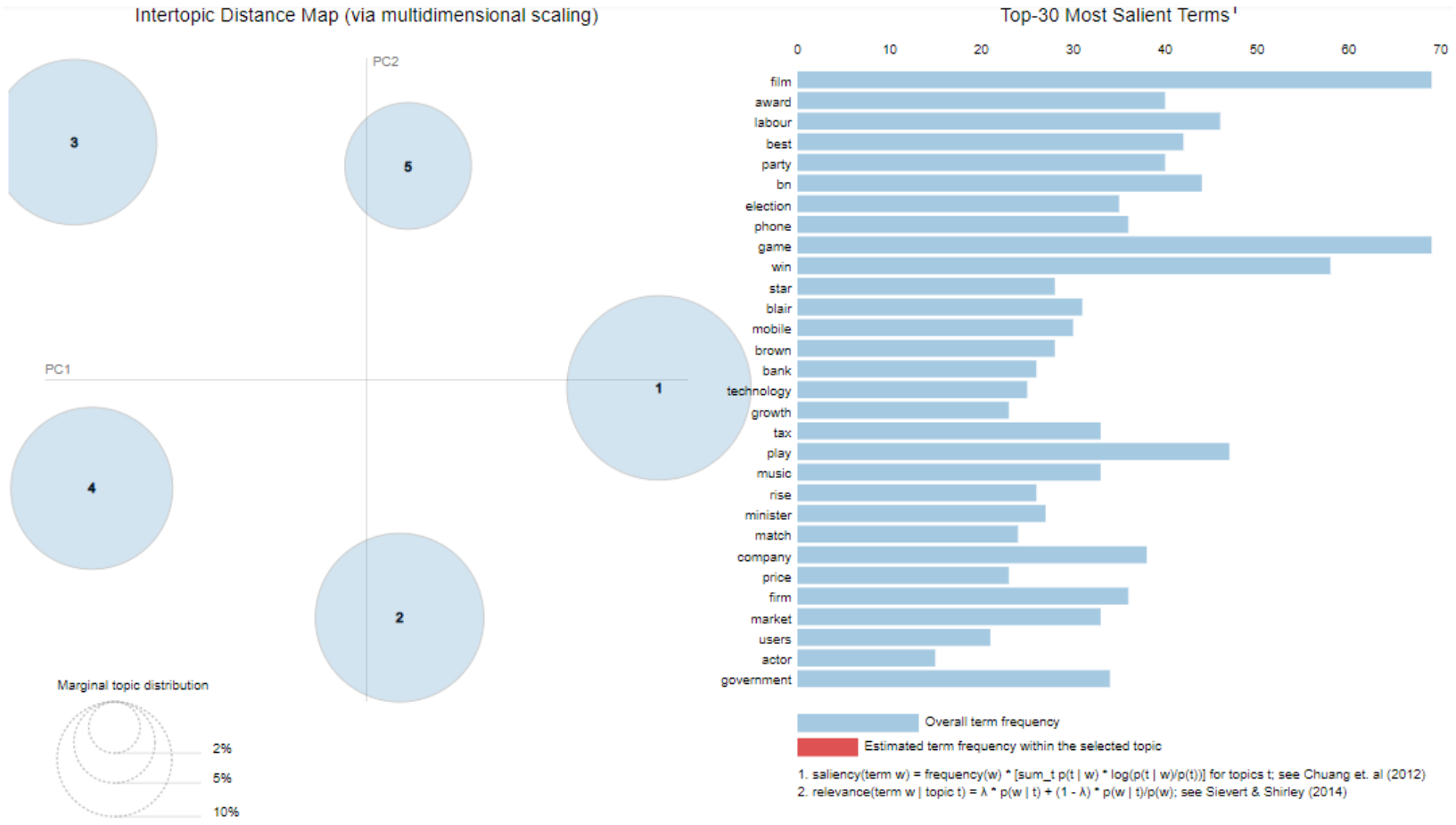




