SUMMARY

ON

TOPIC MODELING ON NEWS ARTICLES

ALMABETTER CAPSTONE PROJECT

-By Yamini Peddireddi

1: Introduction

1.1. Unsupervised Learning

Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.

Below are some main reasons which describe the importance of Unsupervised Learning:

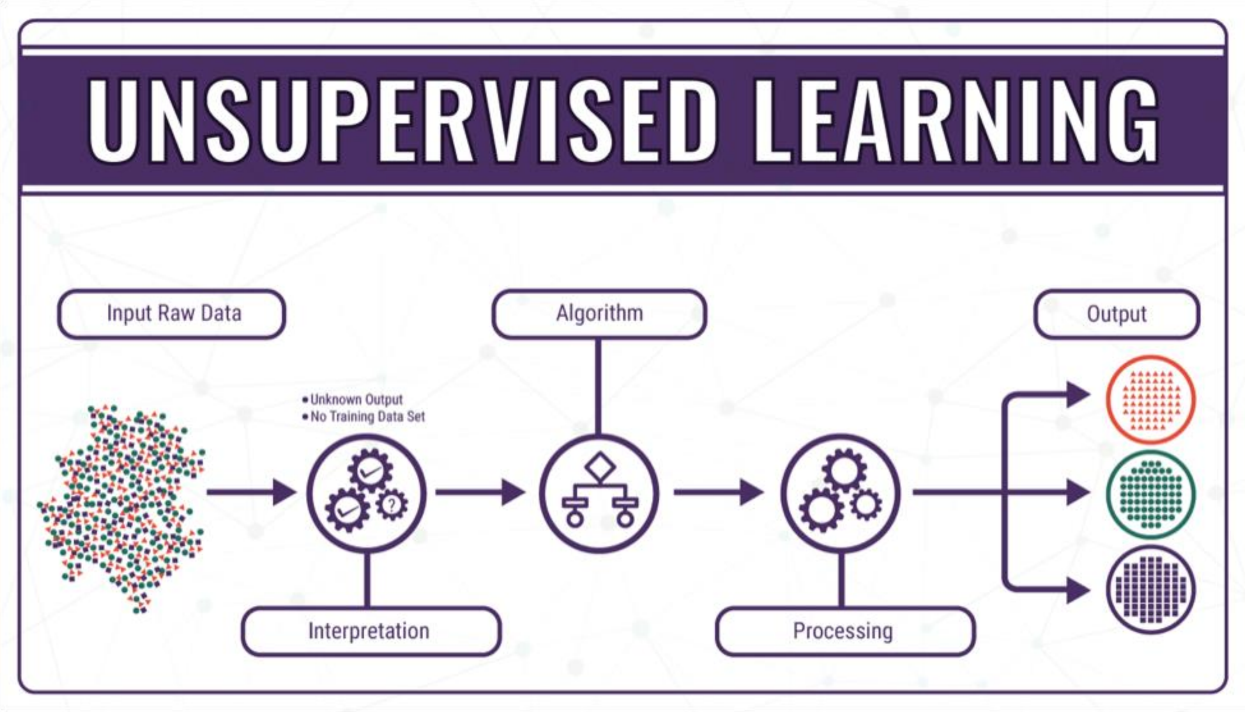
- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabelled and uncategorized data which make unsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

1.2. Popular Unsupervised algorithms

Below is the list of some popular unsupervised learning algorithms:

- K-means clustering
- KNN (k-nearest neighbors)
- Hierarchal clustering
- Anomaly detection
- Neural Networks
- Principle Component Analysis
- Independent Component Analysis
- Apriori algorithm
- Singular value decomposition

1.3. Problem solving



2: Methodology/Approach in solving

2.1. Problem Statement

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

2.2. Steps to problem solving

According to the problem statement, it is clear that the problem here is to help in finding out the topics/themes present in the documents or corpus.

The steps in create such a model or the steps in solving this problem are the following:

- Data Extraction
- Data Cleaning and transformation
- Data Understanding
- Data Analysis
- Data Visualization
- Feature engineering and selection

- Model building
- Model validation
- Some other findings
- Important words from each topic
- Some predictions or topics created using the model

2.3. Challenges

The challenges in creating the model are

- Data collection
- Data limitation
- Privacy and security
- Vague data/ incomplete data
- Chance of misinterpretations when proper model not used
- Feature selection and the good number of features unknown as there are many words in the documents.
- Computation time and cost when very huge dataset is involved

3: Data understanding and analysis

3.1. Data Extraction

There are approximately 2500 text files given for extracting the data from it from the different topic of news articles and find out whether after topic modeling they are correctly tagged to the respective topic news article or not.

