Team 3 – Mini Project 3 - Topic Models

Shyam Reddy Kotha, Sujitha Ravichandran, Shifafatima Khoja

Abstract

This document summarizes the findings after the completion of Mini Project 3. The task is to perform the analysis of how terrorist organizations are portrayed in the media by using topic modeling. The task is to be accomplished using Python by transforming the given data into a corpus that can be used as the basis for the topic modeling process.

1 Introduction

The objective of this assignment is to investigate the portrayal of terrorist organizations in the media using data obtained from the Factiva database. To achieve this, topic modeling techniques will be employed. However, before the data can be analyzed, it needs to be transformed into a corpus and undergo preprocessing to ensure it is ready for further analysis.

The data analysis phase will involve extracting features from the processed corpus, summarizing these features, and creating plots to facilitate exploration of the data. This will provide insights into the characteristics and patterns within the dataset. To ensure robustness, this process will be repeated multiple times to obtain a diverse range of results. Visualizations will be generated to effectively present the final findings of the analysis.

By performing topic modeling techniques on the preprocessed data, we will uncover the underlying topics present in the corpus. This will enable us to identify the main themes and subjects associated with terrorist organizations as portrayed in the media. By conducting multiple iterations of topic modeling, we can account for variations in the results and ensure a comprehensive understanding of the data.

To enhance the understanding and interpretability of the findings, text visualizations will be created. These visual representations will aid in presenting the identified topics, their prevalence, and their relationships with other topics. The visualizations will provide a clear and concise overview of the key findings, enabling a more intuitive understanding of the media portrayal of terrorist organizations.

In summary, this assignment aims to analyze the portrayal of terrorist organizations in the media by employing topic modeling techniques. By transforming the data into a corpus, preprocessing it, and extracting features, we will gain insights into the dataset. Through topic modeling iterations and text visualizations, we will identify and explore the underlying topics, ultimately providing a comprehensive understanding of how these organizations are depicted in the media.

2 METHODOLOGY

To find useful insights from each corpus, the corpus needs to be analyzed and preprocessed first. We used several techniques to preprocess the corpus. To analyze the data, we leveraged NLTK. We started by defining a function to read all text files in a corpus and return the list of the text content of each file. Next, we preprocessed the corpus**.** The first step is to preprocess the text data using the preprocess() function. This involves tokenizing the text into words and sentences using NLTK's tokenization functions. It further removes HTML tags, newlines, digits, punctuation marks, and extra spaces from the text. Stop words (common words) and non-alphabetic characters are also eliminated. The resulting preprocessed tokens are stored in the token's variable.

The next step is to extract relevant features from the preprocessed tokens using the extract feature's function. This includes computing word frequencies using the Counter function to

determine the most common words and their frequencies. The most common words are selected and stored in the most common variable. Additionally, the average sentence length is calculated by dividing the total number of sentences by the total number of tokens. The number of sentences is also counted. The extracted features are stored in the features dictionary.

After extracting the features, the summarize feature's function is used to provide a summary of the extracted information. This includes printing the most common words and their frequencies, the average sentence length, and the number of sentences. This summary allows for a quick overview and understanding of the important aspects of the text data.

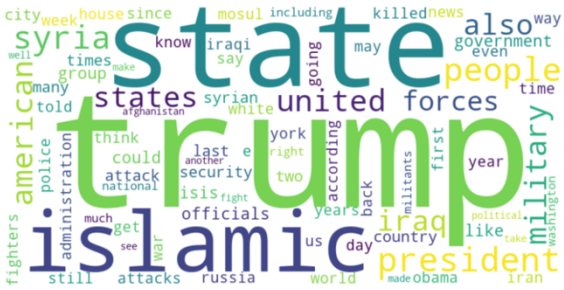


Figure 1

Finally, the plot feature's function is used to visualize the most common words using a word cloud. This step utilizes the Word Cloud library to generate a word cloud image based on the word frequencies. The resulting word cloud provides a visual representation of the most frequent words in the text, allowing for a more intuitive understanding of the data.

3 EXPEREIMENTAL RESULTS

The methodology was applied to perform topic modeling using LDA and evaluate different parameter combinations for the LDA models. LDA is a popular probabilistic modeling technique used for discovering latent topics within a collection of documents.

