Capstone Project

Presentation on

Topic Modeling On News Articles

Presented by : 1. Priyabrata Mohanty
2. Rahul Sharma

Pipeline

- Problem statement
- Objective
- Data summary
- Data cleaning
- Machine learning model building
- Conclusion
- Appendix
    - Data Sources
    - Data Assumptions

Why Topic Modeling?

- 80% of data of the world is **unstructured**. Textual data is biggest example of it. Natural Language Processing is the very best step on structuring it.
- News rooms and editorials spend most of their times trying to figure out the preference of the audience and categories them vividly.
- Machine learning for **topic modeling** comes very handy to save time and work efficiently.

Problem statement

To identify major themes or topics across a collection of BBC news articles. You can use clustering algorithms such as Latent Dirichlet Allocation (LDA).

Data summary

The dataset contains a set of news articles for each major segment consisting of **business**, **entertainment**, **politics**, **sports** and **technology**.

- Number of Business articles: 510
- Number of Entertainment articles: 386
- Number of Politics articles: 417
- Number of Sports articles: 511
- Number of Technology articles: 401

Data cleaning & Processing

- **Merged** different articles together to form whole dataset.

- Removal of **stopwords**, **punctuations** & **unwanted** characters (e.g. “\n”) from data.

- Converted the texts into **lower case** & **tokenization**.

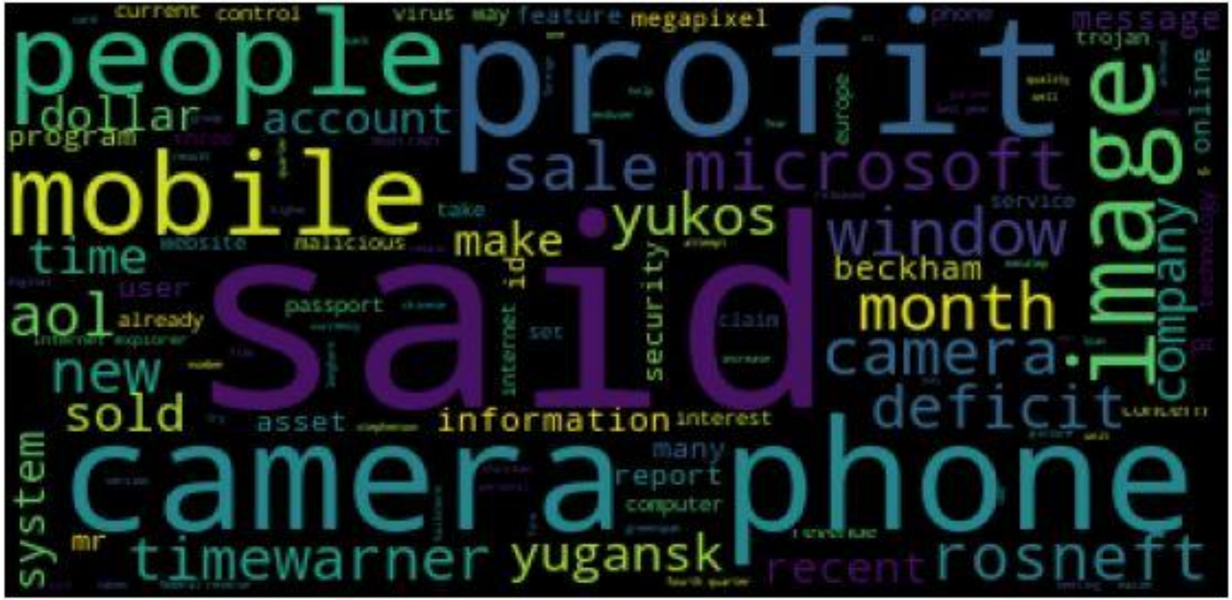
- Added few new columns **number of sentences**,**complex words**, **average length of sentence** to the dataframe.

