Capstone Project - 4

Topic Modeling on News Articles

Team members

Balaji B. Jadhav
Anant M. Patil
Nishigandha Ingale

Content

- Problem definition
- Introduction
- Topic modeling
- EDA on given data
- Model implementation
- Model validation and selection
- Conclusion
- References

Problem definition

- In this project your task is to identify major themes/topics across a collection of BBC news article. With using clustering algorithms such as Latent Dirichlet Allocation (LDA).

Introduction

- In the Web era, people tend to rely on the Web to receive news instead of traditional ways such as newspapers.
- In the past decades, text classification has been a hot topic and received attention from many scholars.
- In areas such as natural language processing, information retrieval, and machine learning, etc.
- A major source for such similarity measurement is the content of the news articles, which is mostly textual.

Introduction

- Therefore, techniques of text classification or text mining were commonly adopted to tackle the news classification tasks.
- Aim of the project is to put a model on given data for classification of news types.
- Data taken for this project is BBC news data.

Topic modeling on Data

- Topic modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents.
- Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic.

EDA

Digging into data we understand that

- There is no null value in the data set.
- News: Text document of news article of various types.
- Type: Type of news article such as Business, Politics, Tech, Entertainment and Sports.

EDA

- We can see there is no null values in the given data set.

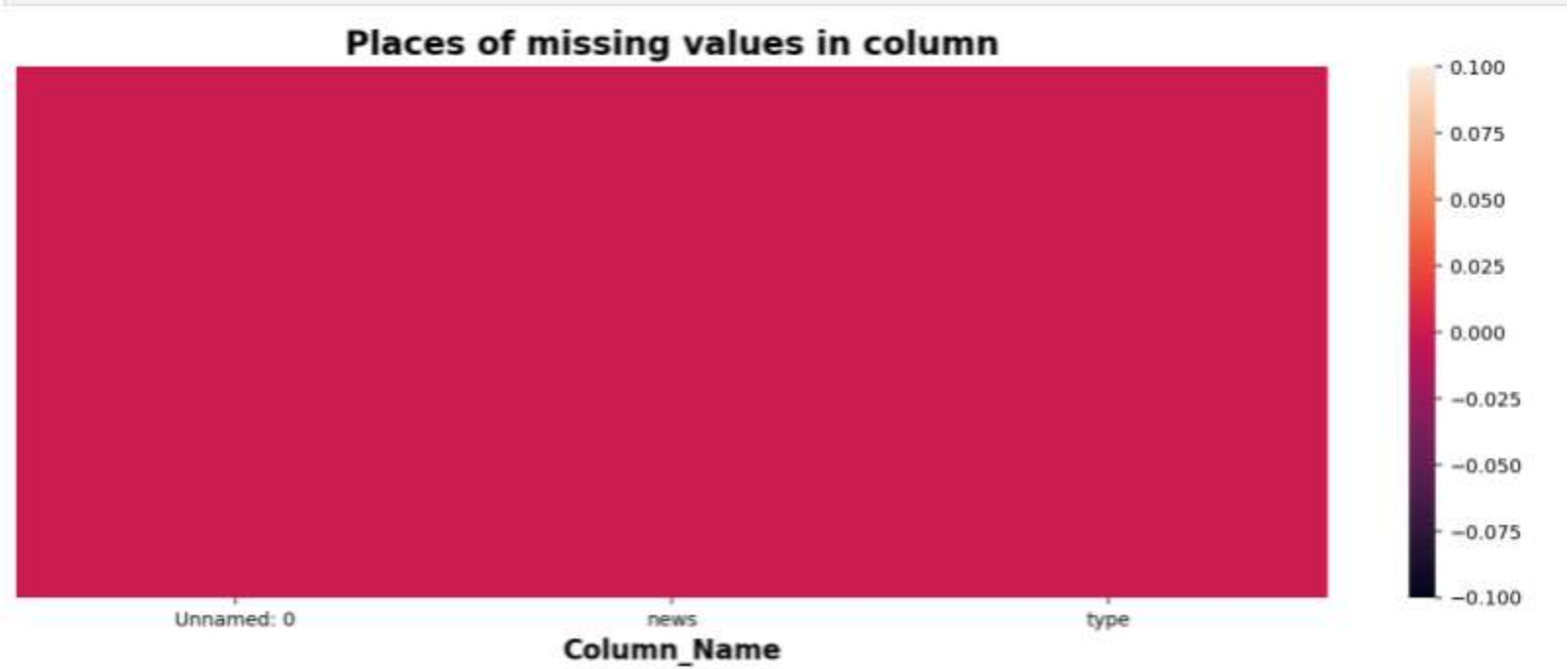


Fig 1: Missing values in data set.

EDA

- Count of news type in data

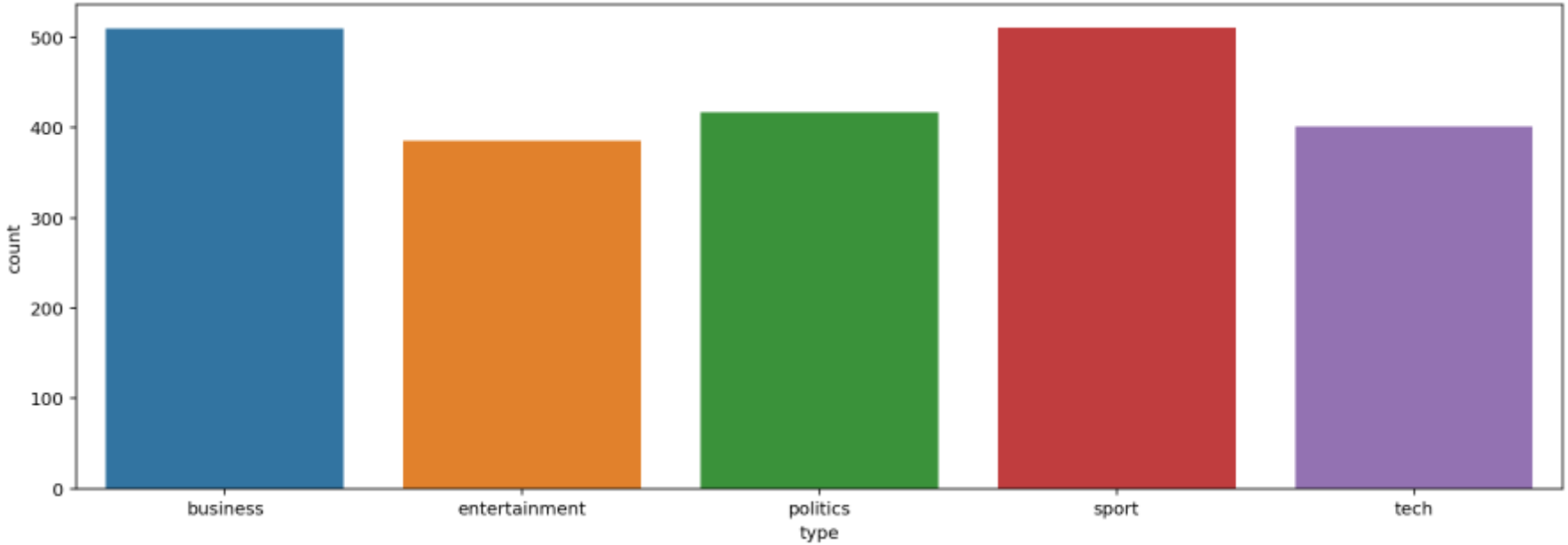


Fig 2: News type count plot.

EDA

- Max news length is of 25000 words.