Here I have created the dataframe with ~500+ news articles for each topic as given in the respective topic folder and their news articles. Then I have combined all these dataframes of different topic news articles into one dataframe and also created another feature column topic to check whether the news is tagged to the correct news topic or not.

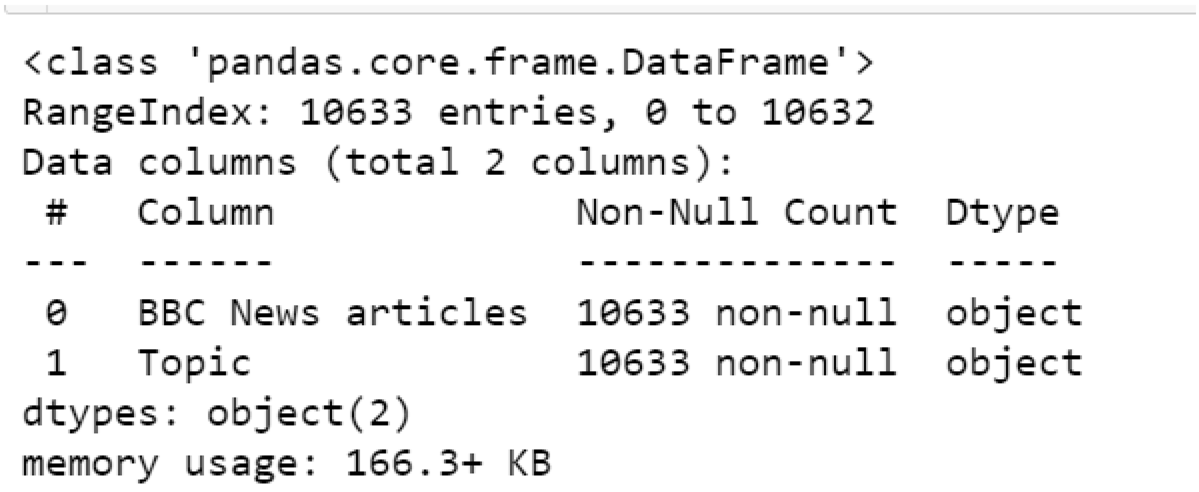
3.2. Data cleaning and transformation

After data extraction from the all the text files, there are 10633 rows and 2 columns. One column consisting of BBC news articles and another the topic of the news articles.

3.3. Data understanding

- The data is further explored to check if there are any missing values as those rows may not be useful.
- There are no missing values in the dataset and also the news articles are clear on what was published.

