

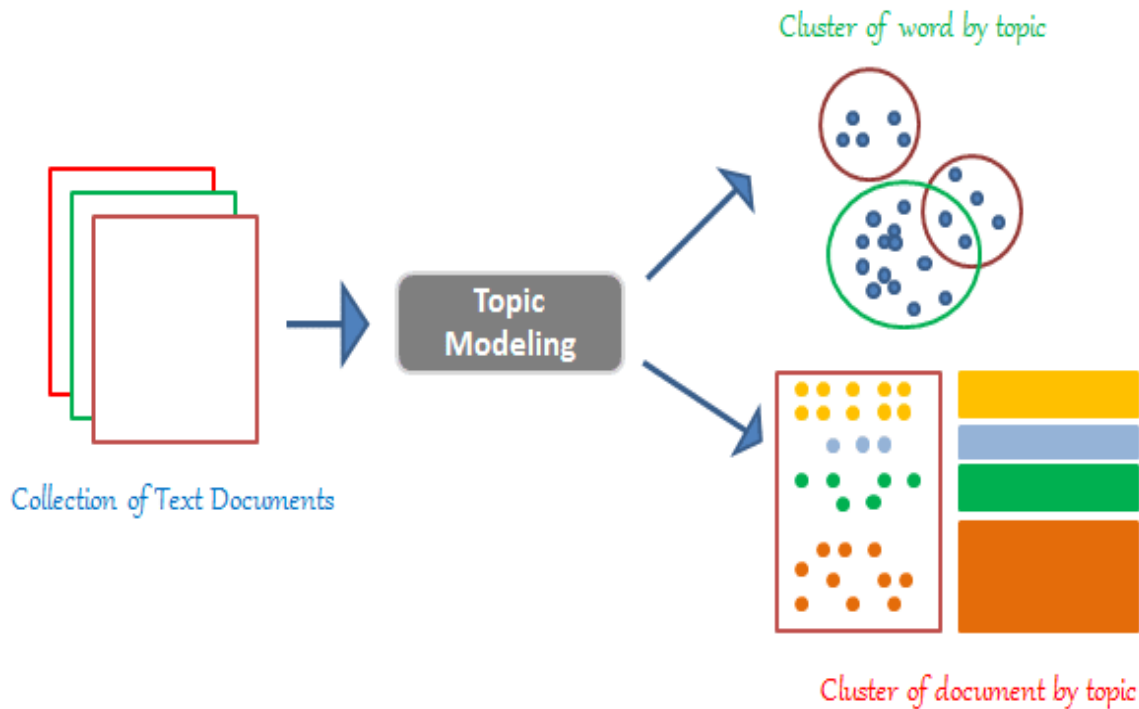
Capstone Project

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.

The screenshot shows the BBC News website interface. At the top is the BBC logo, a 'Menu' dropdown, and a search bar. Below this is a red navigation bar with the word 'NEWS' in large white letters, and a 'Find local news' button with a location pin icon. The navigation bar also contains links to various sections: Home, UK, World, Business, Election 2015, Tech, Science, Health, Education, Entertainment & Arts, Video & Audio, and More.

The main headline is 'Clegg: It's Salmond, Farage or me' by Nick Clegg, dated 15 April 2015, with a '1914' icon. The article text states: 'Nick Clegg says no party will win an outright election victory and claims only the Lib Dems can stop a "lurch to the extremes".' Below the text are links for 'Manifestos reaction', 'Robinson: The coalition choice', 'At-a-glance: Lib Dem manifesto', and 'Courting the hipster vote?'. The article image shows Nick Clegg holding a colorful election manifesto card.

To the right of the main article is a 'Watch/Listen' section with several video thumbnails and titles:

- 'Protester jumps on Draghi desk' (0:16, 15 April 2015)
- 'Southern UK basks in April heatwave' (1:27, 15 April 2015)
- 'Arizona police ram car into armed man' (1:27, 15 April 2015)
- 'Reporter confronts dog trafficker' (0:56, 15 April 2015)
- 'First images of migrant survivors' (2:26, 15 April 2015)
- 'Day 17: Lib Dems and ...' (partially visible)

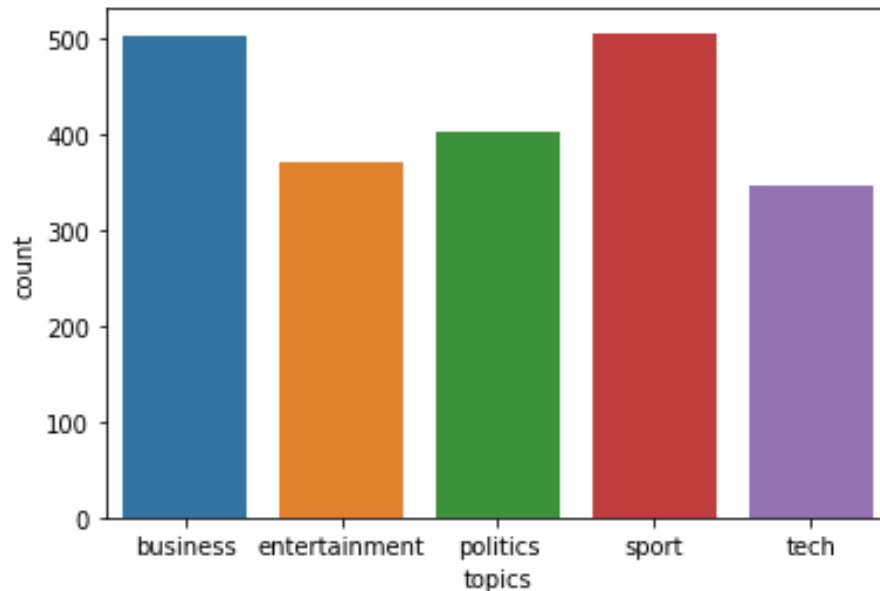
At the bottom of the page are four smaller article thumbnails:

- 'Town's firms wary of praising migrants' (image of a town square)
- 'The place where one street could decide MP' (image of a 'P Town centre' street sign)
- 'Secrets of SNP's social media strategy' (image of a crowd)
- 'What each of the parties is pledging' (image of various political symbols)

Data Summary

- There are total 5 topics -

- Business
- Entertainment
- Politics
- Sports
- Technology



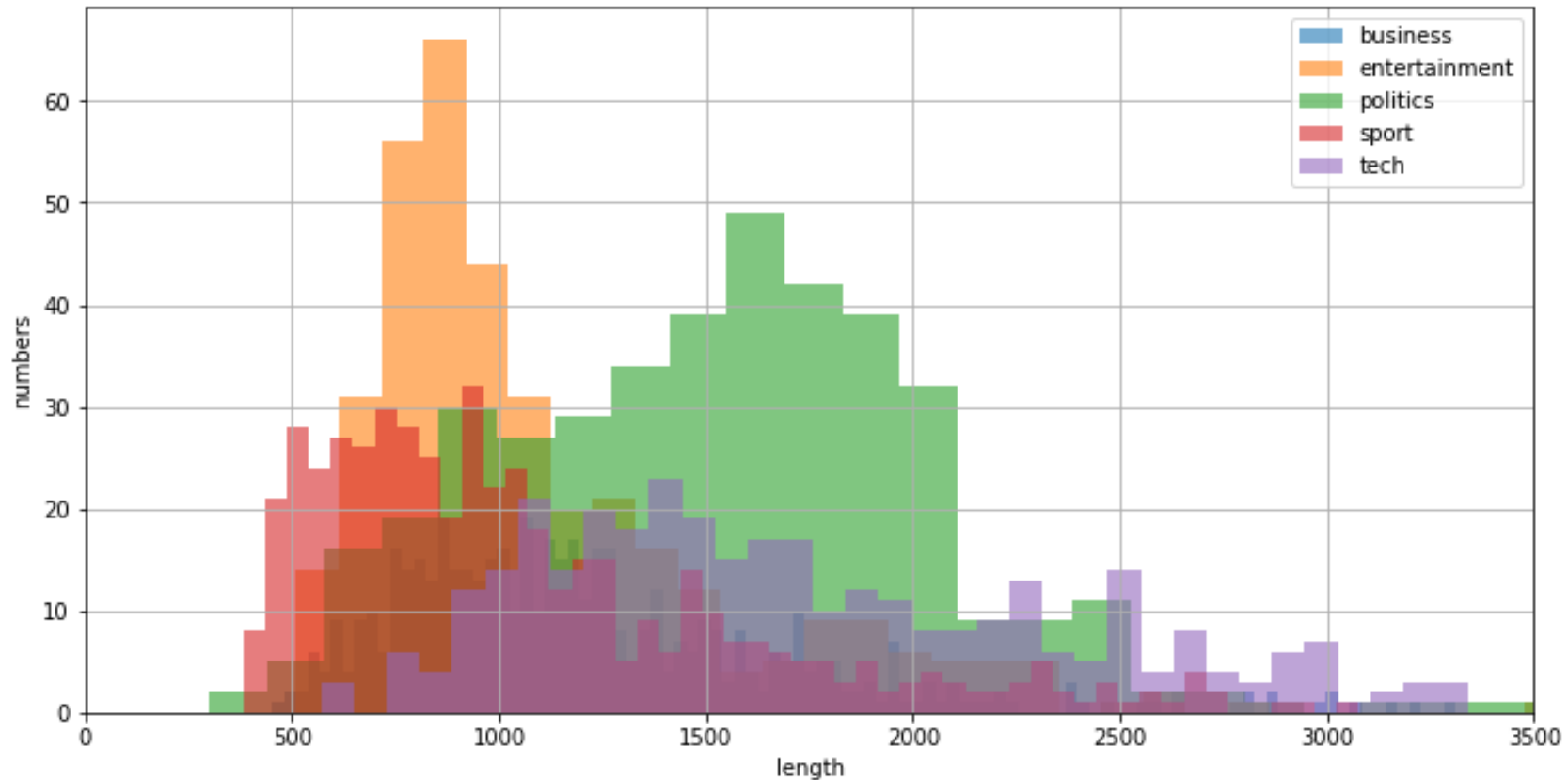
- Dataset consists of total 2225 articles.