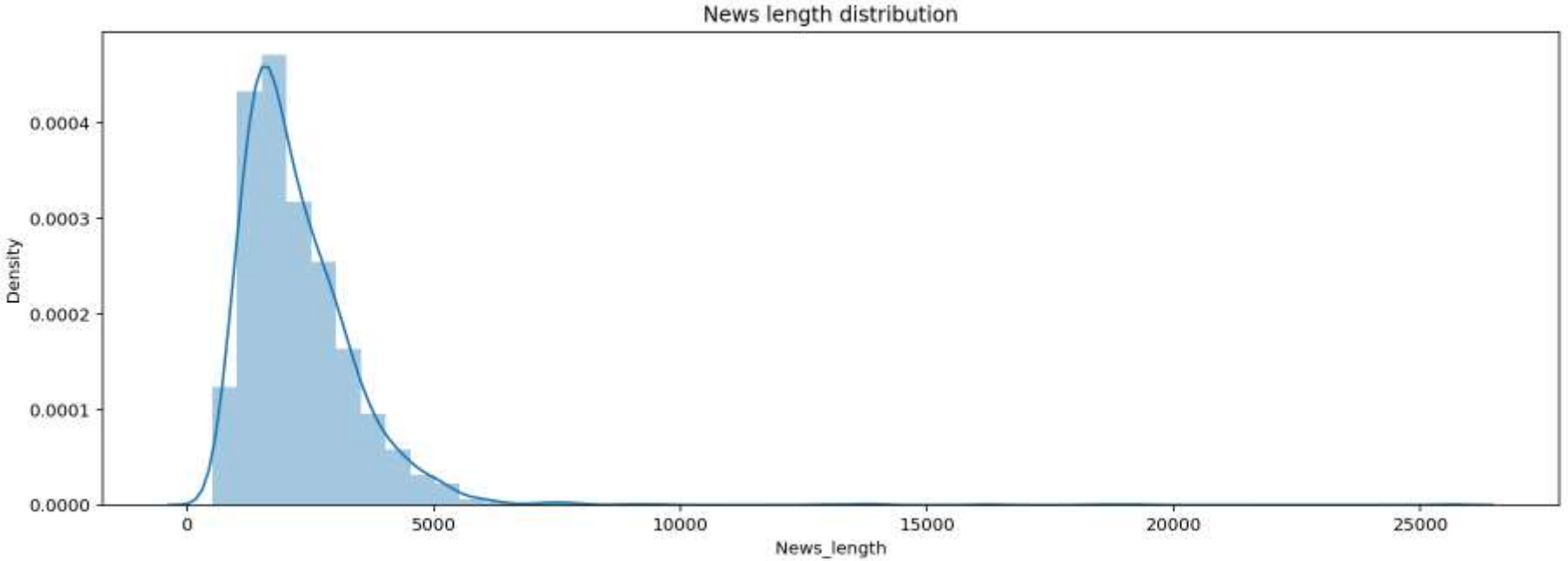


Fig 3: News length distribution.

EDA

- Overall news length in total text.

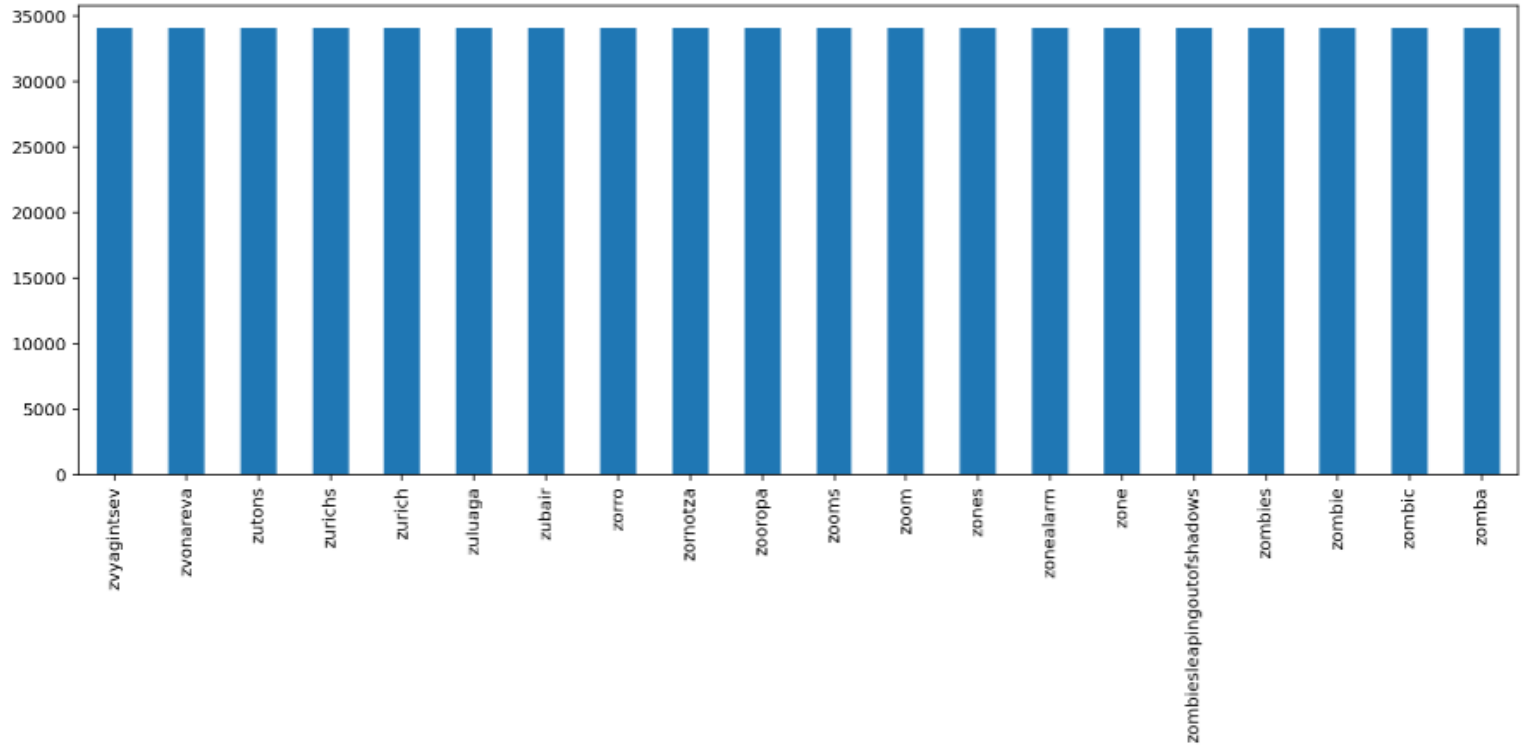


Fig 4: News words (Overall).

EDA

- Overall news words after applying stemming.

