Capstone Project - IV
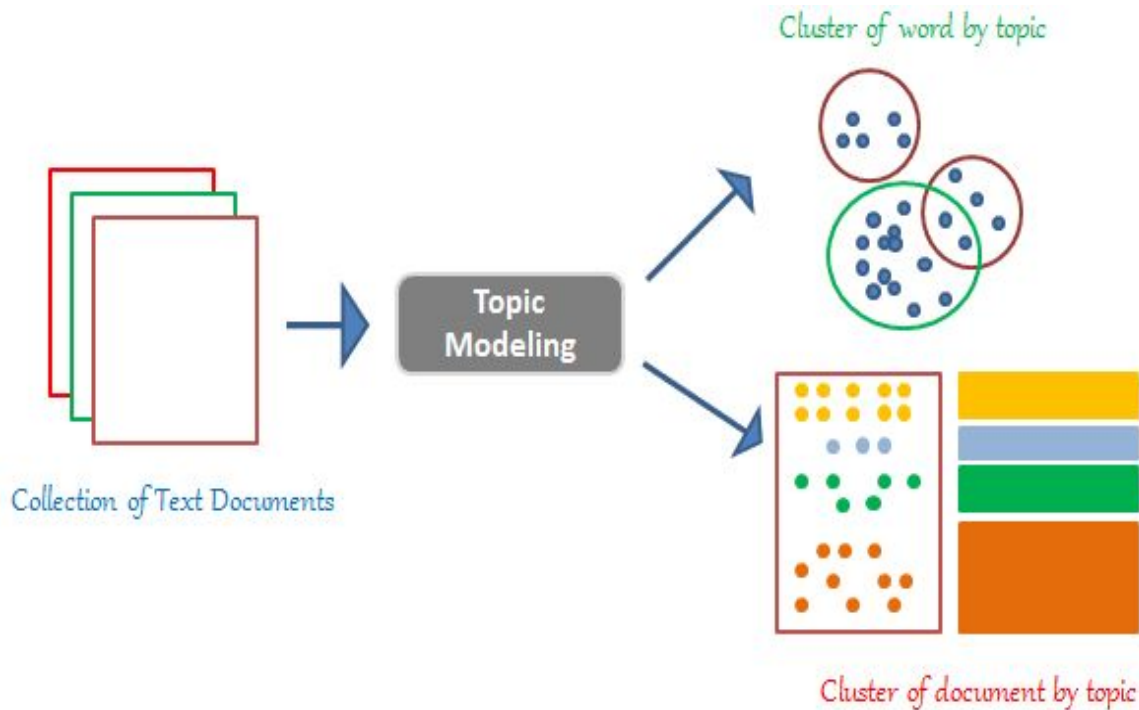
Team 3 : Topic Modeling on News Articles

Content

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

Cluster of word by topic

Topic Modeling

Collection of Text Documents

Cluster of document by topic

AI

Problem Statement

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

Data Summary
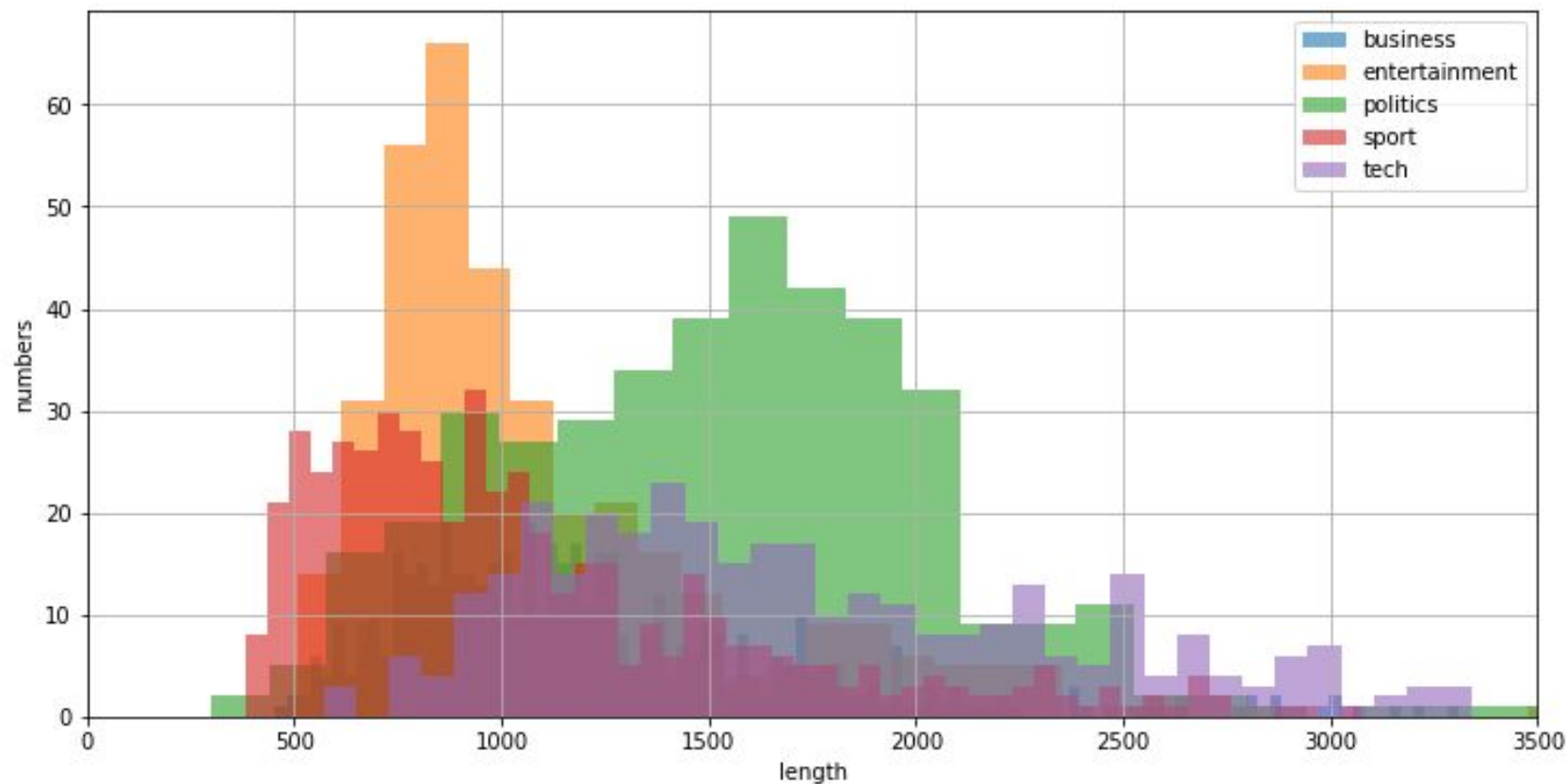
- There are total 5 topics -

  - ➢ Business
  - ➢ Entertainment
  - ➢ Politics
  - ➢ Sports
  - ➢ Technology
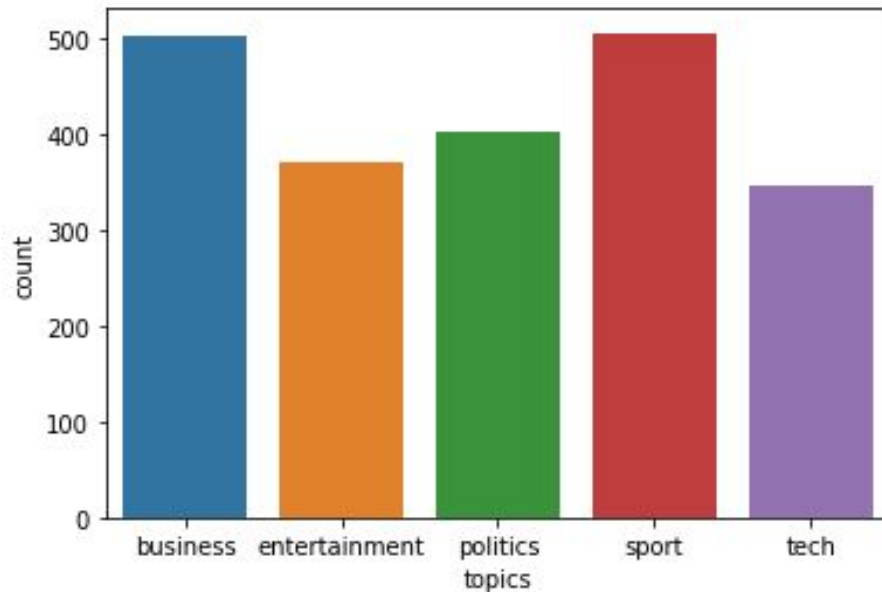
- Dataset consists of total 2225 articles.

