Capstone Project - 4

Topic Modelling On News Articles

Team Member

Harish Kollana

Discussion Points

1. Problem Statement
2. Data Summary
3. Data Cleaning
4. Explorative Data Analysis
5. Topic Modelling

I. LSA - Latent Semantic Allocation

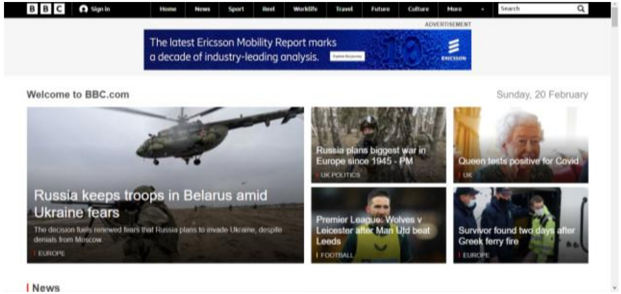
II. LDA - Latent Dirichlet Allocation

6. Challenges
7. Conclusion

The Dilemma

How BBC Works

Users Visits the Pages For News



The latest Ericsson Mobility Report marks a decade of industry-leading analysis.

Welcome to BBC.com

Sunday, 20 February

Russia keeps troops in Belarus amid Ukraine fears

The decision fuels renewed fears that Russia plans to invade Ukraine, despite denials from Moscow.

EUROPE

Russia plans biggest war in Europe since 1945 - PM

UK POLITICS

Queen tests positive for Covid

UK

Premier League: Wolves v Leicester after Man Utd beat Leeds

FOOTBALL

Survivor found two days after Greek ferry fire

EUROPE

News

BBC PORTAL Containing News Articles

BBC: British Broadcasting Corporation

British Broadcasting Corporation (BBC), publicly financed broadcasting system in Great Britain, operating under royal charter.

Problem Statement

In this project our task is to identify major themes/topics across a collection of BBC news articles. By using clustering algorithms such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) etc.

Data Summary

Data Set Name : **bbc**

Data Set Information:
Number of instances: 2225
Number of attributes: 2

Features:
‘news’, ‘topic’

Data Summary

news: News Content

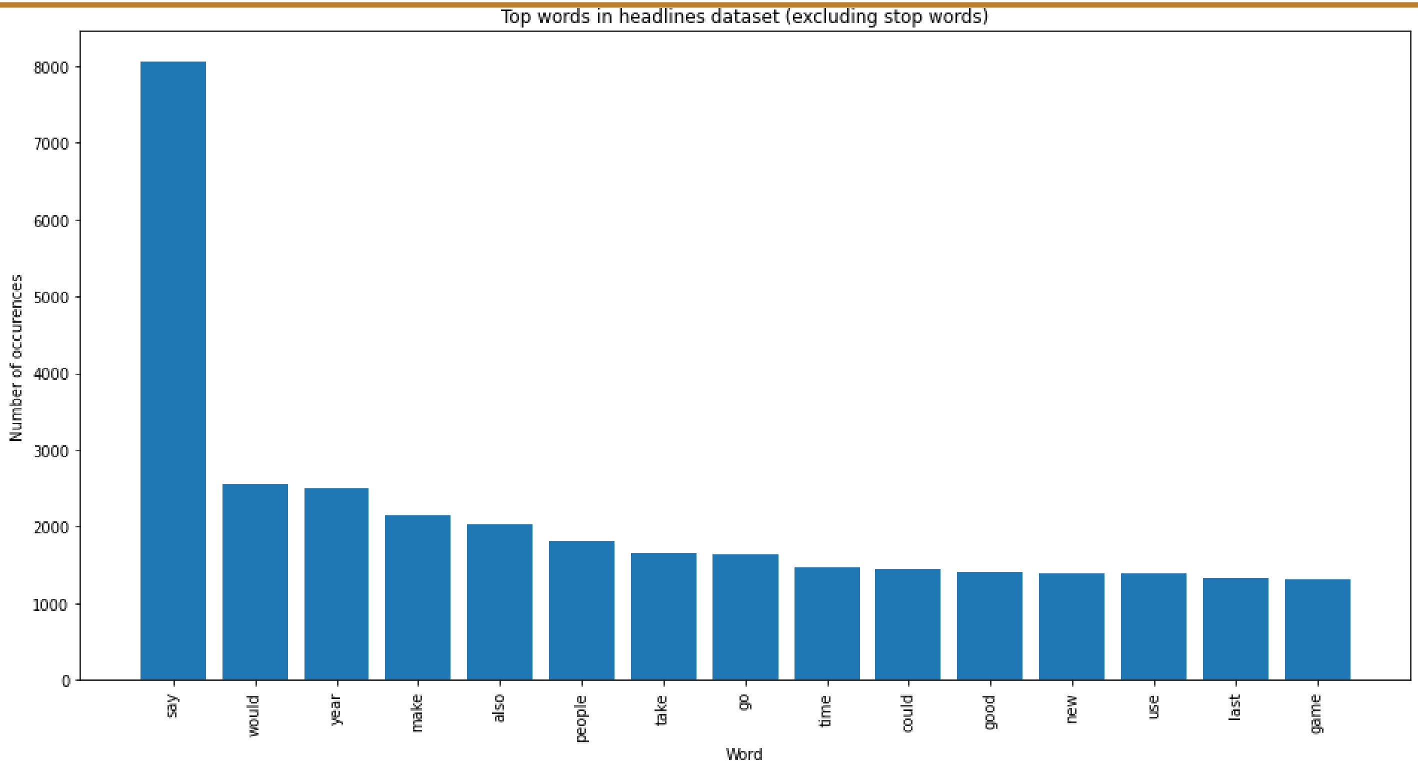
topic : Type of News Category

Data Cleaning

1. Conversion to String Type
2. Removal of Line characters, converting to lower case, removing stop words
3. Lemmatization

