Capstone Project

Topic Modeling on News Articles

By,
Yamini

Steps followed to solve the problem

- Data Extraction
- Data Cleaning and transformation
- Data Understanding
- Data Analysis
- Data Visualization
- Feature engineering and selection
- Model building and selection
- Model validation
- Some other findings
- Important words from each topic

Problem statement

- ✓ In this project your task is to identify major themes/topics across a collection of BBC news articles where we can use clustering algorithms such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) etc.
- ✓ Here the dataset contains a set of news articles for each major segment consisting of business, entertainment, politics, sports and technology. We have to create an aggregate dataset of all the news articles and perform topic modeling on this dataset.
- ✓ Verify whether these topics correspond to the different tags available.

Data Extraction

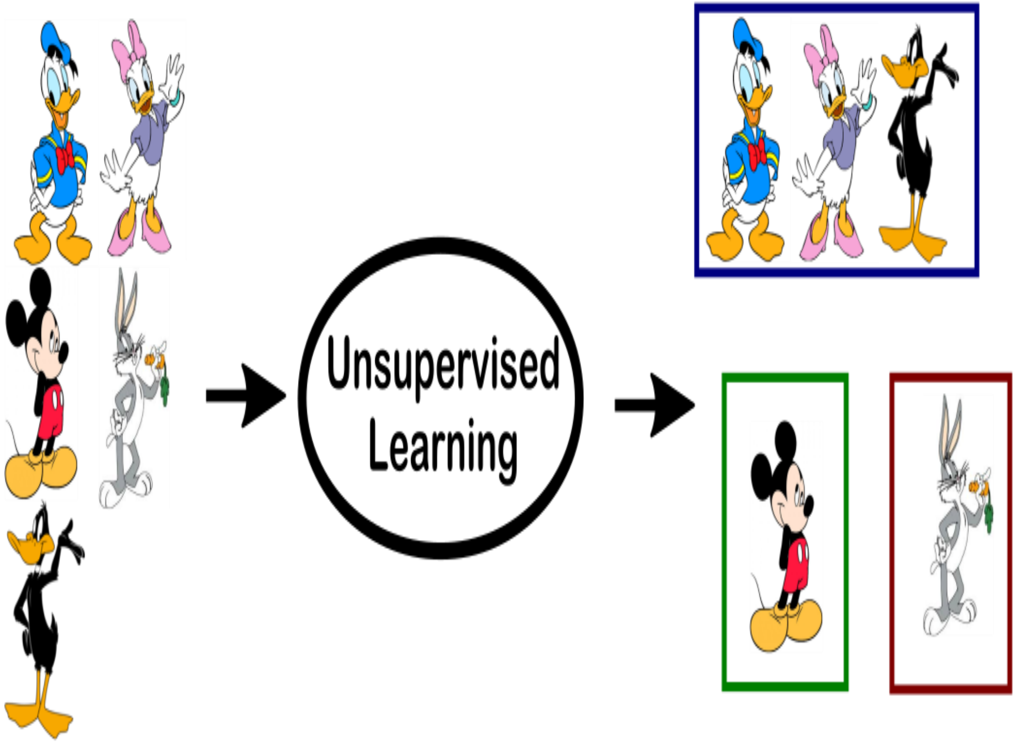
- There are many text files given for different topics of news articles which are to be merged and form a new dataset so that we can find the model which helps in finding the major topics/themes discussed in the new articles.
- I have used glob and pandas libraries to read the text files and to make a new dataframe containing each topic in separate dataframes then concatenated all the dataframes into one dataframe for analysis and further processing.

Data Cleaning and transformation

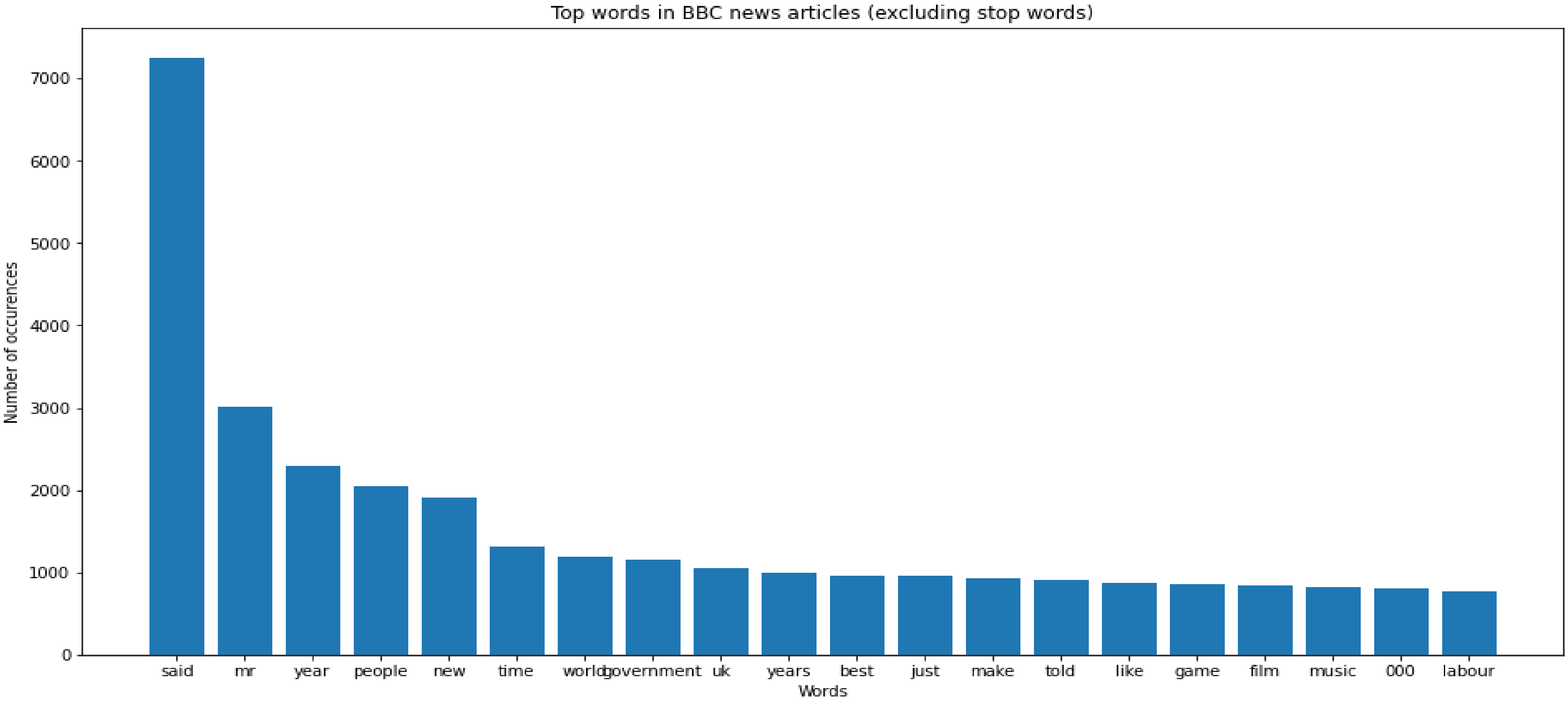
- From all the text data there are headings for each text file and have taken only news which is required for processing by removing the headings from all the text files.
- Renaming the feature names as BBC news articles on final dataframe and placing all the data at one place in one dataframe.
- Creating new Topic columns with their respective news articles topic so that it can be verified after topic modelling if they are tagged to the correct one or not.