Data Pre-processing

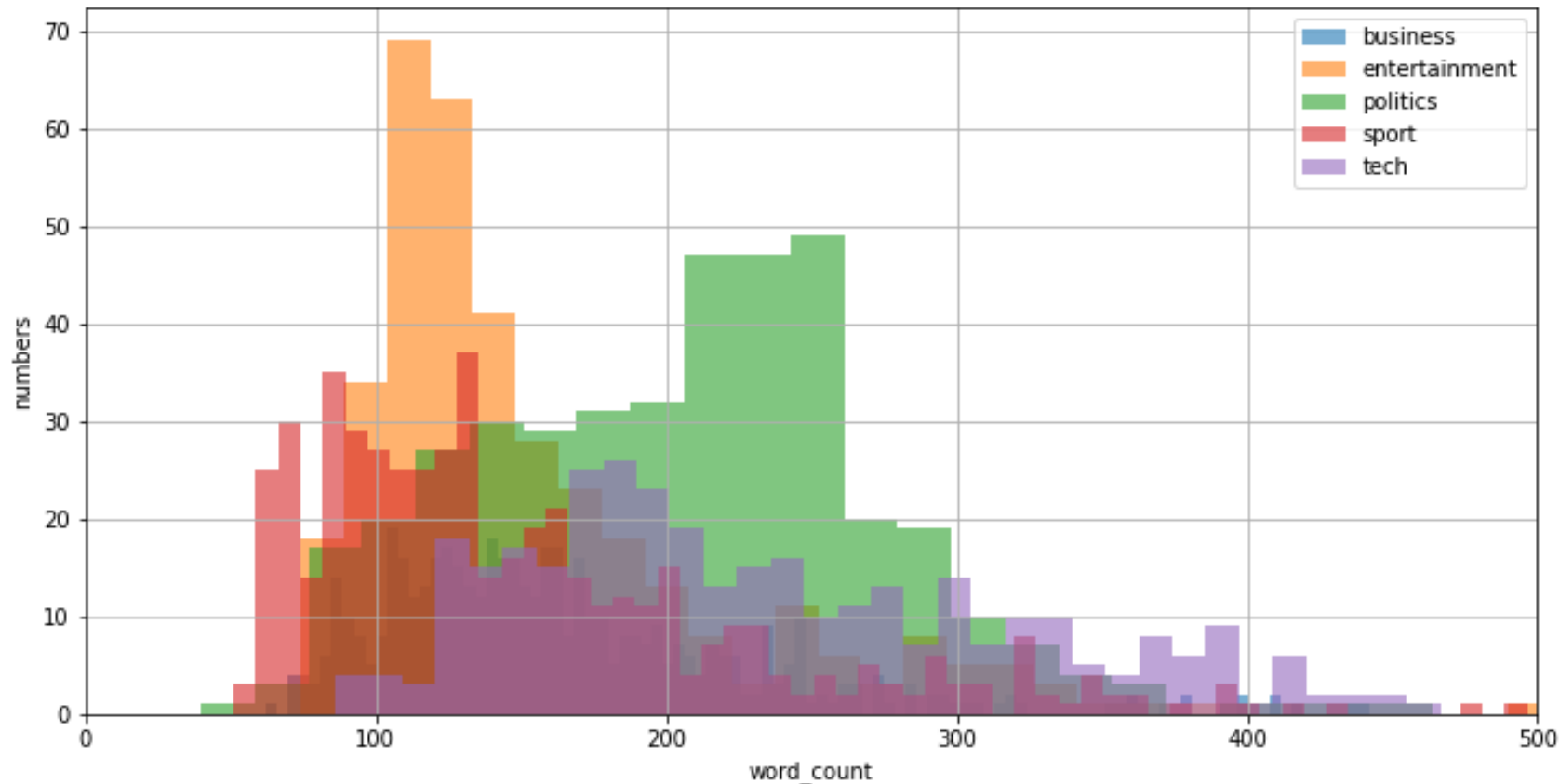
- Remove Html tags and urls
- Convert accented characters to ASCII characters
- Remove punctuations
- Remove numbers
- Split attached words
- Remove small length words
- Remove extra whitespaces
- Spelling corrections
- Lemmatization
- Remove stopwords
- Remove frequent words



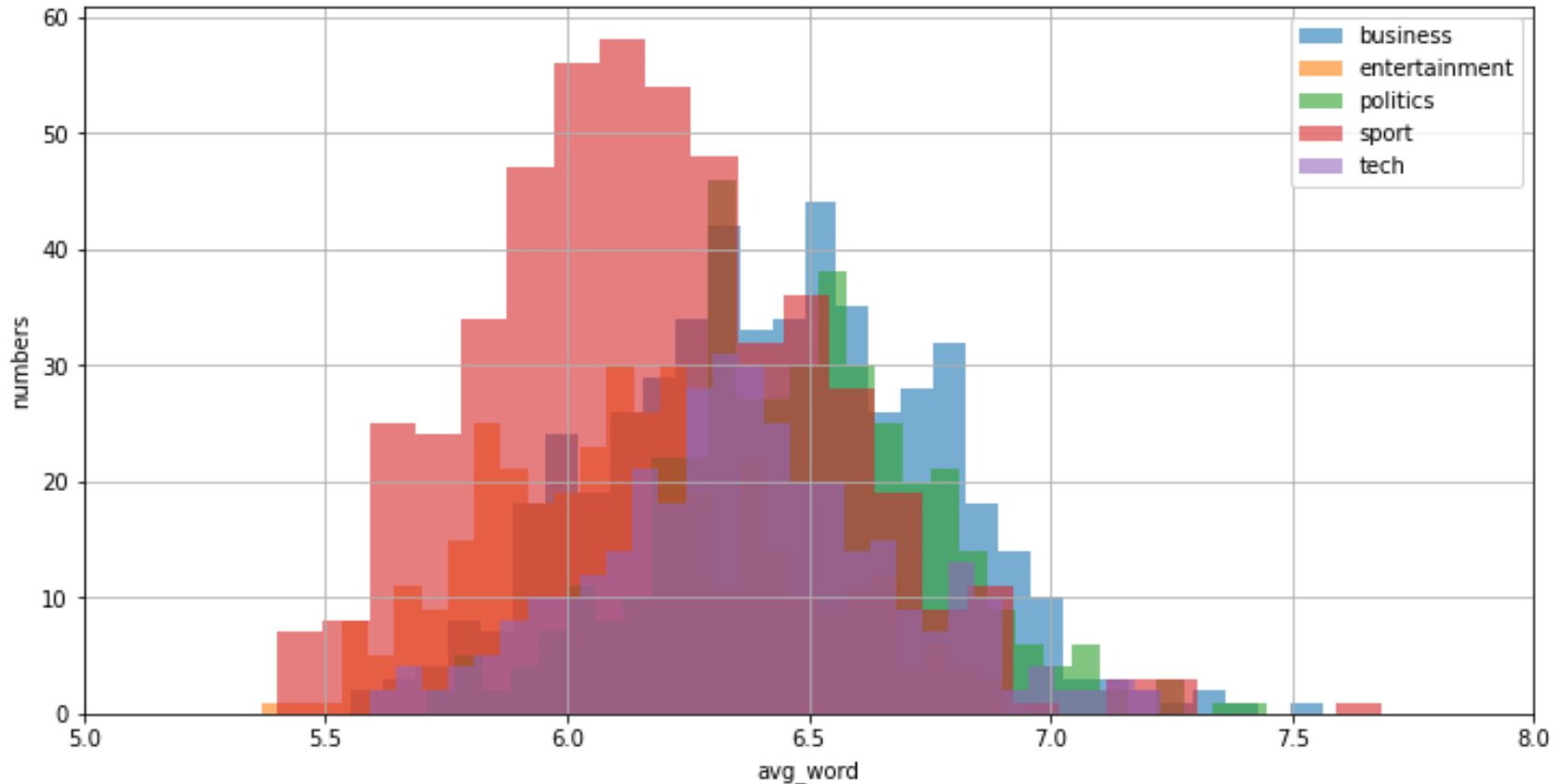
Feature Extraction - Length of documents