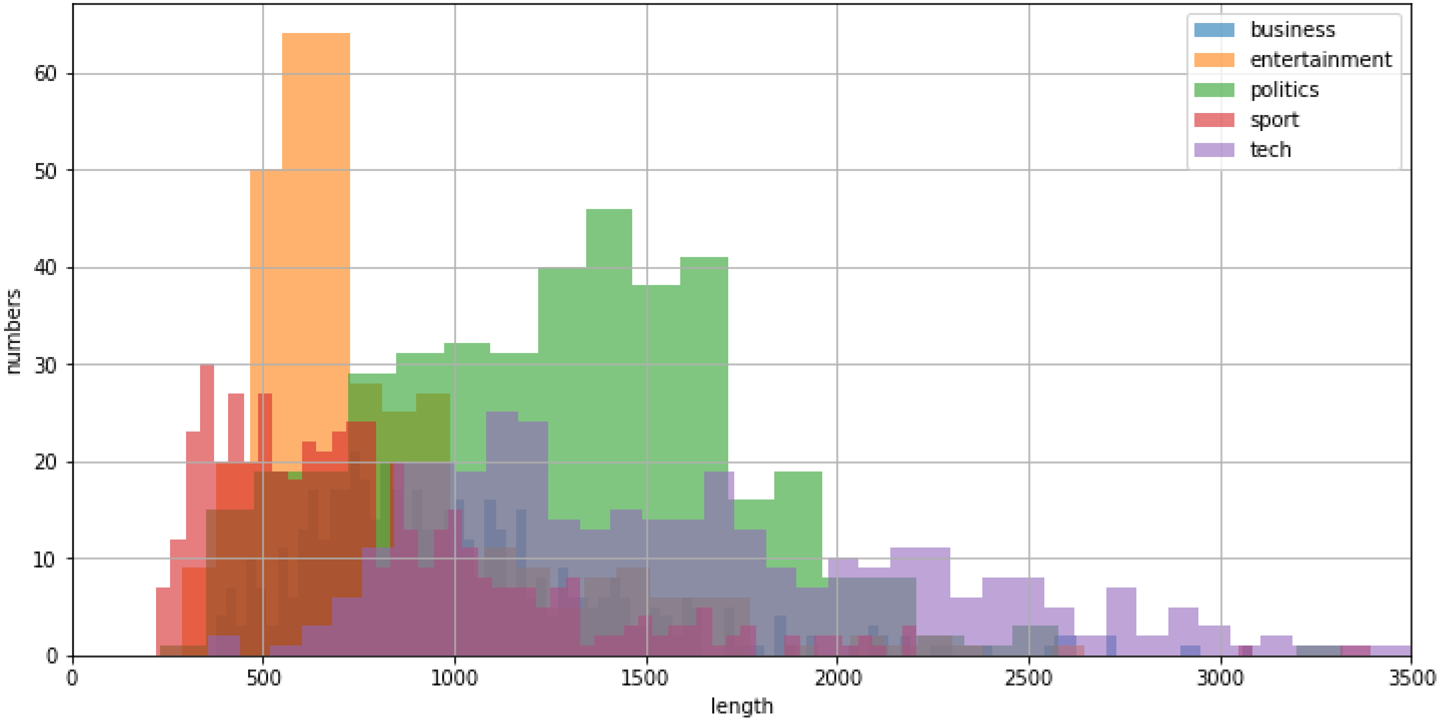
EDA

Top words in headlines dataset (excluding stop words)



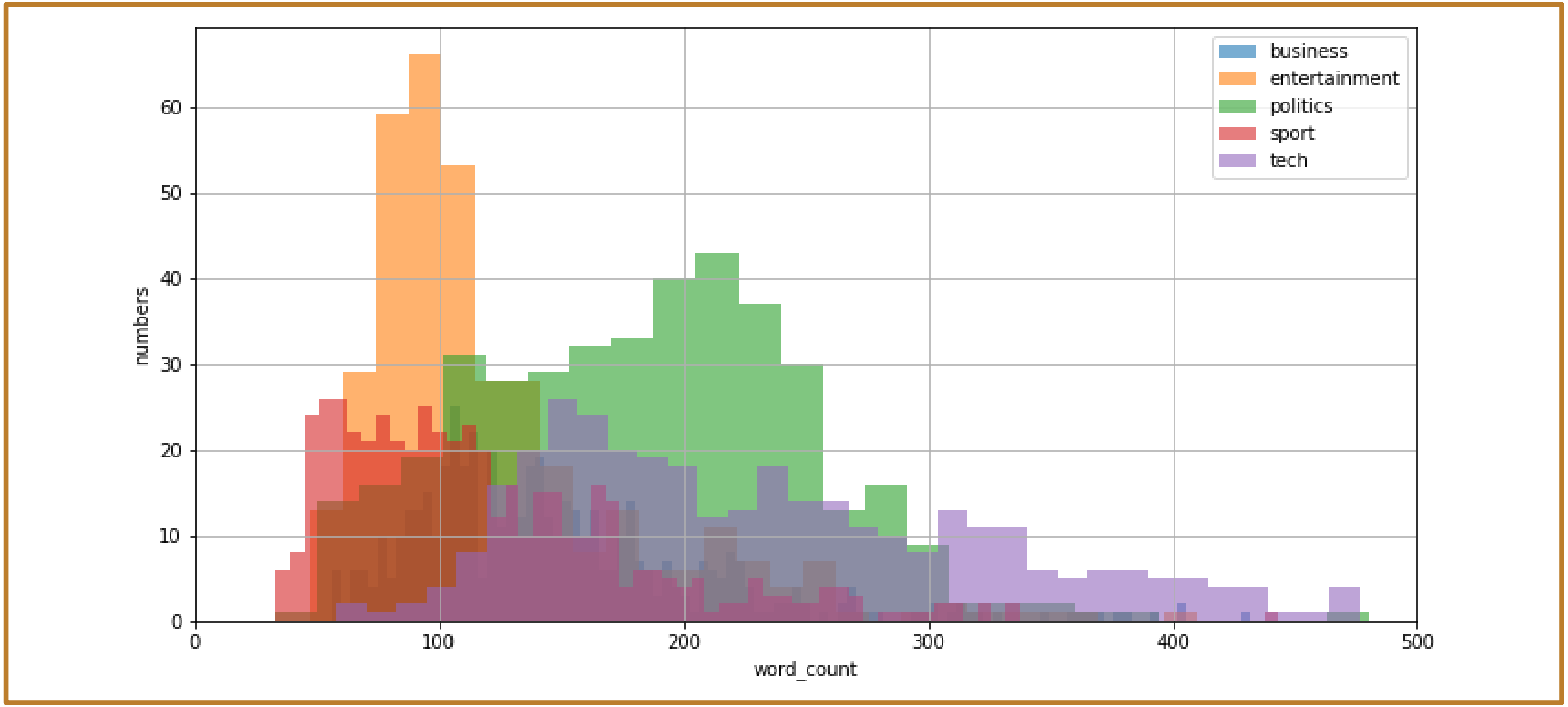
EDA (continued)