3.1 Topic Model using cleaned corpus

Topic modeling is a statistical modeling technique used to uncover hidden themes or topics within a collection of documents or texts. It is a machine learning approach that aims to automatically identify the main topics discussed in a corpus without prior knowledge or supervision. By analyzing the patterns of word usage across documents, topic modeling algorithms can group together words that frequently co-occur and assign them to specific topics. The most widely used topic modeling algorithm is Latent Dirichlet Allocation (LDA). LDA assumes that each document in the corpus is a mixture of various topics, and each topic is a probability distribution over words.

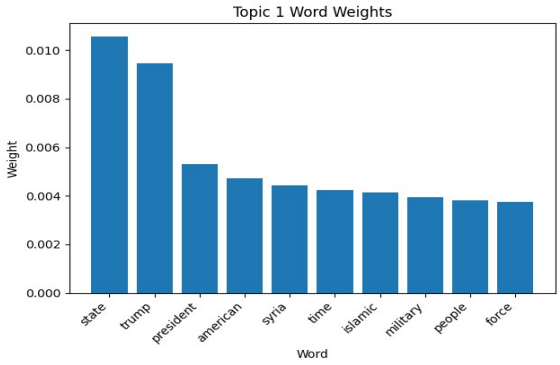


Figure 2

The corpus is preprocessed to prepare it for topic modeling using Latent Dirichlet Allocation (LDA). The preprocessing steps involve removing excess newlines, non-alphabetic characters, and converting the text to lowercase. The corpus is then split into individual documents or paragraphs.

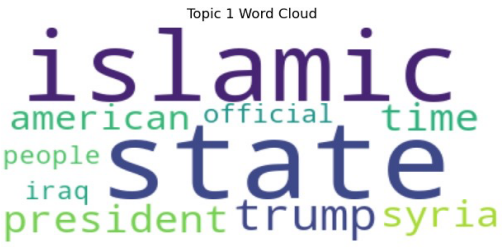
Next, a dictionary and a bag-of-words corpus are created using the preprocessed texts. The dictionary maps each unique word to a unique integer ID, while the bag-of-words corpus represents each document as a list of word-frequency pairs. These preprocessing steps are necessary for LDA, as it requires a numerical representation of the text data. The code then proceeds to create multiple LDA models with different parameter combinations, such as the number of topics, alpha, and eta values. The LDA model is trained on the corpus using the specified parameters and a fixed number of iterations. The coherence score and perplexity of each model are calculated to evaluate their quality and effectiveness.

3.2 Re-run the topic modeling multiple times with different parameters

The topic modeling process is re-run multiple times with different parameters to explore various configurations and obtain a range of results. The parameters that are varied include the number of topics , the alpha value , and the eta value. By re-running the topic modeling with different numbers of topics, we can investigate how the corpus can be best represented by a specific number of themes. Different topic numbers may lead to different interpretations and granularity of topics. It allows for a comparison of the coherence and interpretability of the resulting topics. Additionally, adjusting the alpha and eta values affects the topic-word and document-topic distributions. These hyperparameters control the sparsity of these distributions. Varying the alpha and eta values enables the exploration of different assumptions about the distribution of topics in documents. This helps in finding the most appropriate values that capture the under lying structure of the corpus. By re-running the topic modeling multiple times with different parameter combinations, we can evaluate the impact of these parameters on the coherence of the topics and the overall performance of the model. It allows us to fine-tune the topic model and select the optimal configuration that produces the most coherent and meaningful topics for further analysis.

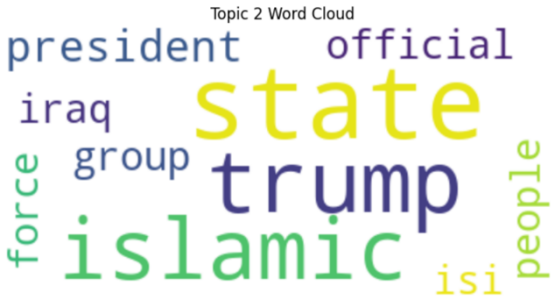
3.3 RESULT DISCUSSION

Topic 1: state, trump, president, islamic, syria, attack, people, iraq, american Weights: 0.009595179, 0.0062024826, 0.00540637, 0.0052916817, 0.004773099, 0.004714872, 0.004599162, 0.0044885534, 0.003929896, 0.003864886