3.4. Data Analysis



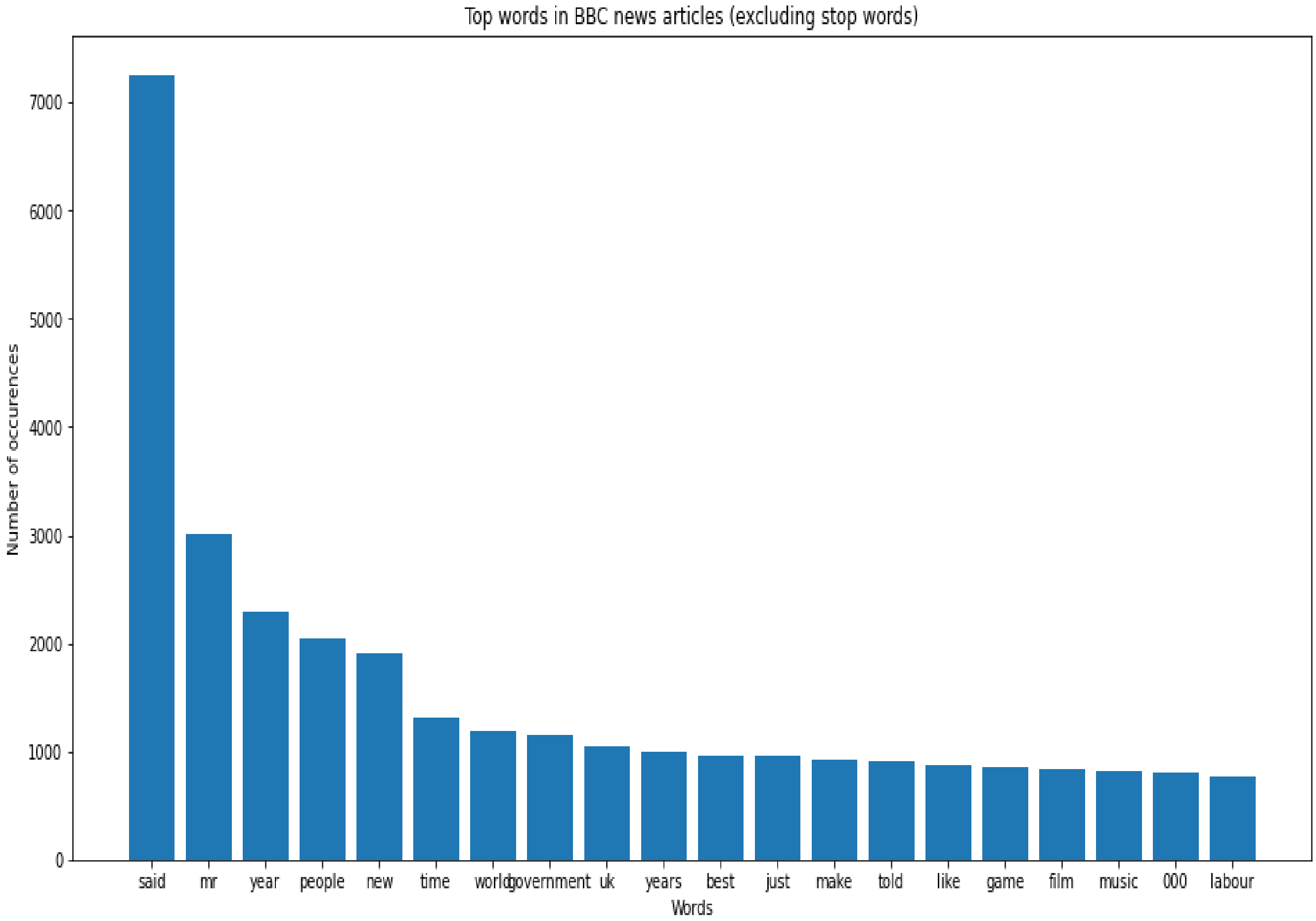
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10633 entries, 0 to 10632
Data columns (total 2 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   BBC News articles   10633 non-null  object
 1   Topic               10633 non-null  object
dtypes: object(2)
memory usage: 166.3+ KB
```

3.5. Feature engineering and selection

- After data cleaning I have created new features from cleaned tweet column using Tfidf vectorizer which extracts important words from the news articles based on the parameter given to it like min_df and max_df which helps in finding the words which are few in number in all the documents and most repeated words in all the documents.
- Now all the features can be used for further processing and analysis.
- After using all the types of feature engineering techniques, I have created good number of features which can be used for analysis and model building.