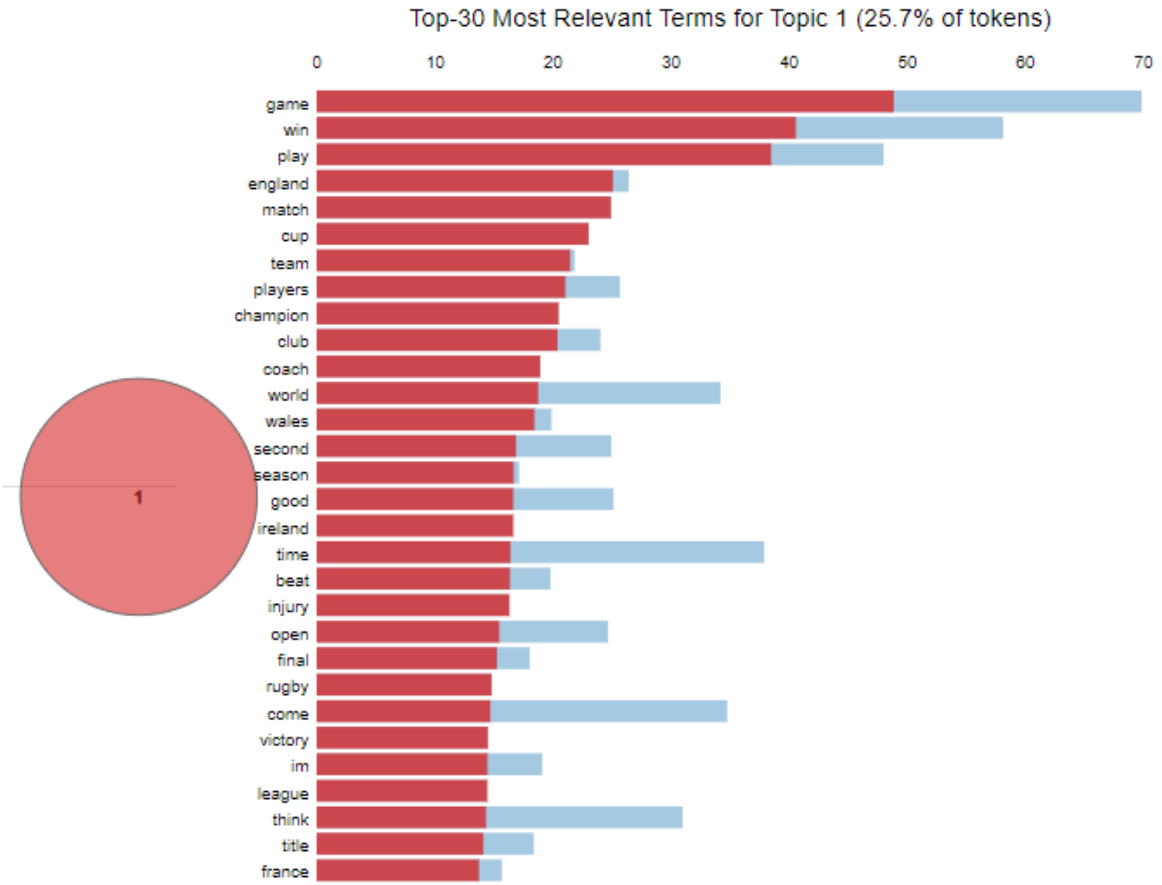
LSA - t-SNE Clustering