Length of News Articles



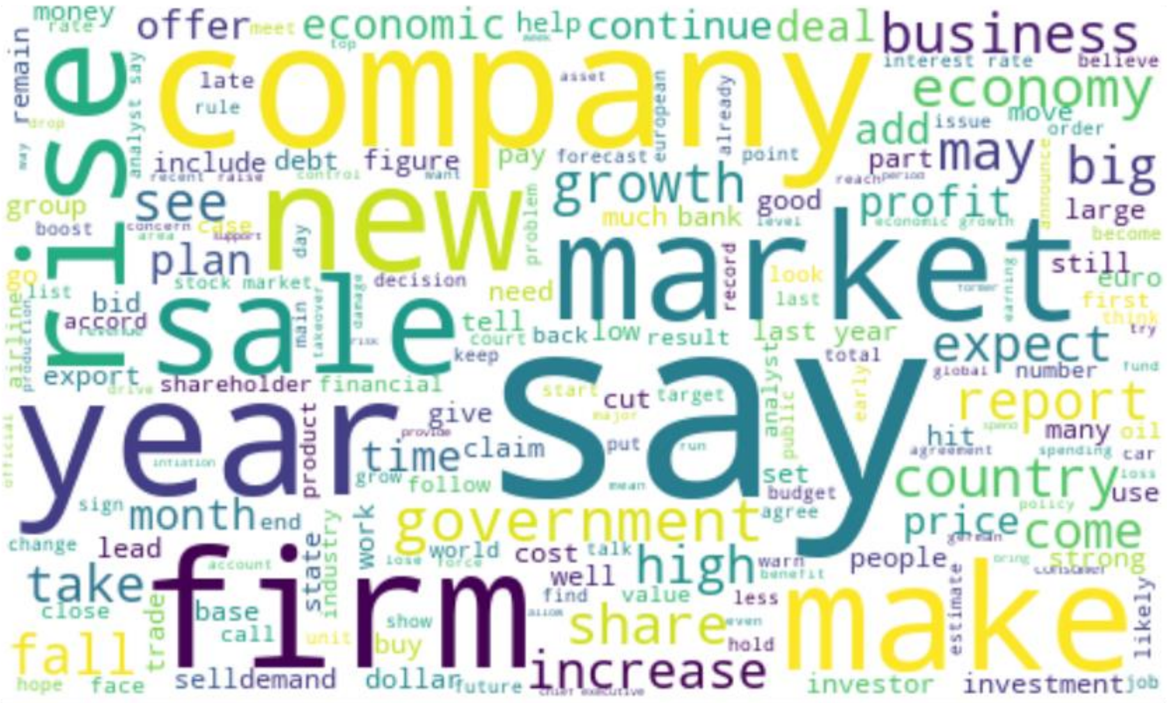
EDA (continued)

Word Count Of News Articles



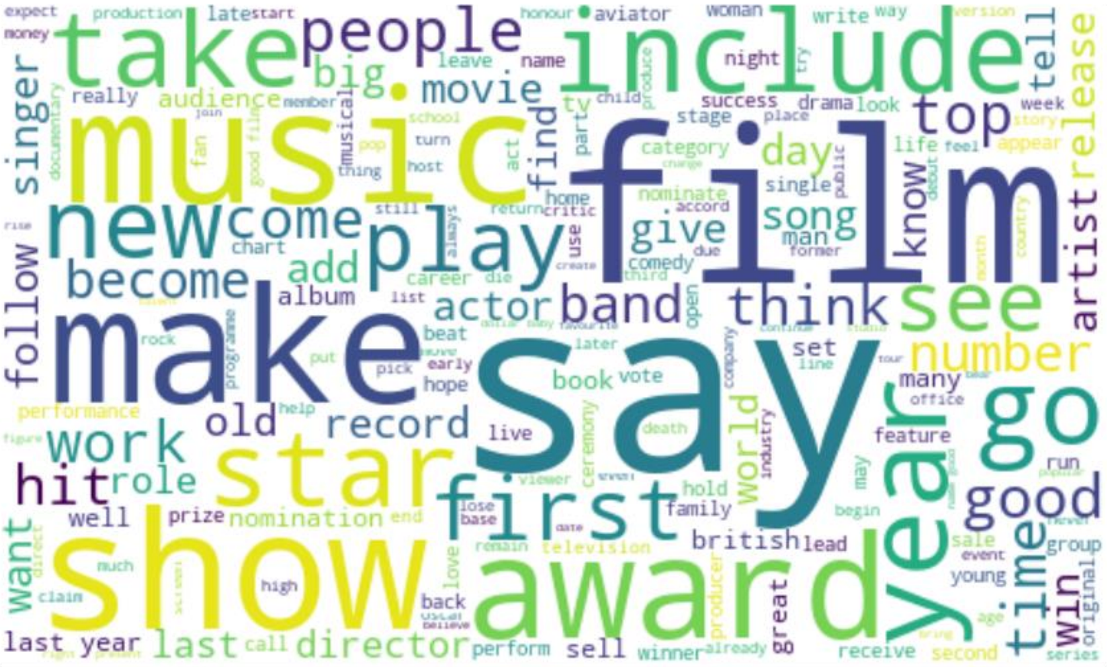
EDA (continued)

Wordcloud of Business



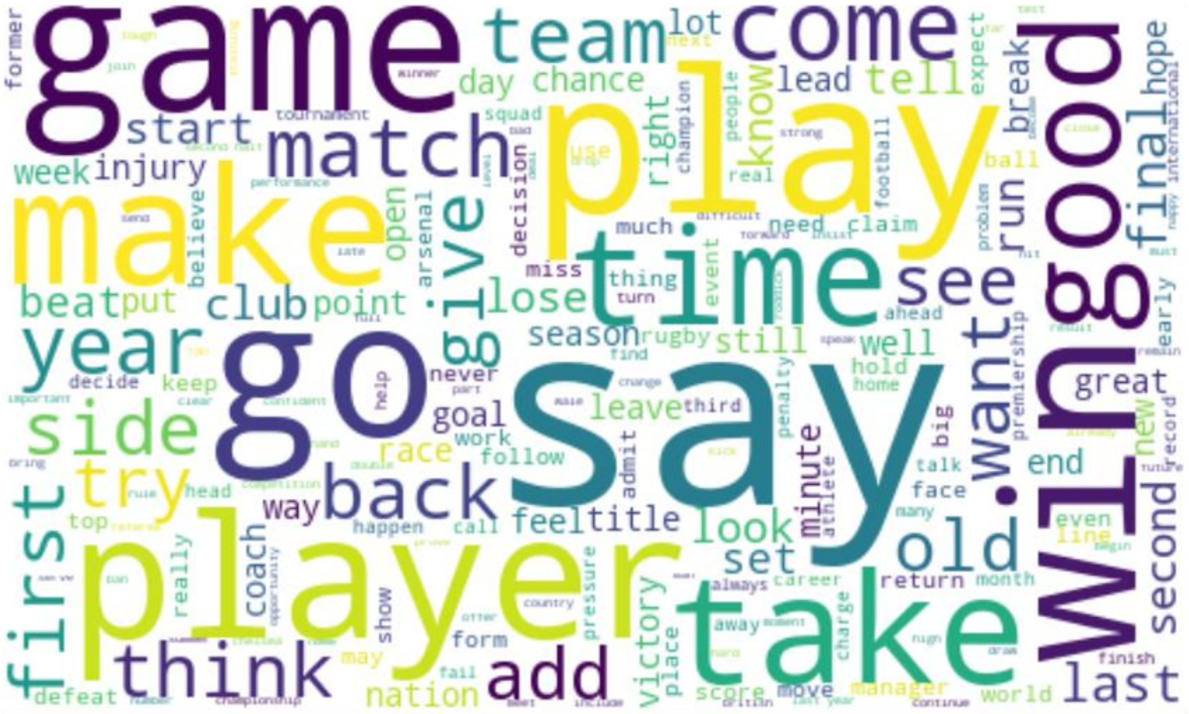
EDA (continued)

