IRWA Final Project - Part 1 Report

Overview

In Part 1 of our final project, we focused on processing a set of tweets related to the 2021 Farmers Protests, aiming to clean and prepare the text data for subsequent indexing and search tasks. Our approach centered on standard Natural Language Processing (NLP) techniques to make the dataset more suitable for information retrieval while also conducting exploratory data analysis (EDA) to gain insights into the content. Below, we outline our methodology and the specific decisions we made at each step of the process.

Pre-Processing

The first step of our implementation involved pre-processing the tweets. Different steps were to be done, to have a clean and workable dataset:

1. **Stop Word Removal**: We used the NLTK library to remove common stop words from each tweet. This was done to reduce noise and retain only the meaningful components of the text that would be beneficial in an information retrieval context.

2. **Tokenization**: We tokenized the text using the NLTK tokenizer, breaking down the tweets into individual words (tokens). This process allows us to analyze the text more effectively by isolating each word for further processing.

3. **Punctuation Removal**: All punctuation marks were removed to ensure consistency and avoid including unnecessary symbols in our tokens. This made our dataset cleaner, particularly since punctuation is not typically useful for understanding tweet content in this context.

4. **Stemming**: We applied the Porter Stemmer, also available through NLTK, to reduce words to their root forms. This step was taken to group together different forms of the same word, thereby reducing the dimensionality of our dataset. For example, "protests" and "protesting" would be reduced to the root "protest."

5. **Hashtag Handling**: We made a specific decision regarding hashtags. We retained the hashtag symbol (#) when preprocessing, treating hashtags as distinctive entities within the tweet content. We thought that hashtags provide a unique context to the discussion and could be useful in distinguishing between topics or trends.

6. **Other Normalization Steps**: Additional processing included converting all text to lowercase to ensure uniformity and removing URLs that were not directly informative for our analysis. Furthermore, we mapped tweet IDs to document IDs as required, which ensures consistency during later phases of the project, such as evaluation and indexing.

Exploratory Data Analysis

To gain a better understanding of the dataset, we performed a comprehensive EDA, employing the following techniques:

1. **Word Count Distribution**: We calculated the word count distribution across all tweets. This gave us insight into the typical length of tweets, which helped us determine if additional processing was necessary.

2. **Vocabulary Size**: We computed the total number of unique words in the dataset after pre-processing, which gave us a measure of the dataset's richness. This metric helped us understand the scope of our vocabulary and informed our decisions on stemming and stop-word removal.

3. **Tweet Ranking by Retweets and Likes**: We ranked tweets based on their retweet and like counts to identify the most influential tweets. This analysis provided valuable context for understanding the spread and popularity of certain topics..

4. **Word Cloud**: We generated a word cloud to visualize the most frequent terms. This helped us quickly identify key themes in the dataset, such as frequently used hashtags and terms closely related to the Farmers Protests.

Conclusion

The pre-processing and exploratory analysis steps completed in Part 1 have been a base point for further work in the project, including indexing and ranking. All our decisions were taken in order to be as efficient as possible when working with the dataset while maintaining the information richness. This foundation ensures that our dataset is both clean and informative, facilitating effective information retrieval in subsequent project phases.

Further mentions:

GitHub URL: https://github.com/enekotrevi/IRWA-2024.git

TAG: IRWA-2024-part-1