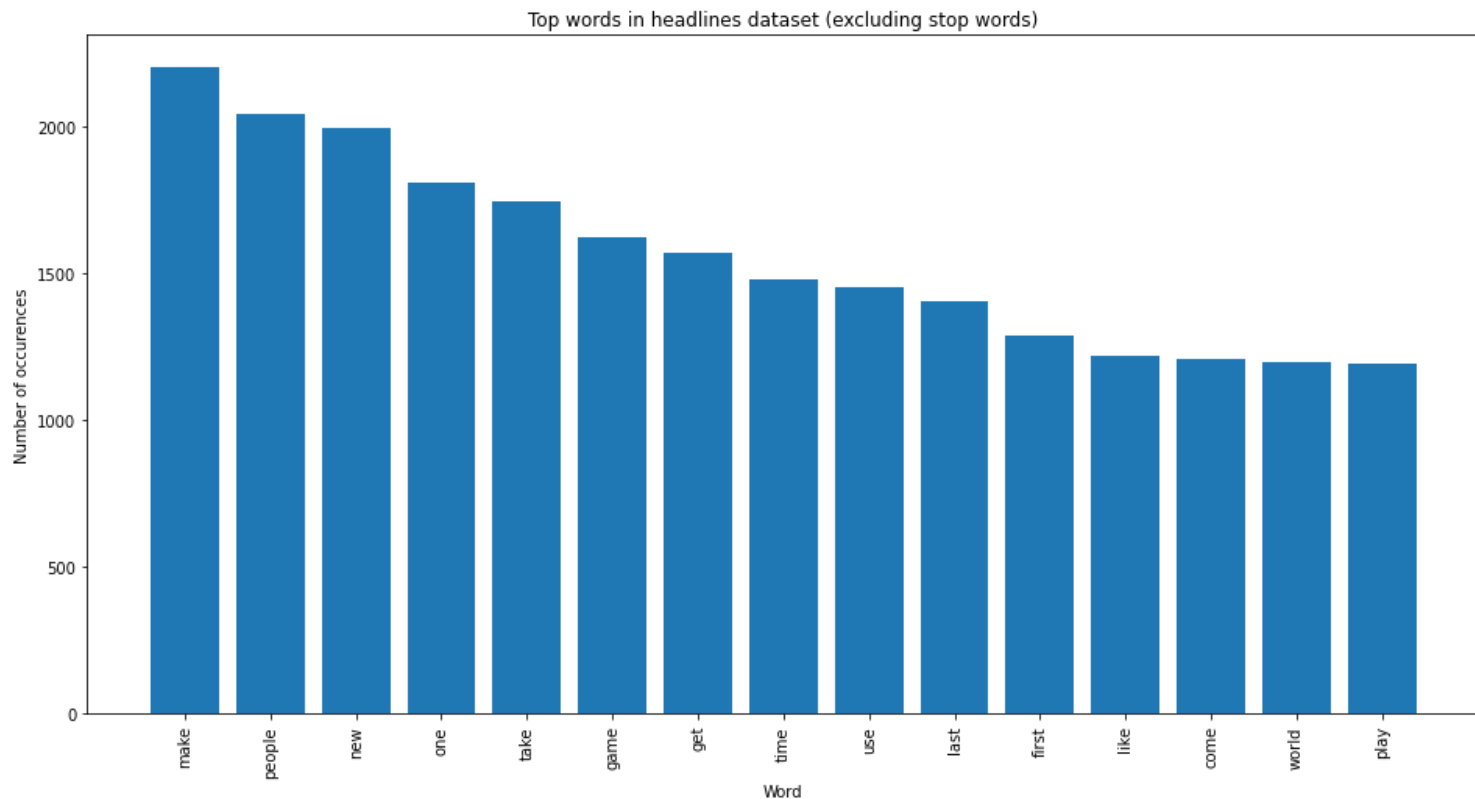
Number of words in documents



Average number of words in documents



Frequent words in all documents







WordCloud - 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)

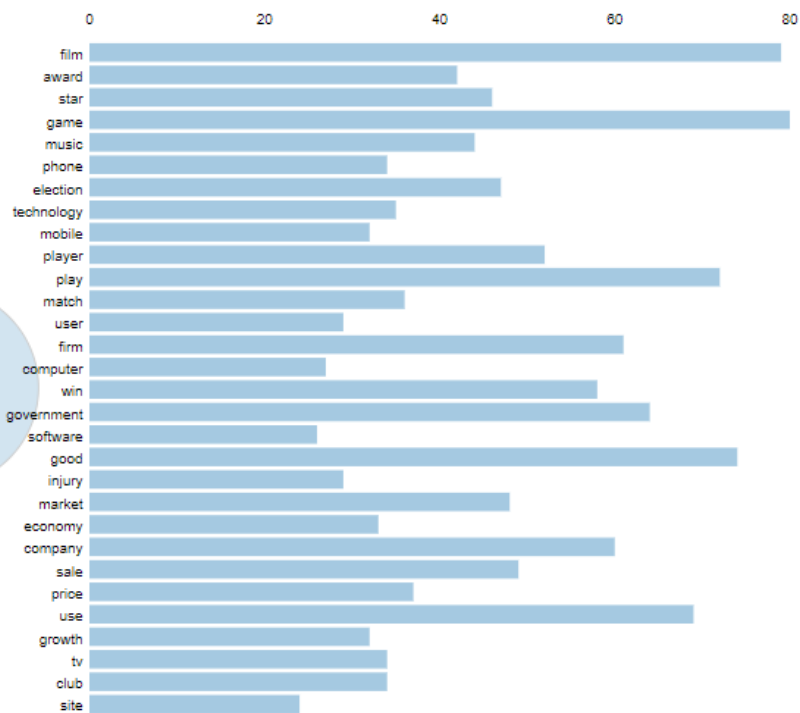
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Salient Terms¹



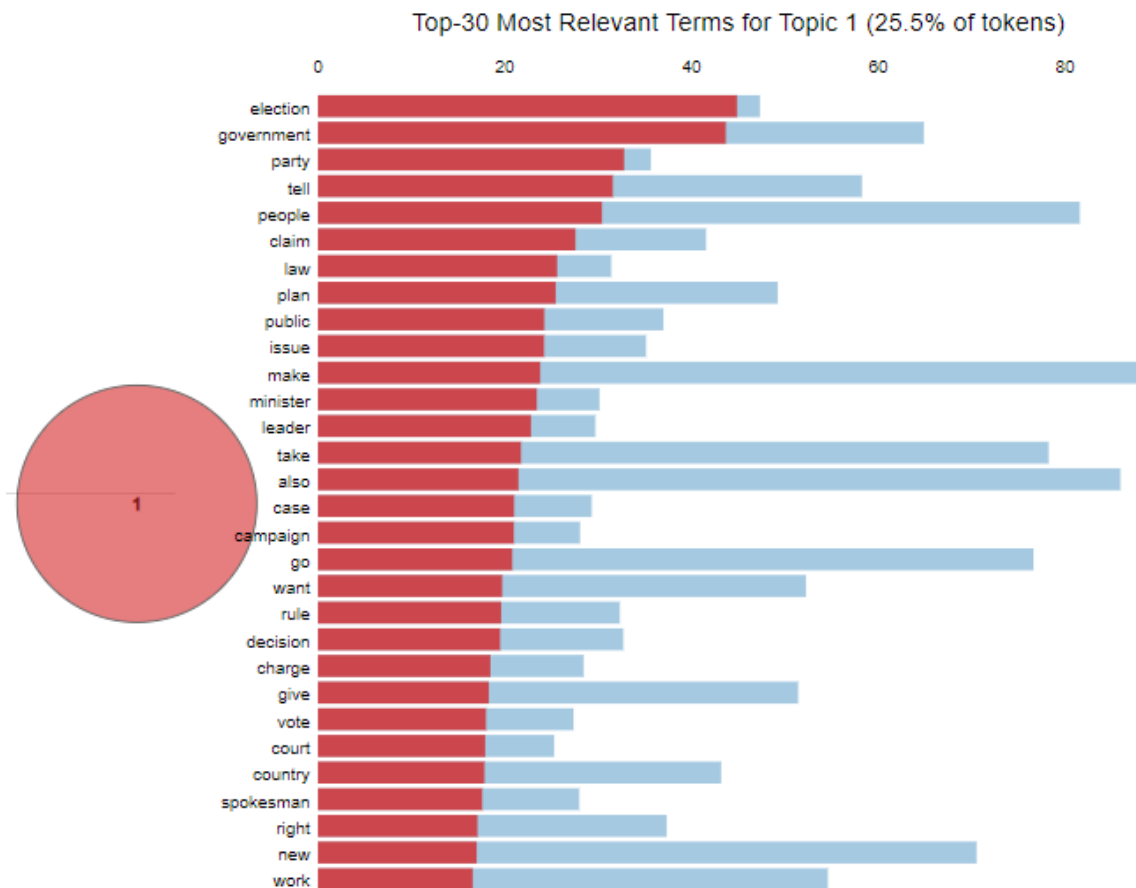
Overall term frequency

Estimated term frequency within the selected topic

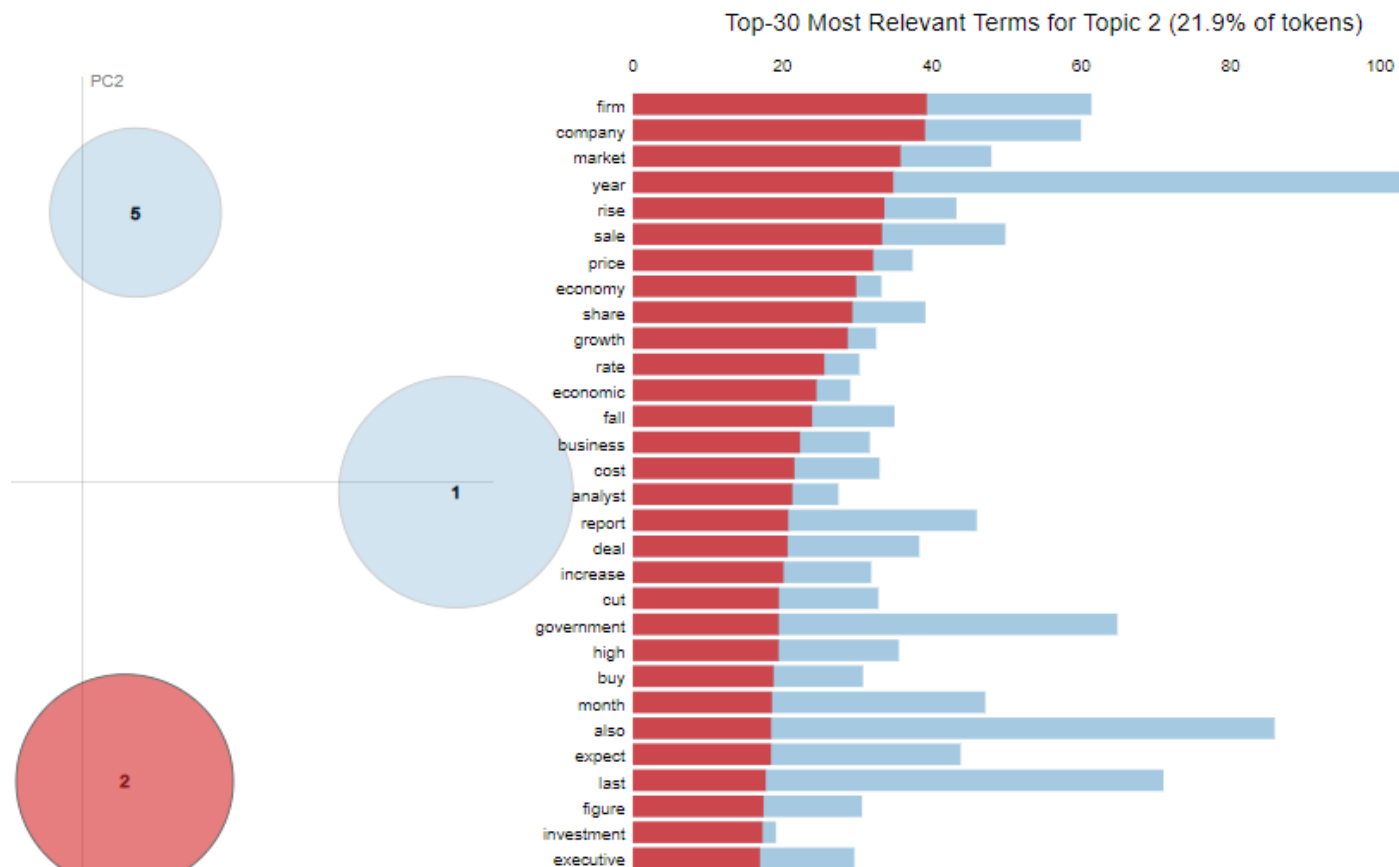
1. $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$ for topics t ; see Chuang et. al (2012)