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
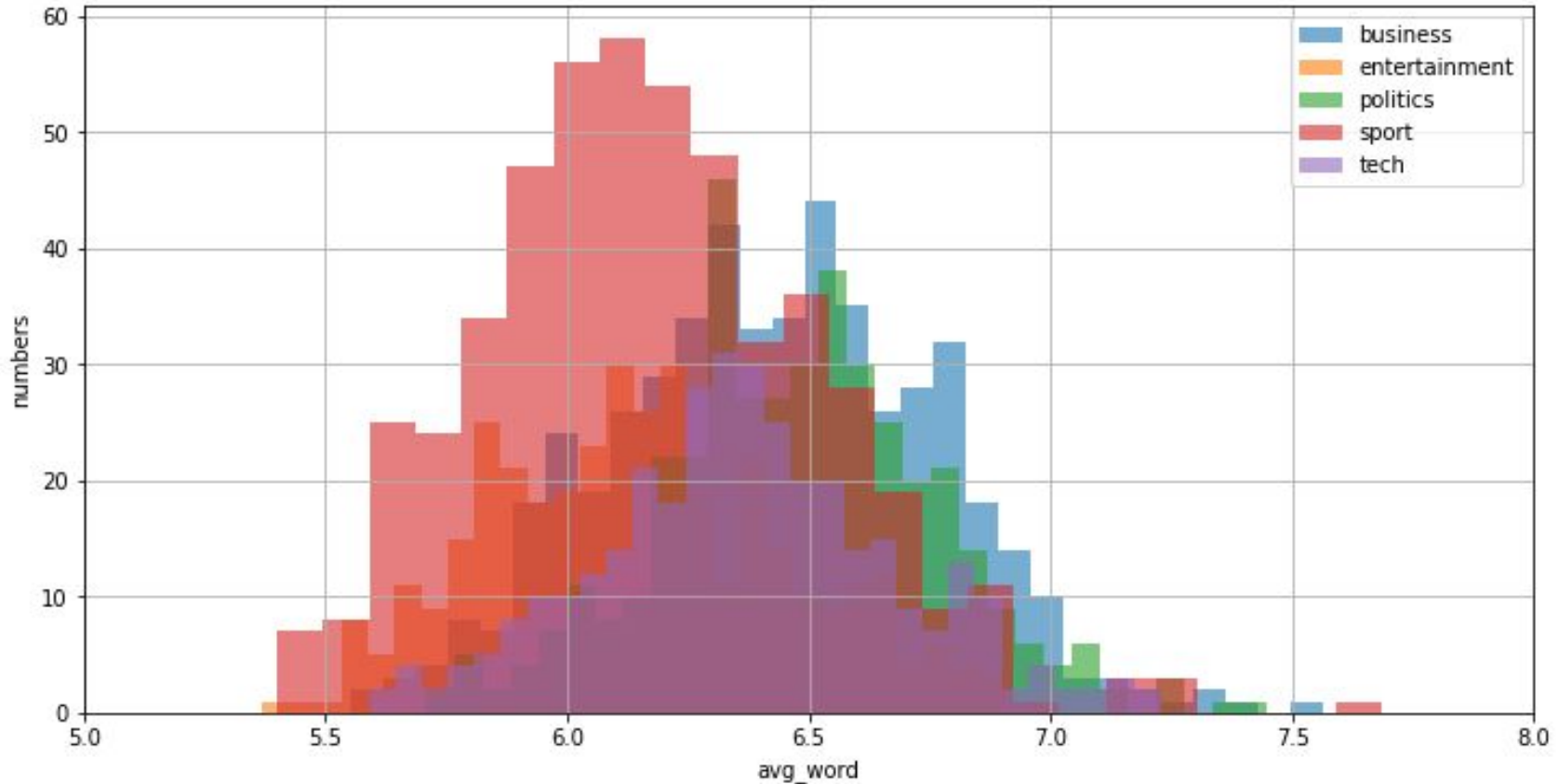
Feature Extraction - Length of documents

Number of words in documents
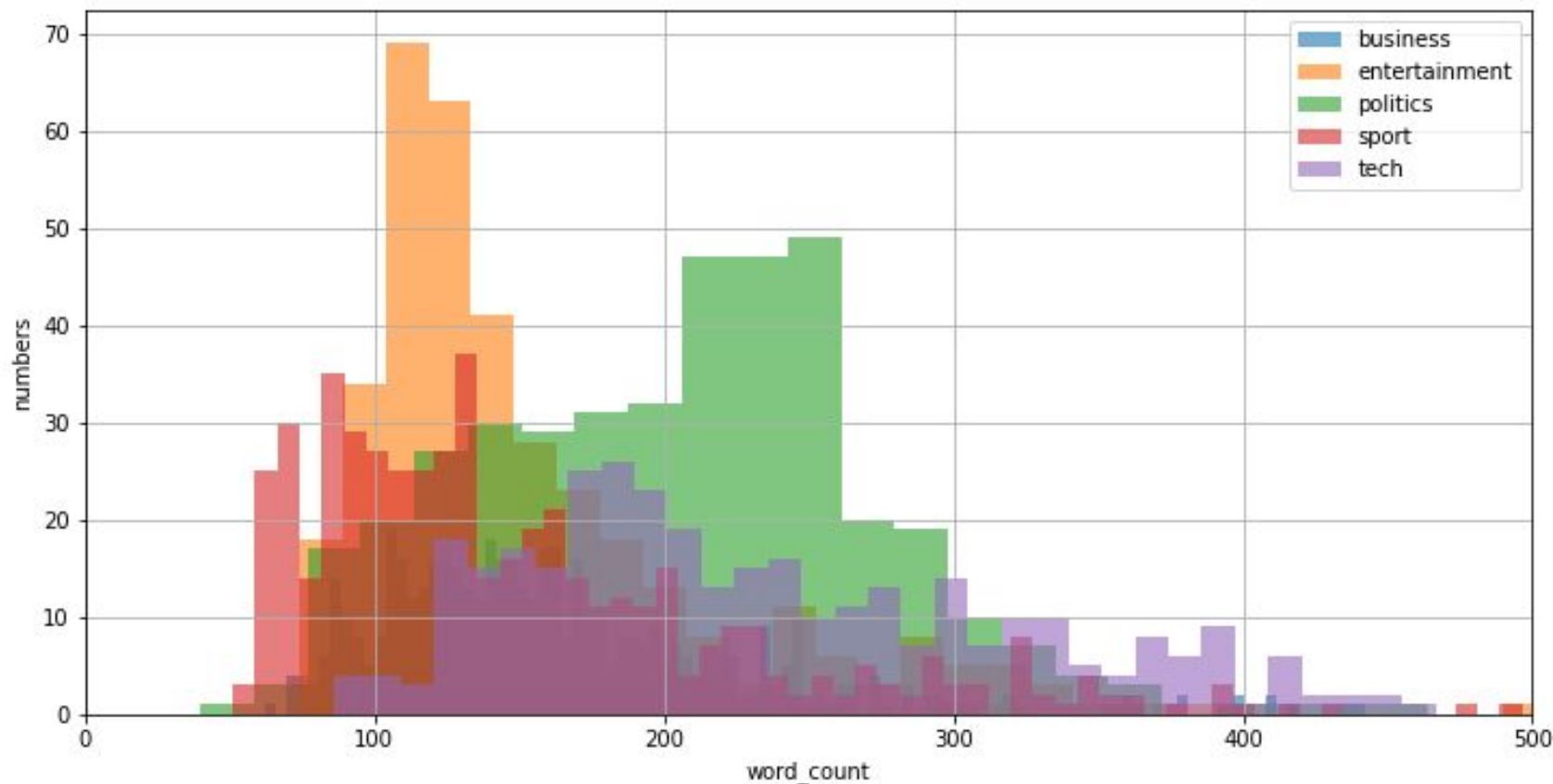
Average number of words in documents

Frequent words in all documents

Top words in headlines dataset (excluding stop words)

WordCloud - Business

WordCloud - Sport

WordCloud - Tech

WordCloud - Entertainment
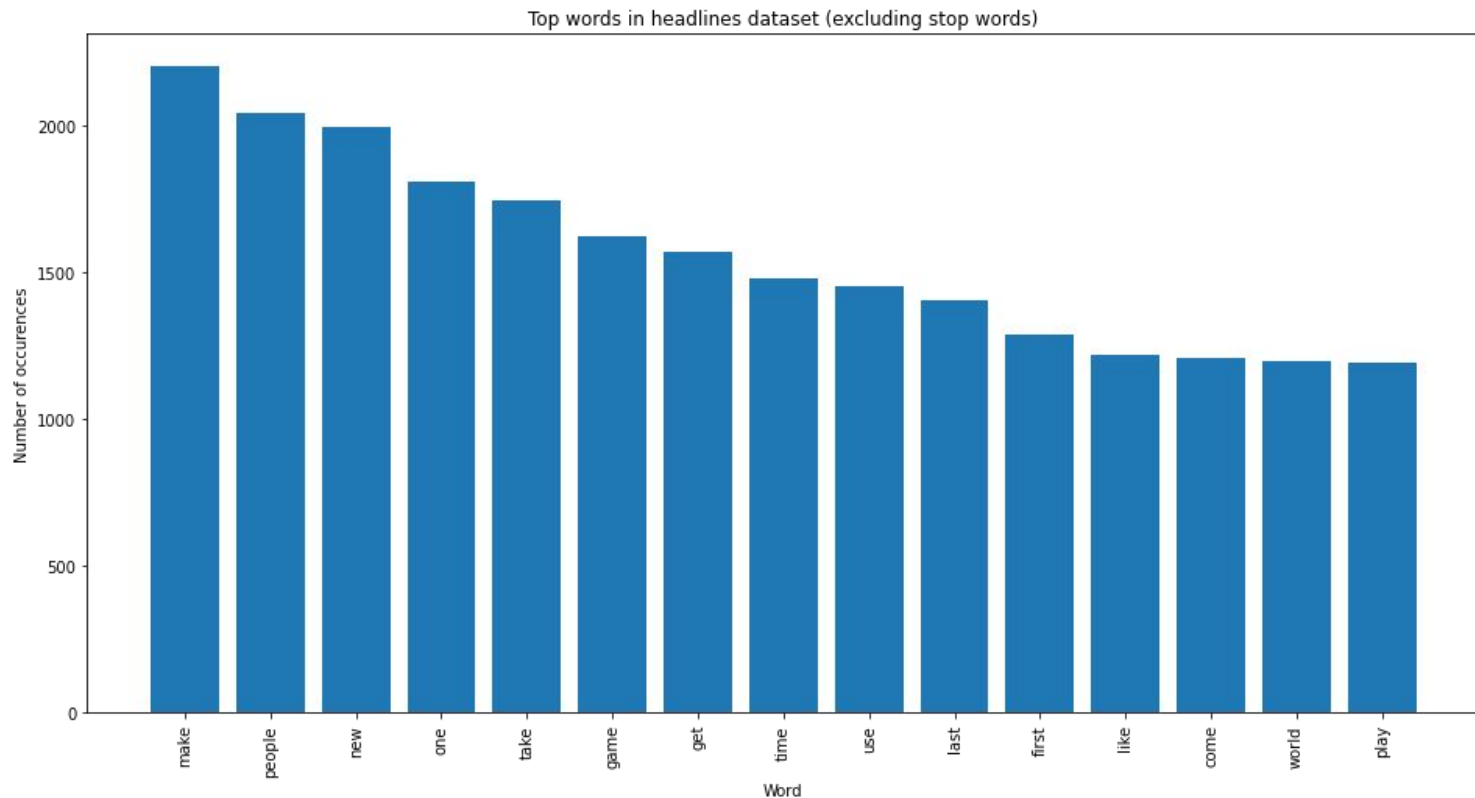
WordCloud - Politics