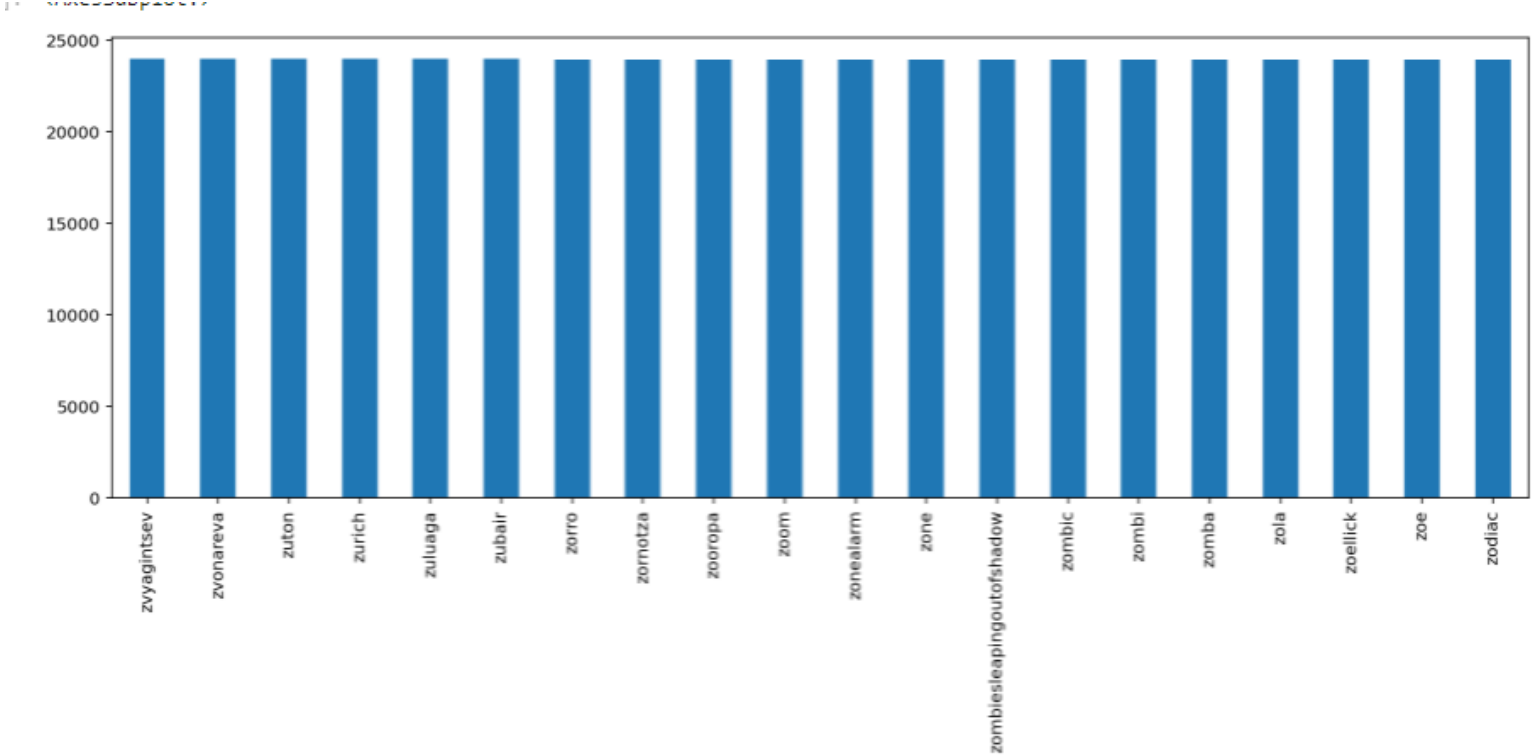


Fig 5: News words after applying stemming.

EDA

- Count of words in Business data after applying stemming.

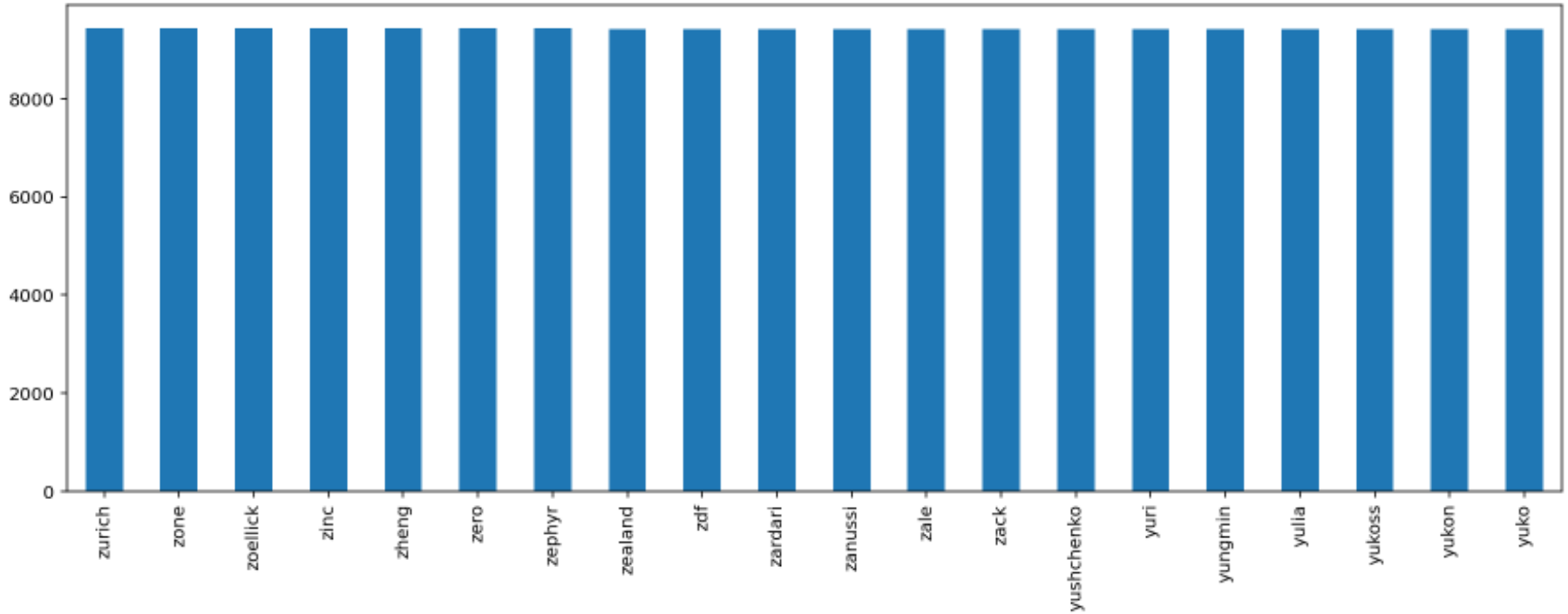


Fig 6:Business data words.

EDA

- Count of words in Entertainment data after applying stemming.

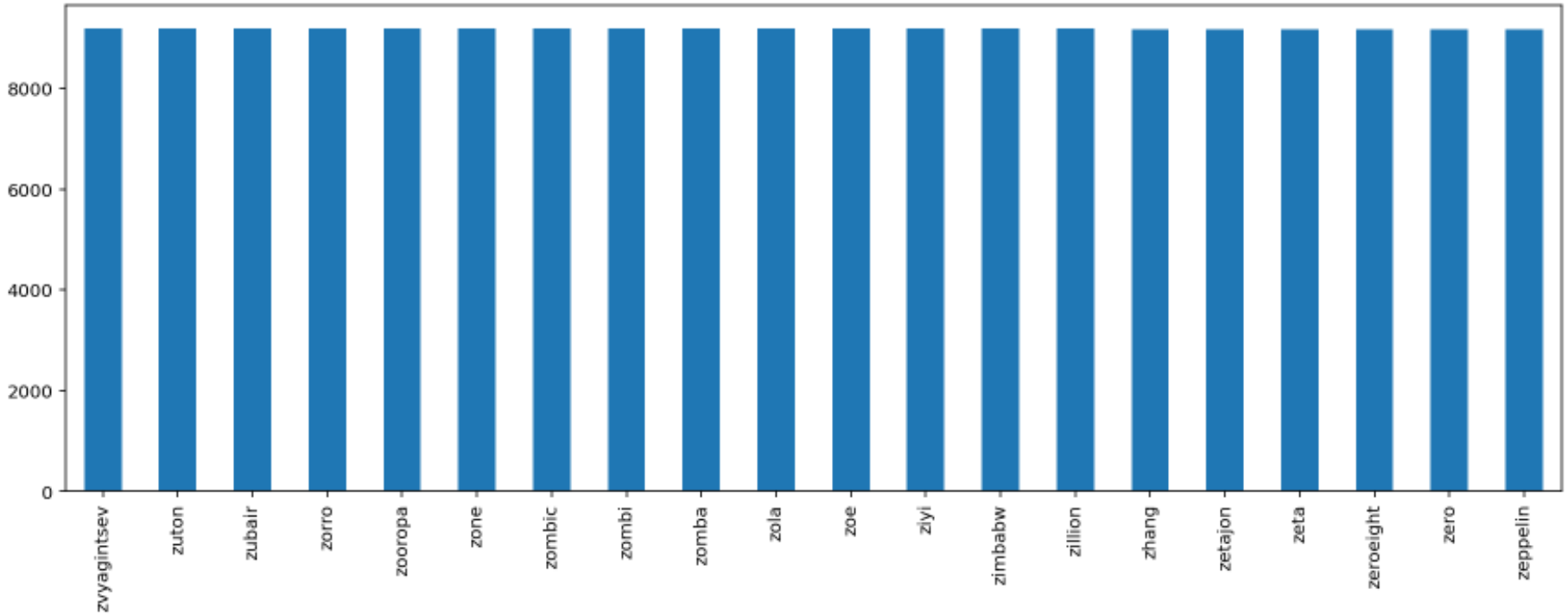


Fig 7: Entertainment data words.