Data Analysis in Unsupervised Learning

- There are a total of 10633 rows with all the kinds of topics in the given news articles and 2 columns.
- There are no missing values
- There are 2 categorical variables in the dataframe created which requires no transformation and further encoding.



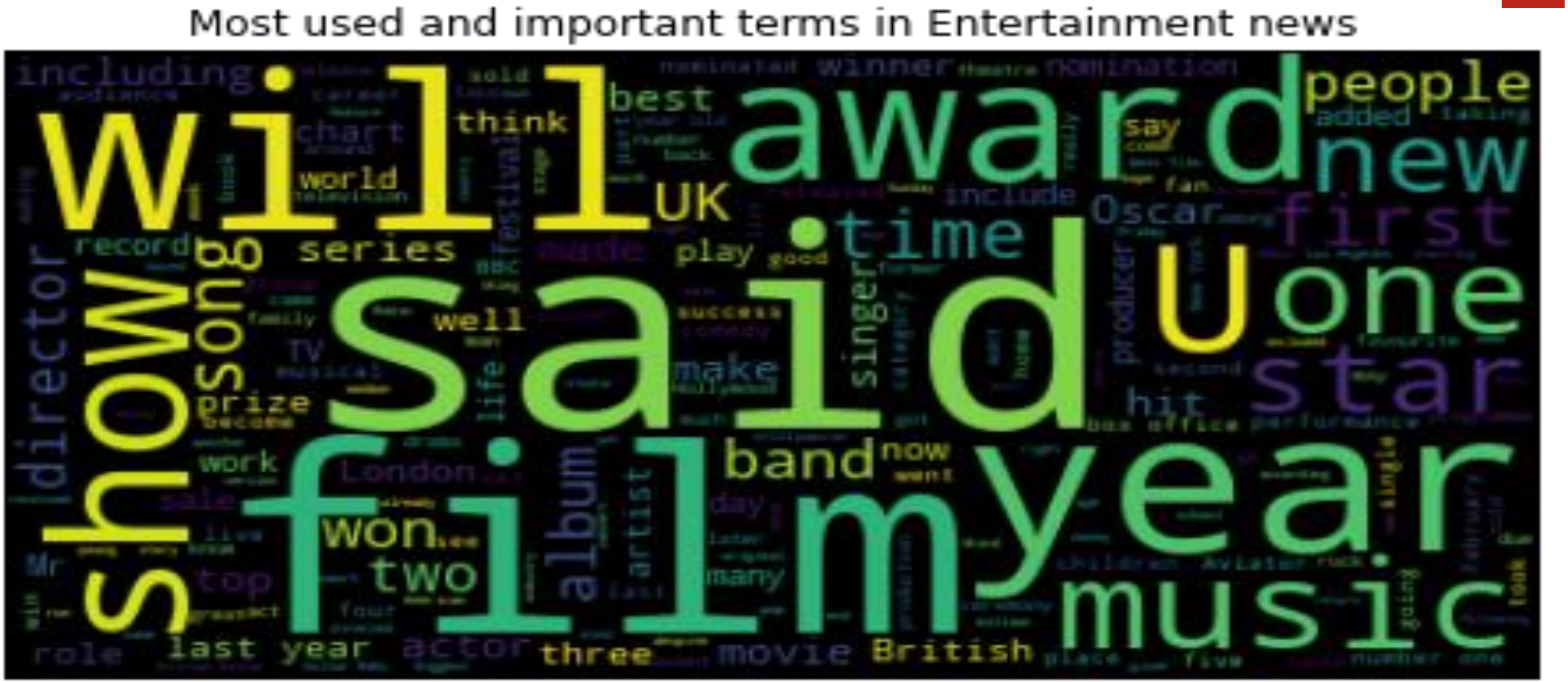
Data Visualization



Most used and important terms in Business news

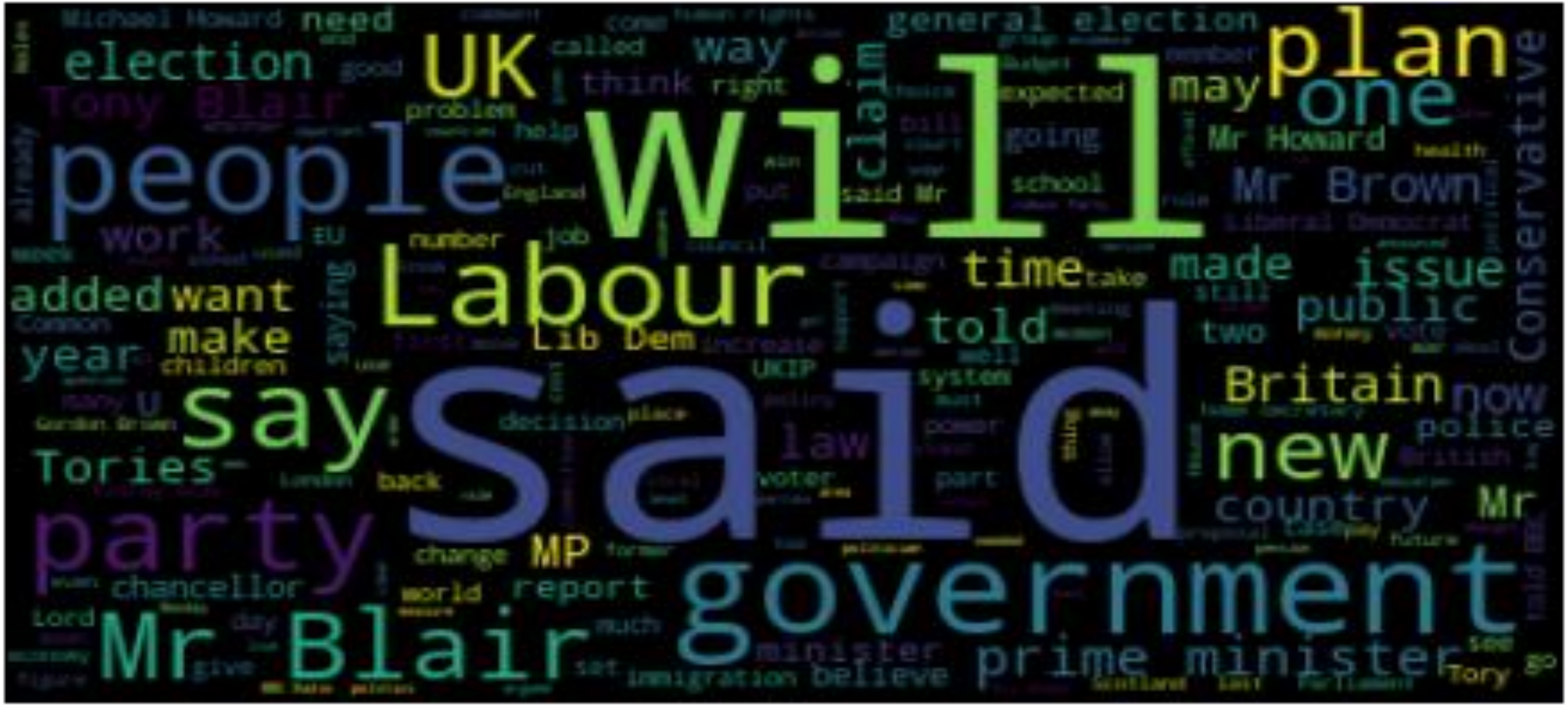
said
will
year
company
U
firm
new
market
government
economy
growth
sale
report
state
one
time
month
analyst
expected
price
share
Bank
China
rise
business
last year
economic
December
many
euro
world
may
profit
country
increase
investment
dollar
cost
two
trade
UK
group
plan
financial
million
still
chief executive
interest rate
make
deal
people
result
now
figure
companies
made
say
Mr
added
investor
cut
number
much
export
countries
India
part
pay
industry
higher
debt
demand
forecast
January
Japan
strong