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)
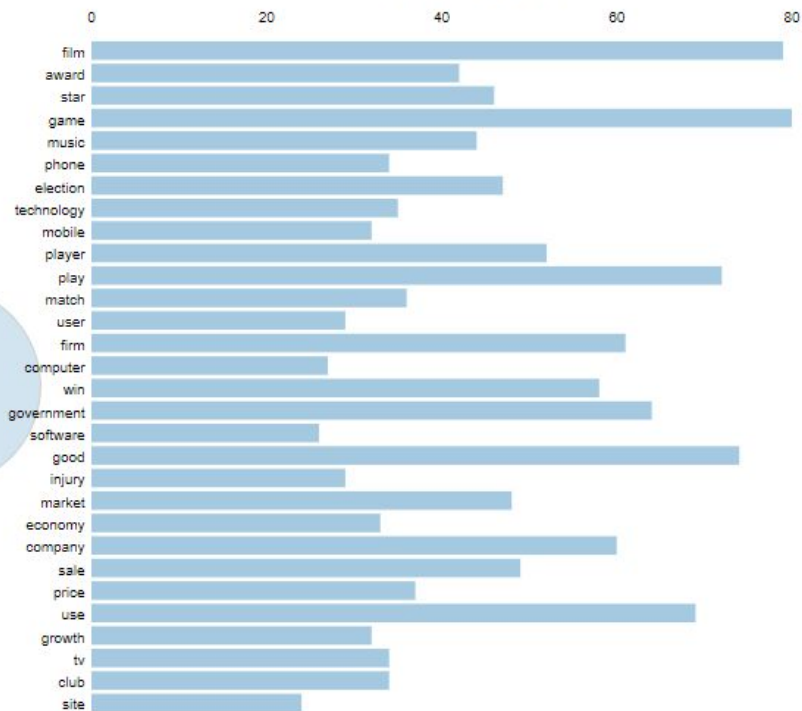
Latent Dirichlet Allocation (LDA)

LDA - Cluster 1 : Politics

Top-30 Most Relevant Terms for Topic 1 (25.5% of tokens)

LDA - Cluster 2 : Business
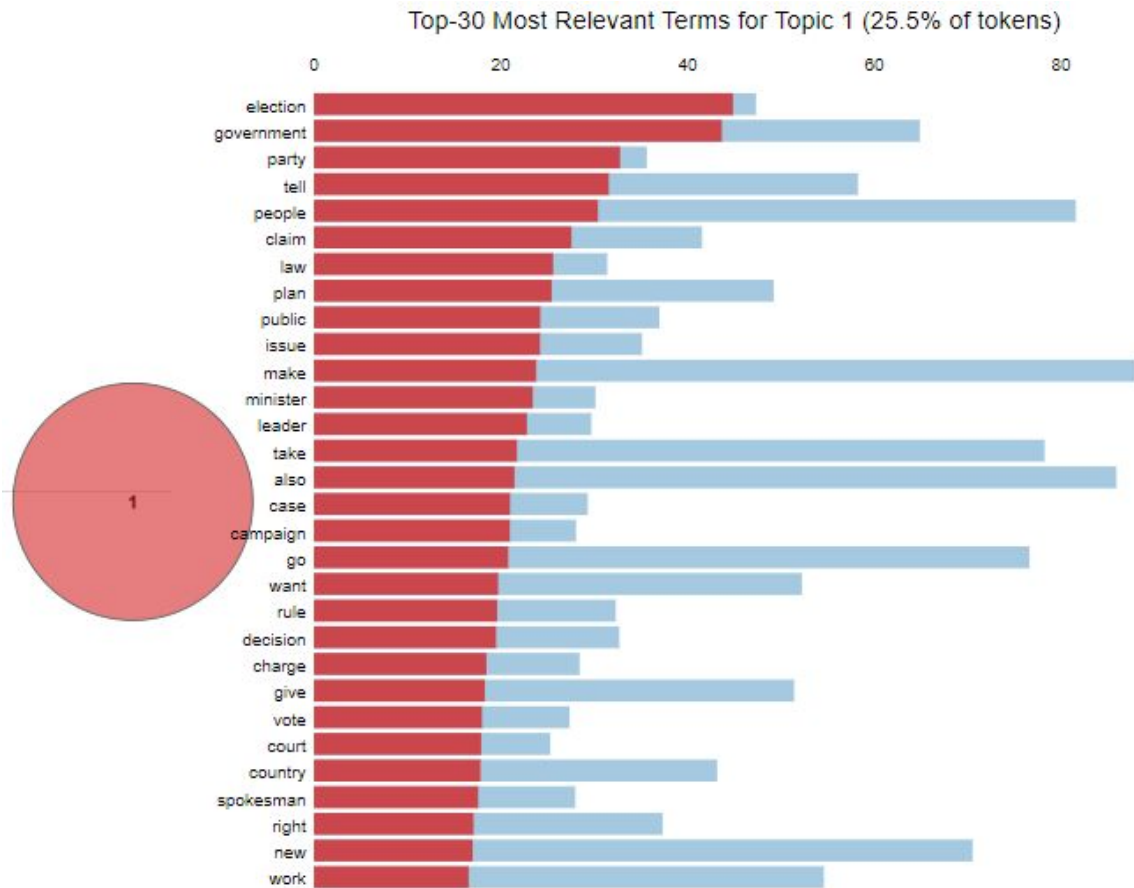
Top-30 Most Relevant Terms for Topic 2 (21.9% of tokens)
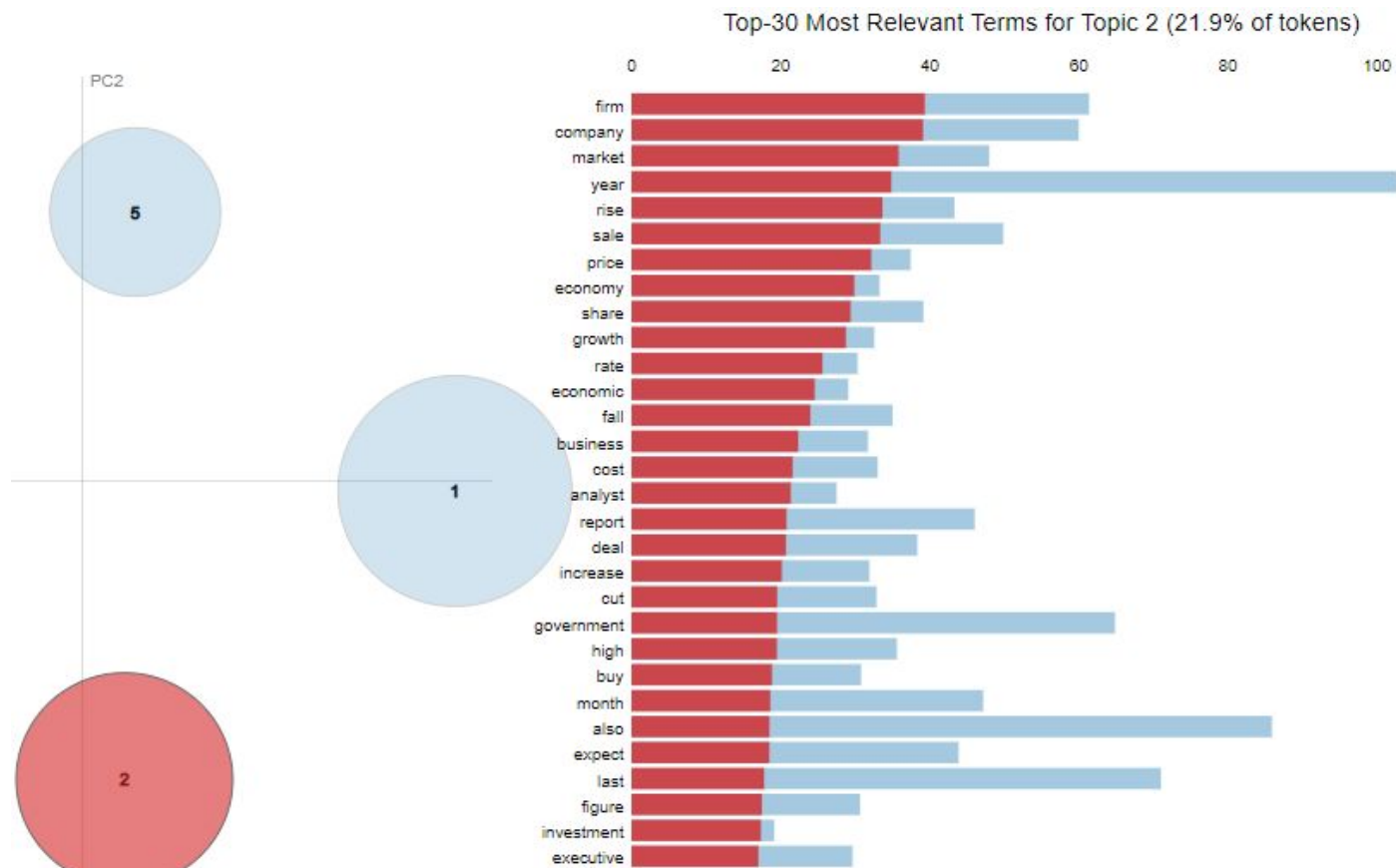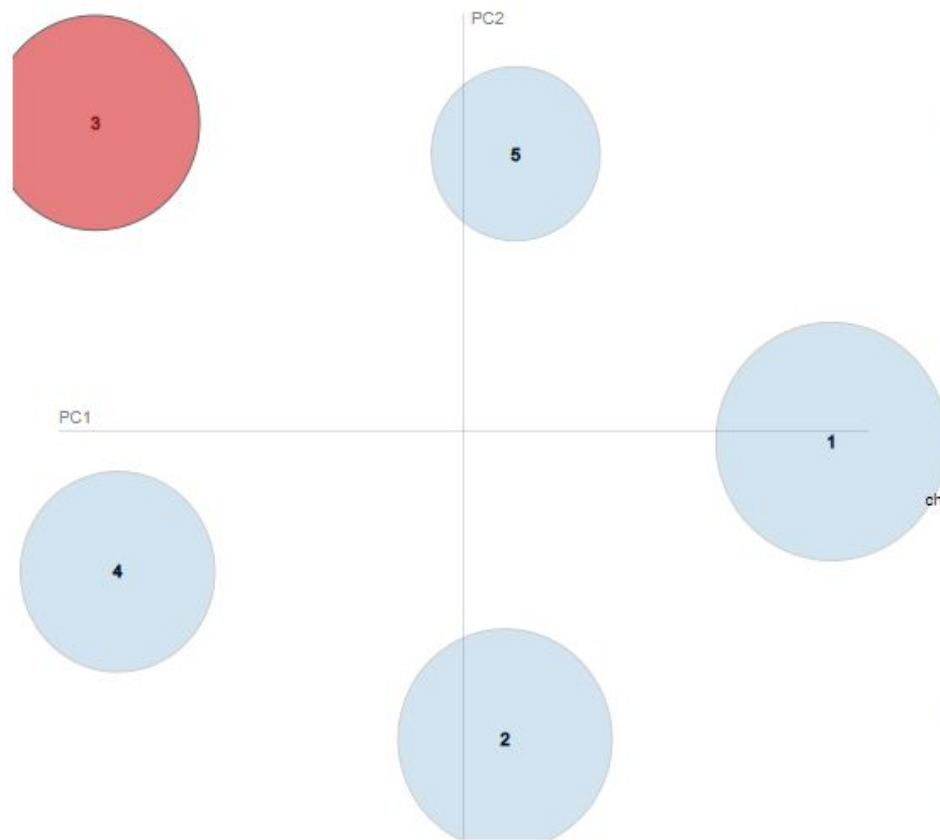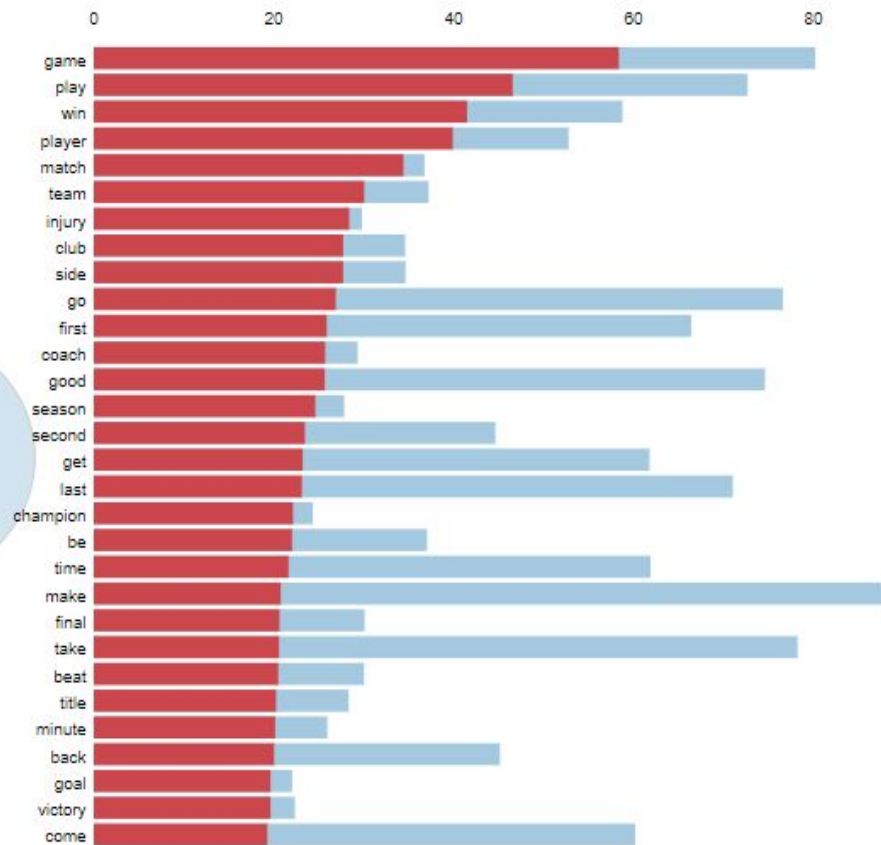
LDA - Cluster 3 : Sport