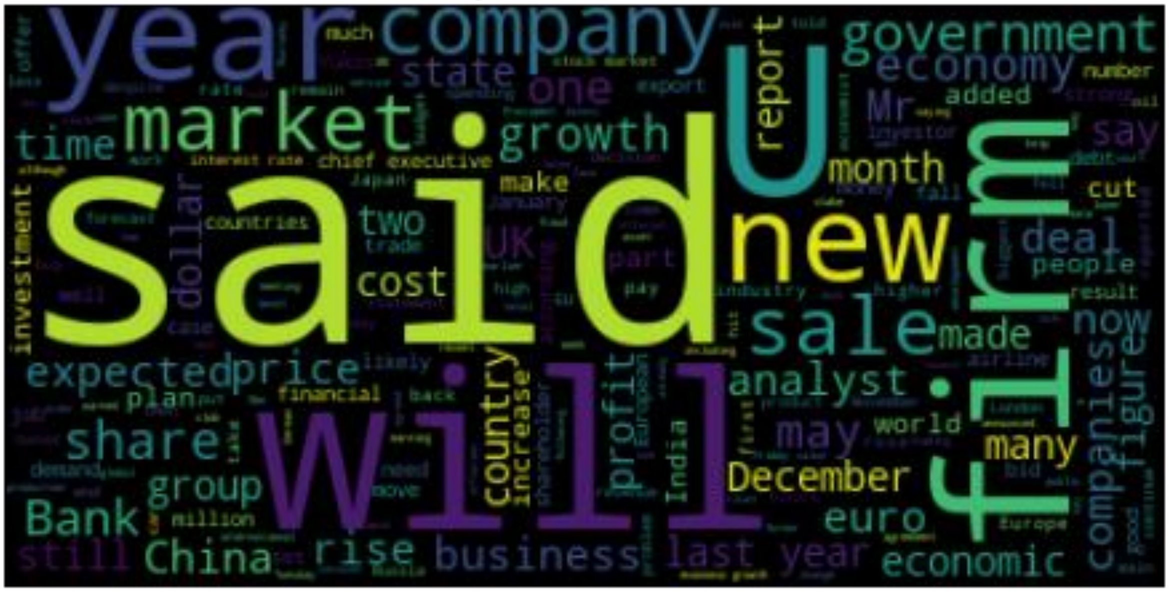
4: Data Visualization

4.1. Plots and insights



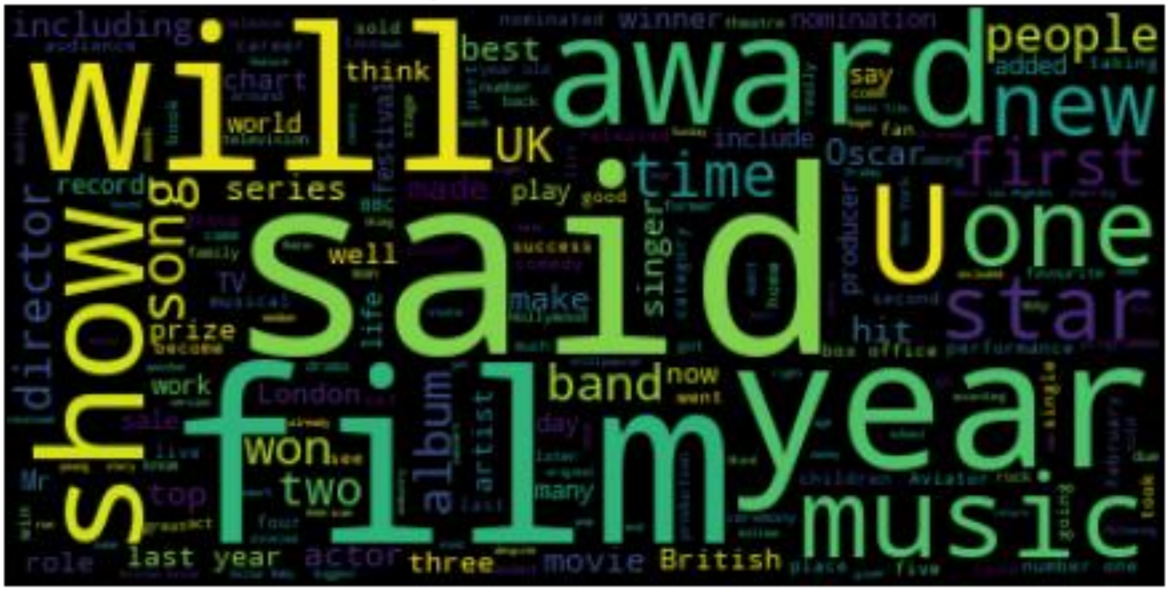
- ➢ From the above plot some of the top words which are mostly used in the documents are known using the Count vectorizer after removing the stop words from them.
- ➢ These are top 20 words in the corpus
- ➢ There are different kind of words in this list may be everyone can find some of their most used or most common things which are done or some of the favourite things can also be found here like music, game, best, film, new, time...etc.
- ➢ **Aren't they the top words that can be under discussion?**

Most used and important terms in Business news



- From the above plot we can find the most used words from the Business news articles.