Exploratory Data analysis

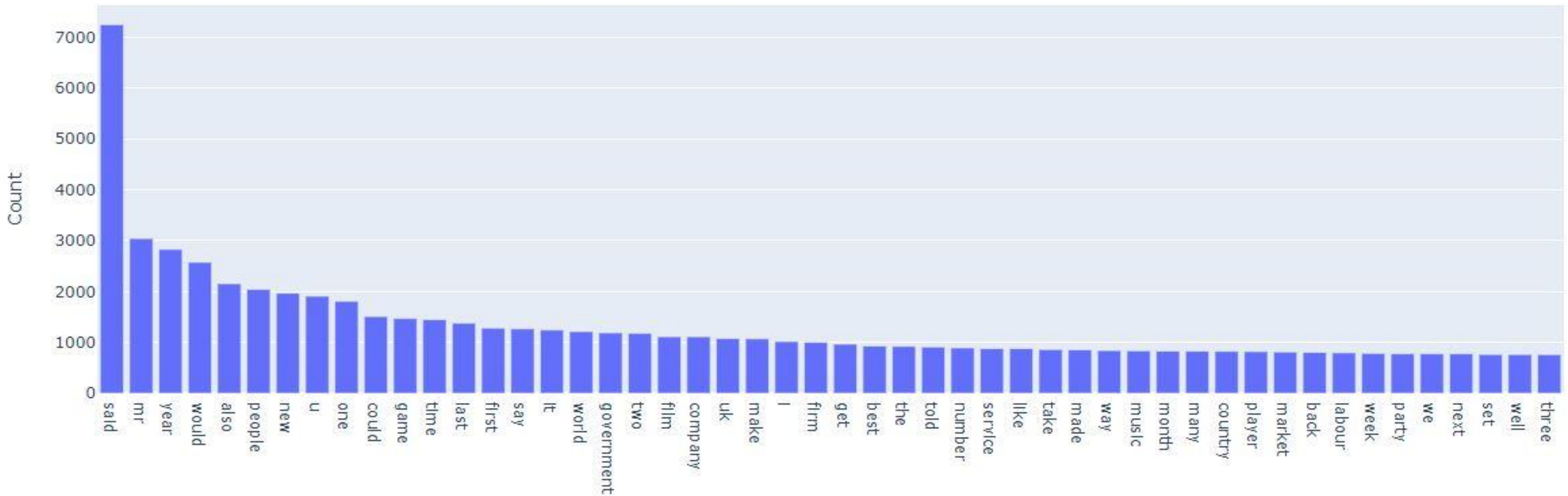
Important Words In The Article

According to above graphs:

- **Said**, **Profit**, **Camera**, **Phone**, **Window**, **People**, **Image**,, **Sale** are the most important words in this article.