Important terms from Business topic in BBC news articles



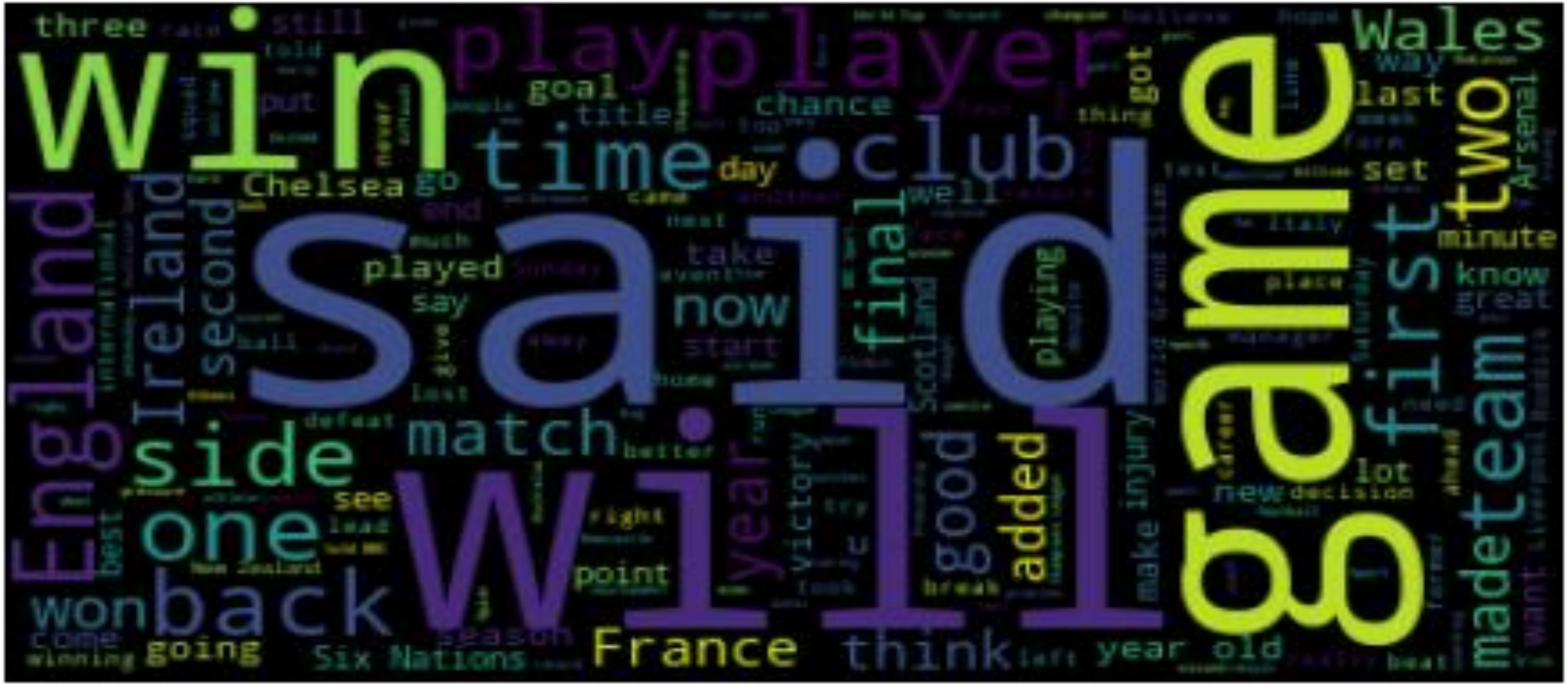
Important terms from Entertainment topic in BBC news articles

Most used and important terms in Politics news



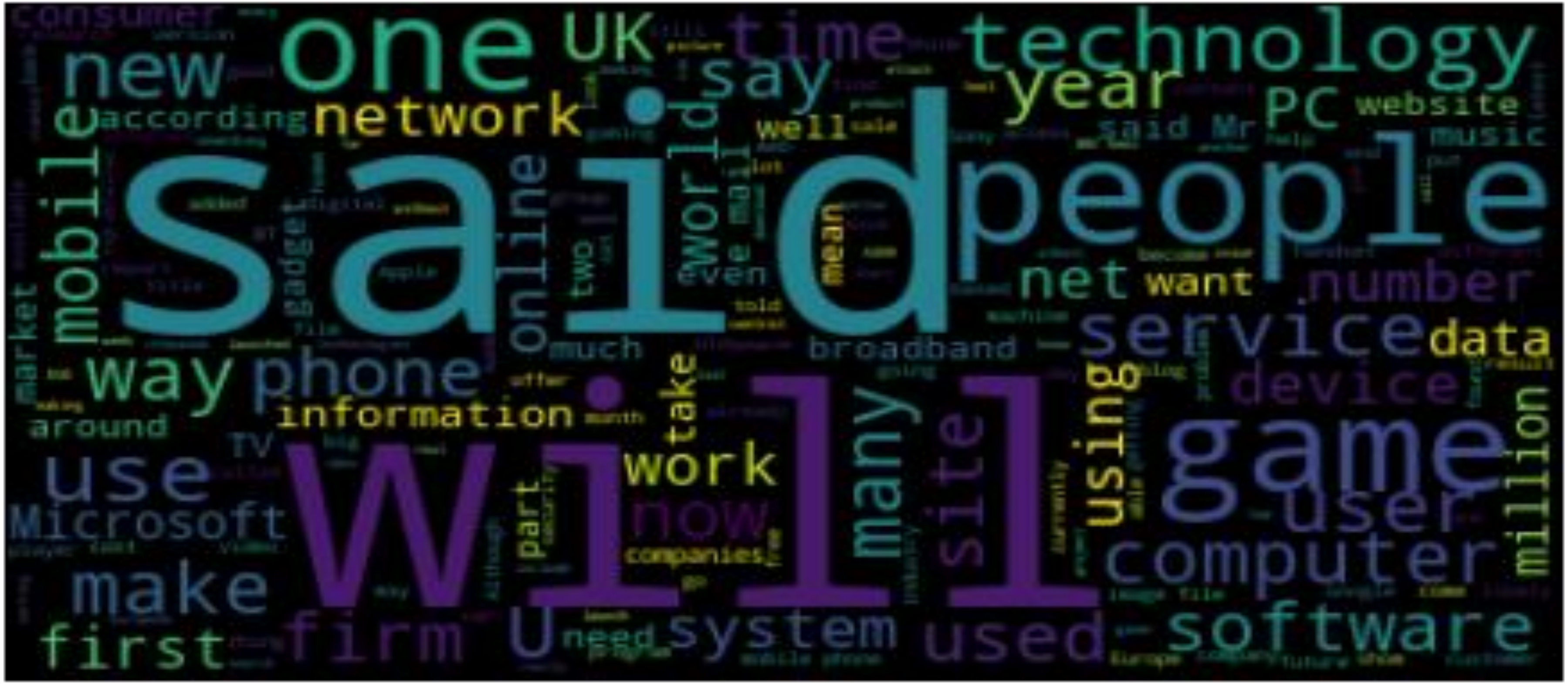
Important terms from Politics topic in BBC news articles