Wordcloud of Entertainment



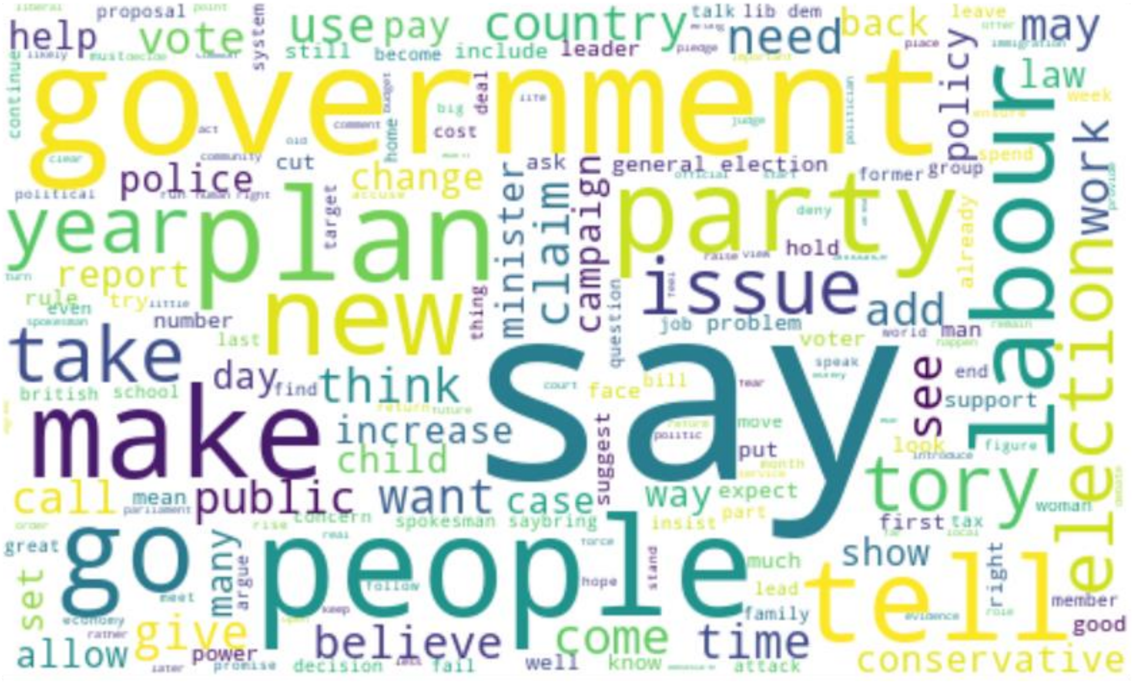
EDA (continued)

Wordcloud of Sport



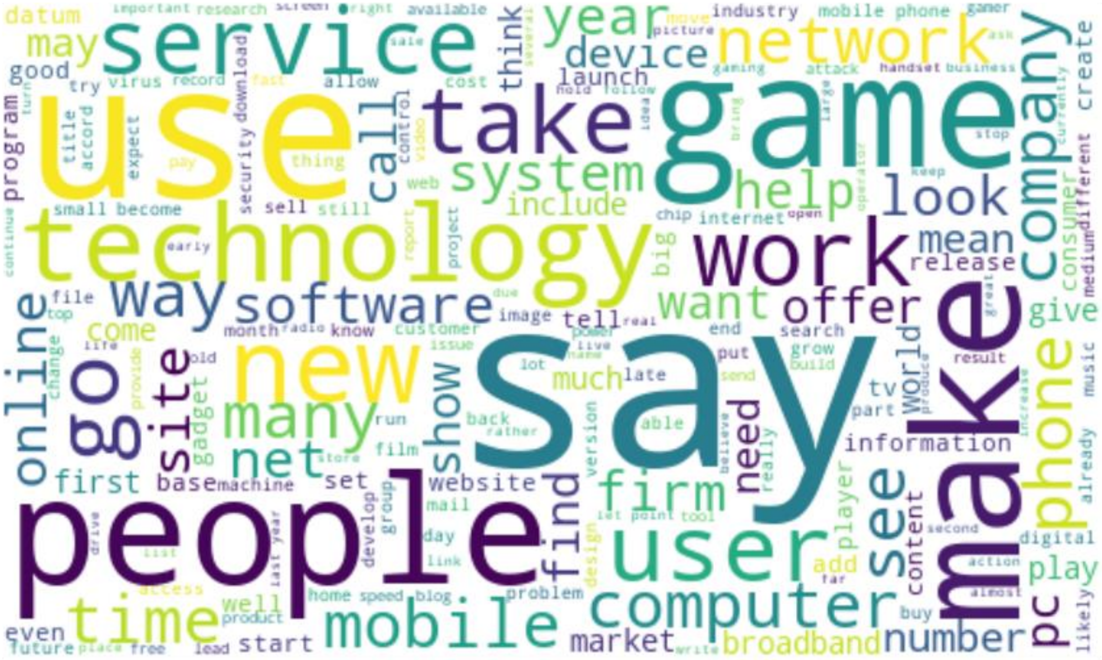
EDA (continued)

Wordcloud of Politics



EDA (continued)