- The Business news mostly consists of the words like firm, company, will, year, said, sale, Bank, share, expected price, cost, market, government, economy, report, euro, country, China, rise, last year, business, growth, figures, deal…etc

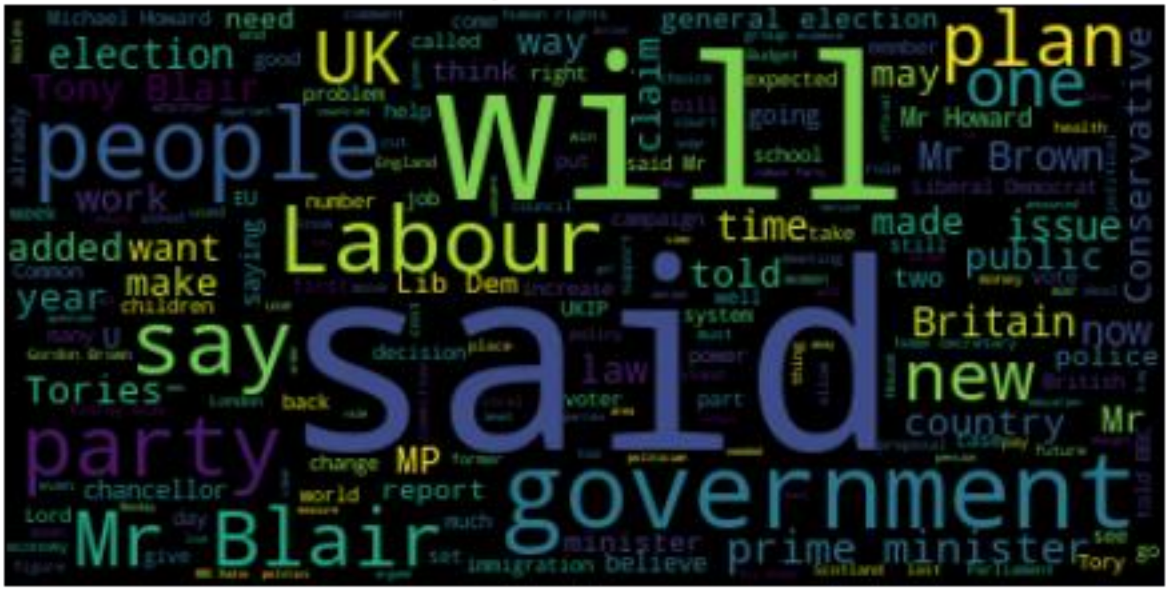
Most used and important terms in Entertainment news



- From the above plot the most used words from Entertainment news can be found

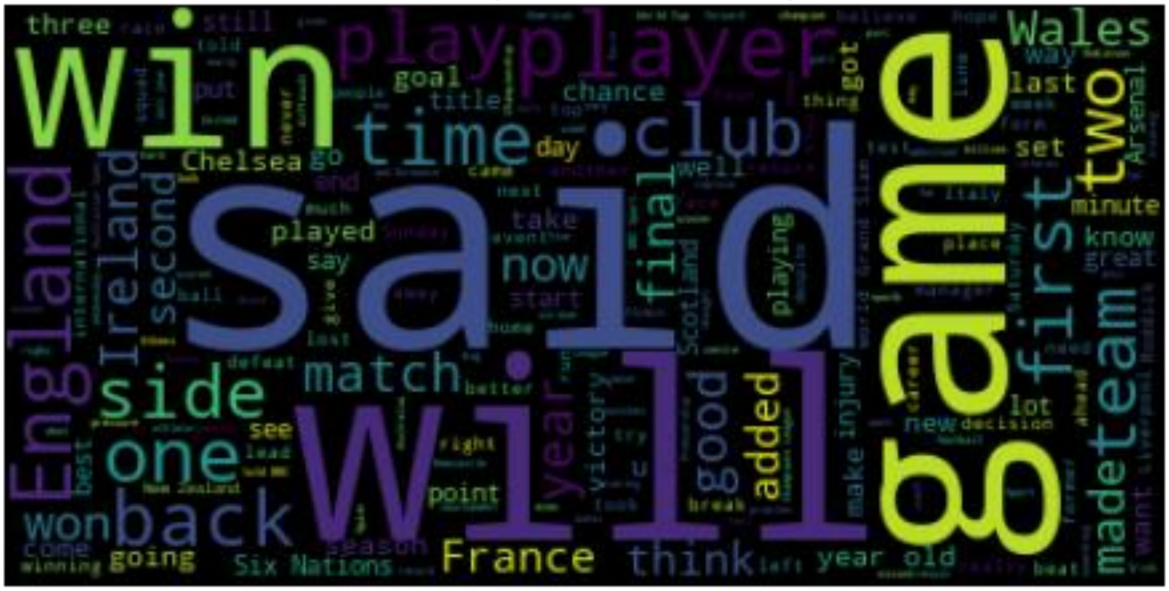
- The most repetitive terms from Entertainment news are award, show, film, year, music, won, song, two, people, new, movie, album, singer, British, last year, work, prize, director, star, hit, Oscar…etc.

Most used and important terms in Politics news



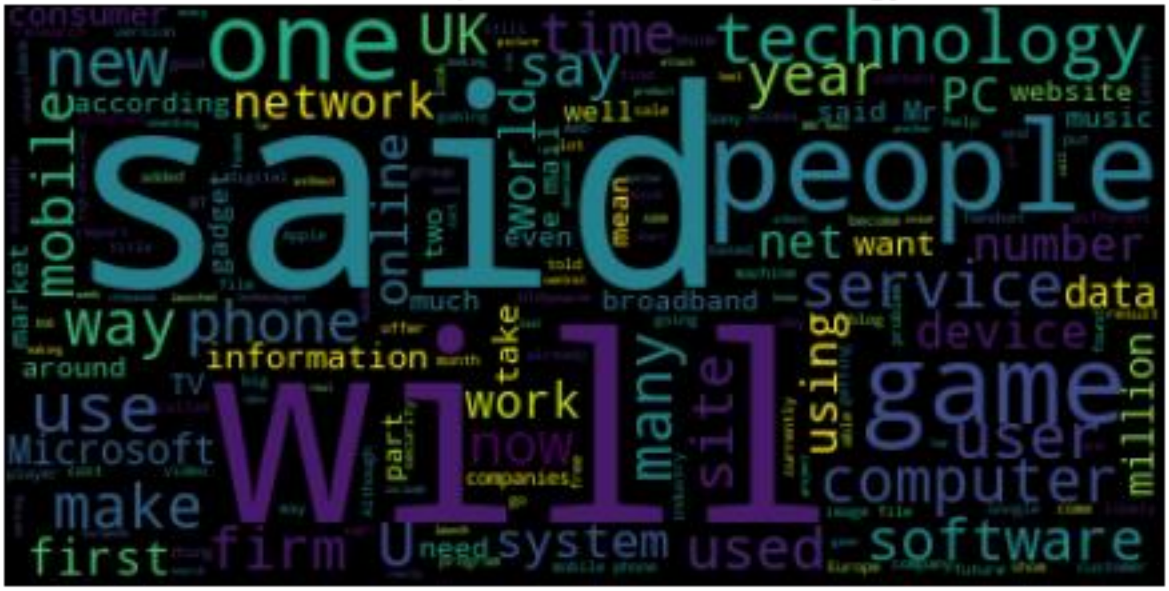
- The most commonly found words from Political news are said, government, Britain, Labour, chancellor, Mr Blair, plan, UK, election, told, issue, public, Mr Brown, may, added, want, make, country, law, report, world, law, policies, conservation, claim…etc.