Most used and important terms in Sports news



Important terms from Sports topic in BBC news articles

Most used and important terms in Technology news



Important terms from Technology topic in BBC news articles

Feature Engineering and Selection

- Some of the feature engineering done to the news articles are used Count vectorizer for getting the important words from it so that we get to know to which topic it belong to.
- Also used some techniques like T-SNE, PCoA are used for multi-dimensional scaling in Topic modelling of news articles.
- T-SNE normally requires more computational time than PCoA.
- While for data exploration and data visualization t-sne is useful as well as PCoA is used at any case.
- The difference between PCA and PCoA (Principal coordinate Analysis) is **PCA** is based on Euclidean distances, **PCoA** can handle (dis)similarity matrices calculated from quantitative, semi-quantitative, qualitative, and mixed variables.

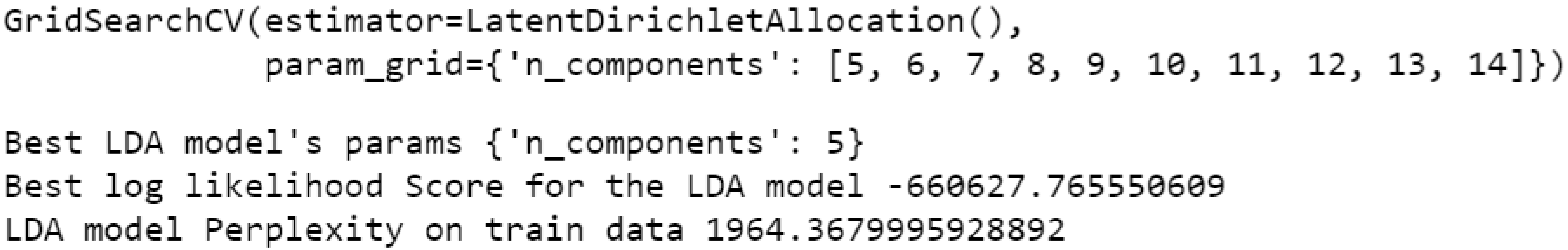
Models that can be used for Topic Modelling

- LDA
- LSA
- pLSA
- lda2vec (Deep learning)
- BERT (Transformers)

- The model I used here for Topic modelling on news articles is LDA(Latent Dirichlet Distribution)
- The **LDA model** does not identify the topic as a distinct word. Instead, it provides a probability of the most likely topic. Thus, users are required to determine a logical topic from the set of words provided by the model and map the topic numbers with the topic names identified by the user. This entire process is called ”Topic Modeling”

Results

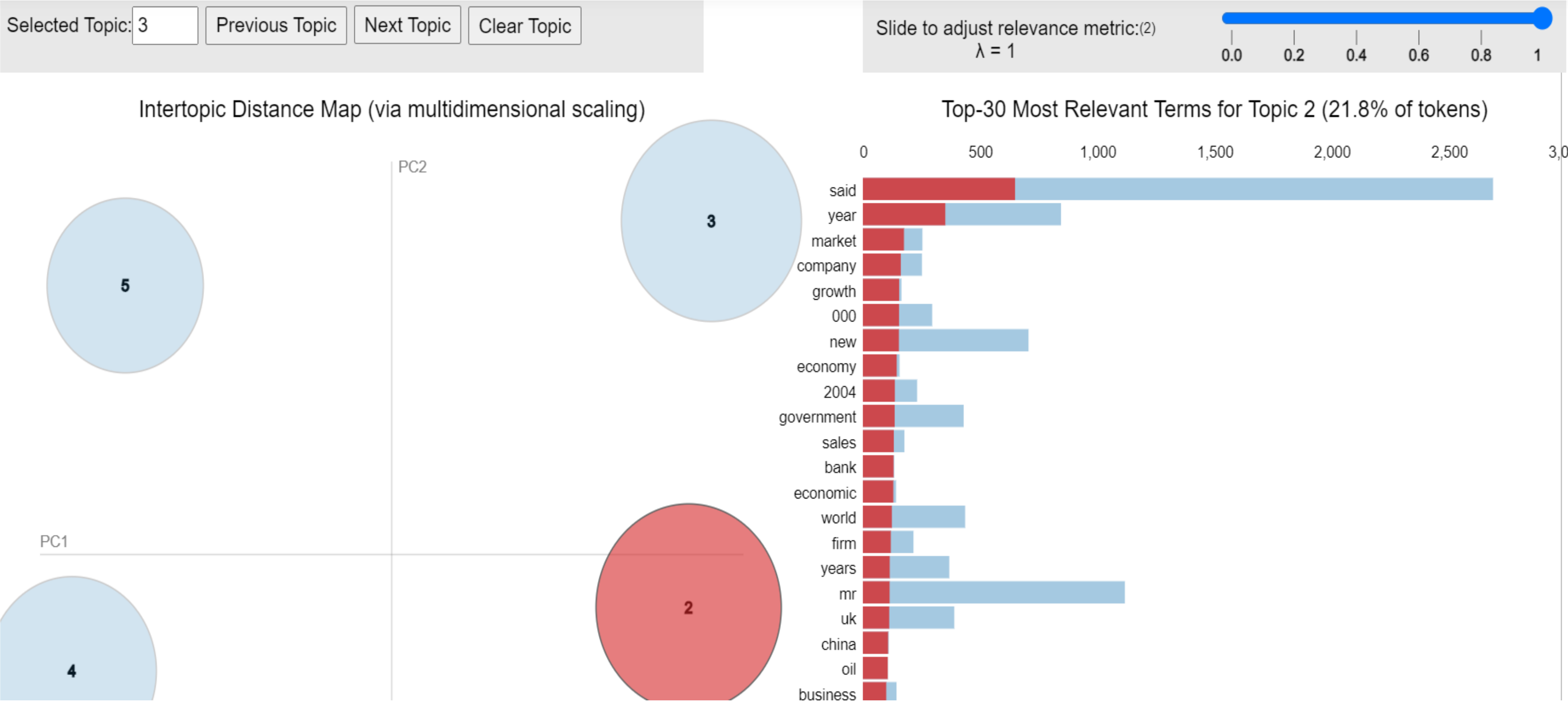
- Here I have used hyperparameter tuning using Grid Search CV for better results the it gives n_components as we require because there 5 different topics of news articles in the data.
- So the task here is to find the major topics/themes in the news articles and verify whether they are clustered in correct group or not.



```
GridSearchCV(estimator=LatentDirichletAllocation(),
             param_grid={'n_components': [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]})

Best LDA model's params {'n_components': 5}
Best log likelihood Score for the LDA model -660627.765550609
LDA model Perplexity on train data 1964.3679995928892
```