2. $\text{relevance}(\text{term } w | \text{topic } t) = \lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

LDA - Cluster 1 : Politics

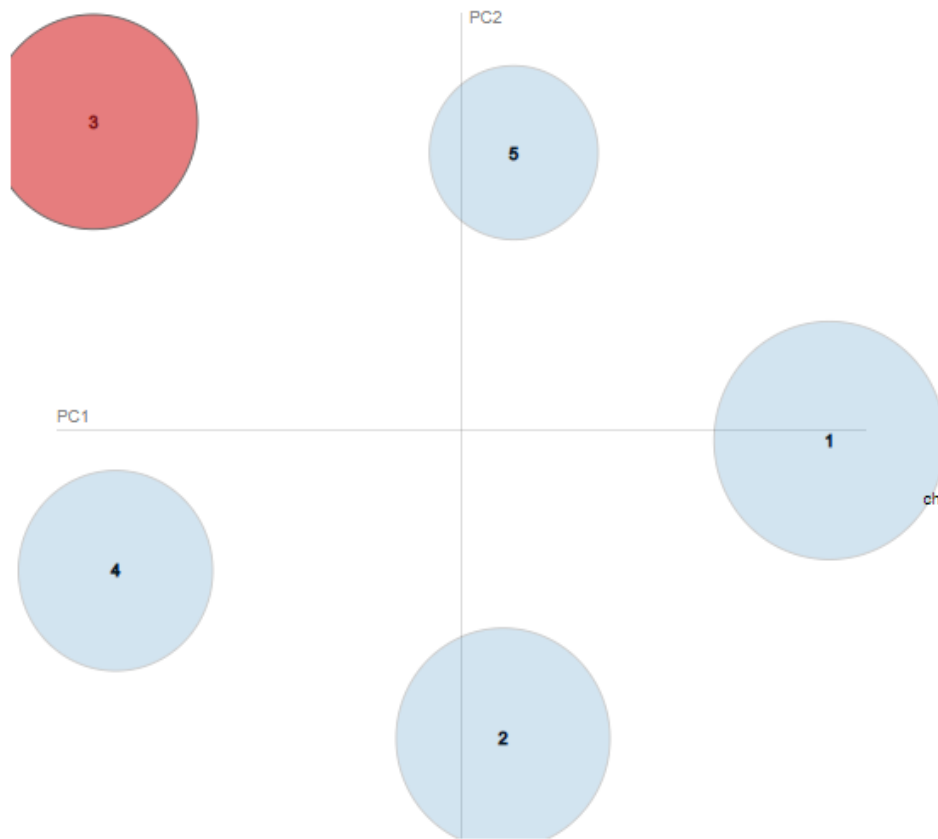


LDA - Cluster 2 : Business

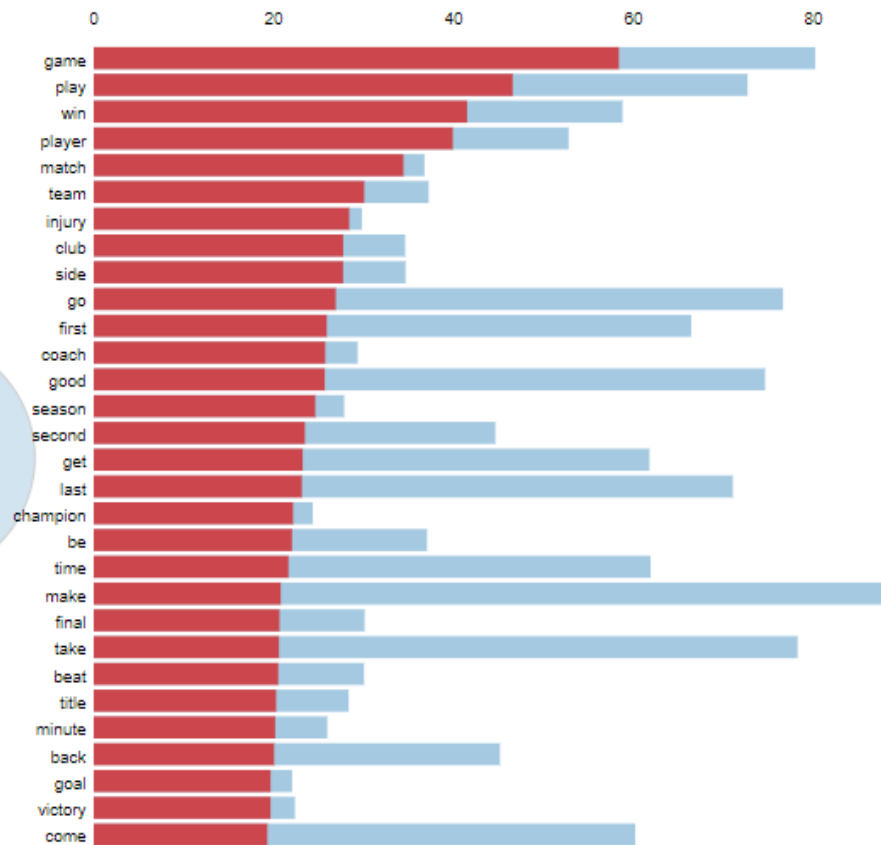


LDA - Cluster 3 : Sport

Intertopic Distance Map (via multidimensional scaling)

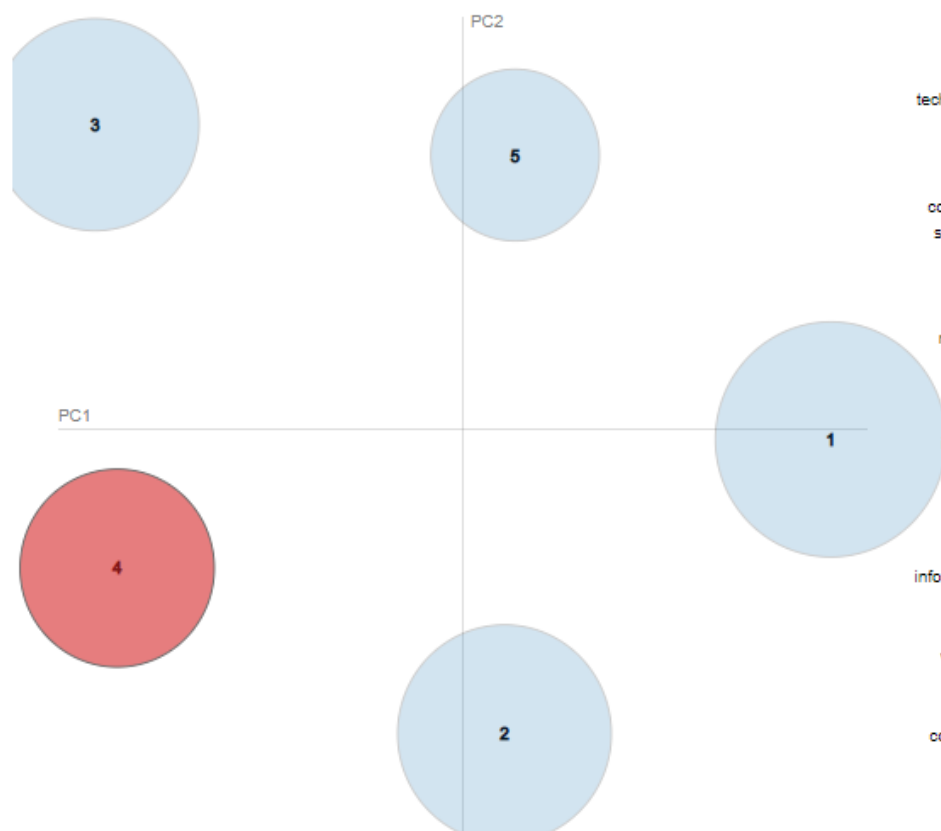


Top-30 Most Relevant Terms for Topic 3 (20.8% of tokens)

