**Topic Modeling to study how terrorist organizations are portrayed in traditional media outlets**

**Mini-Project 3**

**Author:** **Yenimireddy Lokesh Reddy, Pruthvirajsinh Zala,** **Gajula Meher**

**Abstract:-** This study focuses on preprocessing and exploratory data analysis for a corpus of newspaper articles. The first step is to build the corpus by loading the data and splitting it into individual articles. The corpus is then cleaned up by separating metadata from the actual articles, and features are extracted. Summaries and plots are created for exploration. A topic model is created using the cleaned corpus. The model is run multiple times with different parameters and summaries are stored in output files. The results are discussed in the context of the project's overview statement, with relevant model outputs and visualizations used to support the findings.

It is important to note that this project is not expected to have full domain knowledge but to treat the analysis as one would support a journalist. The quality of the post-processed data will also be evaluated as part of the grading for this assignment.

**Keywords*—*** *preprocessing, exploratory data analysis, corpus, newspaper articles, metadata, feature extraction, summaries, plots, topic modeling*

1. **INTRODUCTION**

The explosion of digital content has led to an abundance of unstructured data in various forms, including news articles. Preprocessing and exploratory data analysis of such data can provide valuable insights and knowledge discovery for journalists and researchers alike. In this study, we focus on preprocessing and exploratory data analysis of a corpus of newspaper articles, with the aim of creating a topic model that can help identify and understand the topics that the articles cover.

The study begins with building the corpus, including loading the data as a large character, splitting it into individual articles, and cleaning up the corpus by separating metadata from the actual articles. Features are then extracted, and summaries and plots are created for exploration. Preprocessing is a critical step in this process, involving tokenizing the text and handling nuances such as punctuation marks, headers, tags, and dates. Next, we create a topic model using the cleaned corpus, adjusting parameters and running the model multiple times to store summaries into output files. We discuss the different results in the context of the project's overview statement and support our findings with relevant model outputs and visualizations. It is important to note that full domain knowledge is not expected, and the analysis should be treated as one would support a journalist. This study demonstrates the importance of preprocessing and exploratory data analysis in extracting meaningful insights from unstructured data, and the potential for topic modeling to identify and understand the topics that newspaper articles cover.

1. **METHODOLOGY**

**Diagram

Description automatically generated**

1. Build a corpus: This data that had been downloaded from a large Global News database called Factiva. The data has been collected in 2017 and contains articles from Wall Street Journal and the New York Times. We extracted the .txt files from the data. Once the text files are extracted, we split the articles from the keywords defined ('Document NYTF', 'Document INHT', 'Document WSJ', 'Document J000','Document AWSJ'). A total of 1648 articles are extracted.
2. Clean up the corpus: We checked if the parsed articles are in the right format and if we extracted all of them properly. Since most of the article’s end with “All rights reserved” we separate them else we discard them. Metadata is also extracted, and we now have a corpus of 1620 articles.
3. Pre-Processing: A series of steps were performed such as removing punctuation, converting each word into lower case, tokenization of words, and using the NLTK stopwords list we removed the stopwords and then stemming is performed using the porter stemmer.
4. Extract features: To extract the information a document term matrix is formed, and the top words are visualized using a word cloud.
5. Topic modeling (LDA): We preprocessed the data again for topic modelling using genism, we built bi-gram and tri-gram models and performed lemmatization. We built an LDA model using genism with our input as corpus and set the number of topics.
6. Evaluation of the results: We used pyLDAvis library for visualization of topics and terms within it. The perplexity score measures how well the model predicts new data and the lower the score, the better the model performs. The coherence score, on the other hand, evaluates the semantic coherence between the topics and higher scores indicate better coherence. The results indicate that the model performs well in general, with perplexity scores ranging from -9.59 to -10.77, which suggests that the model predicts new data with relatively low error rates. However, the coherence scores are relatively lower, ranging from 0.44 to 0.50, indicating that the semantic coherence between topics could be improved. The number of topics that produced the best performance based on both metrics is 19, with a perplexity score of -10.09 and a coherence score of 0.50. This indicates that the model was able to produce topics that are both coherent and predictive of new data. However, it is worth noting that the performance does not necessarily improve as the number of topics increases, and there is a trade-off between model complexity and performance.
7. **EXPERIMENTAL RESULTS**

[Link for GitHub Repository](https://github.com/CIS-6397-Textmining-Spring-2023/miniproject3-miniproject3_group7.git)

Word Cloud based on word counts after initial preprocessing.

Text

Description automatically generated

**Selection of Best Number of Topics:**

**Text

Description automatically generated**

As discussed, we calculated the scores while building our topics and found 19 to be the optimal number.

**Approach and thought process while modelling:**

It is important that we don’t stick to score itself and try to model based on our intuition as well, based on the articles it is from various sources depicting terror attacks and discussions around it. So, from a general viewpoint, the topics can be about politics, administration, terrorism, national security and these do tend to have an overlap as well. As a result, we also tried to compute the scores by inputting model numbers from 3 -10 topics.

Table

Description automatically generated

Since the coherence score weren’t optimal, we stuck to build for 19 models.

**Visualization and Inference**

We tried to cluster a total of 19 models, and what we see on the map is shown as below.

Chart, bubble chart

Description automatically generated

We see some overlap and concentration between the clusters. This necessarily doesn’t mean it’s a bad thing, when we are modelling about topics in real life, we do see a lot of overlap between different sections.

Chart

Description automatically generated

We can see the top 30 words associated with each cluster and do not see any words which harmonize our topic properly, there are words like go, think, s, get , say and these are common words. We can have an improved topic modelling when we use a removal list of all common words. Since we have a cluster which depict majority of the tokens, we left it at its current state.

Chart

Description automatically generated

As cluster 4 doesn’t have any overlap we can infer that the words in it will contain only related to it and we can see that the top words are related. Similarly, we see the same with cluster 10 as well but since it only contains 2.4% of tokens, we cannot say that it can be considered as a topic as well. Chart, bar chart

Description automatically generated

1. **CONCLUSION**

To conclude, this study emphasizes the significance of pre-processing and exploratory data analysis in the domain of newspaper articles, followed by topic modeling. By constructing a corpus, refining it, and extracting pertinent features, we managed to generate summaries and visualizations that facilitated us in comprehending the data. Through topic modeling, we successfully identified significant themes and topics that arose from the corpus, which can aid journalists in comprehending the data and enhancing their reporting.

This study underlines the importance of data analysis in journalism and the need for proper preparation of data for analysis. Though the process of pre-processing and analyzing data may be arduous and time-consuming, the insights gained from the analysis can assist journalists in creating informed and impactful stories. In future studies, it would be worthwhile to remove list of common words and explore additional analytical techniques or apply the same methodology to another corpus of data to compare the findings.

**AUTHOR CONTRIBUTIONS**

**Yenimireddy Lokesh Reddy**: Conceptualization, Methodology, Software, Validation, Reviewing and Editing, Supervision.

**Pruthvirajsinh Zala**: Data curation, Investigation, Software, Writing- Original draft, Reviewing and Editing, Resources;

**Gajula Meher**:Reviewing draft, Visualization, Supervision, Editing