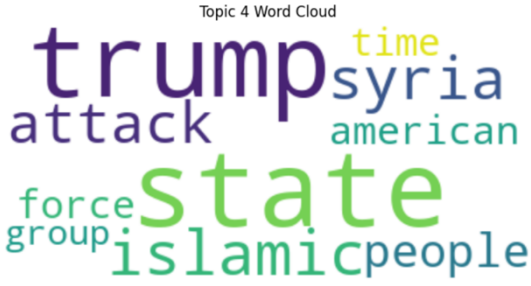
Figure

Topic 2: state, trump, islamic, people, time, american, president, attack, military, official, isi, Weights: 0.00801272, 0.007091743, 0.0052596414, 0.0049599637, 0.004586809, 0.004502214, 0.004329628, 0.004092438, 0.0038517066, 0.0038368294



Topic 3: state, trump, islamic, american, time, force, syria, attack, people, President Weights: 0.0111515, 0.007629762, 0.0073644705, 0.005591276, 0.00450667, 0.004234051, 0.00411387, 0.003943324, 0.0038461557, 0.0037839916

Topic 4: state, trump, islamic, force, people, time, group, american, president, country Weights: 0.009034898, 0.008306639, 0.005643357, 0.004574524, 0.0042806207, 0.0041016527, 0.0038455587, 0.0037025746, 0.0036831729, 0.0036174678



Topic 5: state, trump, islamic, president, american, time, country, official, isi, united

Weights:0.012341319, 0.007926726, 0.0064171185,0.0046068537, 0.0044217044, 0.0041473717, 0.003691047, 0.003572185, 0.0035320064, 0.003514869.

The results obtained from re-running the topic modeling process with different parameters provide valuable insights into the corpus and help in selecting the most appropriate topic model for further analysis. The discussion of coherence scores, parameter variations, and topic-word distributions allows for a comprehensive evaluation of the topic models and facilitates meaningful interpretation and understanding of the corpus.

3.4 Model Outputs and Visualizations

The visualization of topic-word distributions plays a crucial role in topic modeling analysis. It provides a more intuitive and accessible way to understand the topics identified by the LDA (Latent Dirichlet Allocation) model and their associated word importance. Instead of relying solely on numerical values or textual representations, the visual representation allows researchers and analysts to grasp the essence of each topic briefly. By plotting the word weights as bar charts as shown in figure 1, the visualization highlights the most significant words within each topic. The height of each bar represents the weight or importance of a word in the topic, indicating its relevance to the underlying theme. This visual hierarchy helps identify the dominant keywords and provides insights into the main ideas encapsulated by the topic.

Apart from the word weights bar chart, we also implemented word cloud as shown in figure 2 .Word clouds are a popular and visually appealing way to represent the frequency or importance of words in a corpus. They provide a compact and informative overview of the most prominent terms within a given topic. After extracting the topics using the LDA model, the code iterates over each topic and retrieves the top words and their corresponding weights. These word-weight pairs are then converted into a dictionary, where the word serves as the key and the weight as the value. Next, a Word- Cloud object is created, specifying a white background color. The Word Cloud object takes in the word frequencies or weights as input and generates the word cloud visual representation.

4 Conclusion

This assignment allowed us to opportunity to practice our preprocessing skills and perform topic modeling on real data. The results after applying topic modeling techniques on processed data from the Factiva database showed us the most common topics from the corpus with their weight compared to the other topics. After performing the task multiple times, it was clear that the topics found each time were very similar, which was to be expected. Additionally, we were able to visualize the information produced into graphs and word clouds, which helped us analyze our results efficiently. This project helped us to learn more about topic modeling by utilizing it to find out how terrorist organizations are portrayed in the media, which turned out to be a great learning experience and reaffirmed the merits of text mining and its various applications.

5 Authors

Shyam Reddy Kotha: Software, Methodology, Writing - Original Draft

Sujitha Ravichandran: Visualization, Writing - Original Draft, Writing - Reviewing and Editing

Shifafatima Khoja: Data Curation, Writing - Original Draft

6 Appendices

GitHub Repository:

https://github.com/CIS-6397-Textmining-Spring-2023/miniproject3-miniproject3_group-3