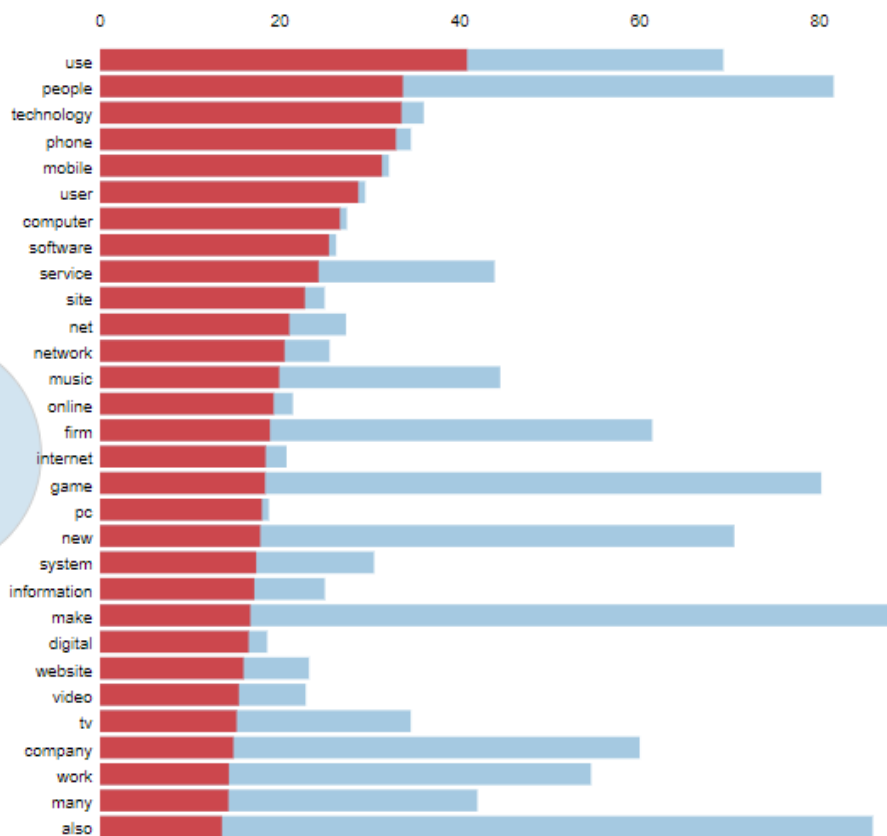


LDA - Cluster 4 : Tech

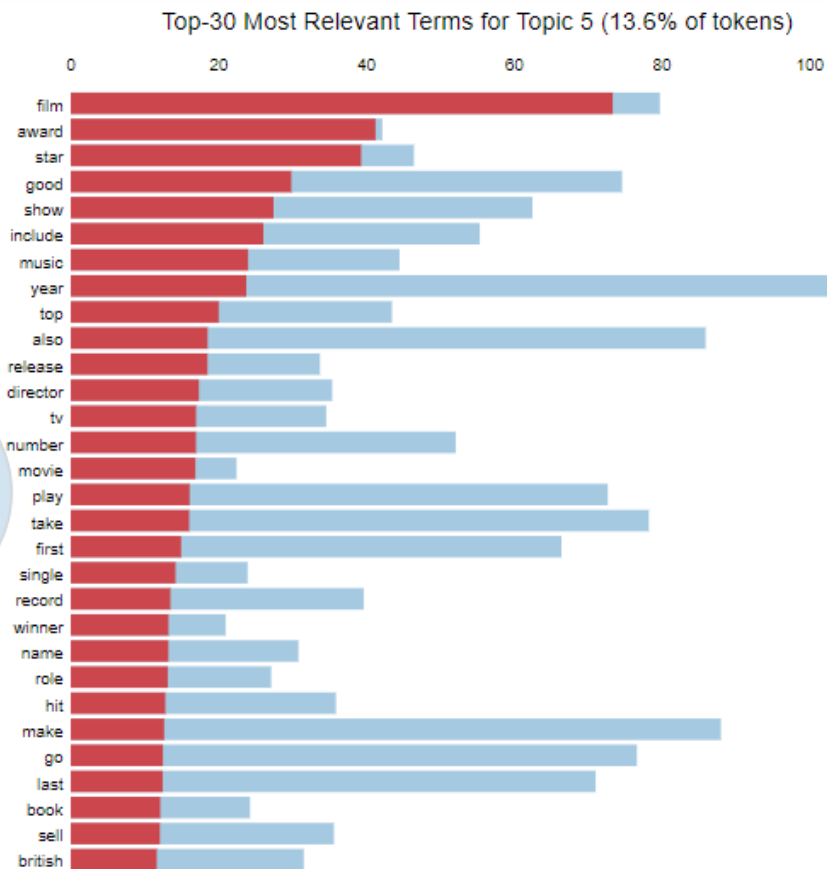
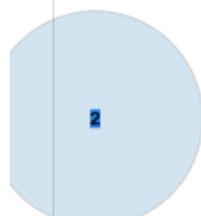
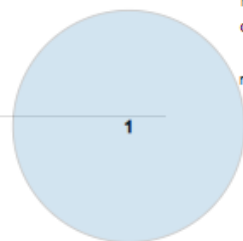
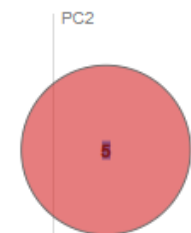
Intertopic Distance Map (via multidimensional scaling)



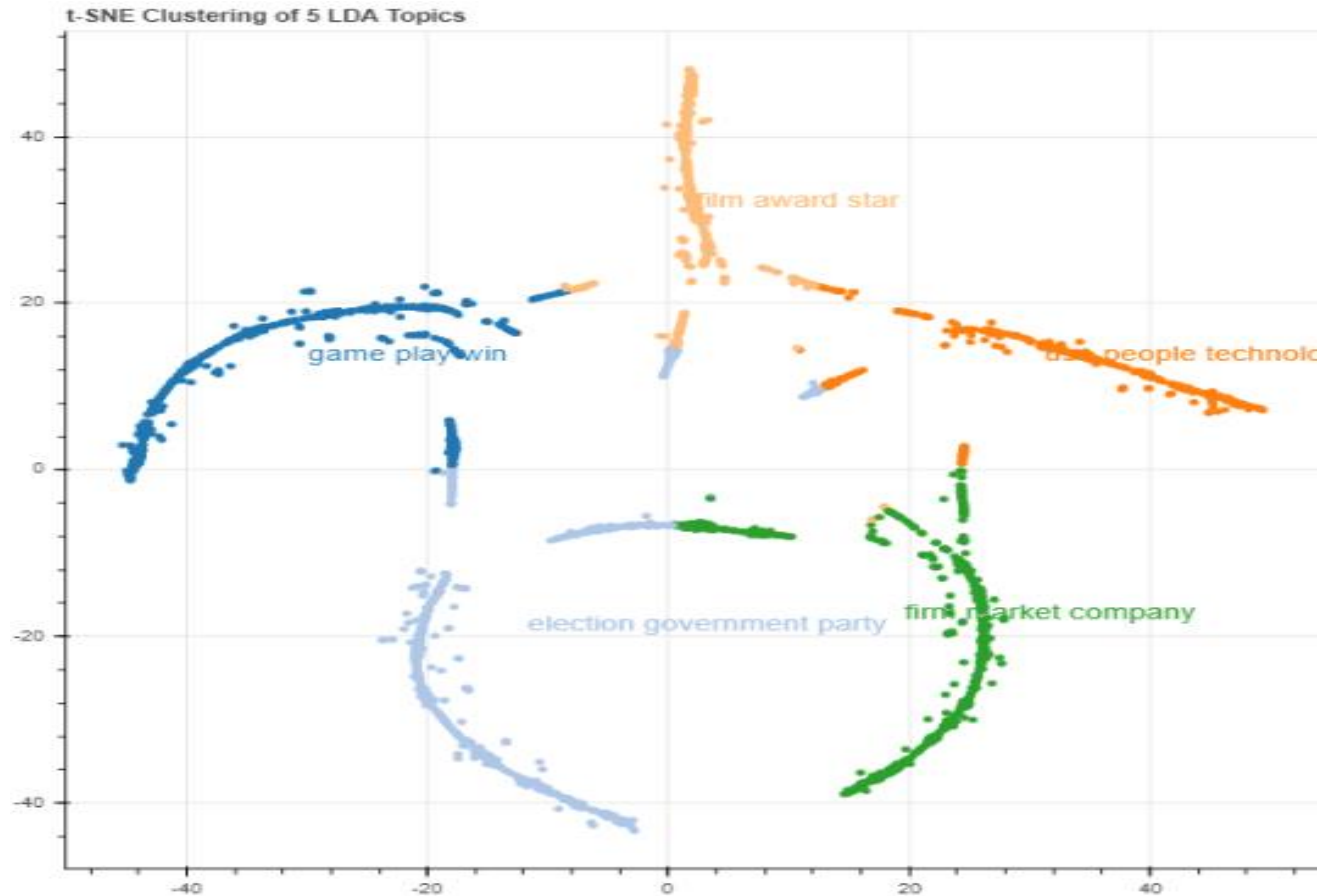
Top-30 Most Relevant Terms for Topic 4 (18.1% of tokens)