Topic modelling Visualization where different topics and probabilities of the terms can be known according to the parameters



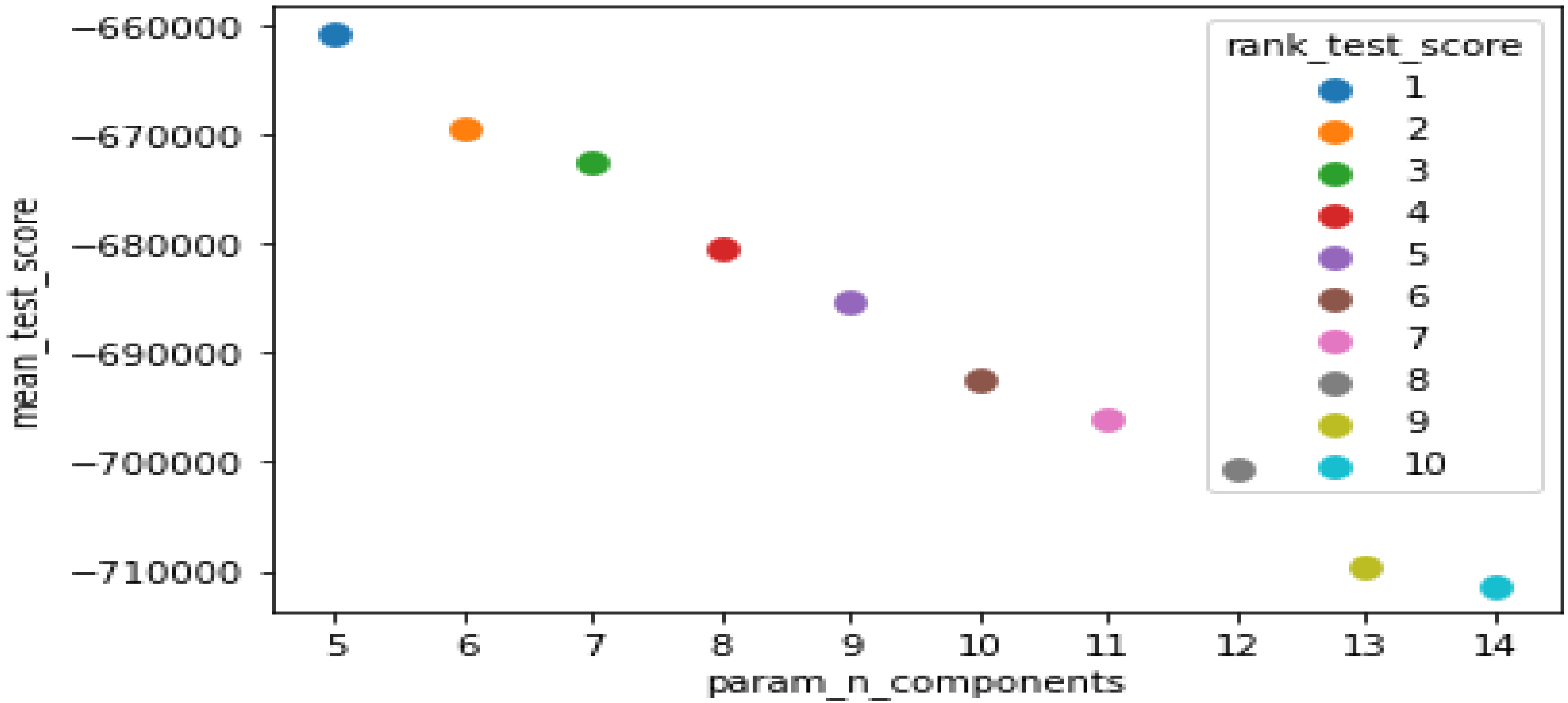
Selected Topic: 3

Slide to adjust relevance metric:(2)
λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 2 (21.8% of tokens)

