Jaume Camps Romaguera - NIA 252021
Arol Garcia Rodríguez - NIA 252718
Albert Garrell Golobardes - NIA 254635

Group 101_5

IRWA 2024 Part 1: Text Processing and Exploratory Data Analysis

1. Introduction

In the first part of our project, we aim to pre-process the dataset of tweets related to the Farmers Protests of 2021 and perform Exploratory Data Analysis (EDA). The dataset contains Twitter data, and the objective is to prepare the dataset for further phases of the project, where this information will be used for building a search engine. This report explains the decisions made in pre-processing the documents and highlights the insights found from the dataset through EDA.

The files used in this part of the project are:

- **farmers_protest_tweets.json**: Contains the raw tweet data, including the tweet text, date, hashtags, likes, retweets, and URLs, needed for the final output of the query.
- **tweet_document_ids_map.csv**: Used to map tweet IDs to document IDs, ensuring consistency across the pre-processing and analysis stages.

2. Document Pre-processing

The initial task was to clean the raw dataset to make it suitable for analysis. The following steps were taken in the build_terms() function to process the tweet content and hashtags separately:

Tokenization: The tweet text was split into individual tokens (words). This step was applied to both the tweet content and the hashtags to process each word individually.

Stop Words Removal: Using the NLTK library, common stop words were removed from both the tweet content and the hashtags to focus on more meaningful words. This reduced noise in the dataset.

Hashtag Handling:

- Hashtags were extracted from the tweet text while keeping the # symbol during extraction.
- Punctuation (except numbers) was removed from the hashtags.

- Stop words were filtered out, and the remaining words were stemmed using the Porter Stemmer algorithm. Finally, any empty strings resulting from the processing were removed.
- After extraction, hashtags were excluded from the tweet text to ensure they were handled separately during the analysis.

Tweet Content Handling:

- Emojis were removed from the tweet text using the *demoji* library, replacing them with spaces to avoid merging words.
- URLs were removed from the tweet content.
- Punctuation was removed, but numbers were retained.
- The tweet content was tokenized, removing individual words like "amp" (which appeared due to HTML encoding of & as &amp;).
- Stop words were removed, and the remaining words were stemmed to reduce them to their root forms.

Punctuation Removal: Punctuation marks were eliminated from both the tweet content and the hashtags, as they do not contribute to semantic meaning in most cases.

Stemming: The Porter Stemmer algorithm was applied to both the tweet content and the hashtags to reduce words to their root forms, grouping similar words (e.g., "running" and "run").

Date Conversion: The tweet dates were converted to pandas datetime objects to facilitate time-based analyses.

Emoji Removal: The *demoji* library was used to remove any emojis present in the tweet content, replacing them with spaces to prevent words from merging during further processing.

We generated **two dictionaries** to store the tweet data:

- **original_tweets_dict**: This dictionary contains the unprocessed data including the tweet text, date, hashtags, likes, retweets, and URL. This information is preserved as is, without any tokenization or processing, to be used when retrieving the full tweet content after a query.
- **processed_tweets_dict**: This dictionary contains the tokenized and processed version of the tweets. The tweet text and hashtags are processed (tokenized and stemmed), making this dictionary essential for building the inverted index and classification tasks in future stages of the project. Both dictionaries are linked via the document IDs to ensure consistency.

Example of both dictionaries, showing the output that will be provided after a query, and the dictionary after all the processing.

original_tweets_dict['doc_1']:

```
{'tweet': "#FarmersProtest #ModiIgnoringFarmersDeaths
#ModiDontSellFarmers @Kisanektamorcha Farmers constantly distroying
crops throughout India. Really, it's hearts breaking...we care about
our crops like our children. And govt. agriculture minister is
laughing on us🚜🌾WE WILL WIN💪 https://t.co/kLspngG9xE",
'hashtags': ['#FarmersProtest', '#ModiIgnoringFarmersDeaths',
'#ModiDontSellFarmers'], 'date': '2021-02-24T09:23:32+00:00',
'likes': 0, 'retweets': 0, 'url':
'https://twitter.com/PrdeepNain/status/1364506237451313155'}
```