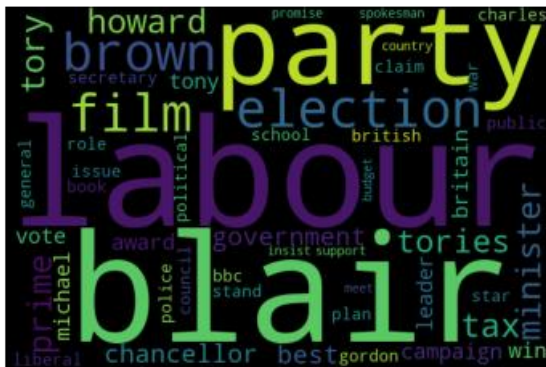


LDA - Cluster 5 : Entertainment

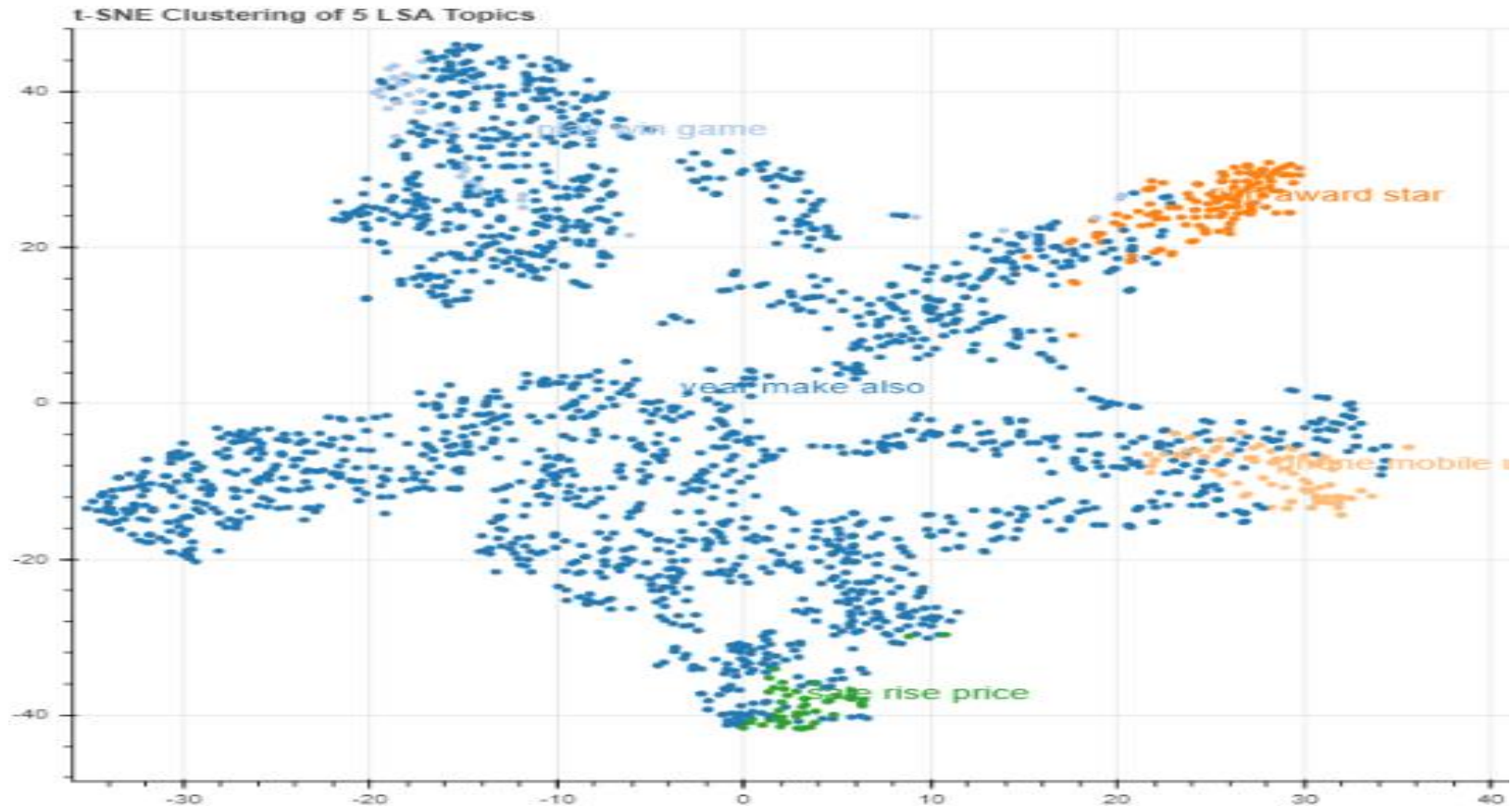


LDA - t-SNE Clustering

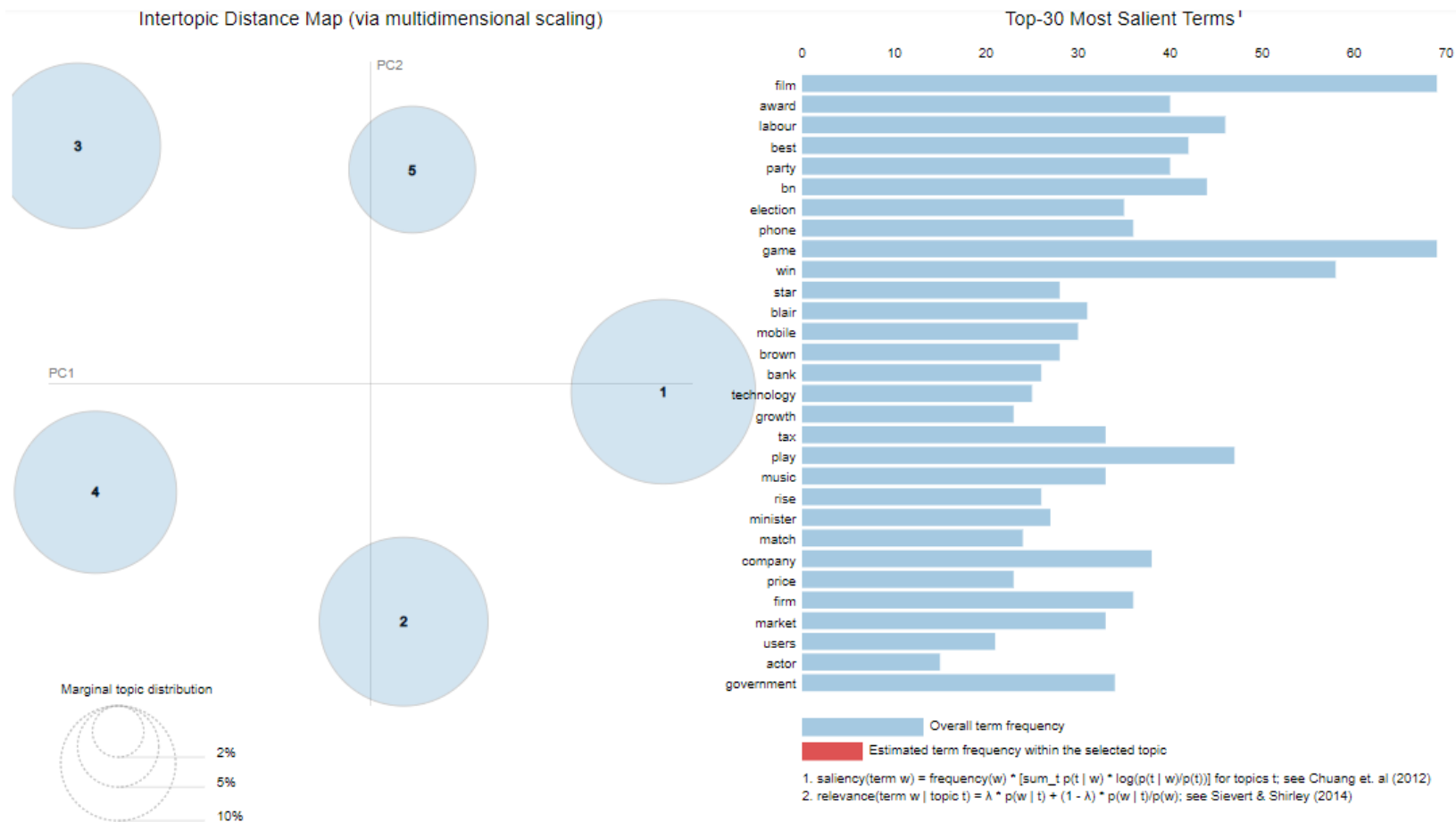




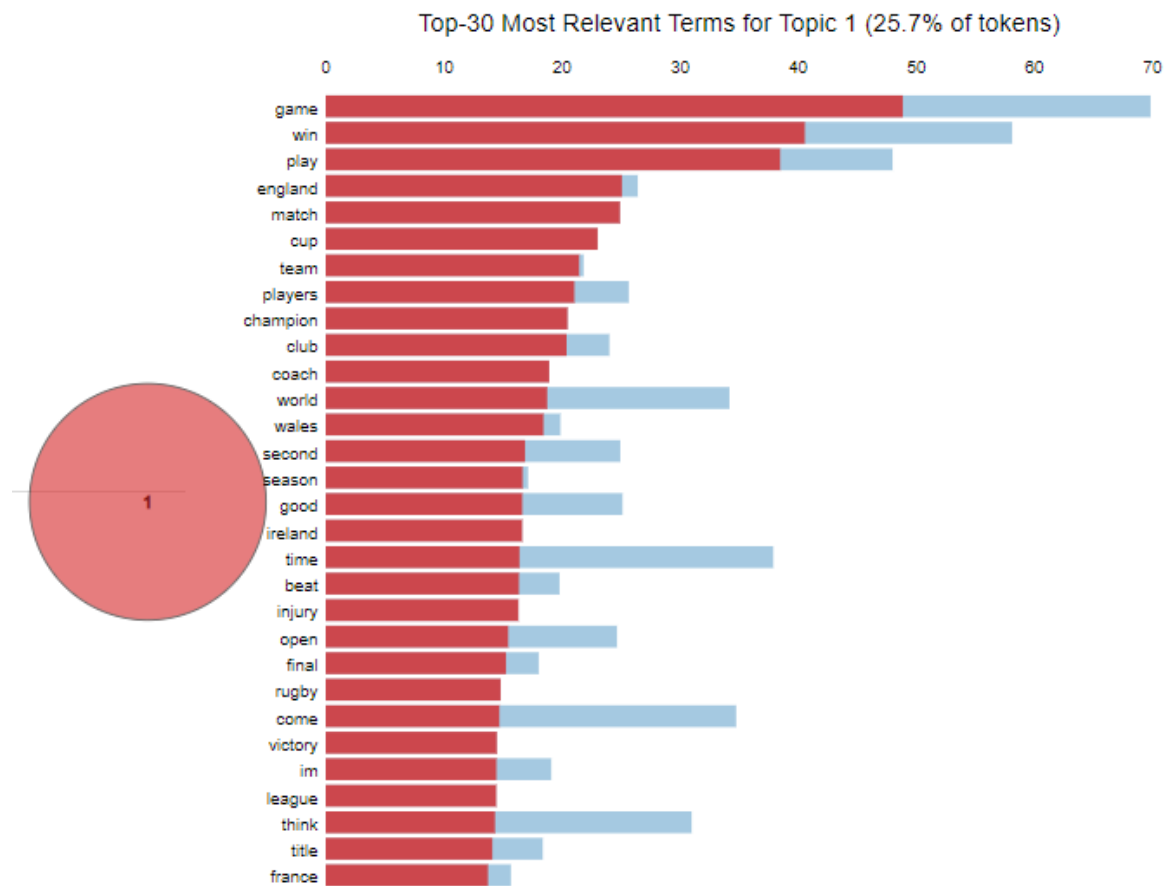
LSA - t-SNE Clustering



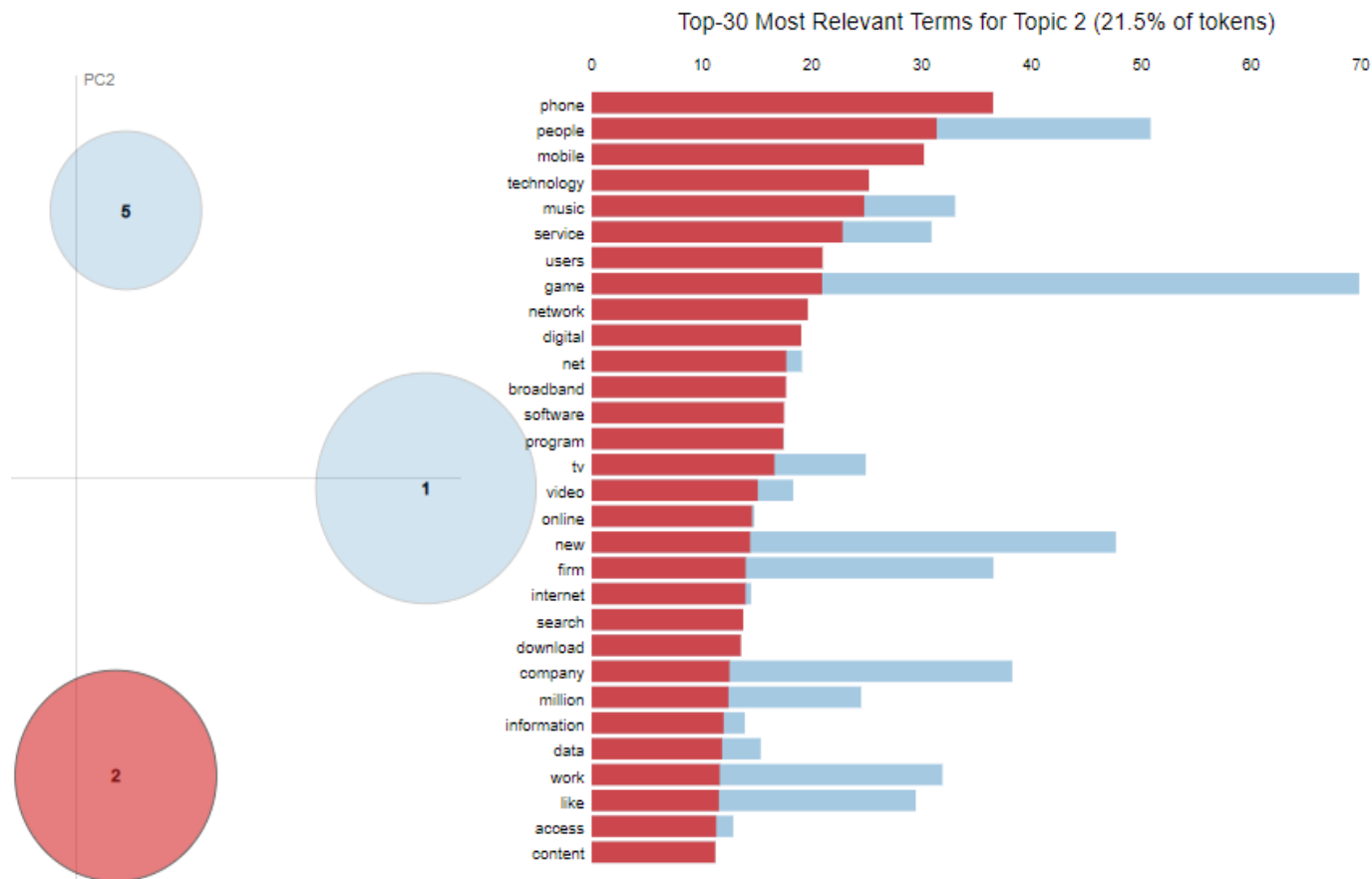