Most used and important terms in Sports news



- ➢ The terminology mostly found in Sports news are win, game, good, added, team, side, won, back, Six Nations, Wales, play, player, goal, chance, club, minute, made…etc.

Most used and important terms in Technology news



- ➢ The terms which mostly seen in the Technology news are will, technology, said, phone, game, system, firm, U, need, software, mobile, work, take, phone, way, gadget, use, technology, PC, website, time, UK, Microsoft, data, using, net, want, number, million, user, online, network, year, according, used, broadband …etc.

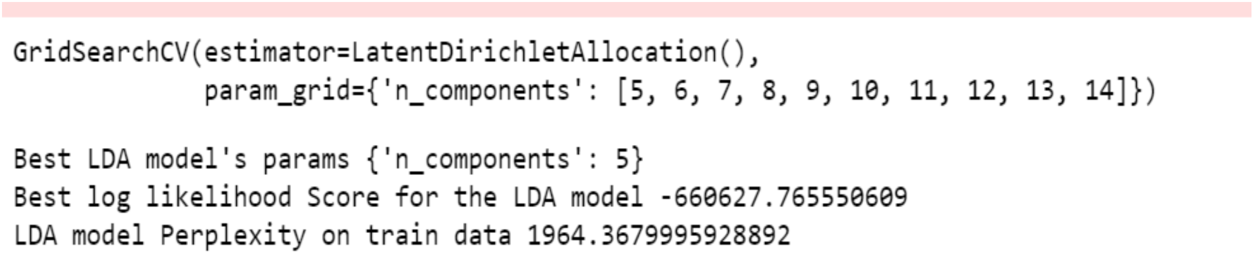
5: Model building, evaluation and selection

5.1. Model building and selection

- There are different models that can be used for topic modeling. Some of the popular models that are used for topic modeling are:
  - a) LSA (Latent Sematic Analysis)
  - b) pLSA (Probabilistic Latent Semantic Analysis)
  - c) LDA (Latent Dirichlet Allocation)
  - d) Ida2vec (Deep learning method uses neural nets)
- The model I used here for this BBC news articles is LDA.
- LDA is found be better for analysis for any type of dataset than using LSA and pLSA because the drawbacks in the other models has made LDA more efficient in solving the problem.

5.2. Model evaluation metrics

- The best parameters and score for this model can found through hyperparameter tuning the model using Grid Search CV.
- A model with higher log-likelihood and lower perplexity is preferred.
- After tuning the model with different parameters, the good parameters are found using the above metrics and results are shown below which is the highest log-likelihood and lowest perplexity that can be obtained on this data.

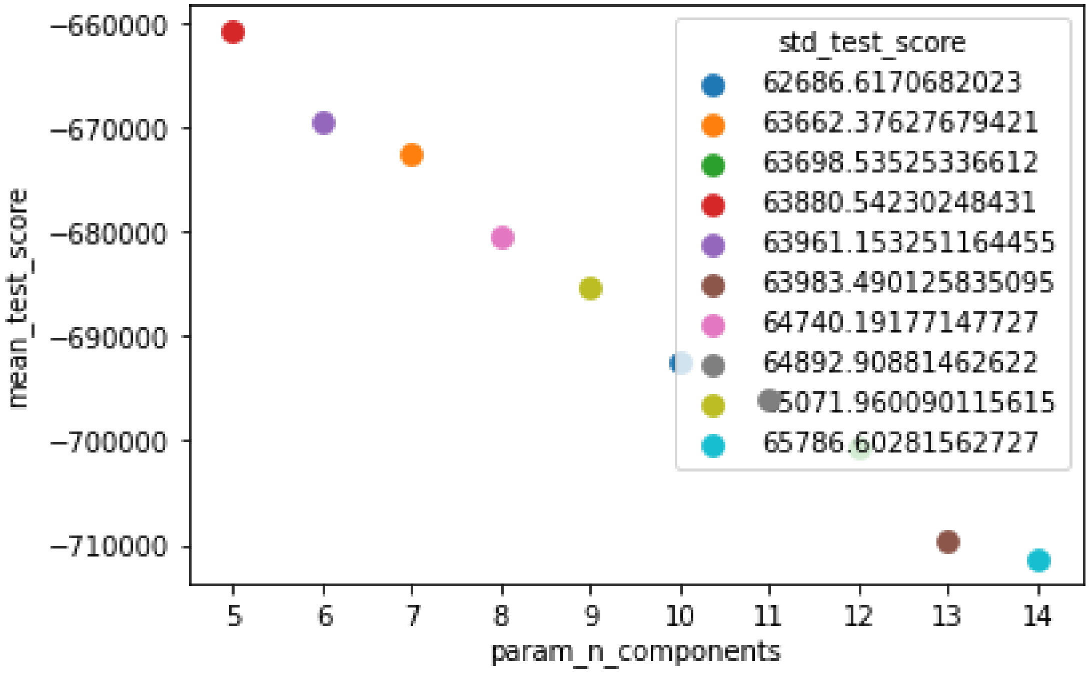