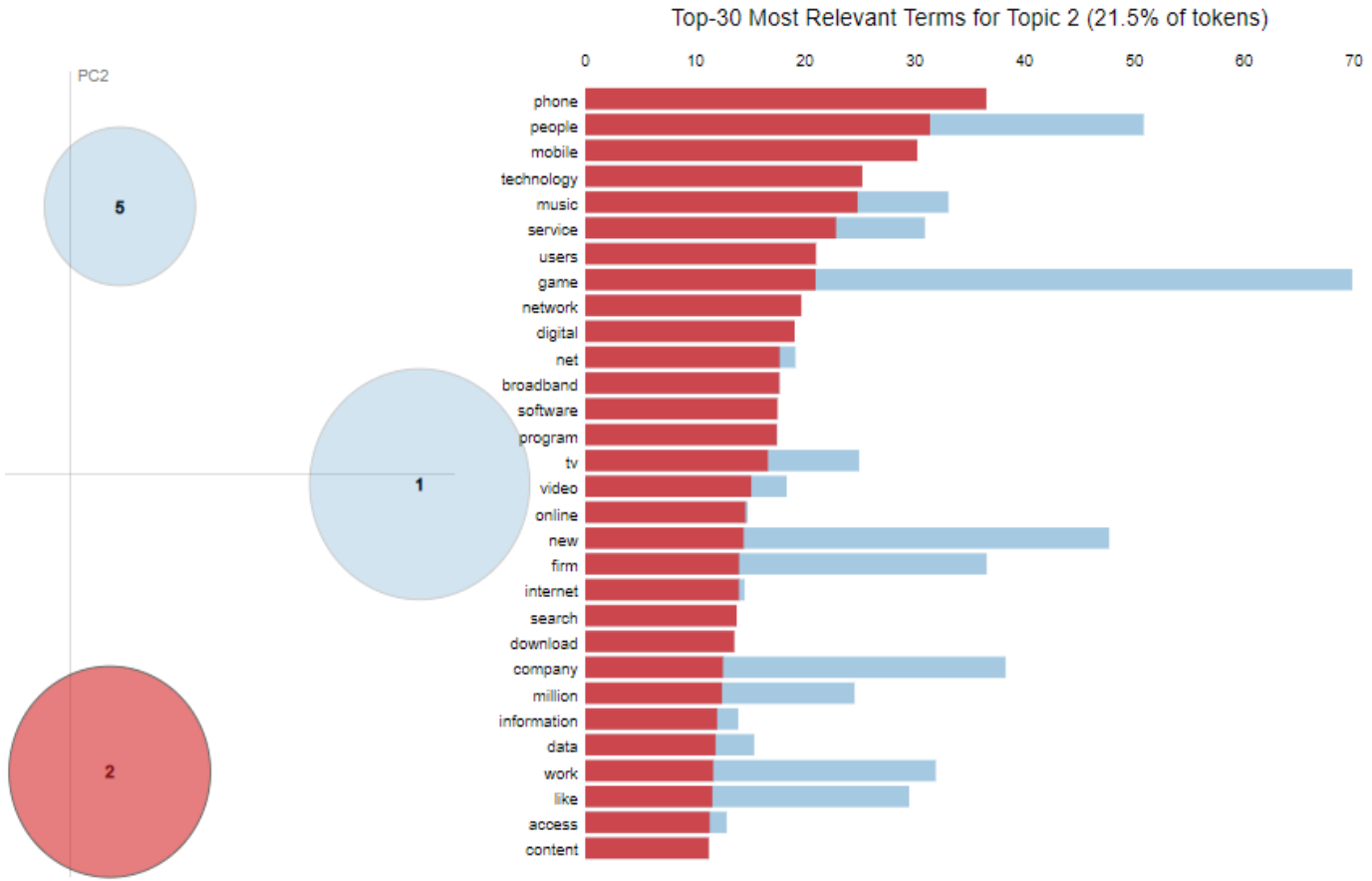
Non-negative Matrix Factorization (NMF)