Non-negative Matrix Factorization (NMF)



NMF - Cluster 1 : Sport

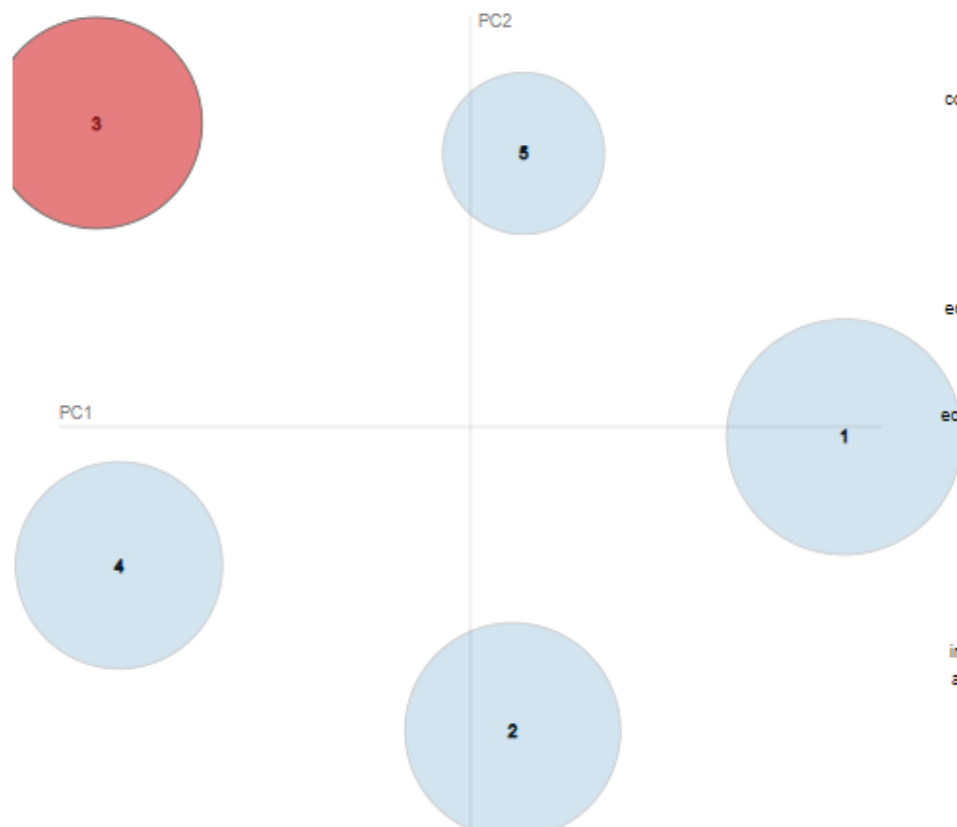


NMF - Cluster 2 : Tech

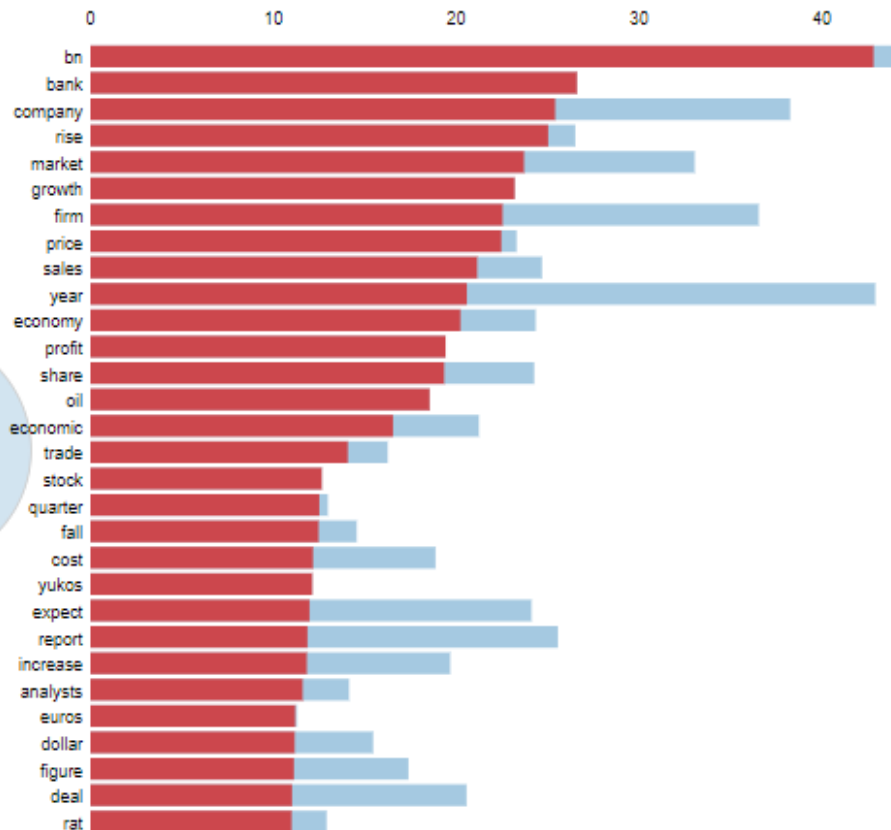


NMF - Cluster 3 : Business

Intertopic Distance Map (via multidimensional scaling)