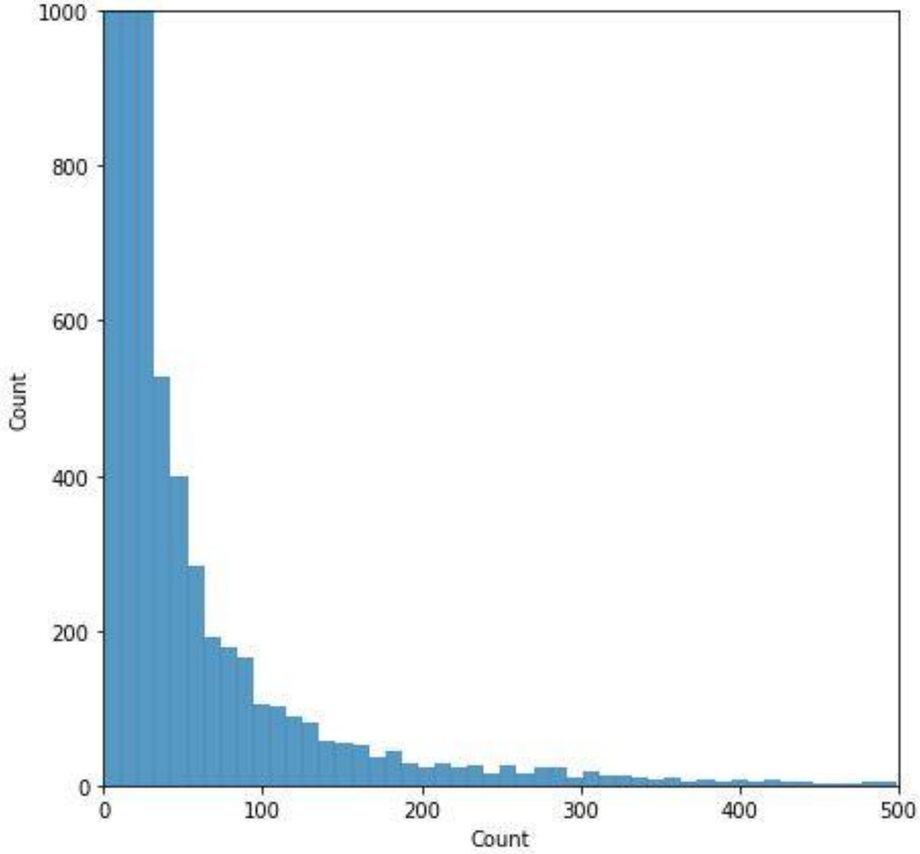
Word VS Frequency Graph



According to above graphs:

- **Said** is the most frequent word with frequency of **7253** (which is obvious - verb)
- Other most frequent words are **Camera**, **Phone**, **Profit**, **Mobile**, **Image**.

Word Counts of articles Distribution



According to above graph:

Most of the articles word counts is in the range of 0-100.

Machine Learning Model Building

Latent Dirichlet Allocation (LDA)

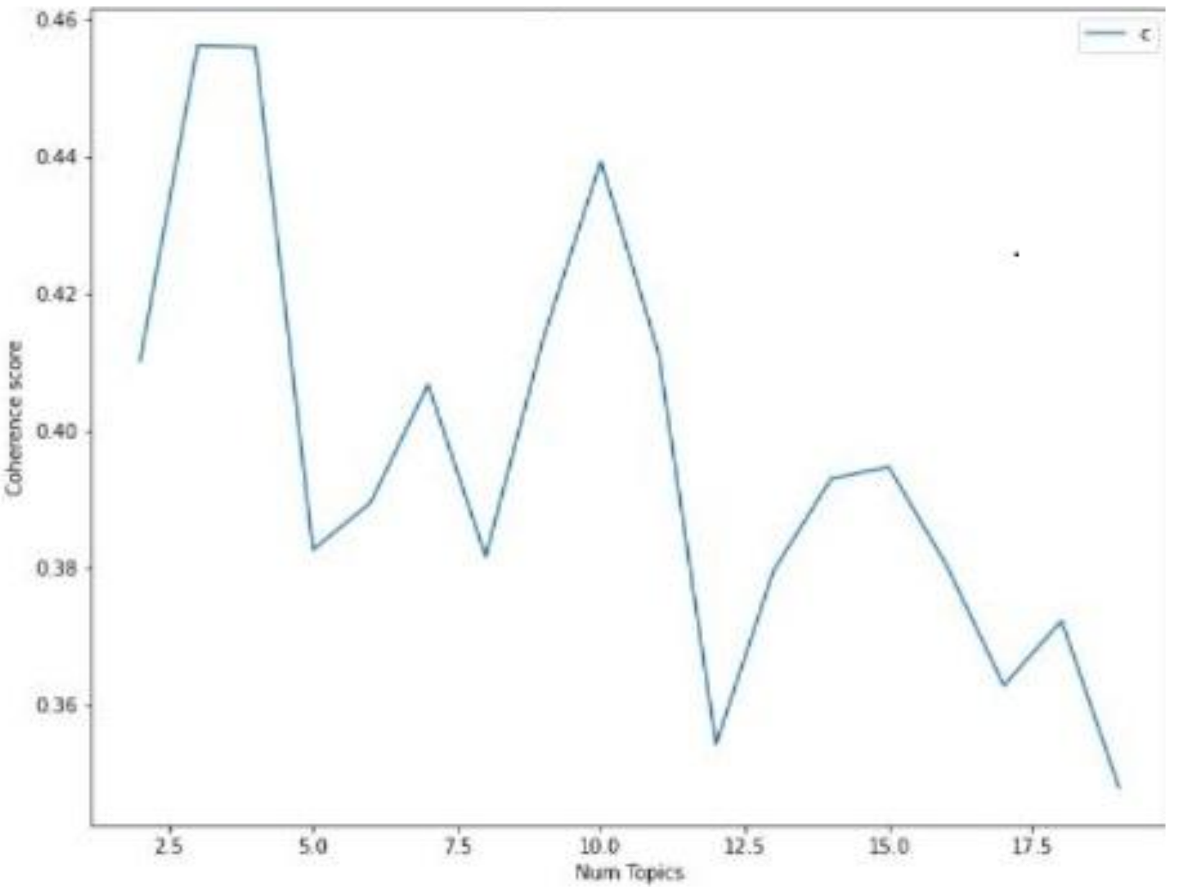
According to the above figure:

- When the number of topics is **10**, most of the topics are **overlapping** with each other.
- The coherence score for Number of topics=10 is 0.**4126**.

Number Of Topics VS Coherence Score

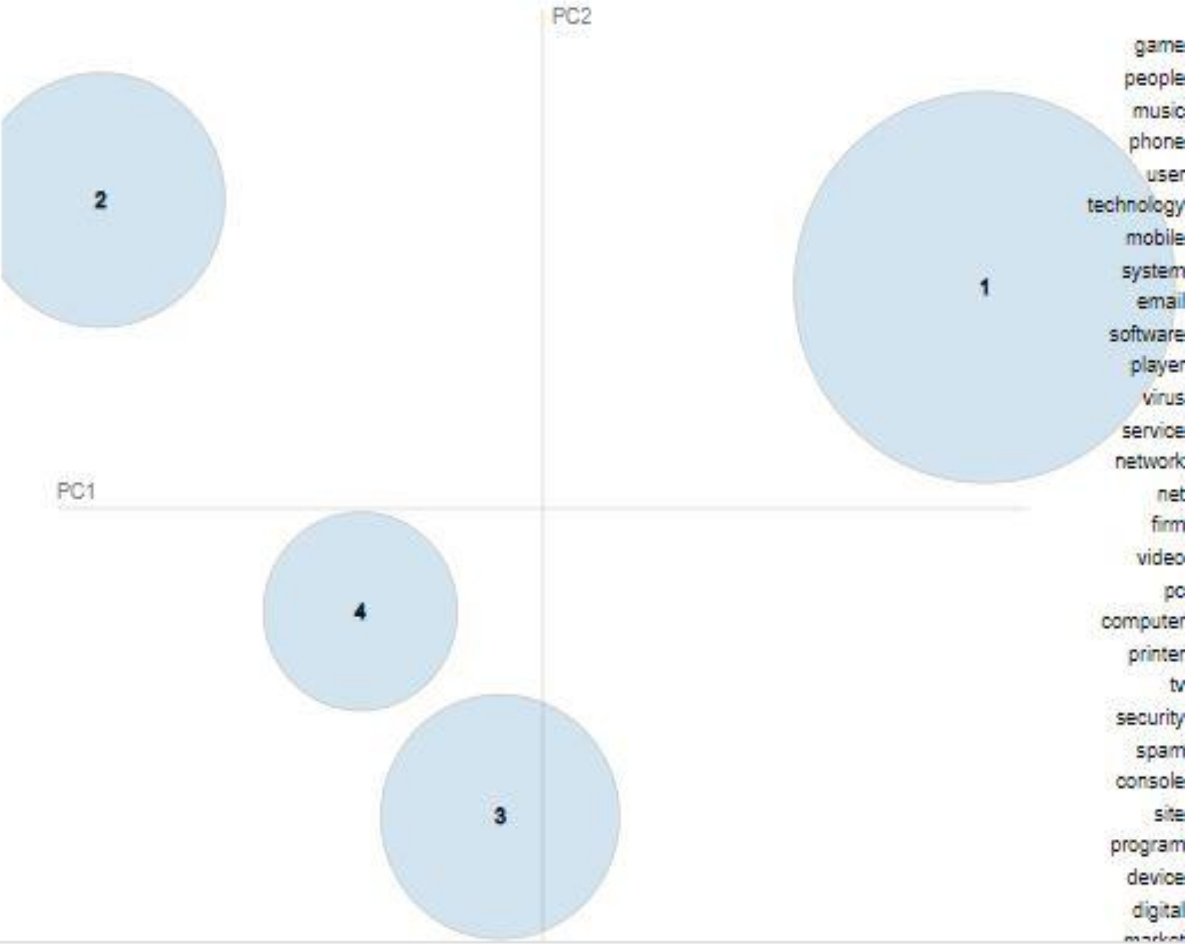
According to the above graph:

- For the number of topics 3 & 4, the coherence score is almost same (**0.456**) & it's the highest coherence score.