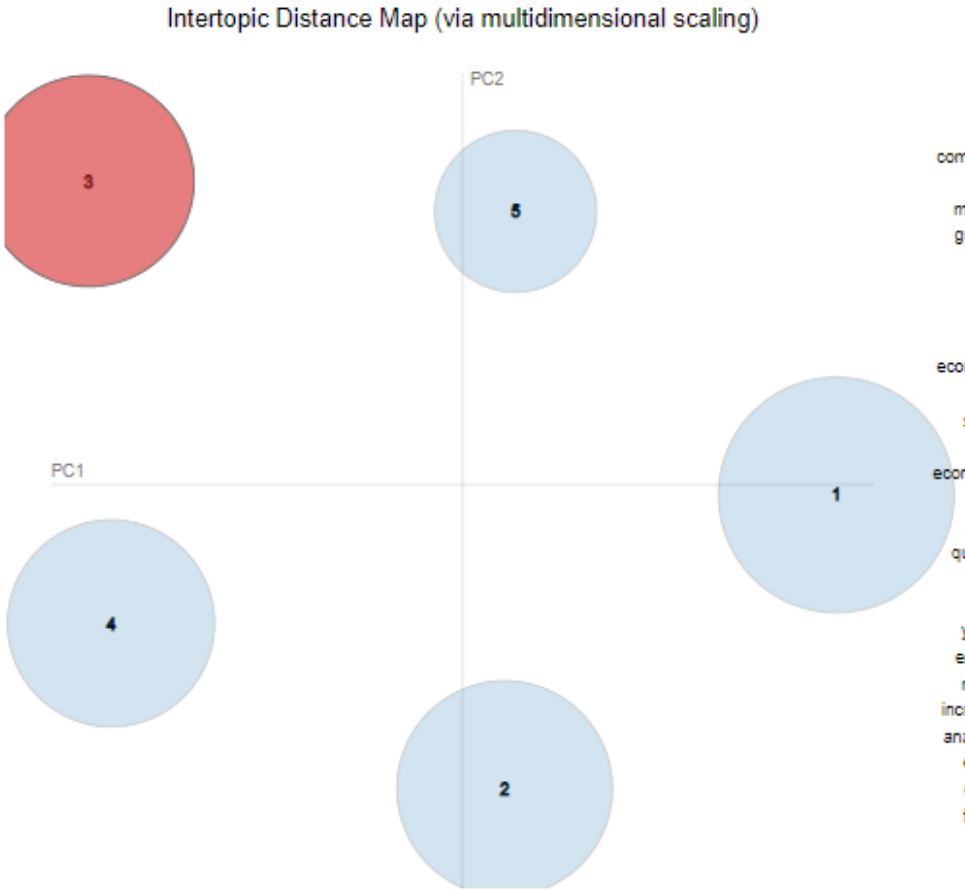
NMF - Cluster 1 : Sport

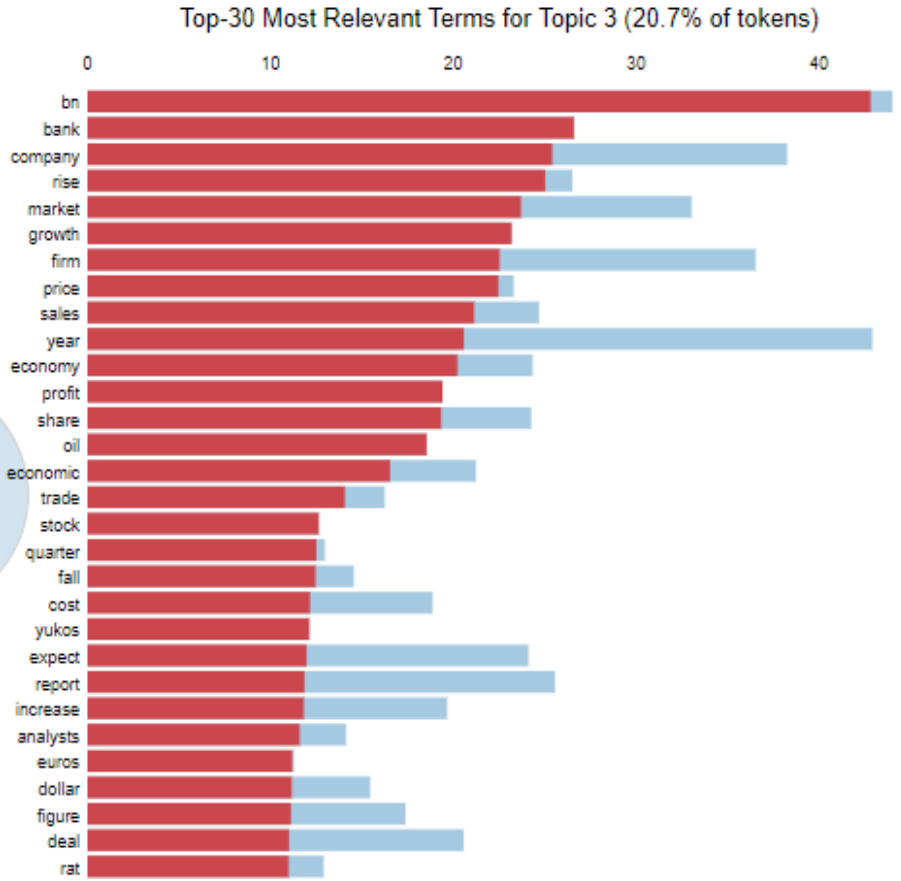


NMF - Cluster 2 : Tech

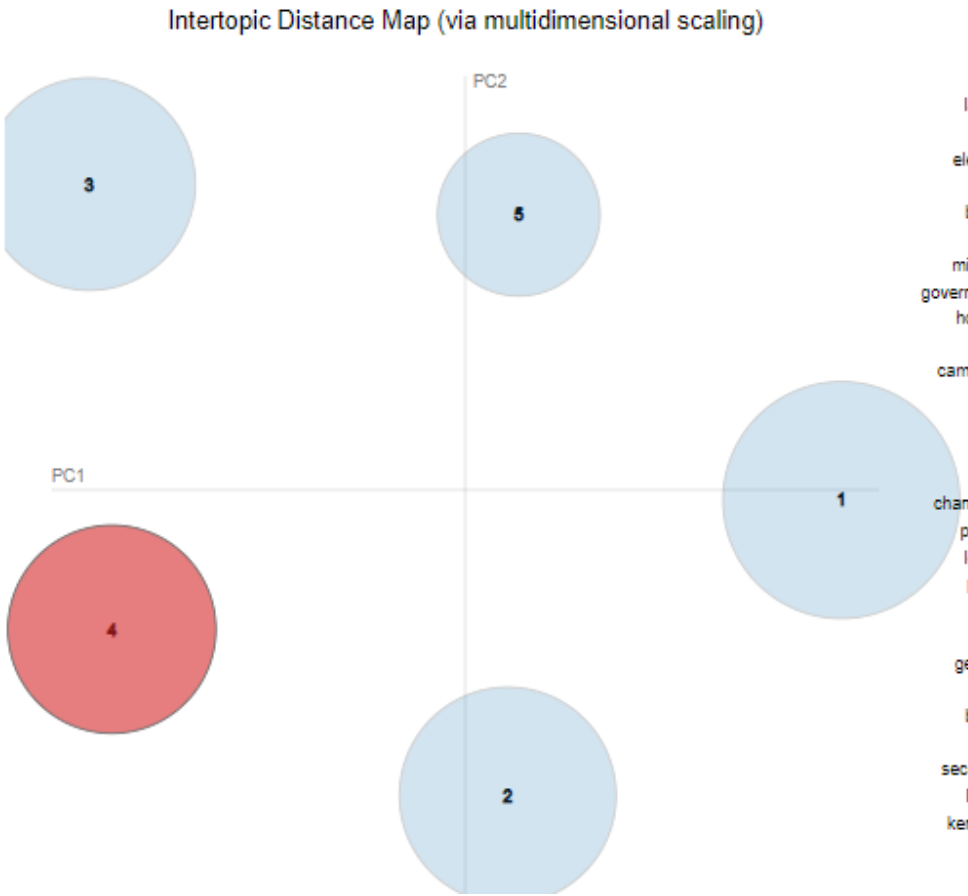