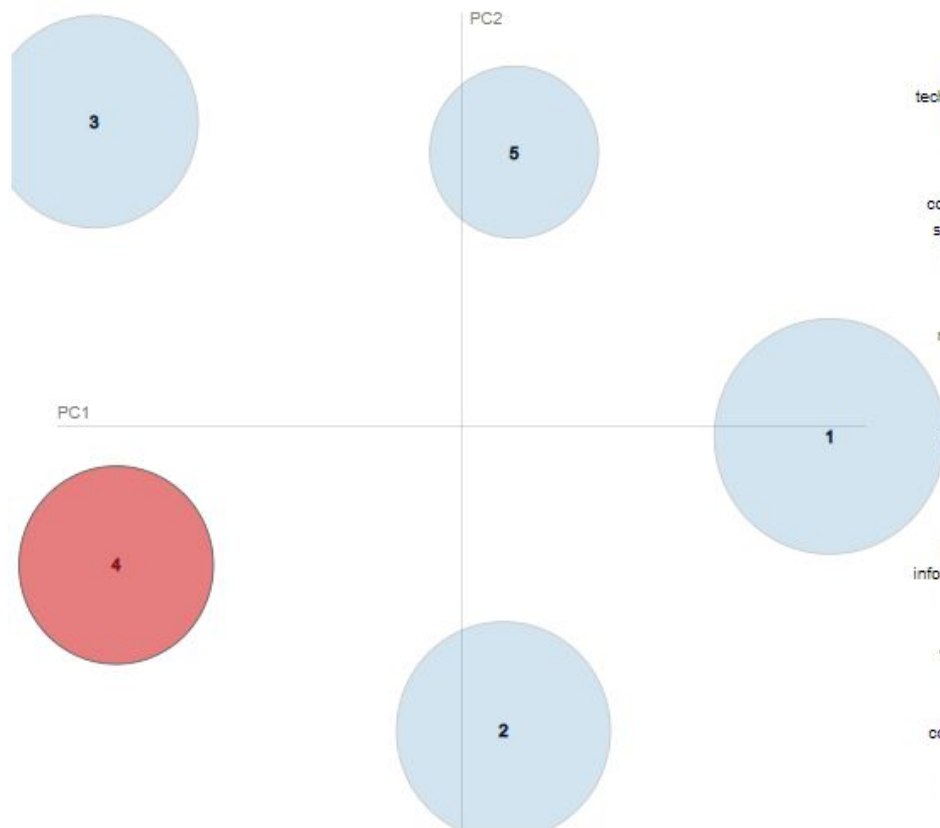Intertopic Distance Map (via multidimensional scaling)