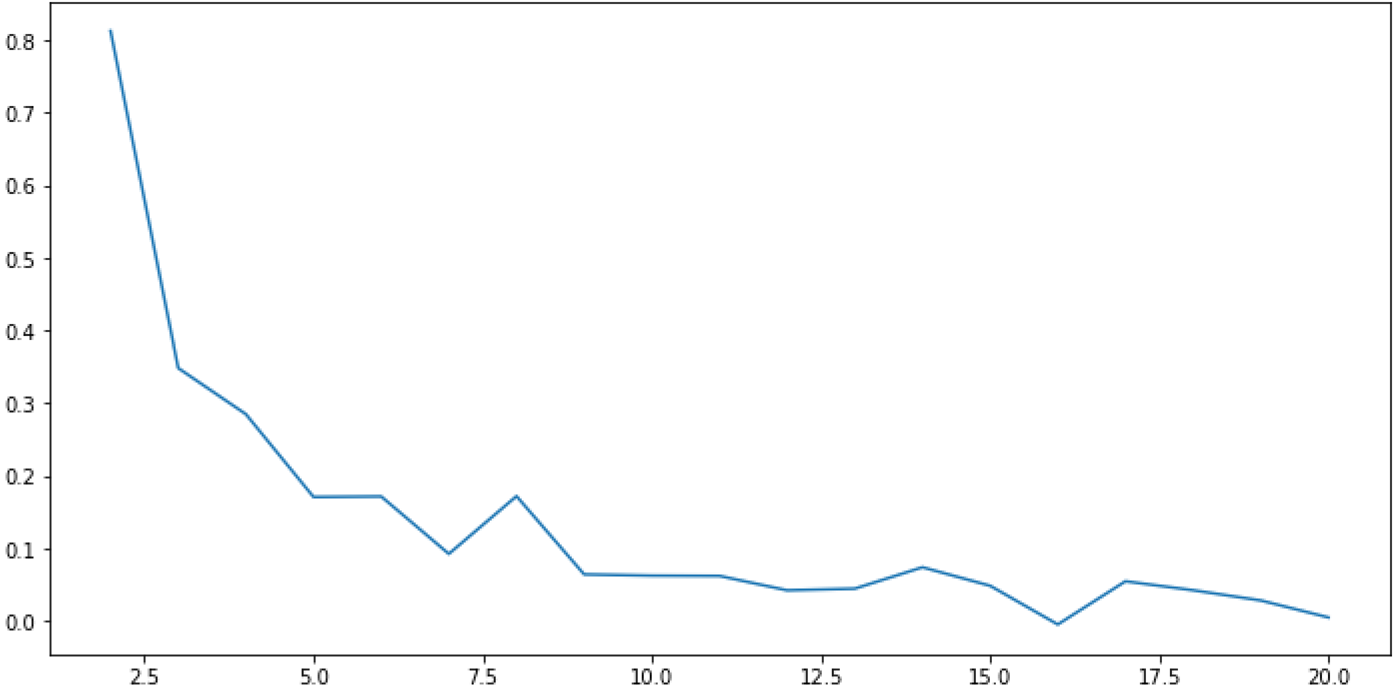
Wordcloud of Tech



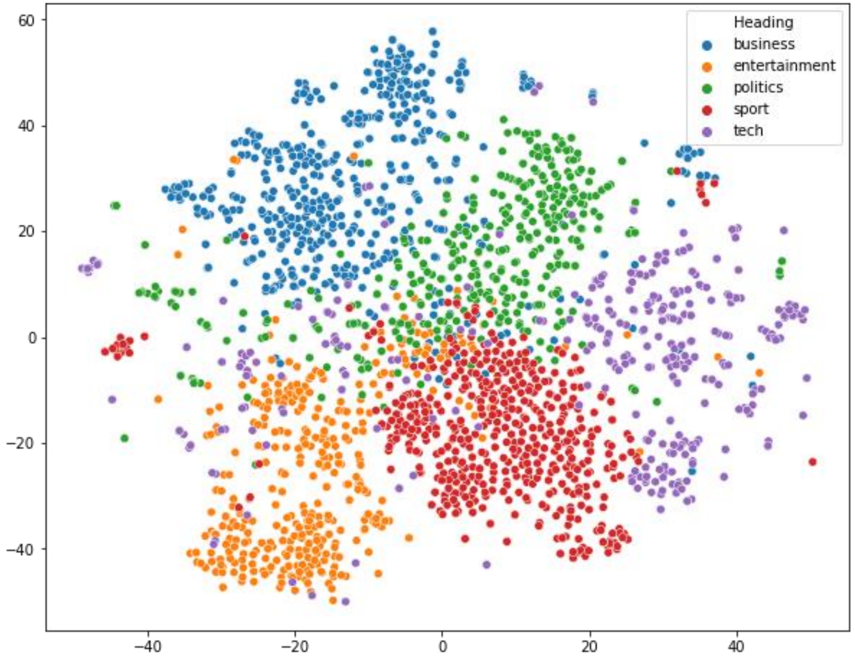
Topic Modelling

1. LSA - Latent Semantic AllocationRandom Forest Model
2. LDA - Latent Dirichlet Allocation

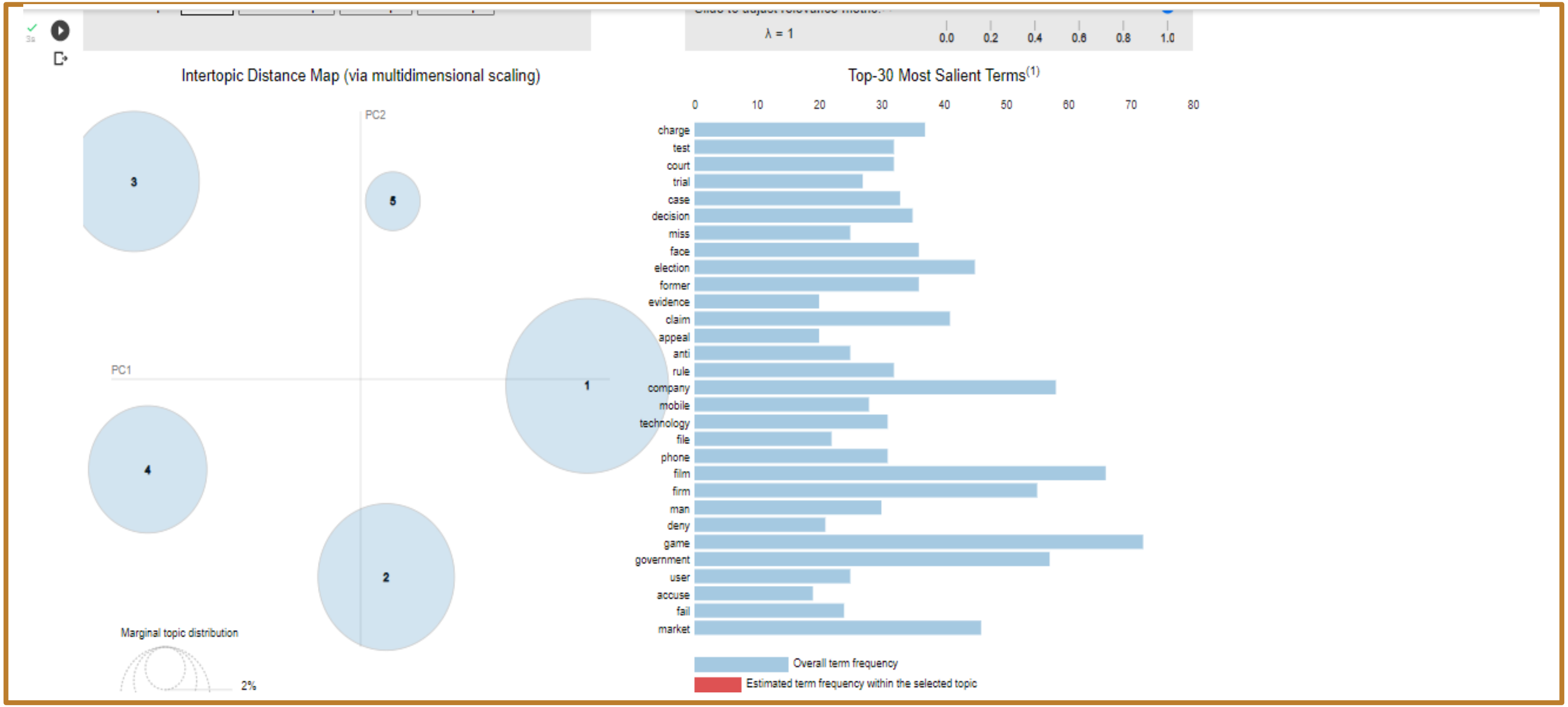
silhouette_score by Latent Semantic Allocation