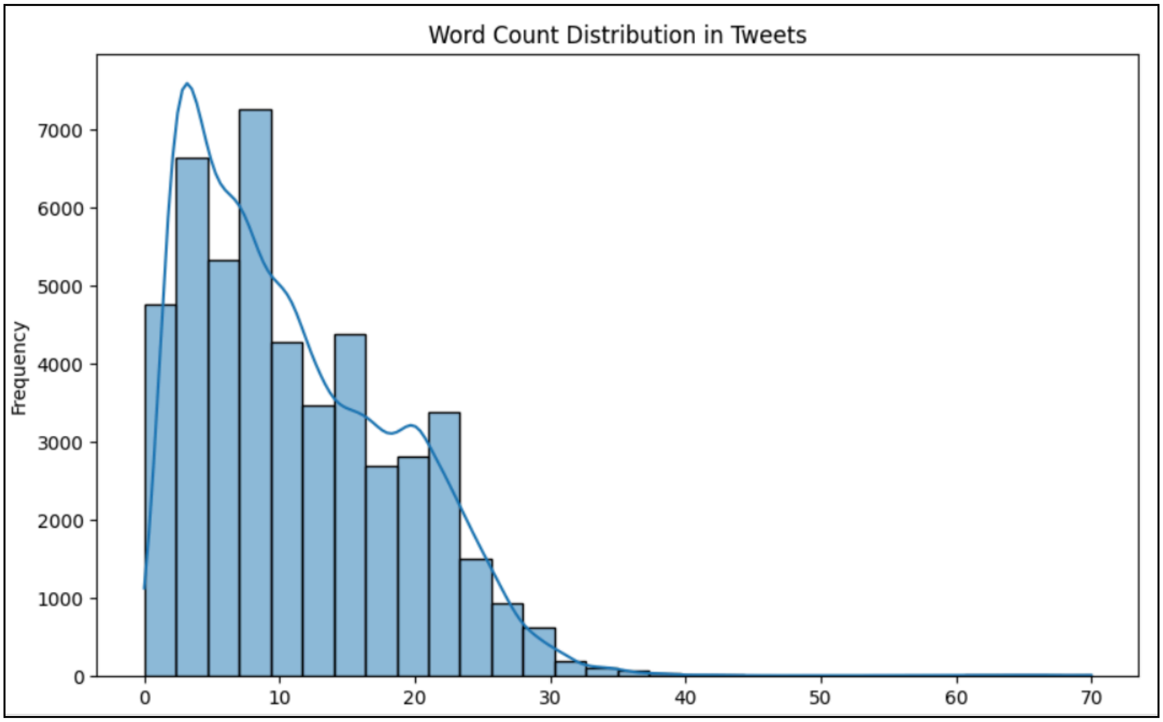
processed_tweets_dict['doc_1']:

```
{'tweet': ['kisanektamorcha', 'farmer', 'constantli', 'distroy',
'crop', 'throughout', 'india', 'realli', 'heart', 'breakingw',
'care', 'crop', 'like', 'children', 'govt', 'agricultur', 'minist',
'laugh', 'us', 'win'], 'hashtags': ['farmersprotest',
'modiignoringfarmersdeath', 'modidontsellfarm']}
```

3. Exploratory Data Analysis (EDA)

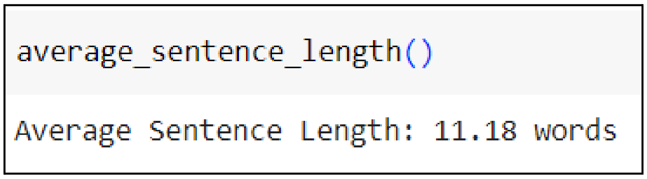
To explore and gain a deeper understanding of the dataset of tweets related to the Farmers Protests of 2021, which we extracted from the json file, we conducted various types of analysis on the data. Here is a brief description of all the analysis done:

Word Count Distribution in Tweets: A histogram showing the distribution of word counts across tweets. This facilitates our understanding of the common tweet length and allows us to analyze the average length of the tweets in the dataset.



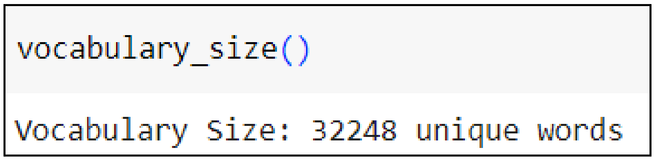
As we can see, word distribution follows somehow a poisson distribution. The most frequent word length is 9, and the average length is computed below.