EDA

- Count of words in Politics data after applying stemming.

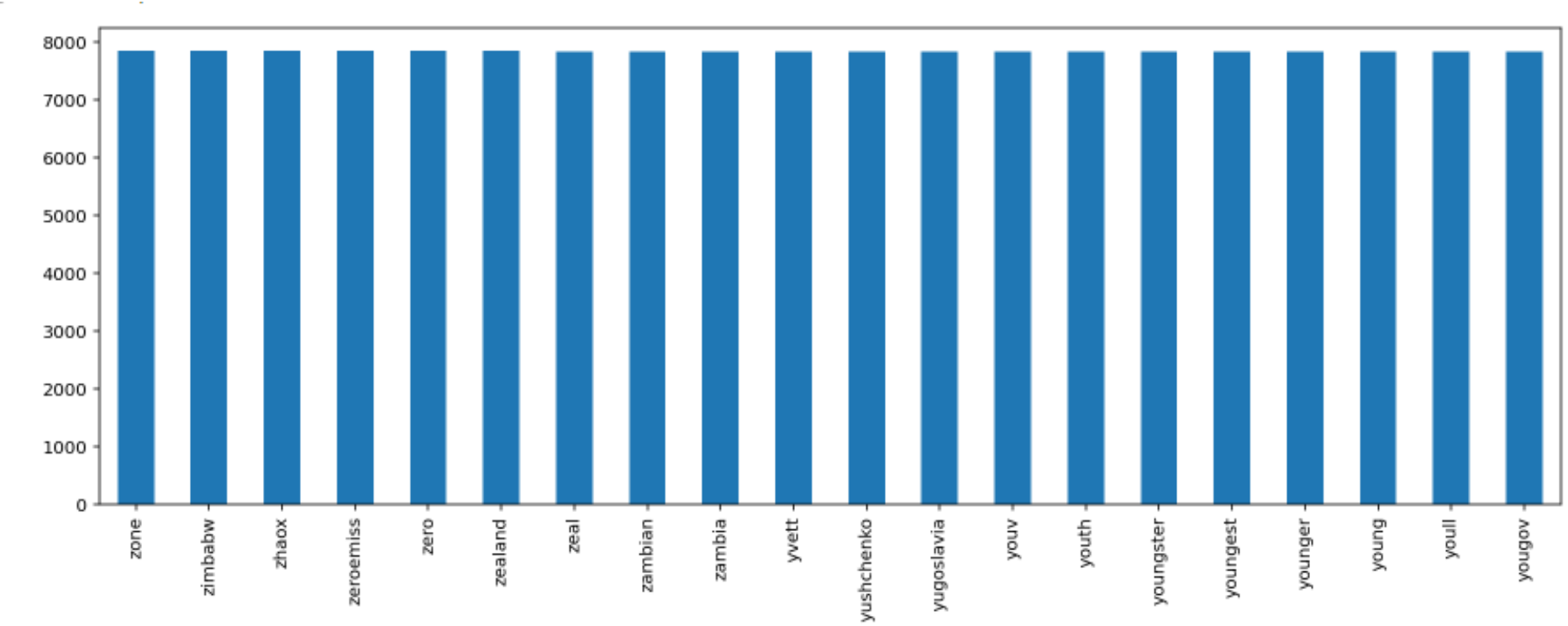


Fig 8: Politics data words.

EDA

- Count of words in Sports data after applying stemming.

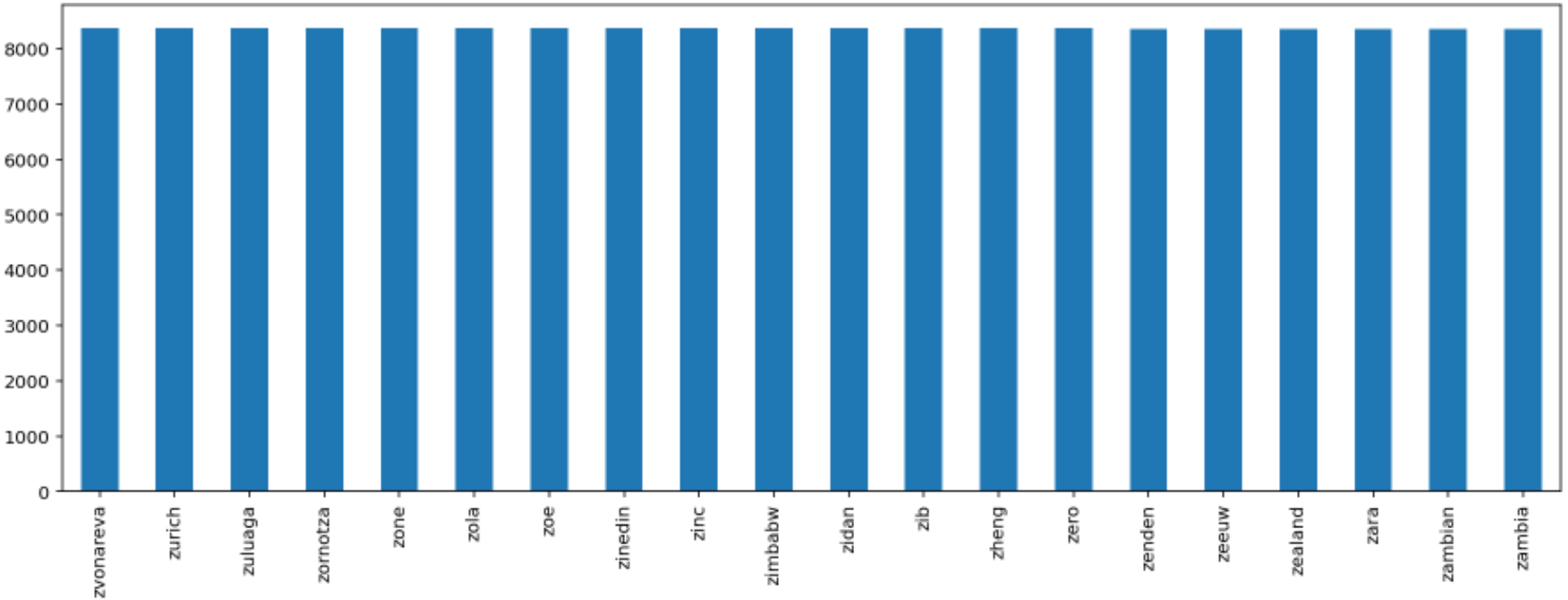


Fig 9: Sports data words.

EDA

- Count of words in Tech data after applying stemming.

