Report Part 1: Text Processing and Exploratory Data Analysis

1. INTRODUCTION

In this project, the final objective is to build a search engine. As covered in theory, the foundation for developing an efficient search engine lies in properly processing the data. Therefore, the aim of this first phase was to preprocess textual data and explore it adequately, laying the groundwork for search engine development. The dataset used in this project consists of tweets, which provide a rich source of information for text-based data analysis. By preprocessing the text and conducting exploratory data analysis (EDA), the objective was to better understand the dataset's characteristics and prepare it for future tasks, such as indexing and query processing.

2. METHODOLOGY

Now let’s focus on the methodology we used at data preparation, text processing and exploratory data analysis phases, explaining what we have done in each phase and why we have made some decisions.

2.1. Data Preparation

The process began by importing essential libraries, including:

- NLTK (Natural Language Toolkit): For natural language processing tasks, such as tokenization and stemming.
- NumPy and Pandas: For data manipulation and statistical analysis. The dataset of tweets was accessed from a shared Google Drive folder and loaded into the environment for further processing.

2.2. Text Processing

The text processing phase involved several steps to clean and normalize the data:

- Language: First we filter by language and select only english tweets.
- Lowercase: Letters are reduced to lowercase.
- Tokenization: Tweets were broken down into individual words or tokens to facilitate further analysis.
- Stopword Removal: Commonly used words that do not add significant meaning to the content, such as 'the,' 'is,' and 'and,' were removed from the dataset.
- Stemming: Words were reduced to their base forms using the Porter stemming algorithm to ensure consistency in word representation.
- Special Character Removal: Non-alphabetic characters and extra spaces were stripped out to clean the text.

2.3. Exploratory Data Analysis

The EDA phase aimed to uncover patterns and insights within the dataset through the following techniques:

- Frequency Analysis: The most frequently occurring terms were identified to understand the predominant topics discussed in the tweets.
- Entity Recognition: Named Entity Recognition (NER) was applied to identify key elements in the text.
- Sentiment Analysis: The overall tone of the tweets was assessed to gain a perspective on public opinion.

3. RESULTS

3.1. Frequency Analysis

The analysis revealed the most frequently used terms in the dataset, reflecting common topics and themes. Words related to protests, agriculture, and social issues appeared prominently, indicating the dataset's focus.

3.2. Entity Recognition

Named Entity Recognition (NER) was applied to automatically identify and classify key elements in the tweets, such as names of people, organizations, locations, dates, and other relevant terms. This helped structure the information contained in the unstructured data.

- People and Organizations: Frequently mentioned names of public figures, activists, and organizations involved in the topics discussed in the tweets were detected.
- Locations: Various geographical locations were recognized, facilitating the analysis of how tweets related to specific areas or regions.
- Dates and Events: Identifying dates and references to events contributed to contextualizing the tweets and understanding the time frame in which they were made. Using entity recognition improved the understanding of the content of the tweets and enabled the extraction of relevant information for the development of the search engine.

3.3. Sentiment Analysis

Sentiment analysis revealed that the dataset contained a mixture of positive, negative, and neutral tweets, providing information about public opinion trends regarding the discussed topics. The average sentiment value of the tweets was calculated to obtain an overall view of the predominant tone, and was -0.05, indicating that the top 10 most retweeted tweets tended to be neutral.

4. CONCLUSIONS

The first phase of the project successfully prepared the dataset for further search engine development. Text processing and exploratory data analysis revealed valuable information about the nature of the data, including common topics, relevant entity recognition, and sentiment trends. These findings provide a solid foundation for building an effective search engine, enabling the retrieval of relevant information based on user queries. The next steps will involve implementing indexing techniques, query processing, and ranking algorithms to further develop the search engine. Additionally, fine-tuning the preprocessing pipeline can improve search performance by enhancing text normalization and feature extraction.

5. OTHER ASPECTS

GITHUB Repository: *https://github.com/davidobrero/IRWA-2024-G102-10*
All related with Part 1 can be found inside the folder IRWA-2024-PART-1 containing:

- This report (IRWA-2024-PART-1-REPORT)
- The code (IRWA-2024-PART-1-CODE)

It is possible to open the code with Google Colab, connecting the data to Google Drive.