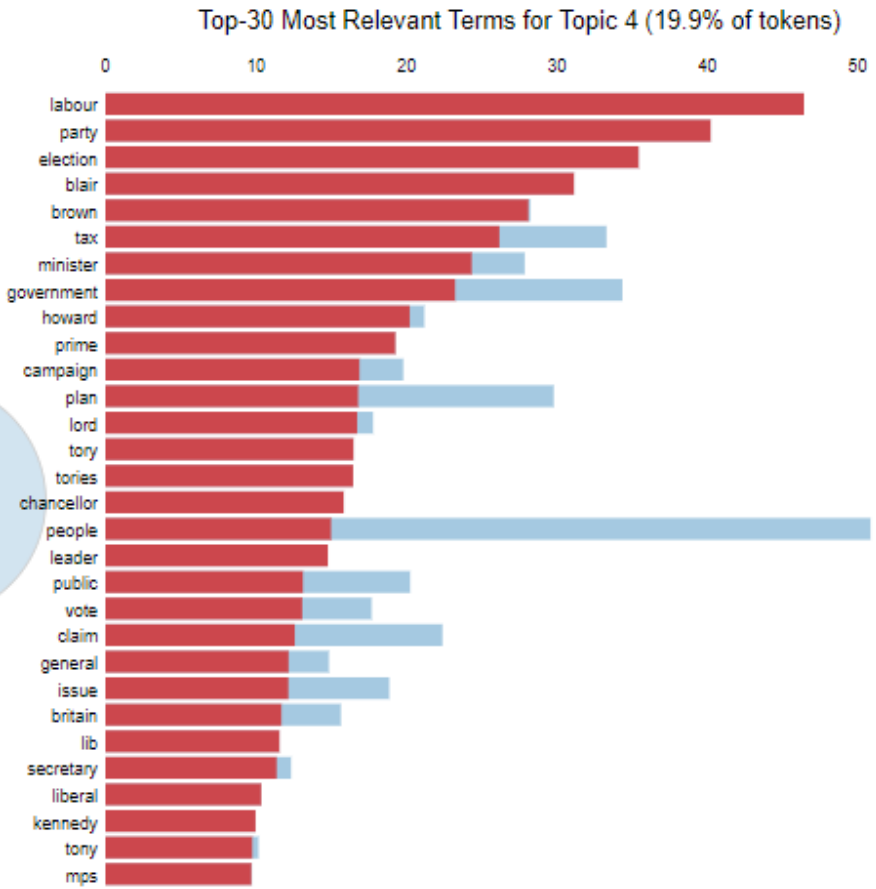
NMF - Cluster 3 : Business



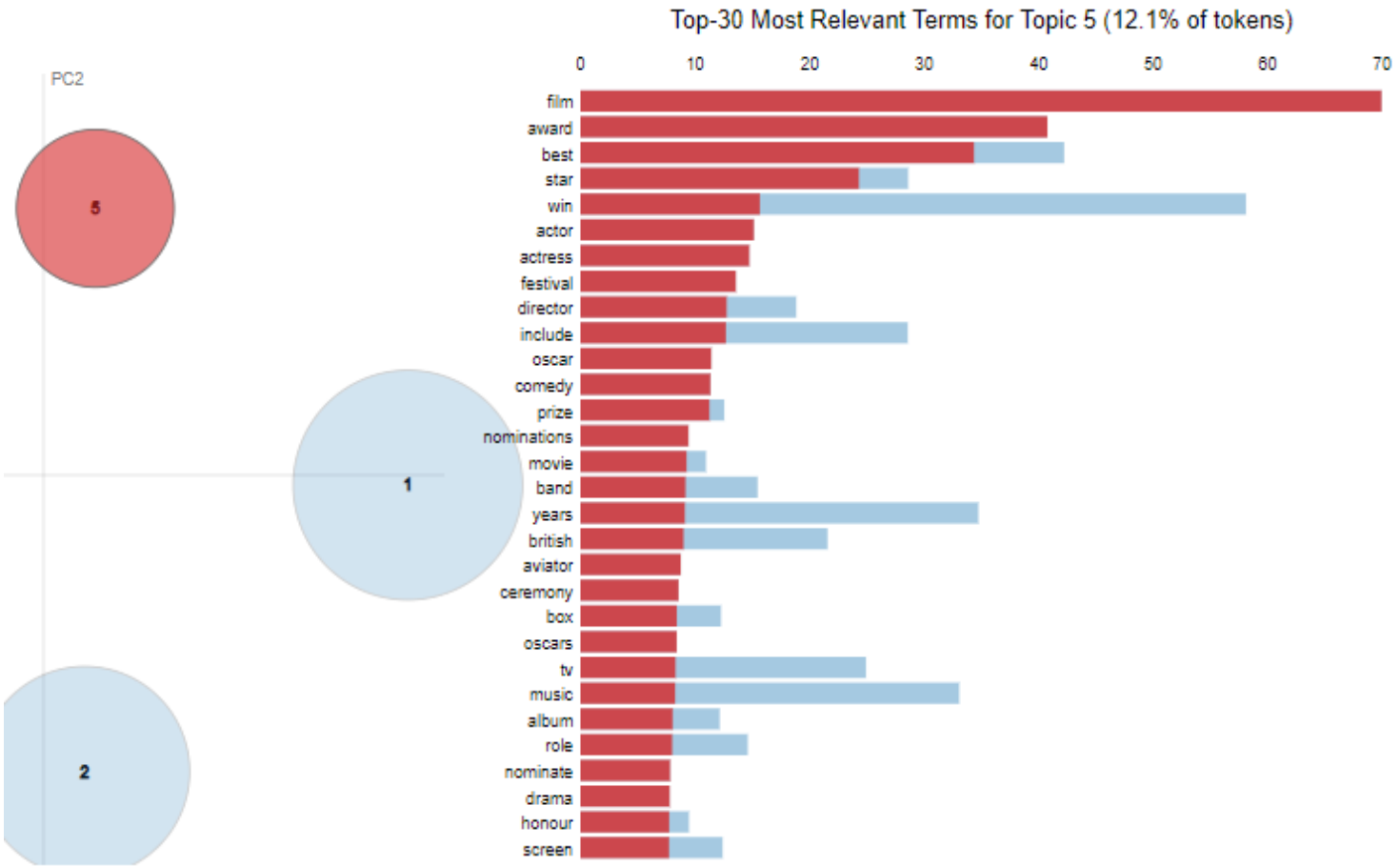


NMF - Cluster 4 : Politics





NMF - Cluster 5 : Entertainment



Challenges

- Must read 2000+ text files and formulate a Dataset to work with.
- Some text pre-processing technique took too much time to execute (autocorrect)
- 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