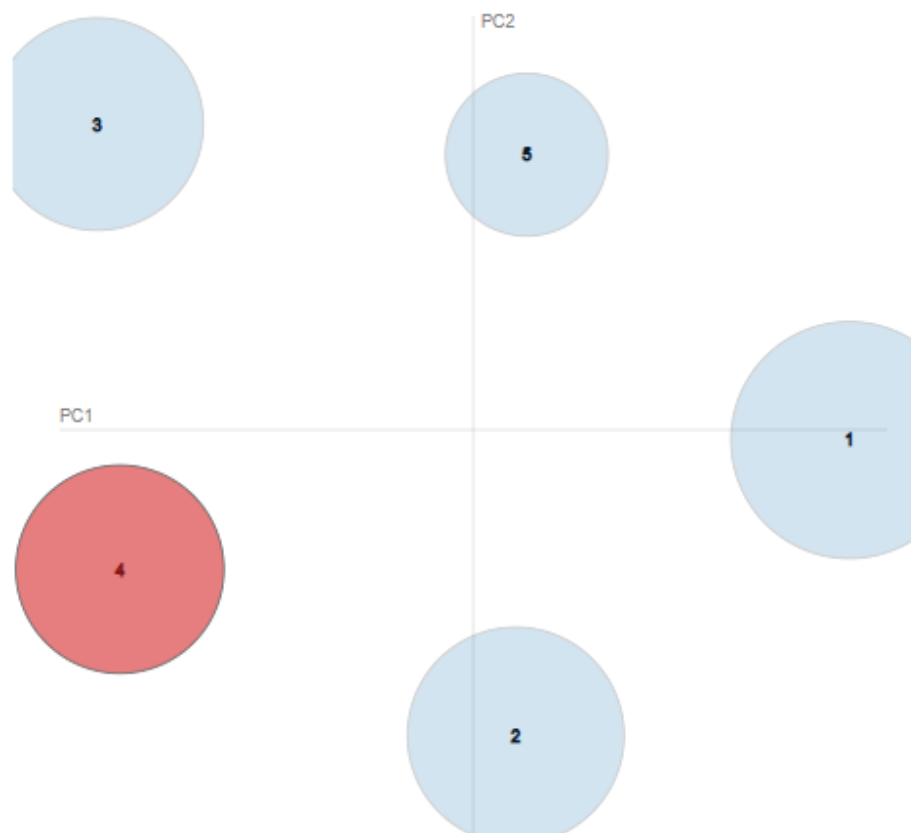


Top-30 Most Relevant Terms for Topic 3 (20.7% of tokens)

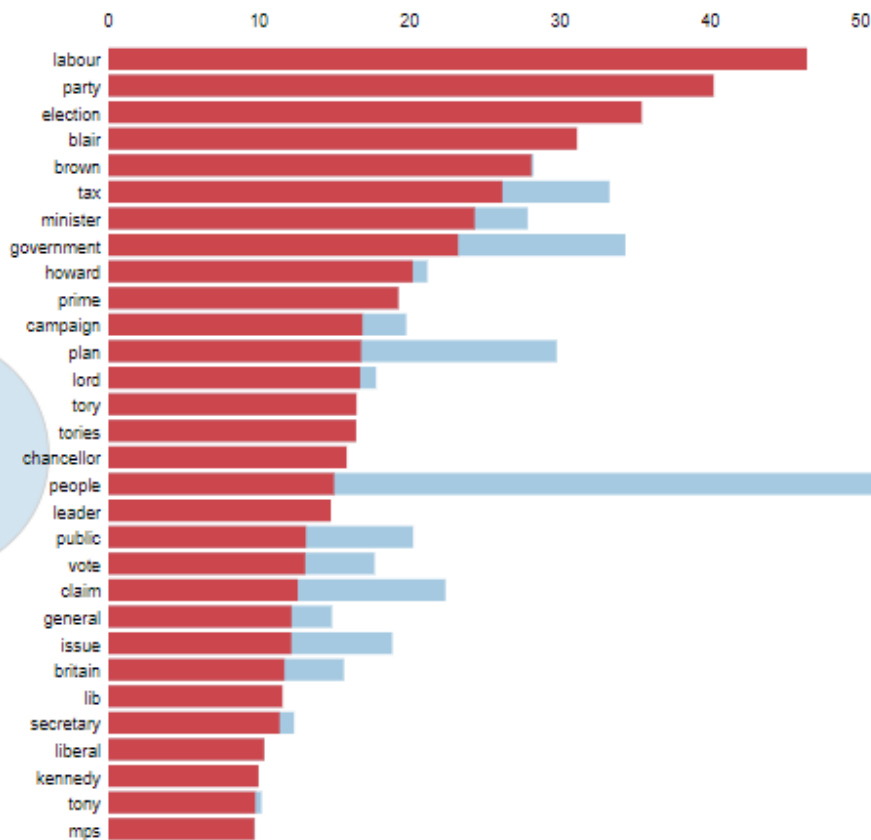


NMF - Cluster 4 : Politics

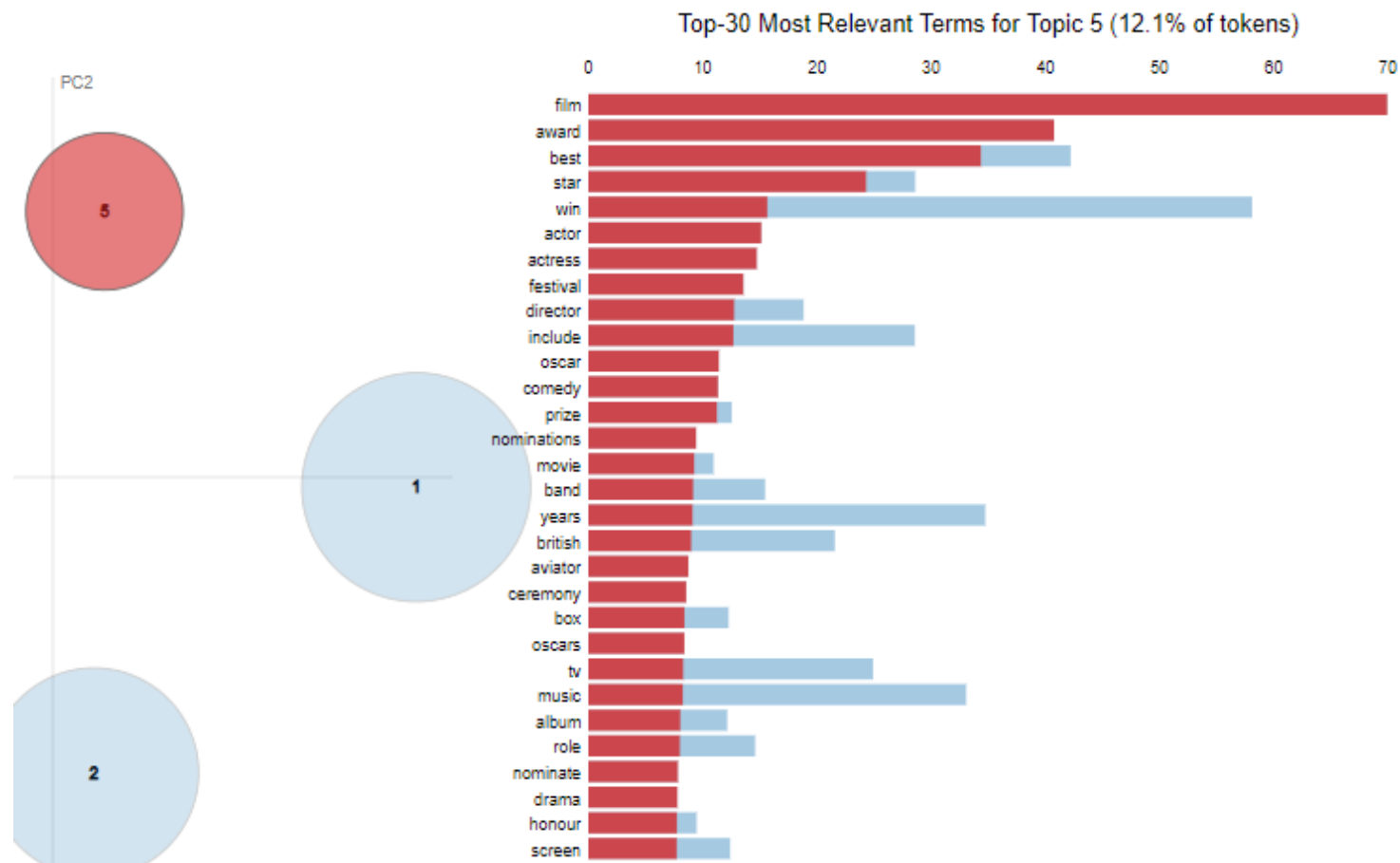
Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 4 (19.9% of tokens)



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