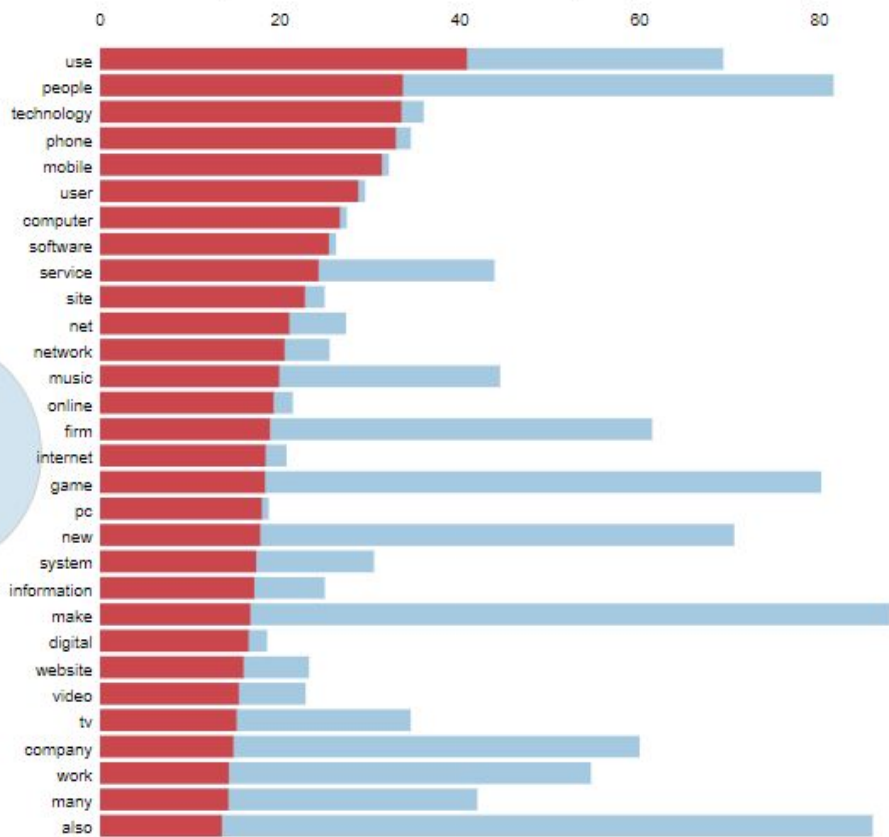Top-30 Most Relevant Terms for Topic 3 (20.8% of tokens)
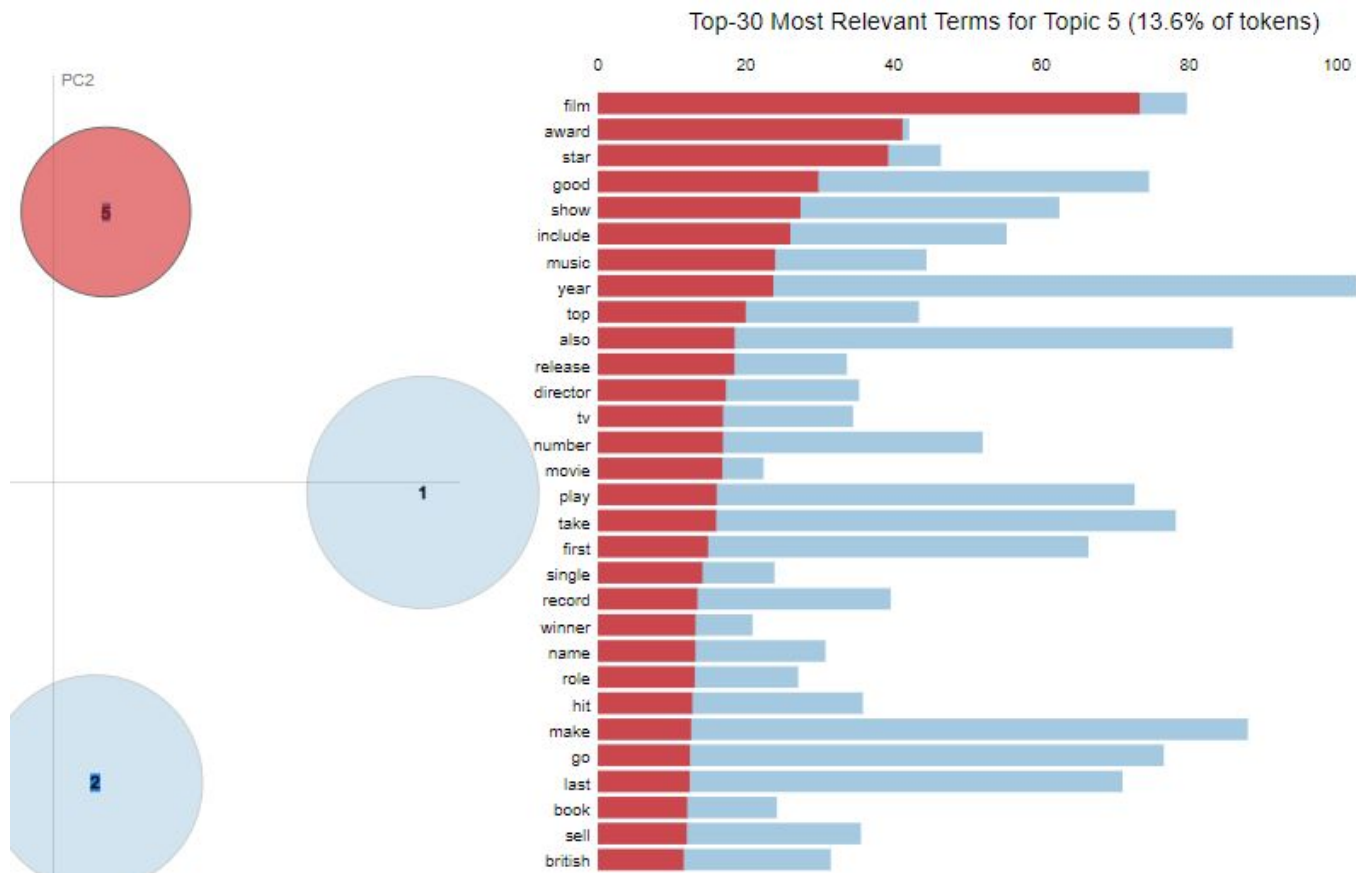
LDA - Cluster 4 : Tech
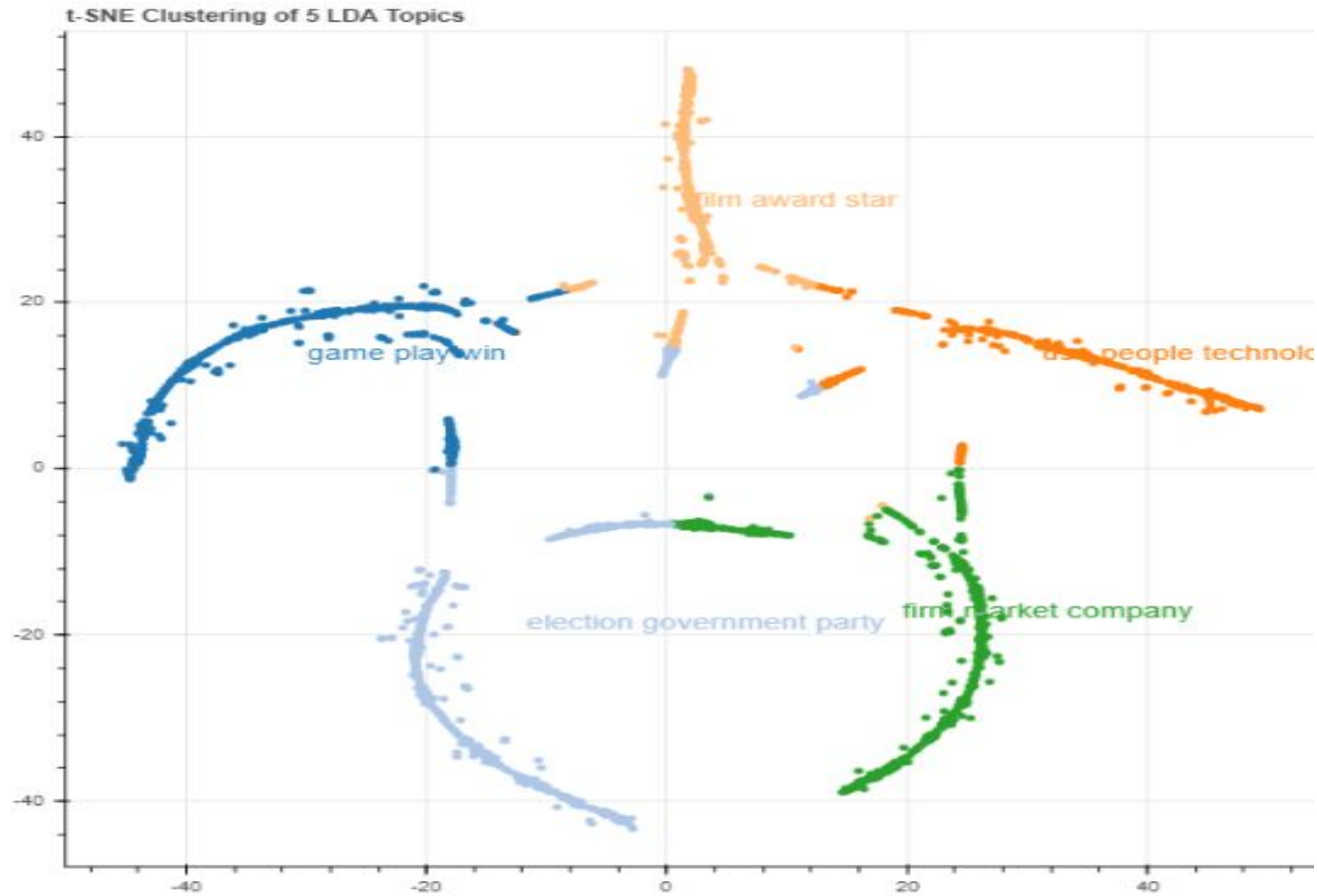

Intertopic Distance Map (via multidimensional scaling)
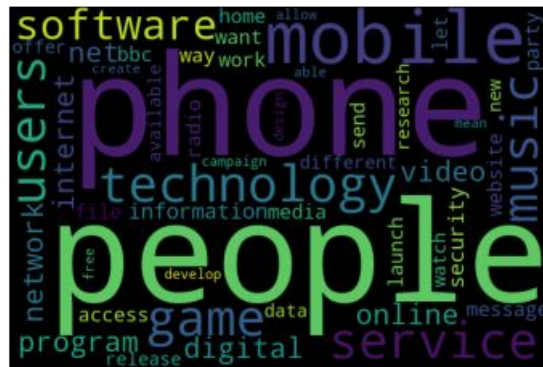
Top-30 Most Relevant Terms for Topic 4 (18.1% of tokens)

LDA - Cluster 5 : Entertainment

Top-30 Most Relevant Terms for Topic 5 (13.6% of tokens)

film, award, star, good, show, include, music, year, top, also, release, director, tv, number, movie, play, take, first, single, record, winner, name, role, hit, make, go, last, book, sell, british

LDA - t-SNE Clustering

t-SNE Clustering of 5 LDA Topics

film award star

game play win

use people technolo...

election government party   film market company

Latent Semantic Analysis (LSA)
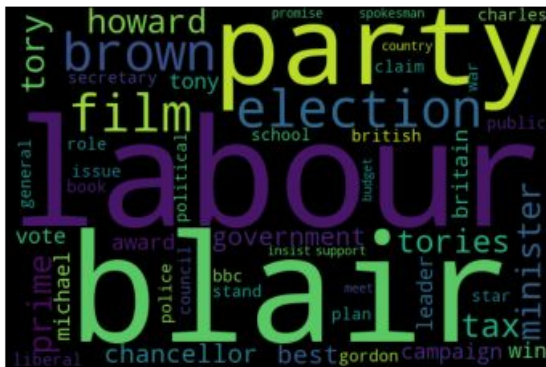
LSA - t-SNE Clustering

t-SNE Clustering of 5 LSA Topics

Non-negative Matrix Factorization (NMF)