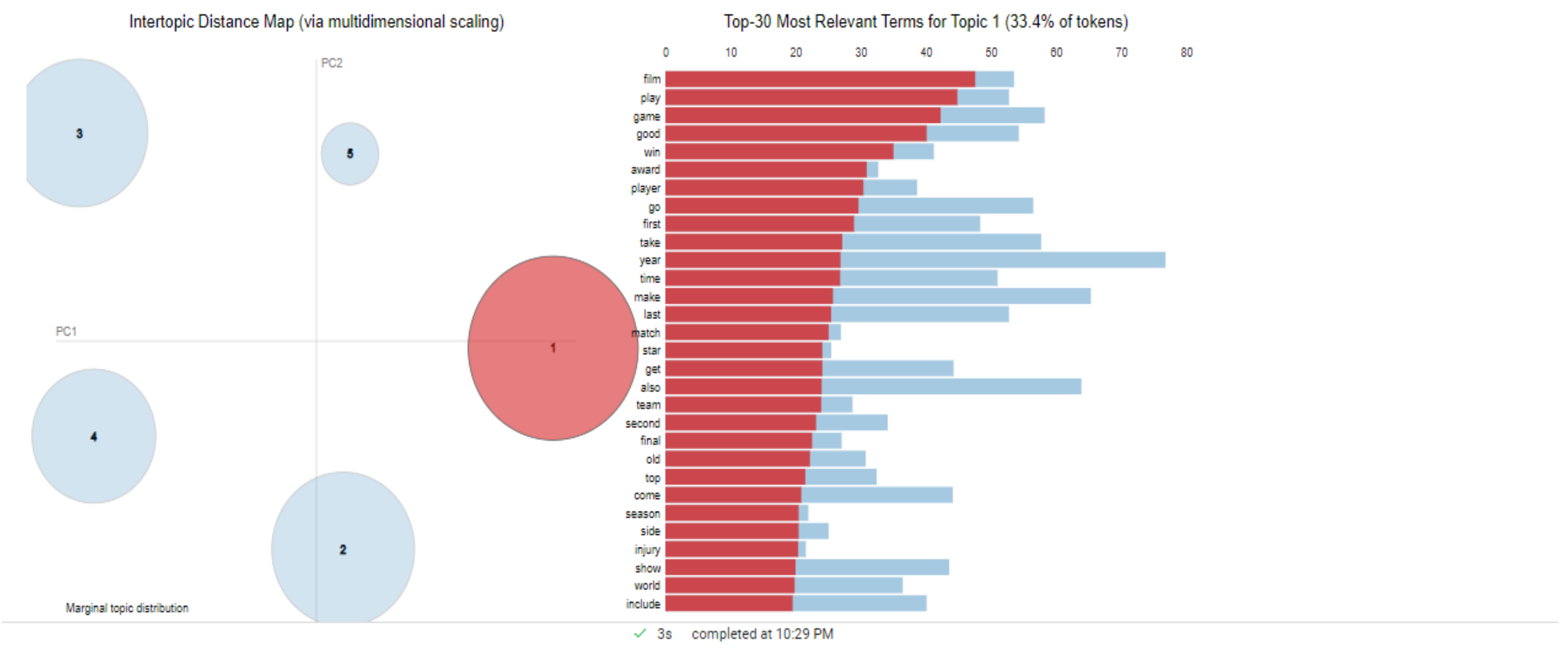
Scatter Plot Of Topics by Latent Semantic Allocation