According to this graph:

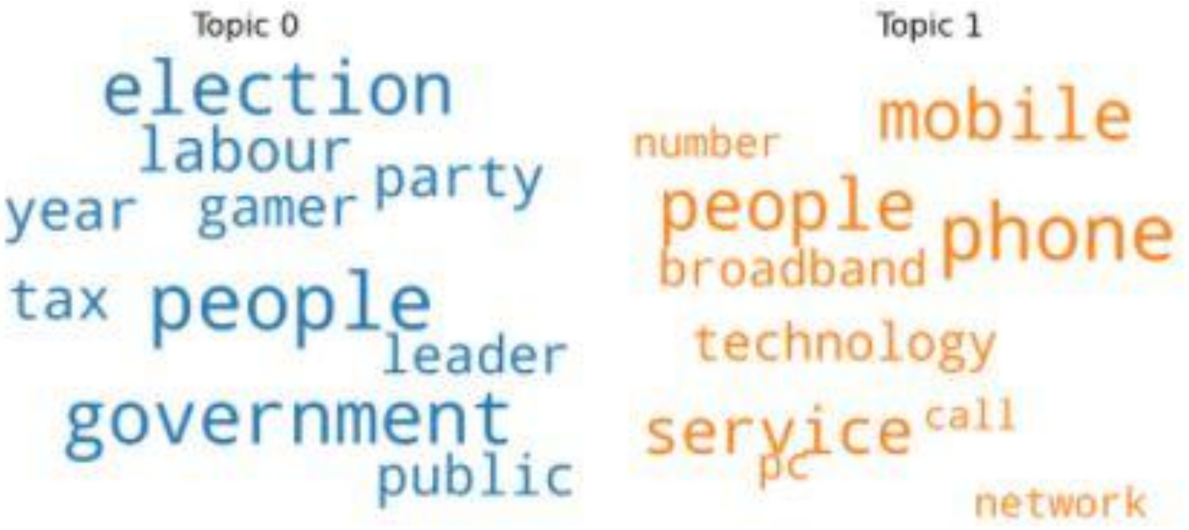
- The topics are not overlapping with each other when the number of topics is 4.



Wordcloud of Top N words in each topic

According to the above graph:

- These are the **Top 10** words for the **Topic-0** & **Topic-1**.



Wordcloud of Top N words in each topic

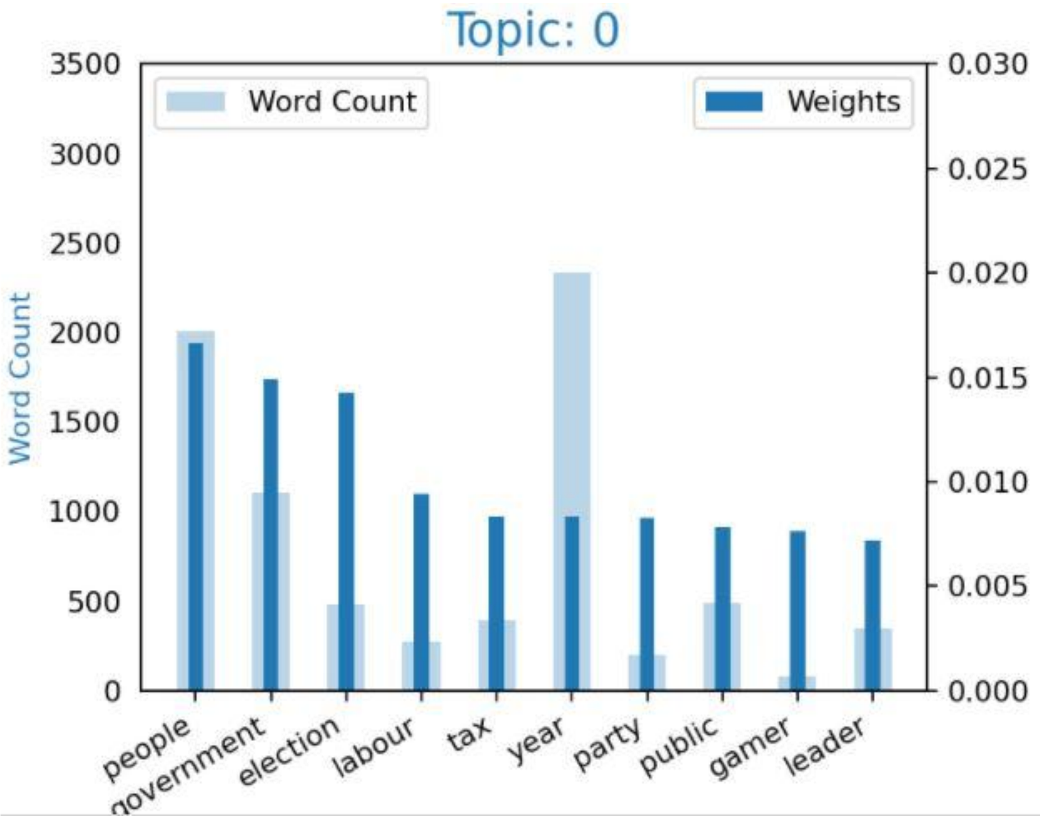
According to the above graph:

- These are the **Top 10** words for the **Topic-2** & **Topic-3**.

Words VS Frequency Graph

According to the above graph:

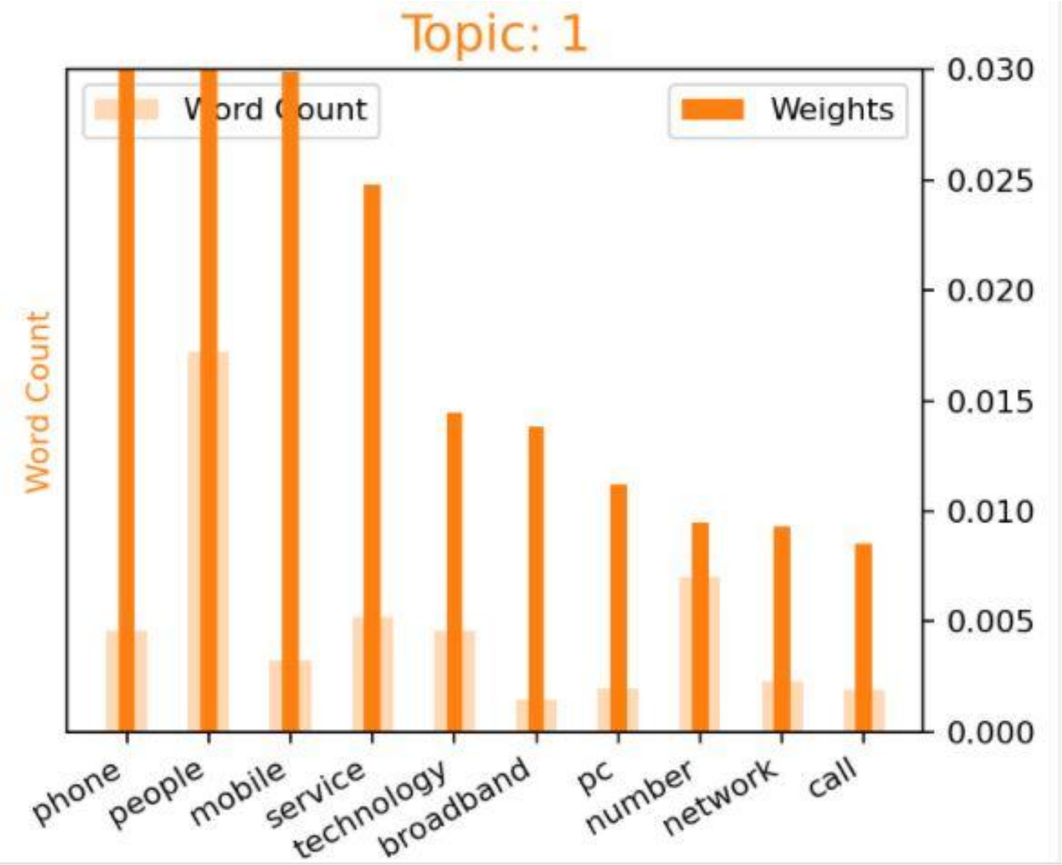
- These are the **Top 10** words for the **Topic-0.**
- These words are related to topic **government** & **Year** is the most frequent word.