NMF - Cluster 1 : Sport

Top-30 Most Relevant Terms for Topic 1 (25.7% of tokens)

NMF - Cluster 2 : Tech

Top-30 Most Relevant Terms for Topic 2 (21.5% of tokens)
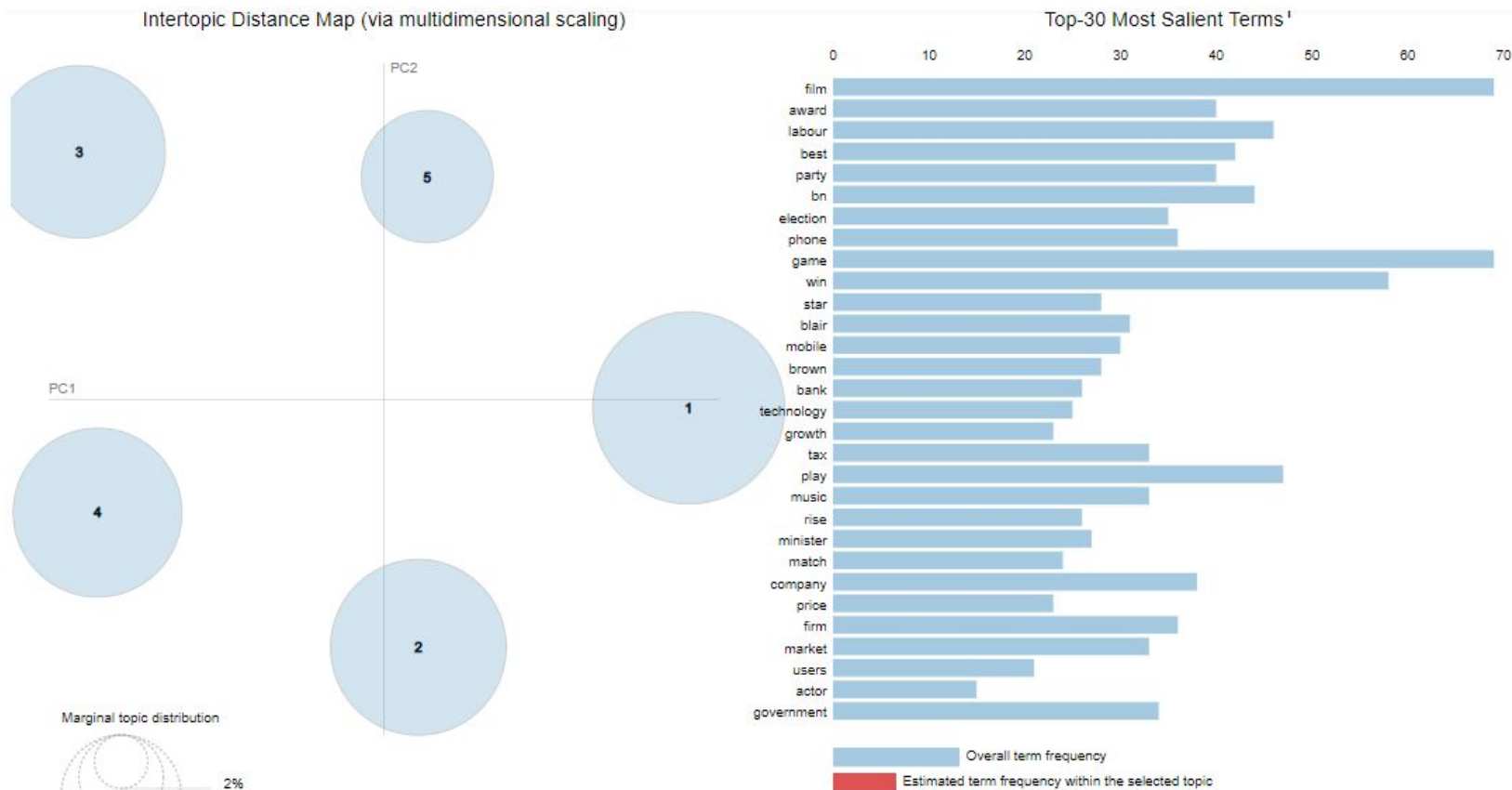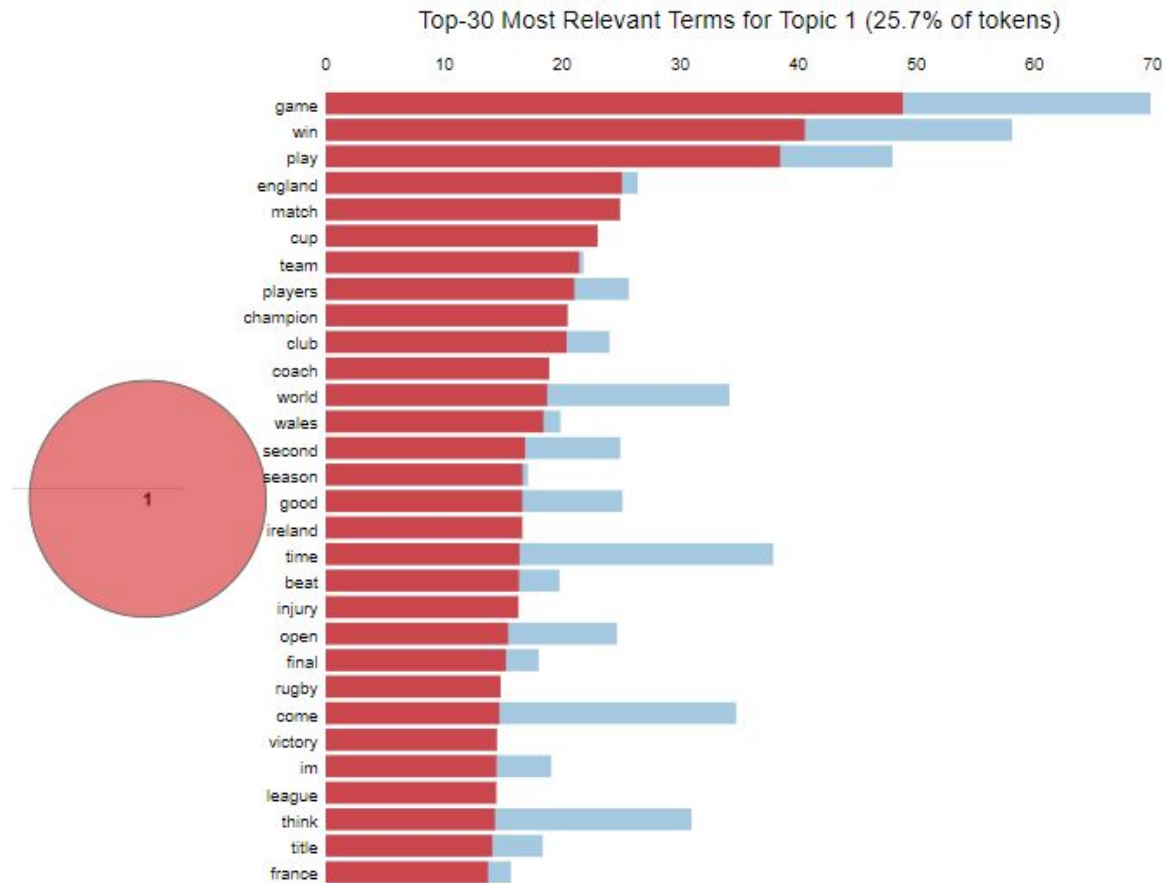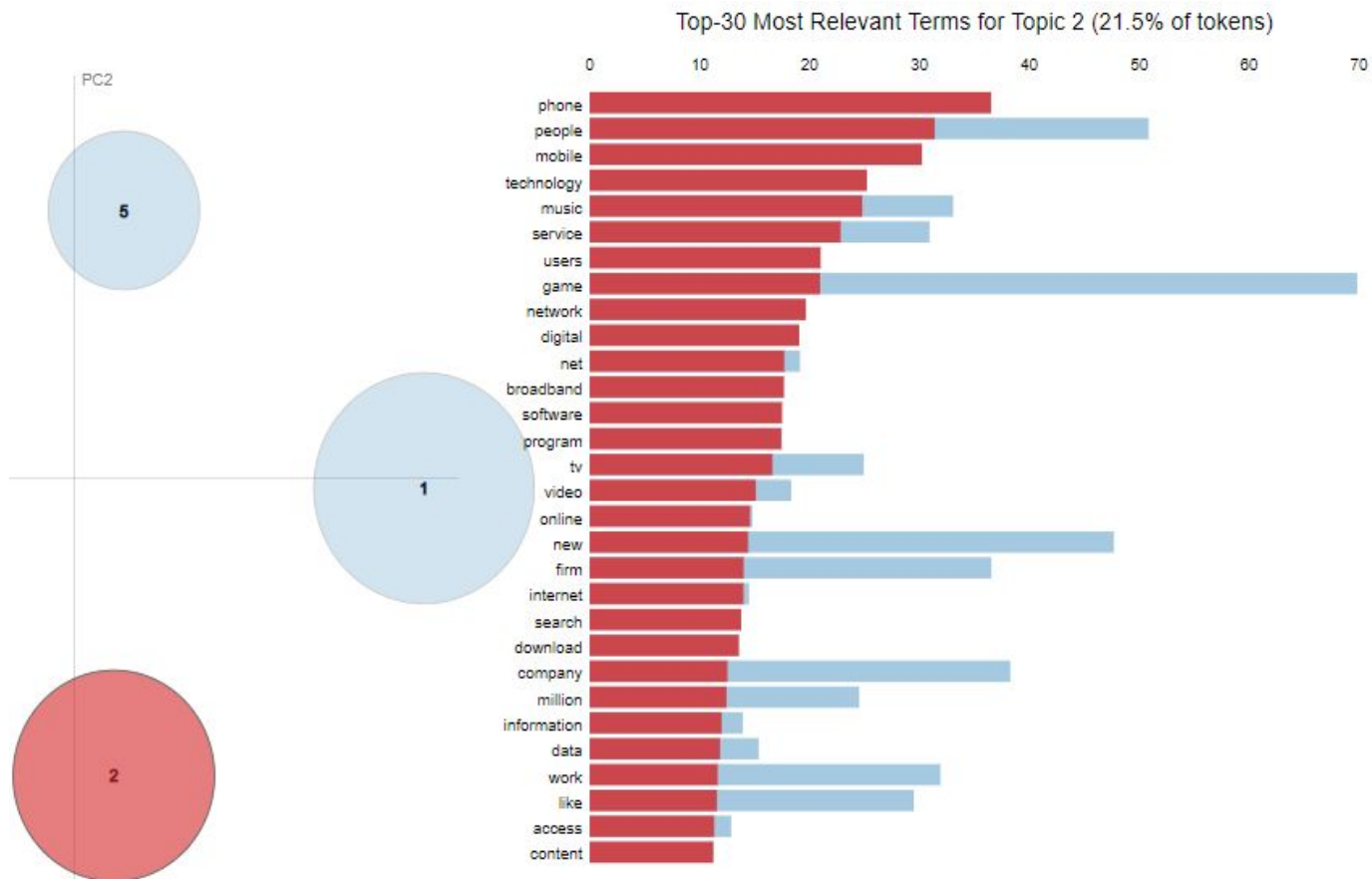
NMF - Cluster 3 : Business