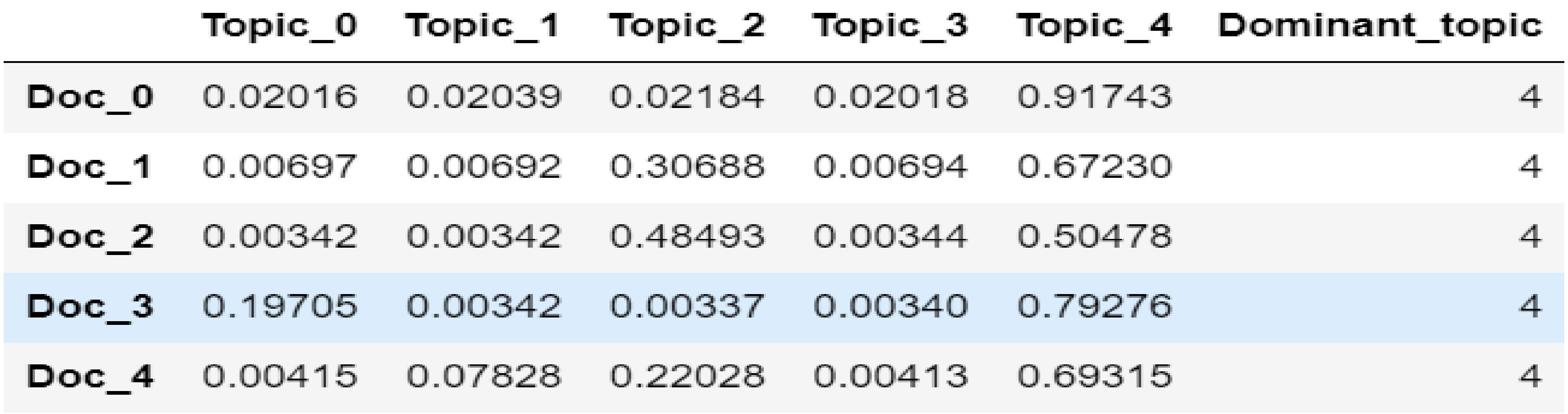
Note: Much more interactive visualization in the codebook



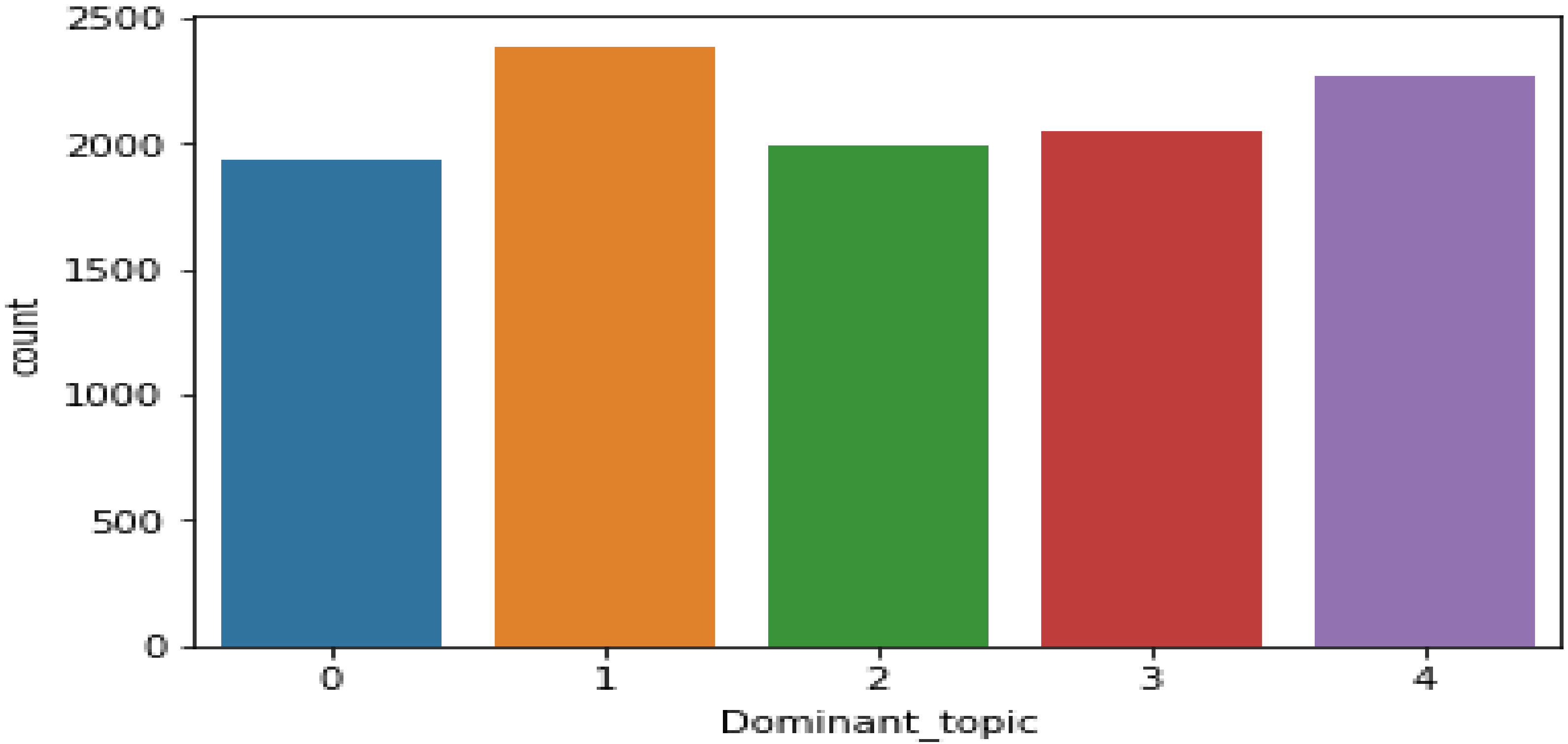
Best scores on choosing 5 components for the news articles