PyLDAvis Panel by Latent Dirichlet Allocation

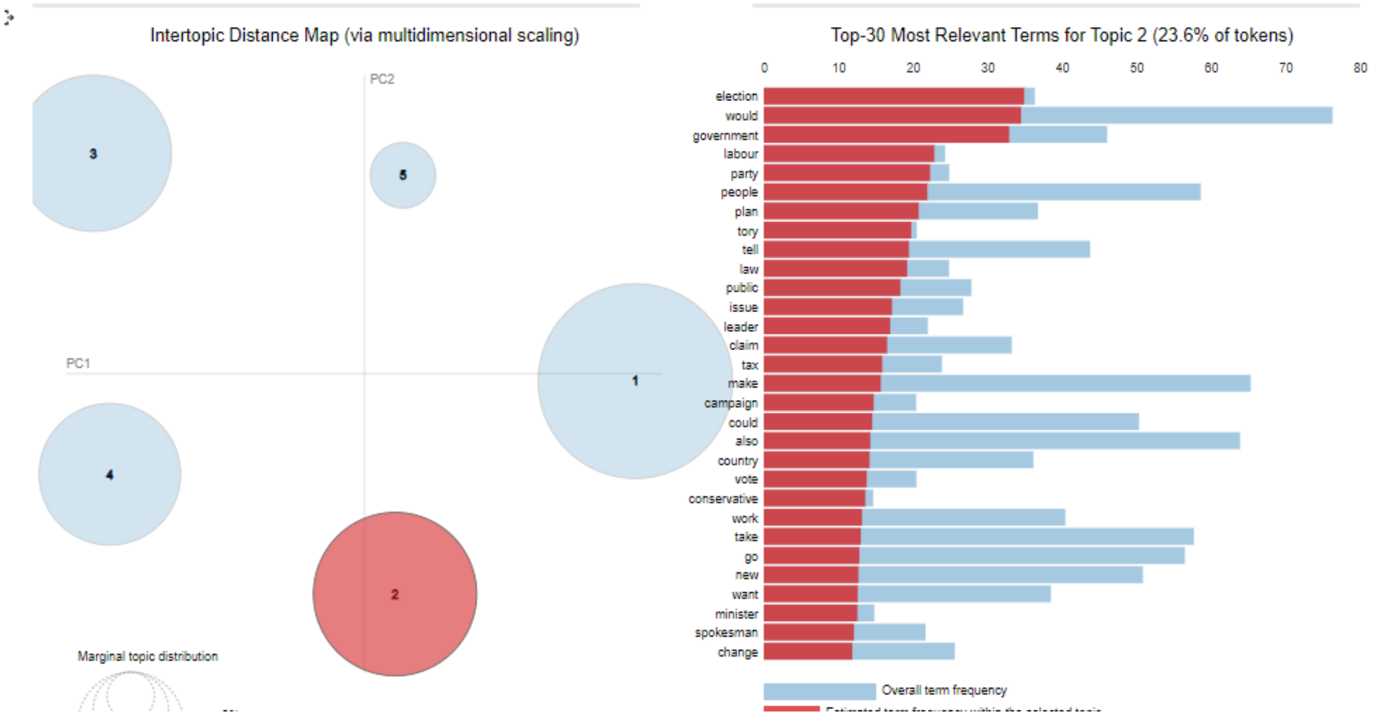


LDA Cluster 1 : Entertainment

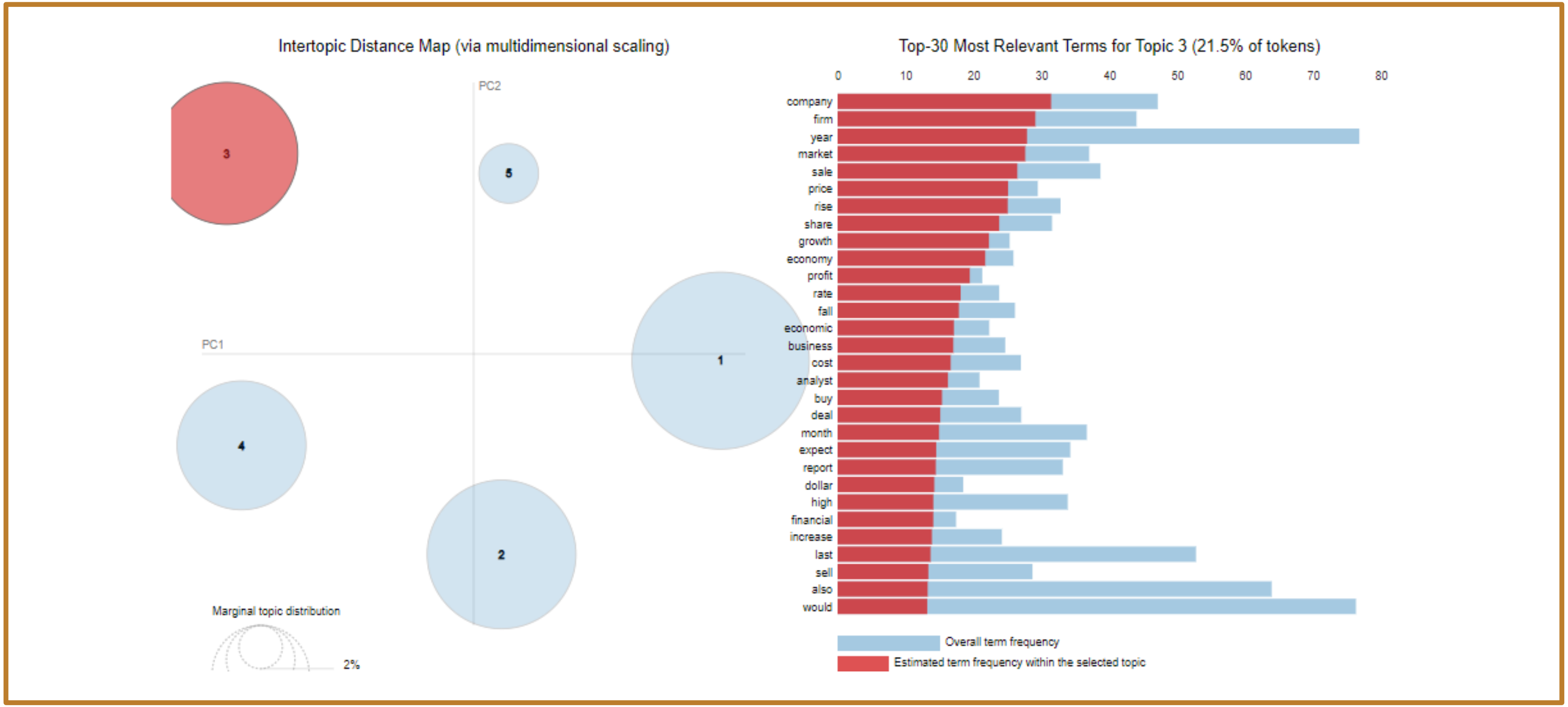


✓ 3s completed at 10:29 PM

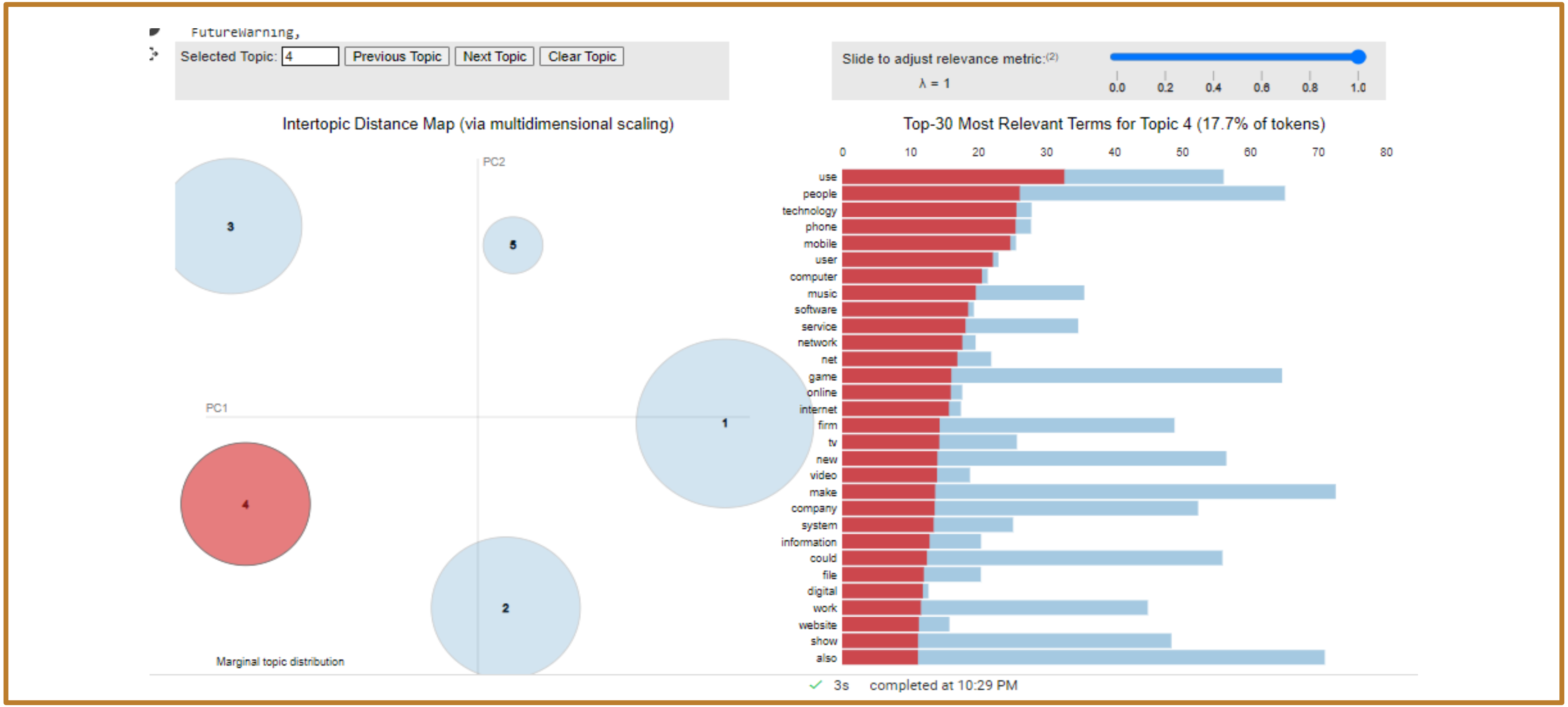
LDA Cluster 2 : Politics



LDA Cluster 3 : Business



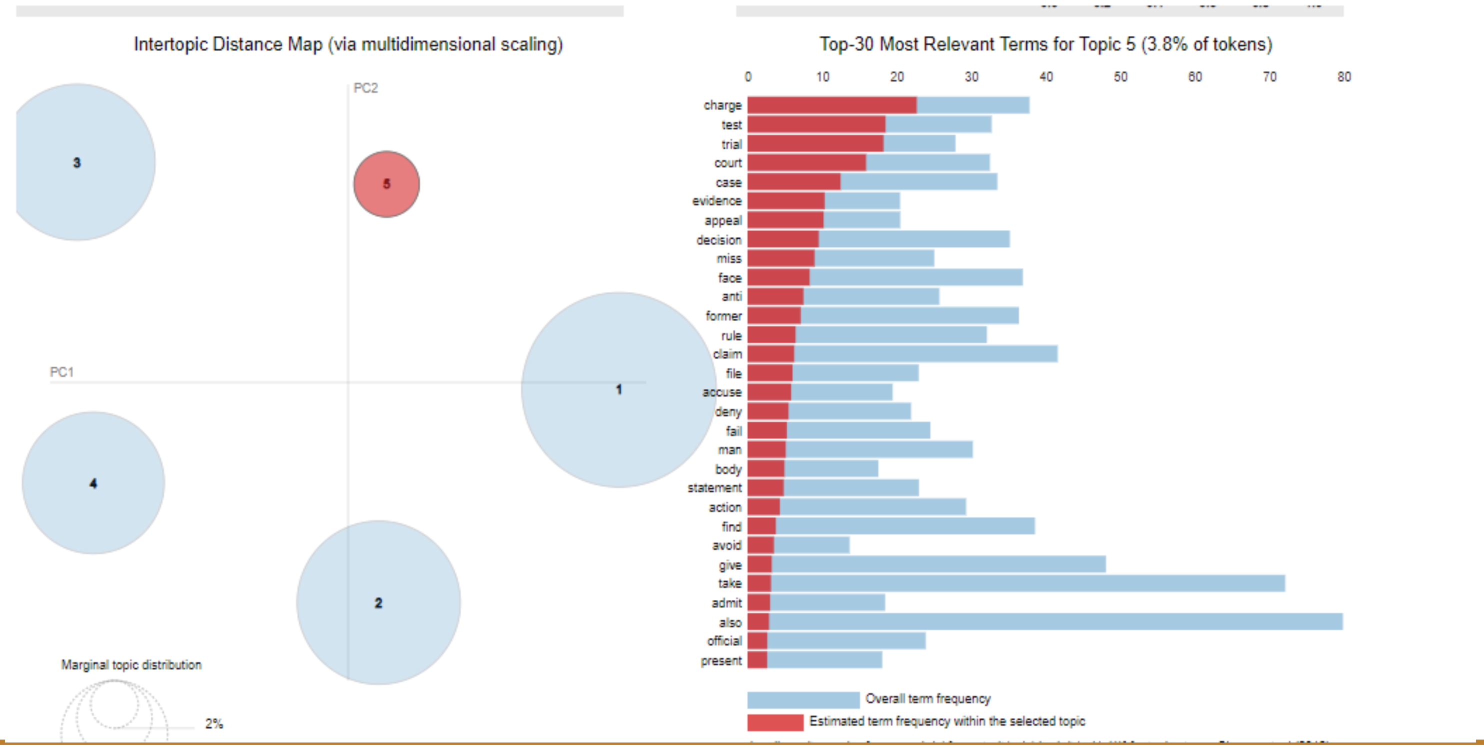
LDA Cluster 4 : Tech



```
FutureWarning,
```

✓ 3s completed at 10:29 PM

LDA Cluster 5 : Sport



Challenges

1. Data Cleaning.
2. Difficulty in Algorithm Implementation.

Conclusion

In this Notebook we had analyzed the BBC articles using LDA and LSA Topic modelling techniques and found lsa seems more impactful on segregation of topics.

In future we can use one of model to predict the user input of text query to type of news. We can recommend news articles to the users by following these methods.

Thank You