Intertopic Distance Map (via multidimensional scaling)
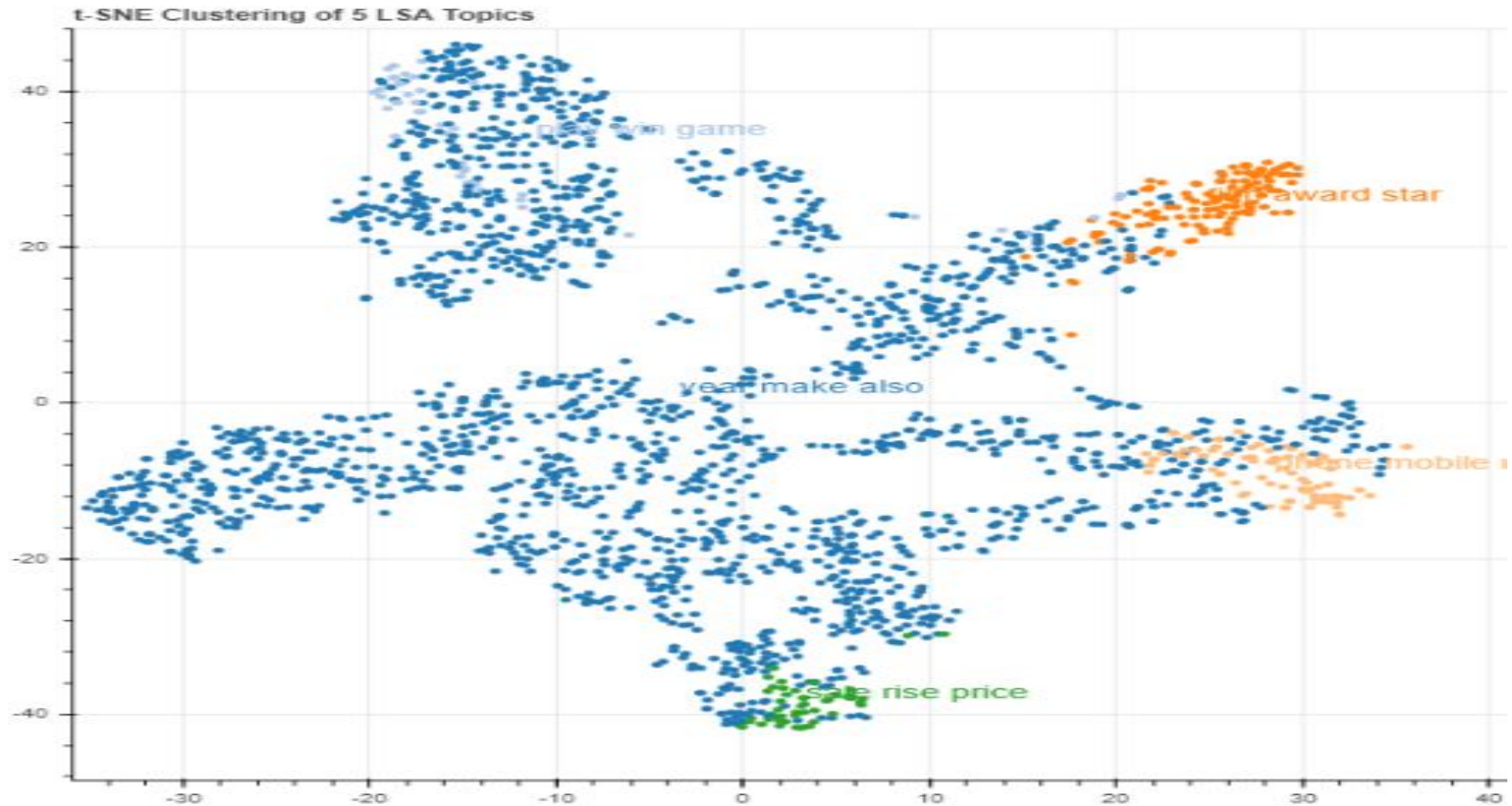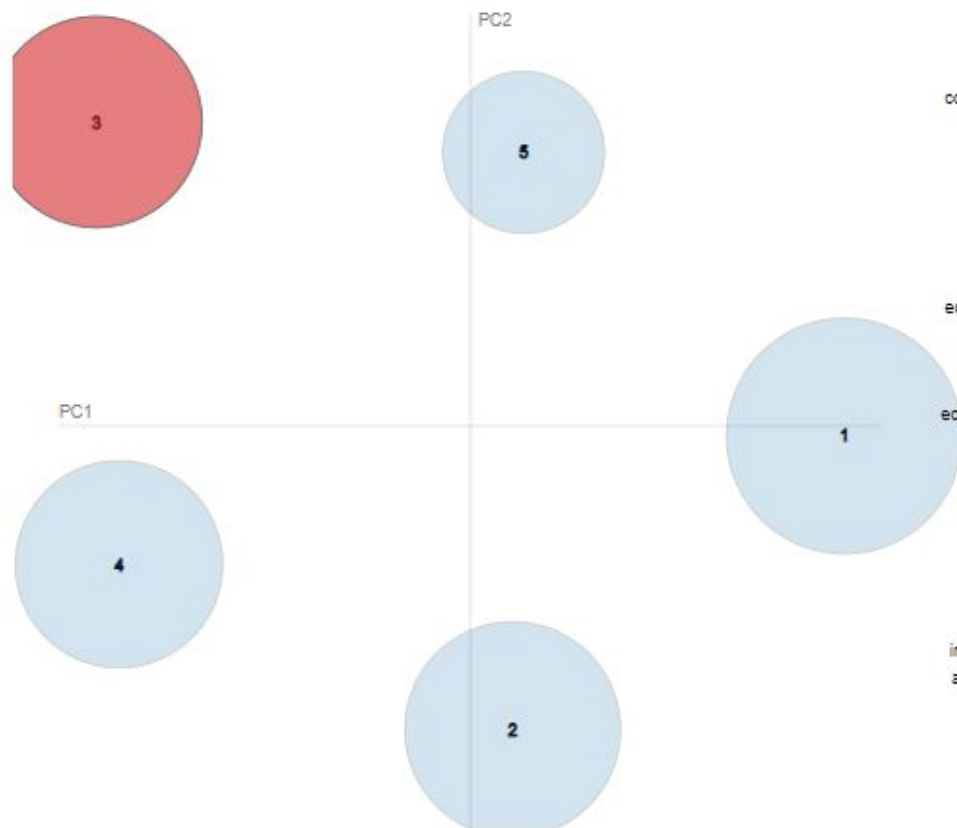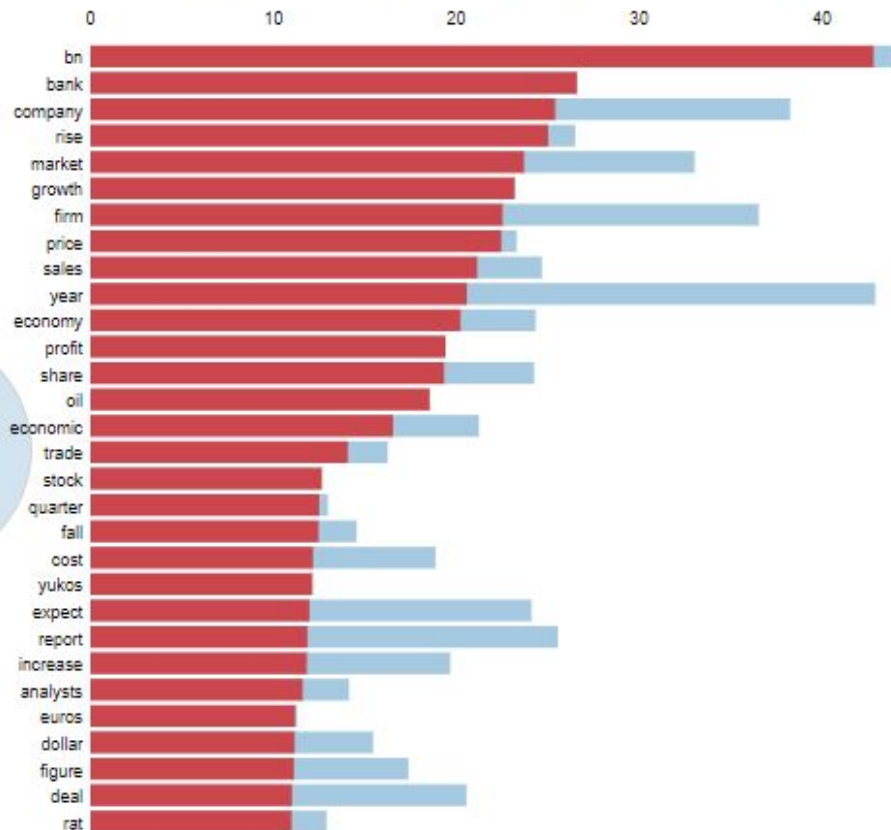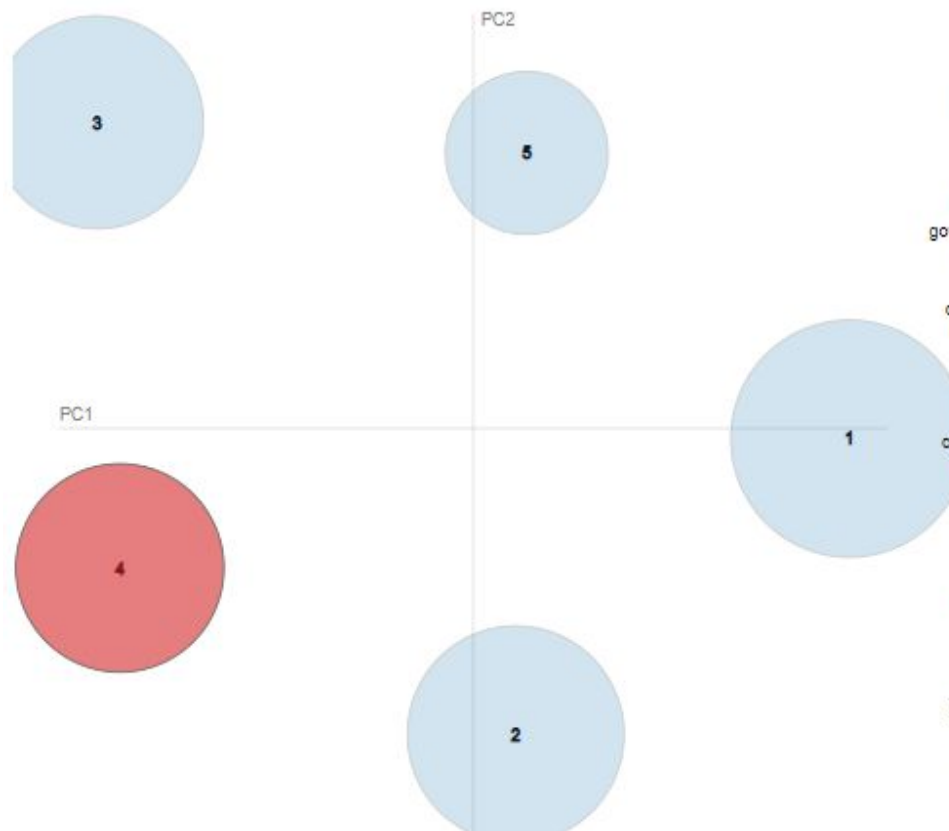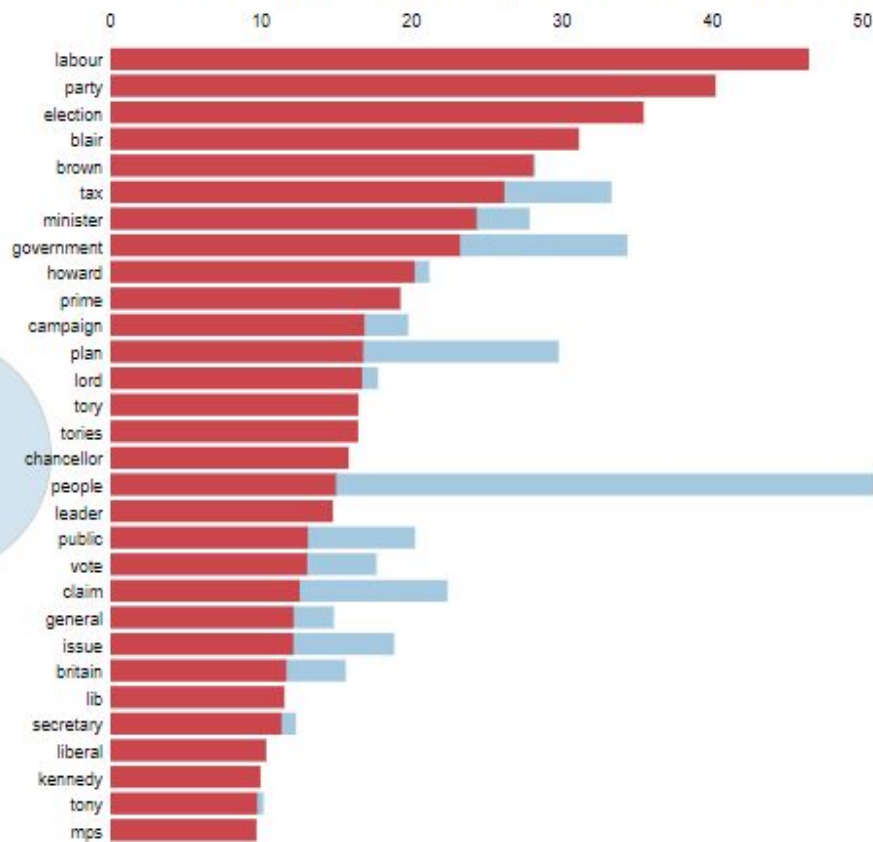