```
GridSearchCV(estimator=LatentDirichletAllocation(),
             param_grid={'n_components': [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]})

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
```

6: Case studies

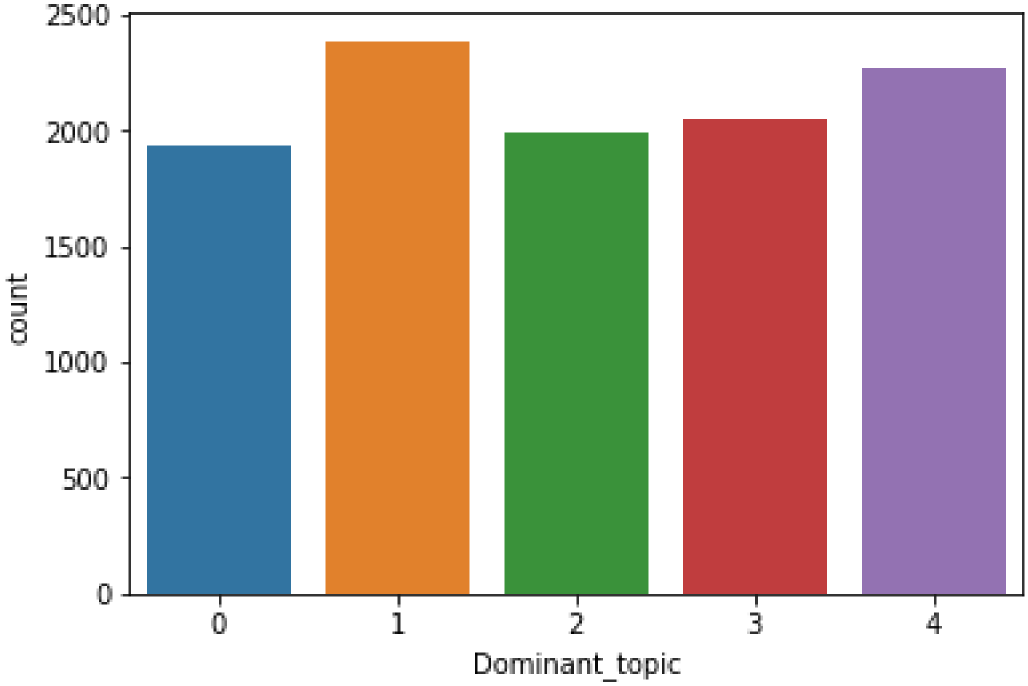
1. Find the test score results obtained when LDA is used on this BBC news articles data.



Observations:

- It can be said from the above plot that the test scores are good when the model uses 5 components or topics to extract from all the available documents.
- This plot is obtained from the results attribute present in the grid search cv class.

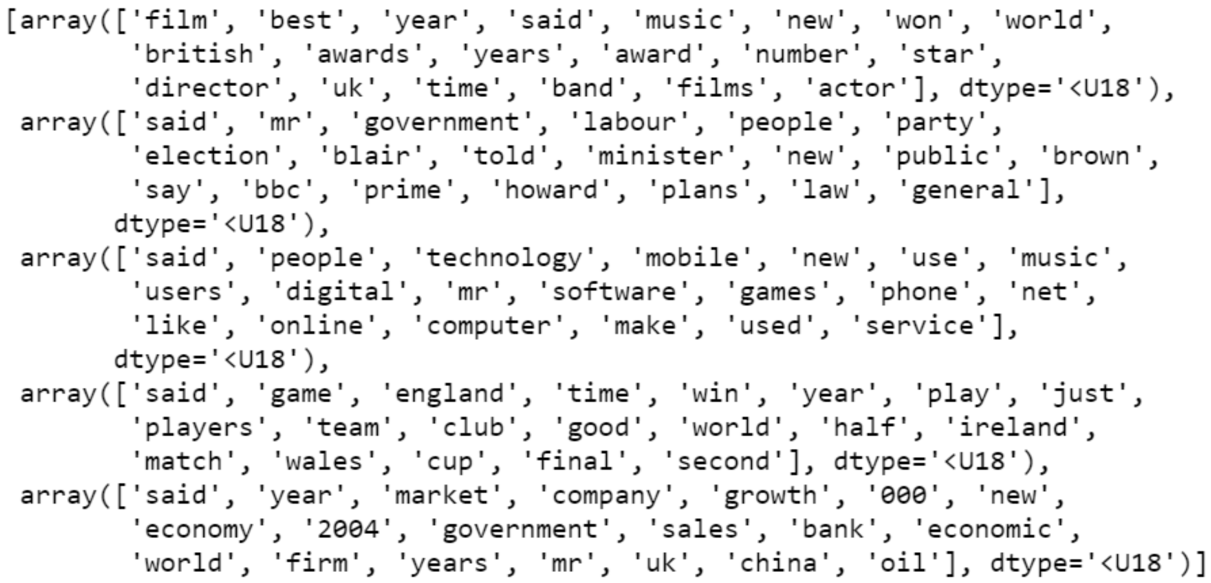
2. Find the dominant topic from all the topics based on the words attached to it or from topic modeling.



Observations:

- After topic modeling the probability of each word on the topic is represented in the dataframe format using which this plot is obtained
- From the above plot it can be seen all the topics have good number of words to distinguish from one another and also there are no minority topic column.
- It can be seen that the topic 1 is more or little higher than the other topics saying it is having more weightage than the other topics with the more word's probability towards Topic 1.

3. Find the top 20 words from each topic after topic modeling



```
[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')]
```

- The top 20 words from each topic obtained after topic modeling using LDA are present above.
- From the above representation of all the terms from each topic it is easy to distinguish and can say from which topic they belong to and also can say that the model has correctly identified the topics/themes from BBC news articles.

7: Conclusion

- Here the one and only most important feature that gives predictions or using which the model divided into different topics is using BBC news articles column which contains the cleaned news articles from the raw data given for analysis.
- I have used LDA here which I have found as the best and easy model that can be for topic modelling with less disadvantages compared to LSA and pLSA models.
- After analysis it is found that the identification of topics was correct and the model has distinguished between different topic news articles.