Words VS Frequency Graph

According to the above graph:

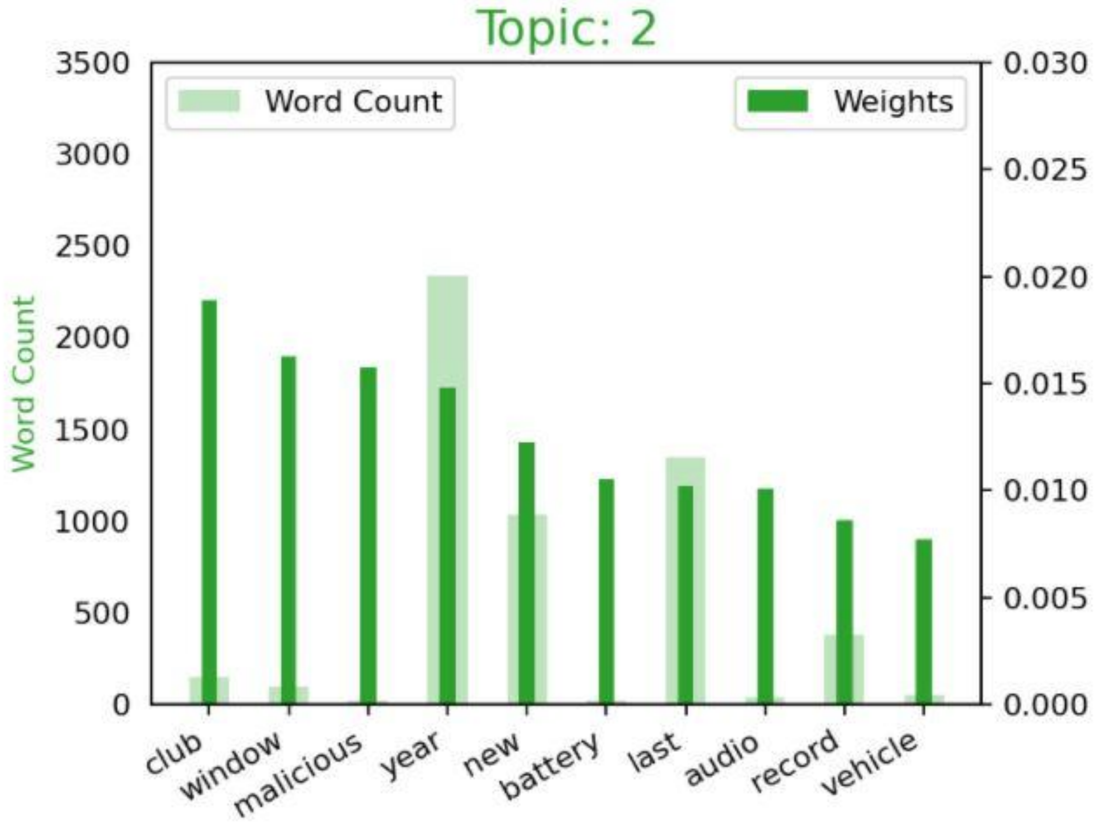
- These are the **Top 10** words for the **Topic-1.**
- These words are related to topic **Technology** & **Phone, People, Mobile** are the most frequent word.



Words VS Frequency Graph

According to the above graph:

- These are the **Top 10** words for the **Topic-2.**
- These words are related to topic **Entertainment** & **Club, Window** are the most frequent word.