Top-30 Most Relevant Terms for Topic 3 (20.7% of tokens)
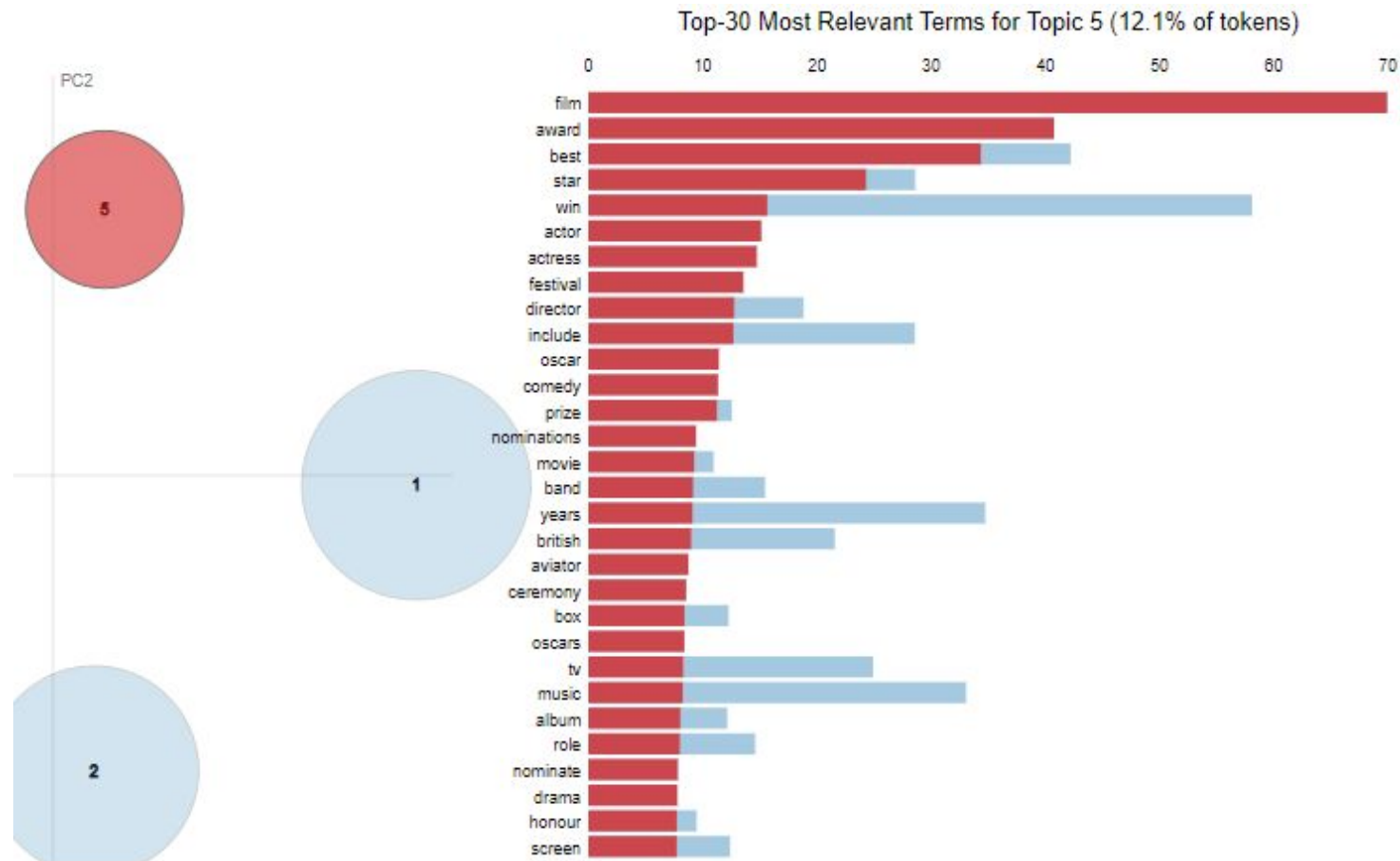
NMF - Cluster 4 : Politics

NMF - Cluster 5 : Entertainment

Top-30 Most Relevant Terms for Topic 5 (12.1% of tokens)

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