Average Sentence Length: The code calculates the average number of words per tweet, providing a useful insight to understand the dataset.



```
average_sentence_length()

Average Sentence Length: 11.18 words
```

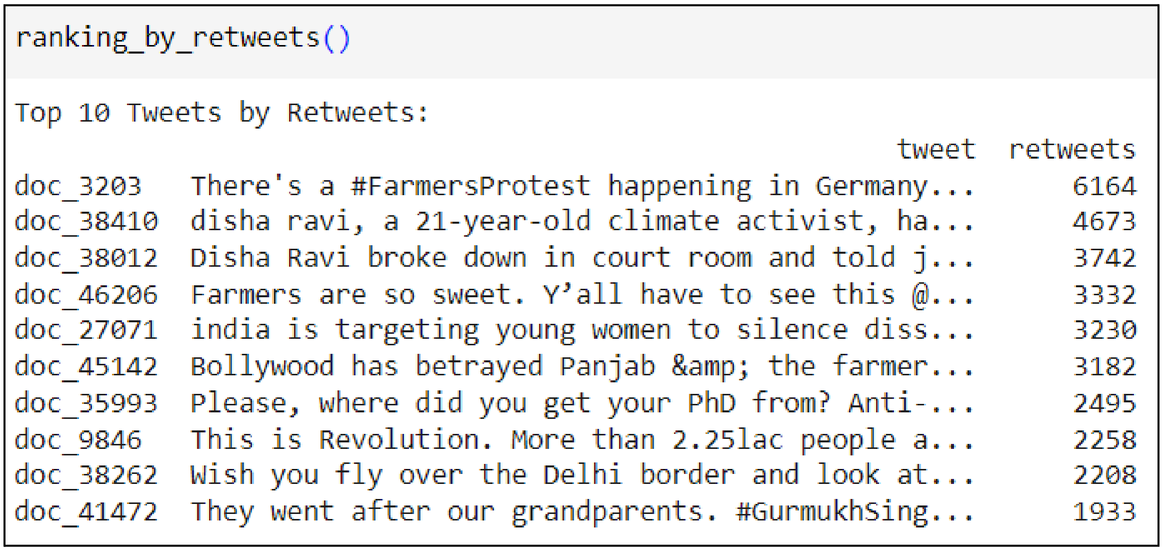
Vocabulary Size: The total number of unique words in the processed tweets. This gives an idea of the richness of the language used in the tweets.



```
vocabulary_size()

Vocabulary Size: 32248 unique words
```

English vocabulary has around 170.000 words, but taking into account that the analysis uses stemmed words, 32.248 words might represent a significantly larger number of original words before stemming.

Ranking of Tweets by Retweets: A table of the top 10 tweets ranked by retweet count, highlighting the most popular or influential tweets, and identified by their doc_id.

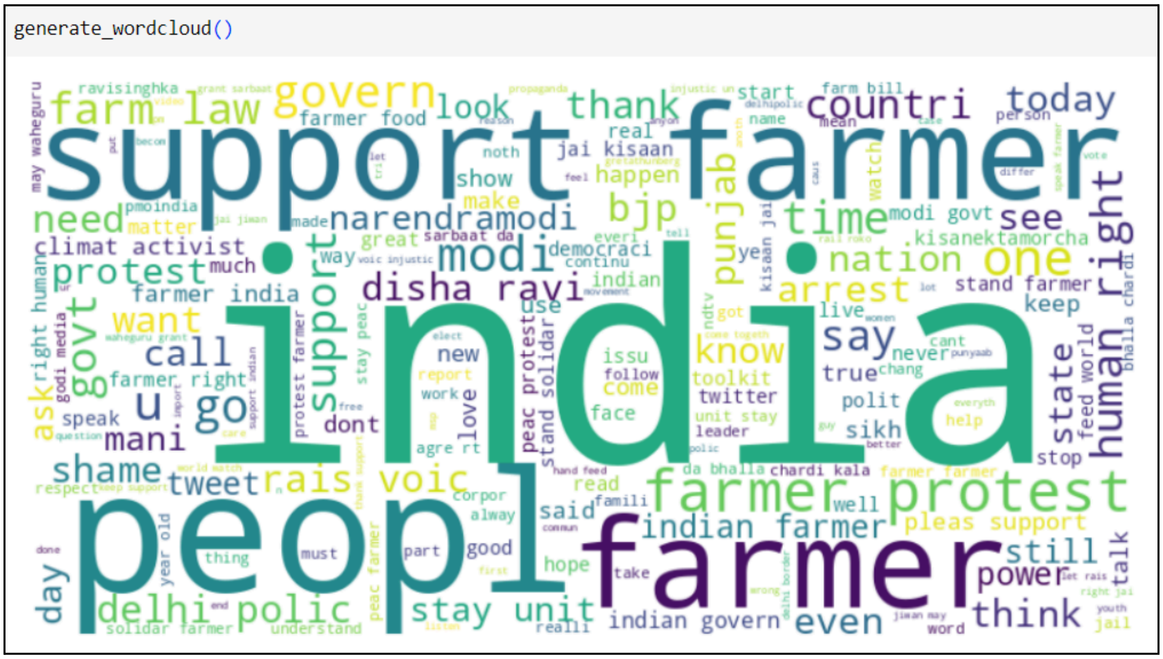


```
ranking_by_retweets()
```

```
Top 10 Tweets by Retweets:
                                                       tweet  retweets
doc_3203   There's a #FarmersProtest happening in Germany...      6164
doc_38410  disha ravi, a 21-year-old climate activist, ha...      4673
doc_38012  Disha Ravi broke down in court room and told j...      3742
doc_46206  Farmers are so sweet. Y’all have to see this @...      3332
doc_27071  india is targeting young women to silence diss...      3230
doc_45142  Bollywood has betrayed Panjab &amp; the farmer...      3182
doc_35993  Please, where did you get your PhD from? Anti-...      2495
doc_9846   This is Revolution. More than 2.25lac people a...      2258
doc_38262  Wish you fly over the Delhi border and look at...      2208
doc_41472  They went after our grandparents. #GurmukhSing...      1933
```