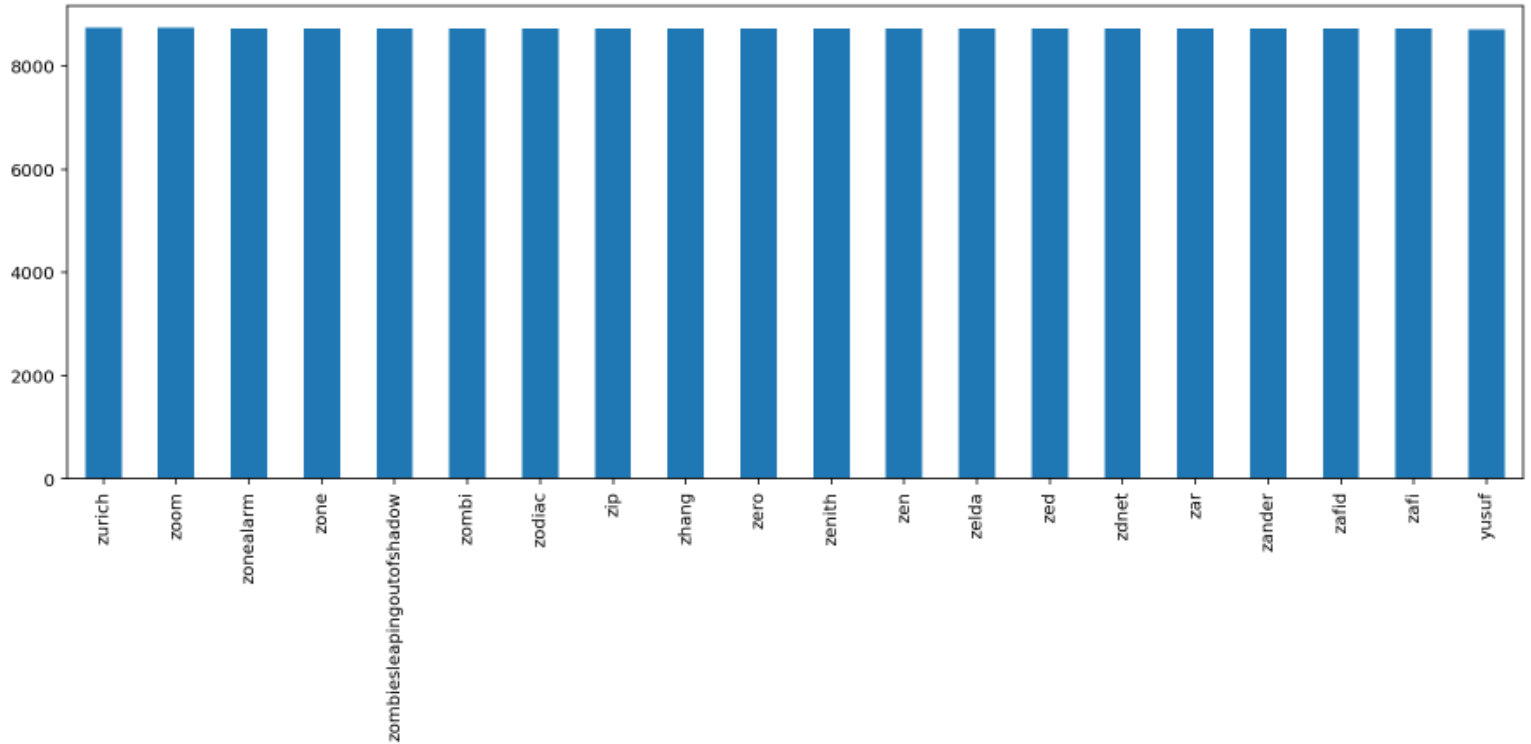


Fig 10: Tech data words.

EDA

- Maximum frequency words in Business data.

Fig 11: Words in Business data.

EDA

- Maximum frequency words in Entertainment data.

Fig 12: Words in Entertainment data.

EDA

- Maximum frequency words in Politics data.

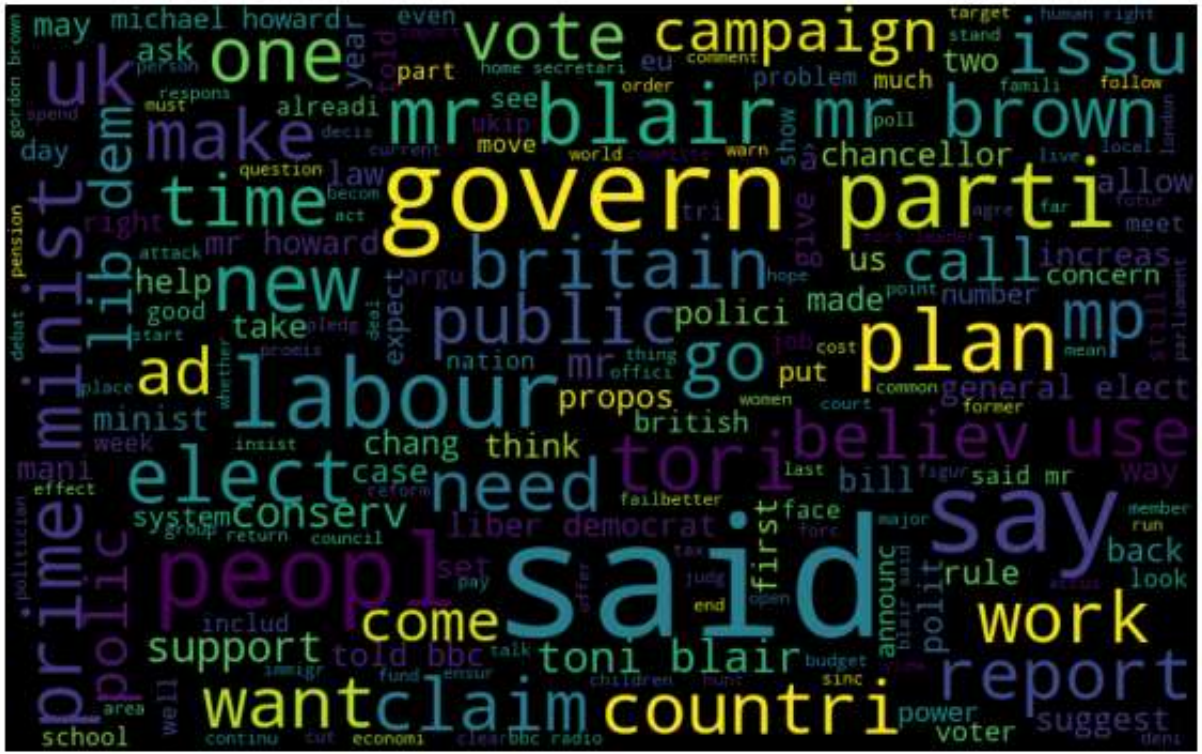


Fig 13: Words in Politics data.

EDA

- Maximum frequency words in Sport data.

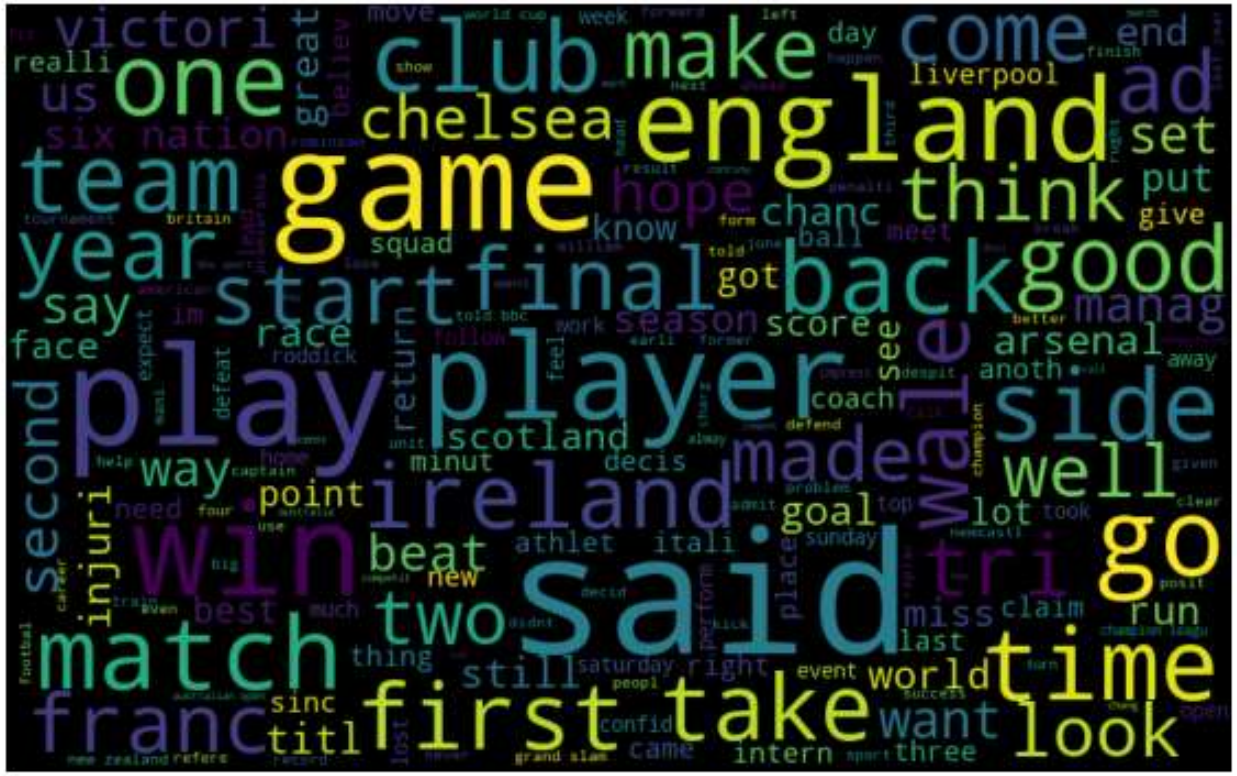


Fig 14: Words in Sport data.