The topic modelling results on 5 documents

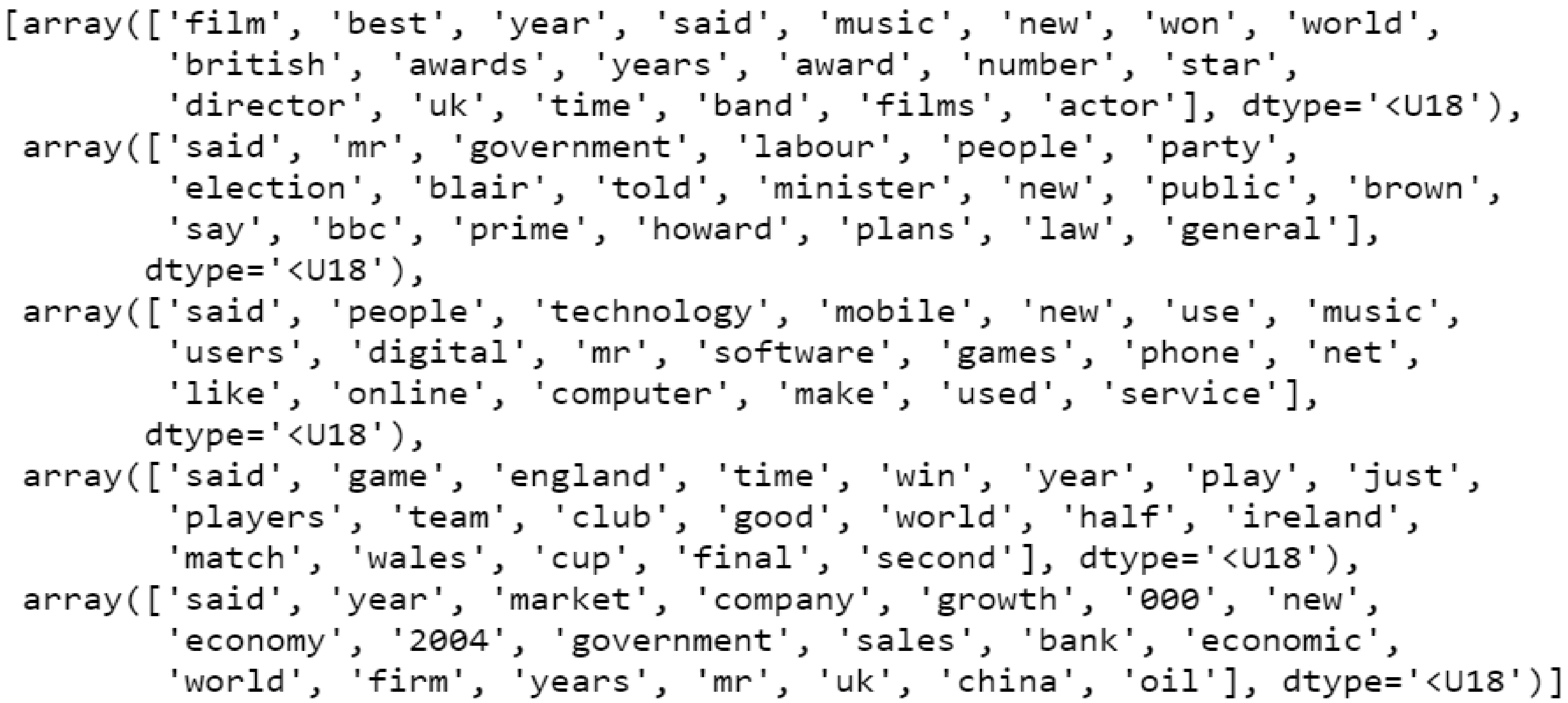
- It can be seen from the graph that from the 5 news articles it says that it belongs to topic 4 and there is some probabilities for other topics too.



	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Dominant_topic
Doc_0	0.02016	0.02039	0.02184	0.02018	0.91743	4
Doc_1	0.00697	0.00692	0.30688	0.00694	0.67230	4
Doc_2	0.00342	0.00342	0.48493	0.00344	0.50478	4
Doc_3	0.19705	0.00342	0.00337	0.00340	0.79276	4
Doc_4	0.00415	0.07828	0.22028	0.00413	0.69315	4



Topic 1 is dominant topic with more news articles on Business news



```
[array(['film', 'best', 'year', 'said', 'music', 'new', 'won', 'world',
        'british', 'awards', 'years', 'award', 'number', 'star',
        'director', 'uk', 'time', 'band', 'films', 'actor'], dtype='<U18'),
 array(['said', 'mr', 'government', 'labour', 'people', 'party',
        'election', 'blair', 'told', 'minister', 'new', 'public', 'brown',
        'say', 'bbc', 'prime', 'howard', 'plans', 'law', 'general'],
       dtype='<U18'),
 array(['said', 'people', 'technology', 'mobile', 'new', 'use', 'music',
        'users', 'digital', 'mr', 'software', 'games', 'phone', 'net',
        'like', 'online', 'computer', 'make', 'used', 'service'],
       dtype='<U18'),
 array(['said', 'game', 'england', 'time', 'win', 'year', 'play', 'just',
        'players', 'team', 'club', 'good', 'world', 'half', 'ireland',
        'match', 'wales', 'cup', 'final', 'second'], dtype='<U18'),
 array(['said', 'year', 'market', 'company', 'growth', '000', 'new',
        'economy', '2004', 'government', 'sales', 'bank', 'economic',
        'world', 'firm', 'years', 'mr', 'uk', 'china', 'oil'], dtype='<U18')]
```

- Top 20 keywords which represent each topic.
- The topics can be easily found according to the terms it has divided into five topics.

THANK YOU 😊