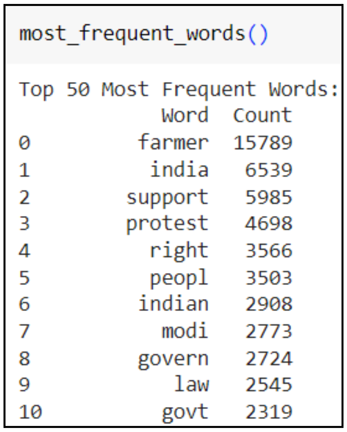
In thi case, influential tweets can be shown or ranked better than other tweets as they are considered more popular (they have reached more people). In that way, we can take into account the retweet count for our future search algorithms.

Word Cloud for Most Frequent Words: A word cloud representing the most frequently used words in the processed tweets.

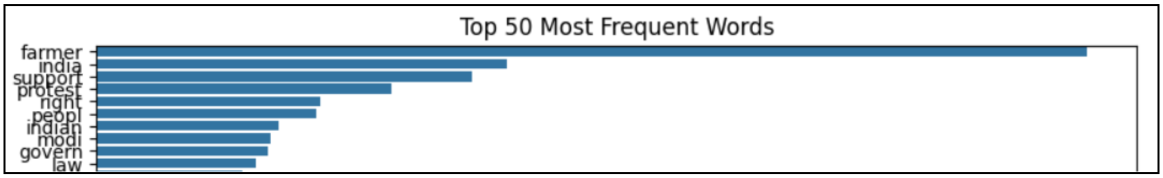


```
generate_wordcloud()
```

In the same way, we provide the top 50 words and their count; here are the top 10:



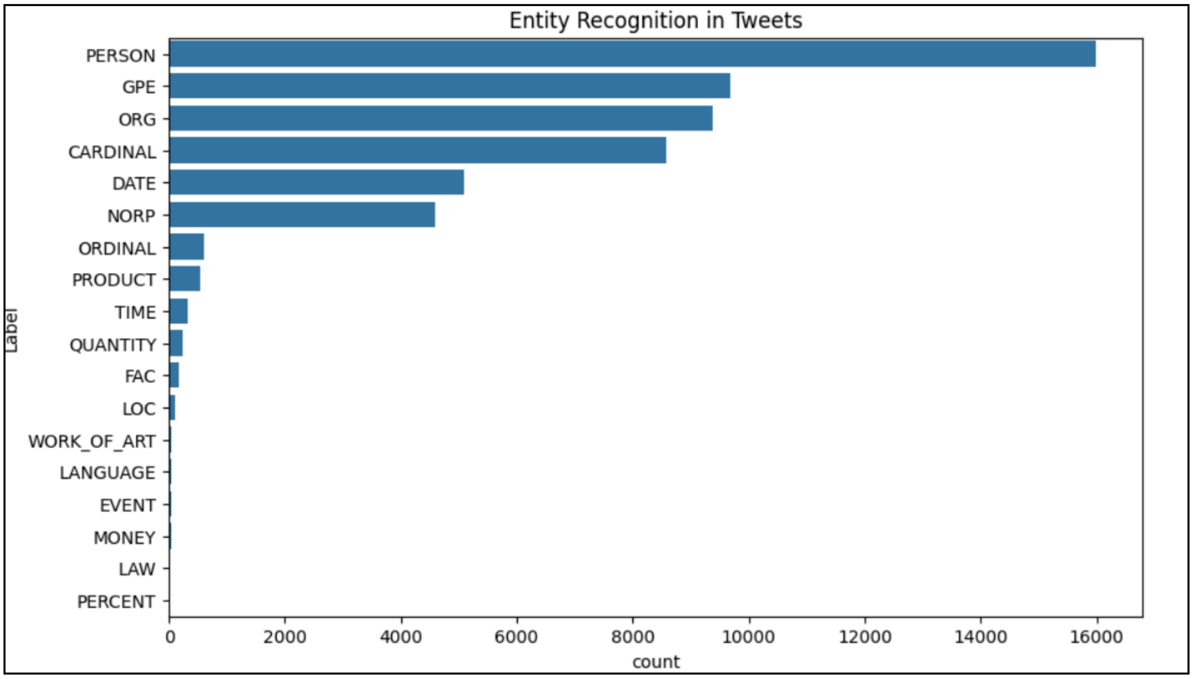
```
most_frequent_words()

Top 50 Most Frequent Words:
          Word  Count
0       farmer  15789
1        india   6539
2      support   5985
3      protest   4698
4        right   3566
5        peopl   3503
6       indian   2908
7         modi   2773
8       govern   2724
9          law   2545
10        govt   2319
```