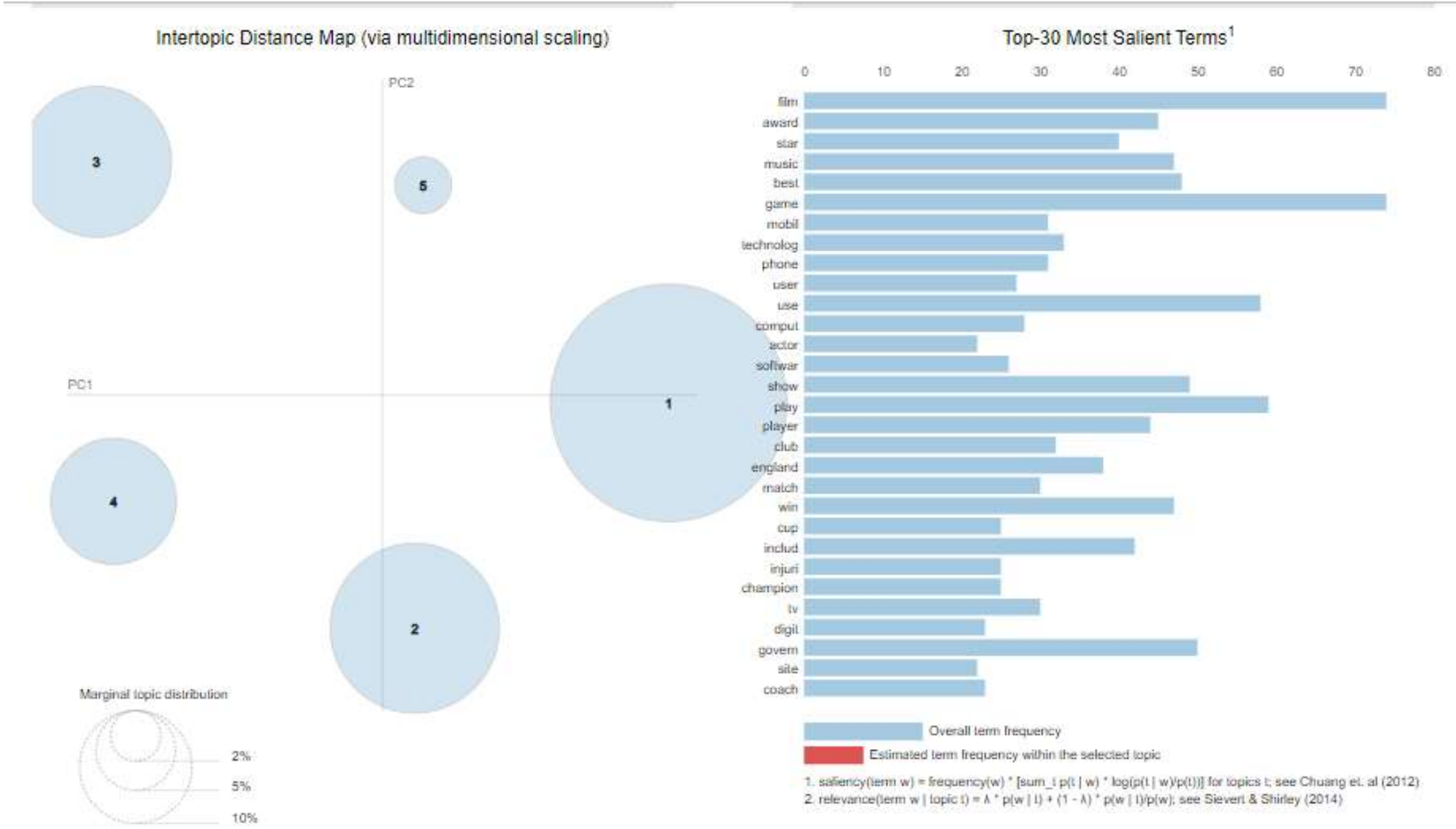
EDA

- Maximum frequency words in Tech data.

Fig 15: Words in Tech data.