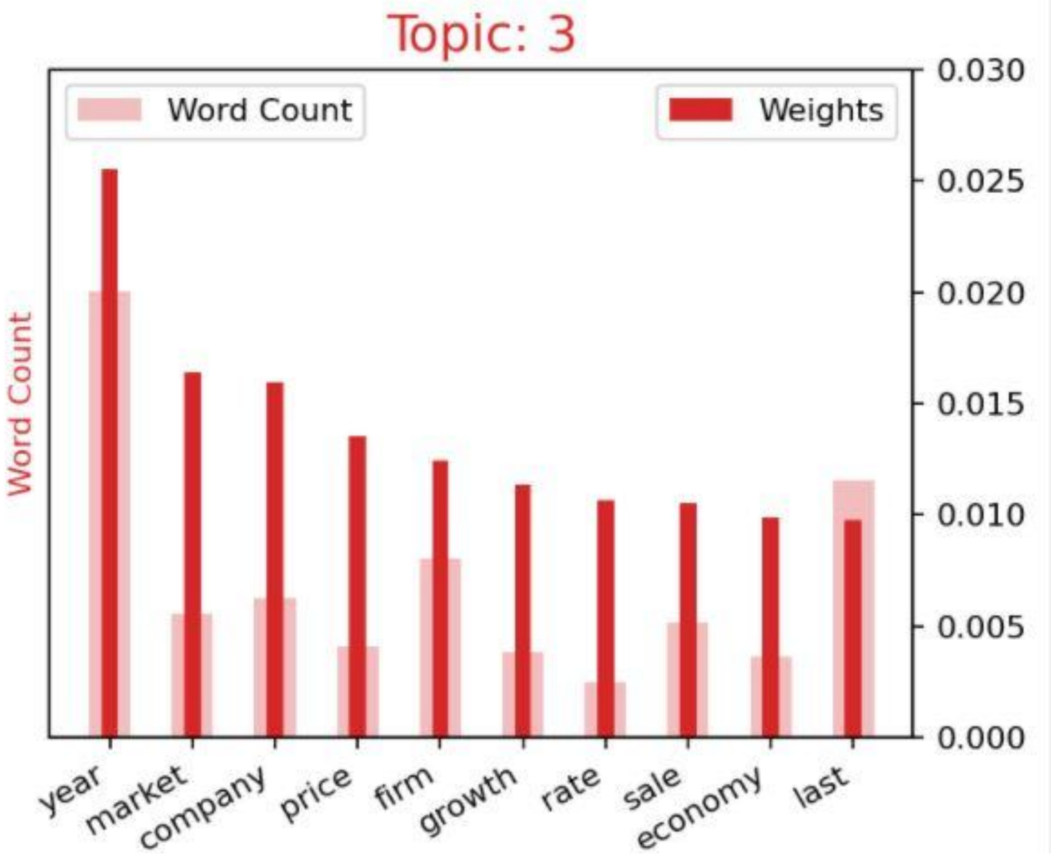


Words VS Frequency Graph

AI

According to the above graph:

- These are the **Top 10** words for the **Topic-3.**
- These words are related to topic **Business** & **Year, Market, Company, Price** are the most frequent word.



Conclusion

- The data was highly unstructured. So removed **unwanted characters**, **punctuations**, **stopwords**.
- Articles were distinctly analysed by aggregating their sentence counts and words.
- Coherence score were showing varying results for each additional run time. The best it did with 5 topics was **0.5**. But in the last run we had to keep it at 4 clusters because 5 was making overlapping between topics.
- The results were different for every runtime.

Conclusion

- The **People**, **Music**, **Technology**, **Player**, **Game** were the most important word for the **Topic-0**.
- The **Software**, **Phone**, **User**, **System**, **Computer** were the most important word for the **Topic-1**.
- The **Game**, **User**, **People**, **Firm**, **Mobile** were the most important word for the **Topic-2**.
- The **New**, **Year**, **Time**, **Price**, **Sale** were the most important word for the **Topic-3**.

Appendix - Data sources

- Here is a snapshot of data dictionary:

  We used the past **BBC News** articles based on 5 major

  Segments.

  1. Business
  2. Entertainment
  3. Politics
  4. Sports
  5. Technology

Appendix - Data Methodology

We conducted a thorough analysis & ML model building for Topic Modeling. The process includes:

- **Data cleaning** – Removal of stopwords, unwanted characters.
- **EDA** - Understanding data using different visualization methods.
- **ML Modelling** – Built & tested the machine learning models to do the topic modeling.

Thank You !