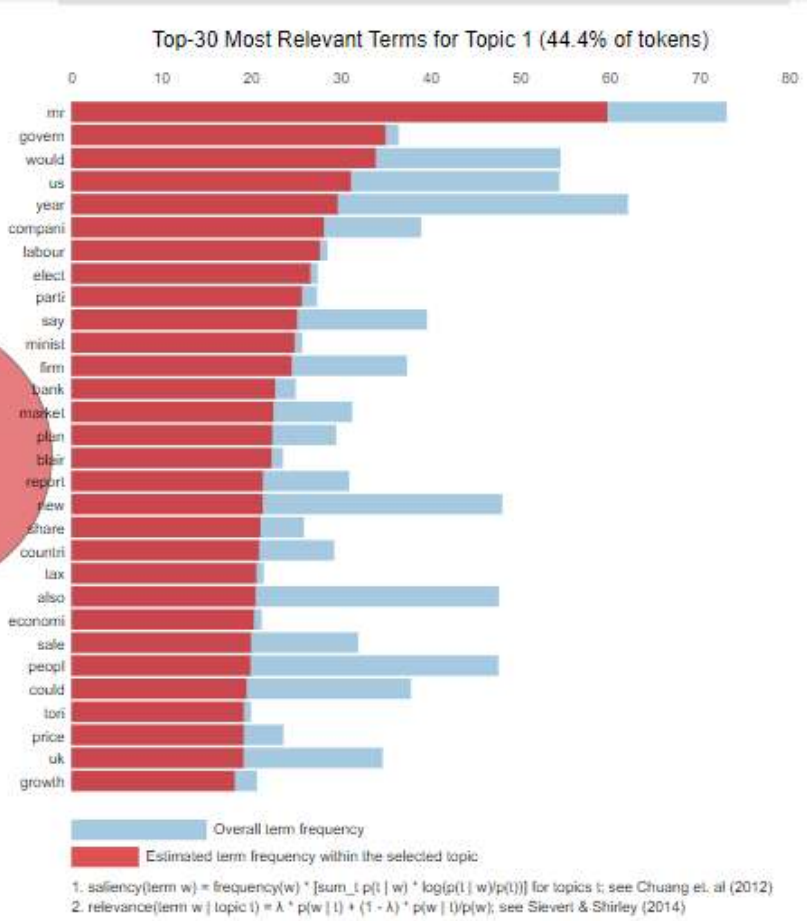
EDA

Lda clusters



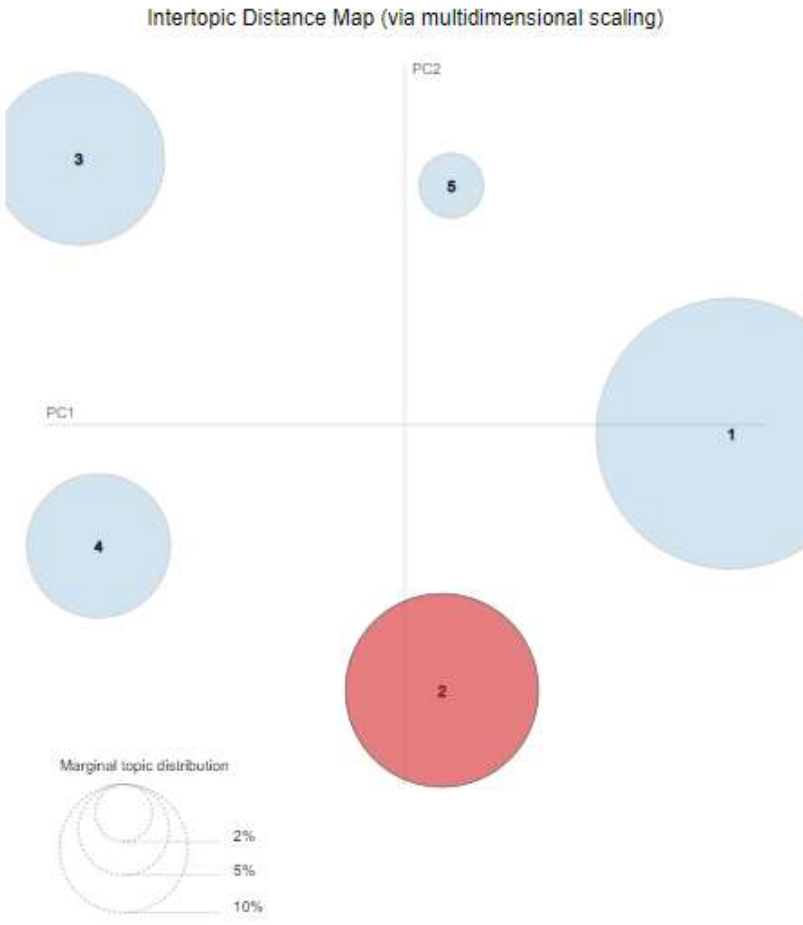
EDA

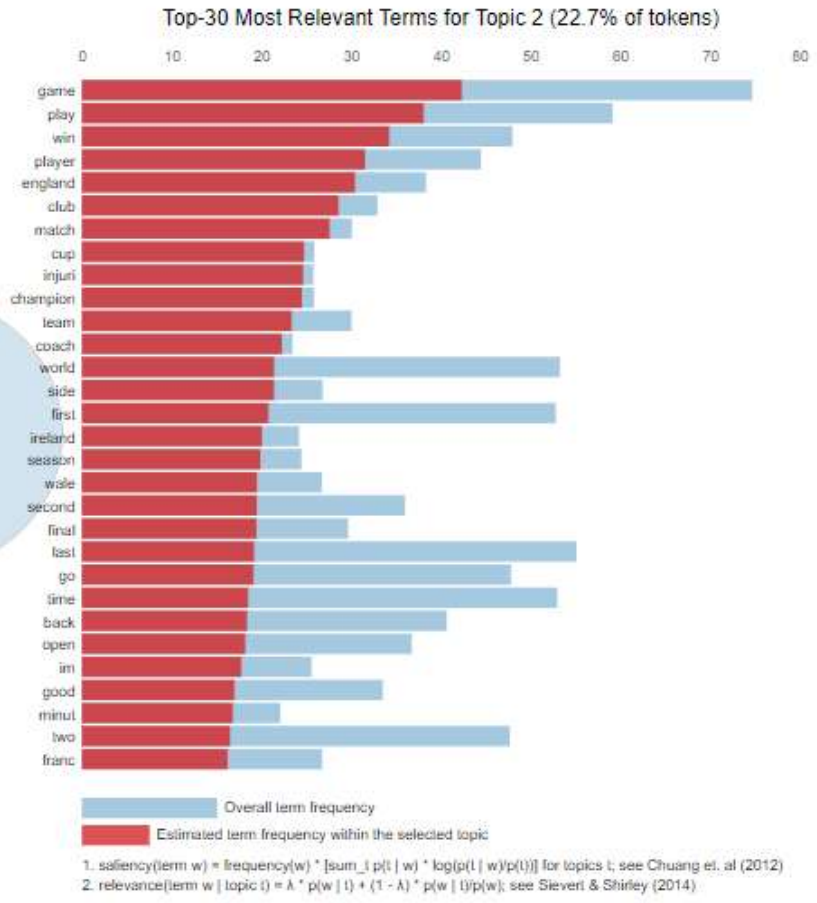
Lda cluster 1 - Politics



EDA

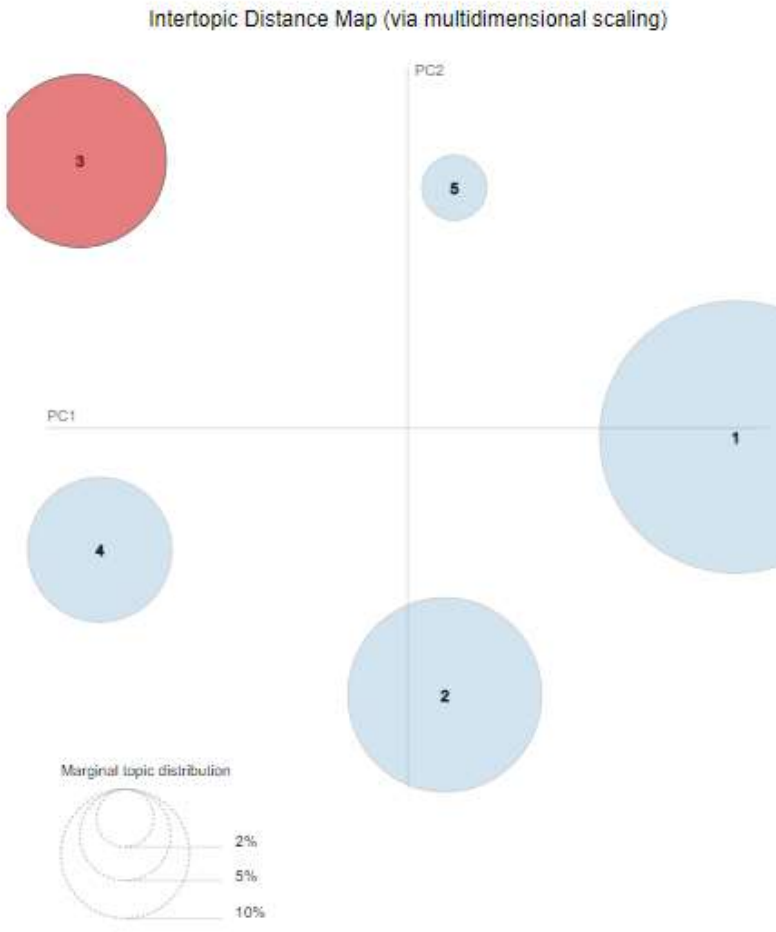
Lda cluster 2 - Sports

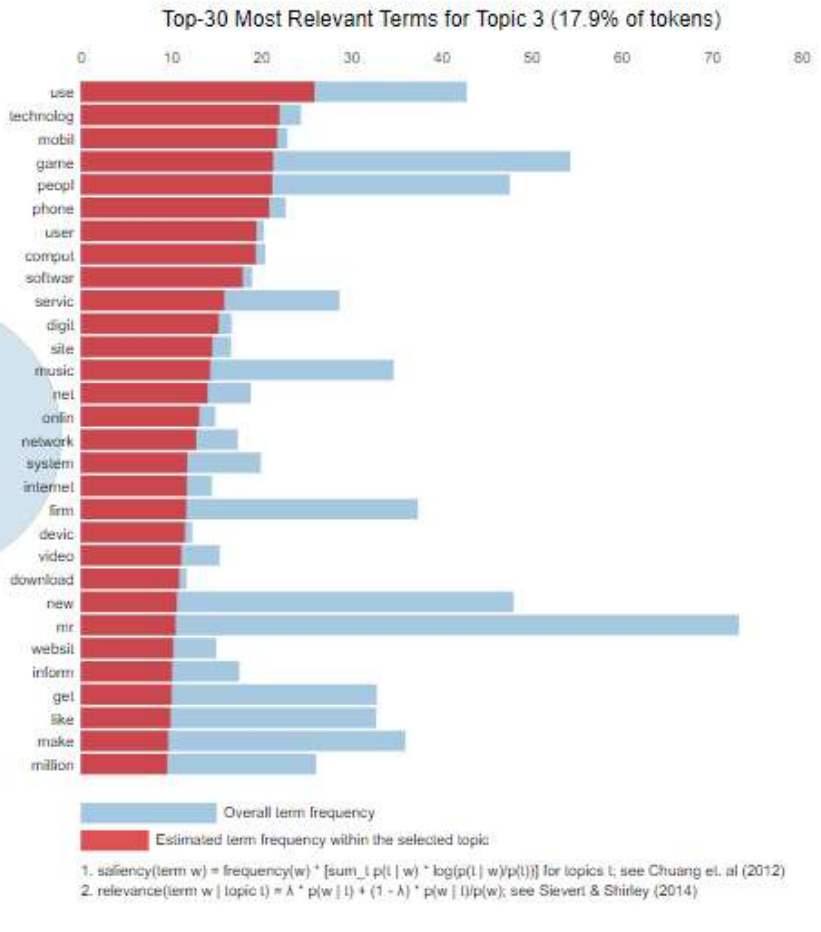




EDA

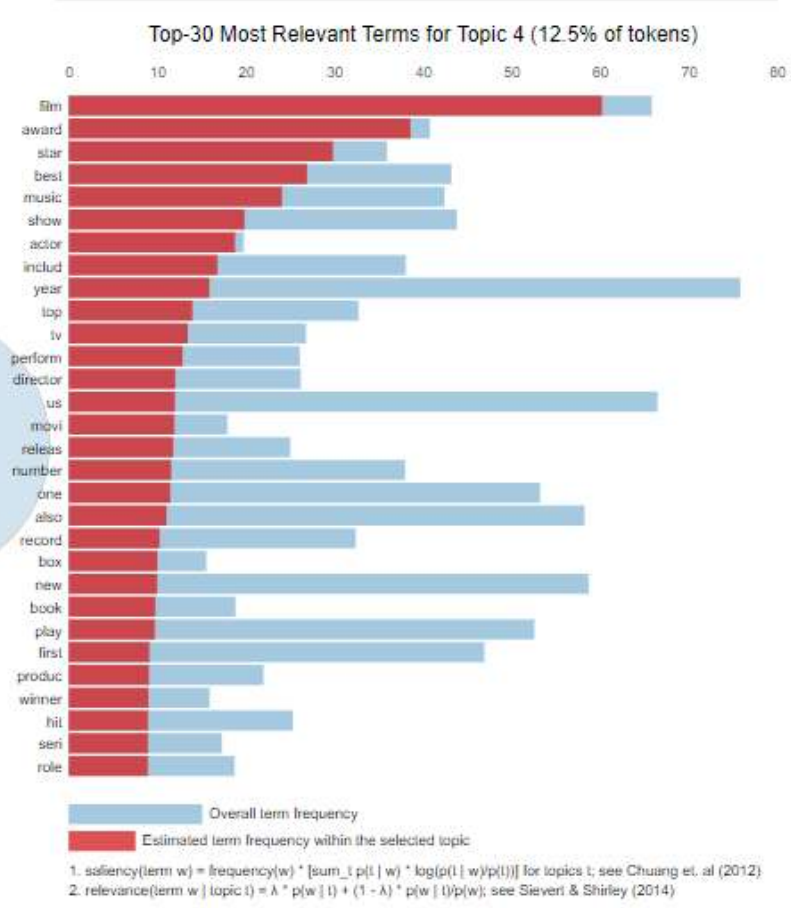
Lda cluster 3 - Tech





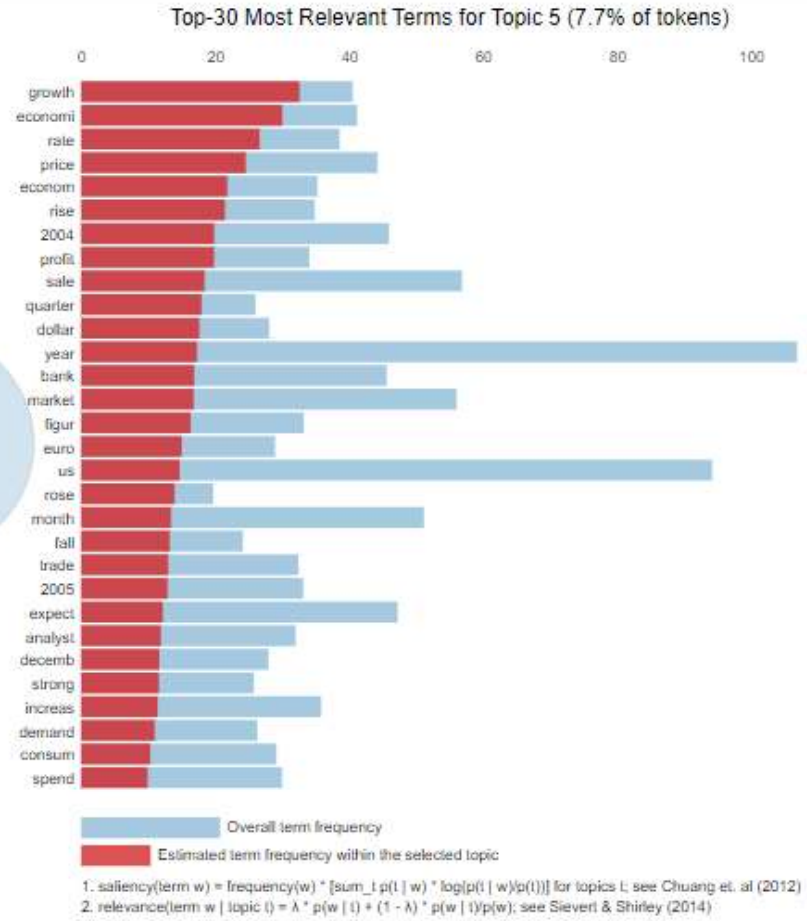
EDA

Lda cluster 4 – Entertainment



EDA

Lda cluster 5 – Business



Model implementation

- After implementing various models on the given data such as Logistic Regression, Decision Tree Classifier, Naïve Bays Classifier, KNN Classifier, Random Forest Classifier. Train and test accuracy are as follows.

Model	Train	Test
Logistic Regression	0.98	0.95
Decision Tree Classifier	1.00	0.86
Random Forest Classifier	1.00	0.91
Naïve Bays Classifier	0.95	0.92
KNN Classifier	0.96	0.95

Table 1 Result table of models.

Model implementation

- We get maximum accuracy with KNN classifier and Logistic regression with train and test accuracy as
- KNN Classifier Train accuracy: 0.96, Teat accuracy: 0.95
- Logistic Regression Train accuracy: 0.98, Test accuracy: 0.95
- As per the working of KNN classifier we would like implement it on News classification.

Model	Train	Test
KNN Classifier	0.96	0.95
Logistic Regression	0.98	0.95

Table 2 Result table of Complex models.

Conclusion

- The news is almost every second used in different sources of media in soft and hard.
- Due to impact of social media everyone is referring to online news platform. News classification became important aspect.
- We get maximum accuracy with KNN classifier and Logistic regression with train and test accuracy as 0.96,0.95 and 0.98,0.95 respectively.
- As per the working of KNN classifier we would like implement it on News classification.

References

[1] “Content Enrichment Using Linked Open Data for News Classification” by Hsin-Chang Yang, Yu-Chih Wang.

[2]” Classification of News Articles using Supervised Machine Learning Approach” by Muhammad Imran Asad, Muhammad Abubakar siddique, Safdar Hussain, Hafiz Naveed Hassan, and Jam Munawwar Gul.

[3]” Text classification of BBC news articles and text summarization using text rank” by Abhishek Dutt and Kirk Smalley.

Thank You...