We don’t have any surprises here, as the top words are farmer, India, support and protest, all related to the tweets that we are analyzing, the Farmers Protests of 2021 in India.

That is, word cloud shows visually the most frequent words in the whole dataset, and with the above table we can see its absolute frequency value.

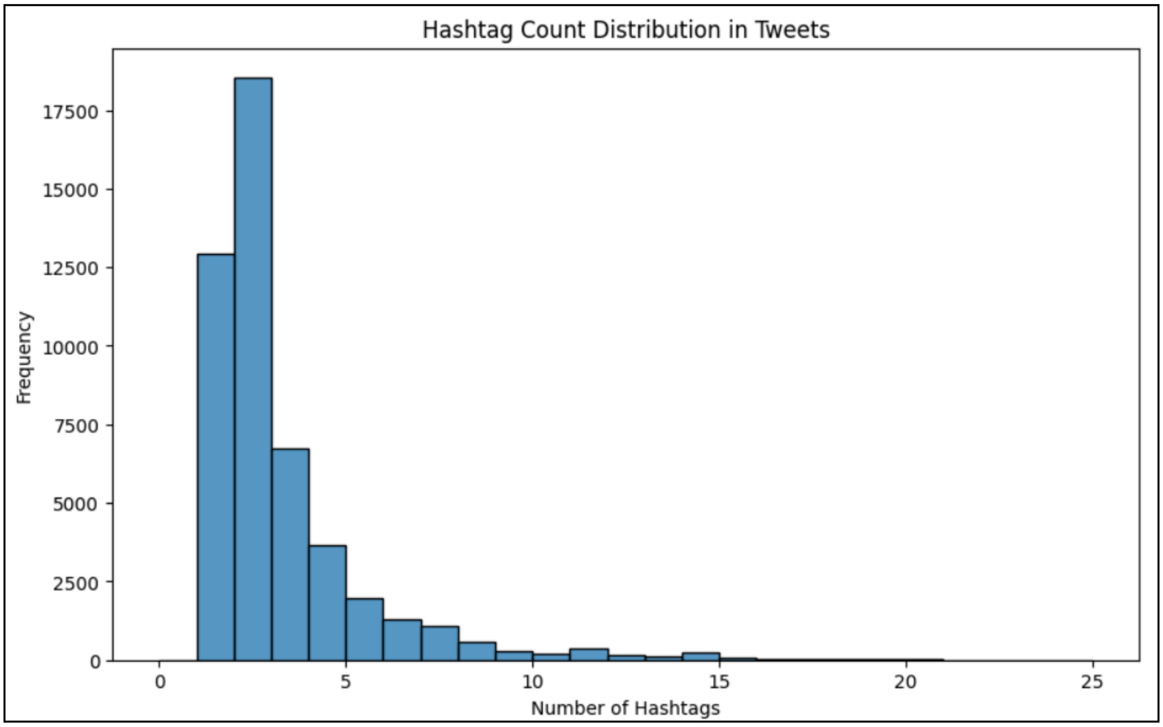
Entity Recognition: An algorithm that analyzes the words of the tweets and classifies them by an entity. Here is the the distribution by type:



The number one entity recognized in the tweets are persons. This could be expected, as normally these kind of protests are addressed to specific people, most likely being part of the government and organizations.

In the same way, GPE and ORG refer to Geopolitical Entity and Organization, respectively. These results suggest that the dataset is rich in references to individuals, locations, and organizations, indicating a focus on people (e.g., protesters) and their interactions (e.g., protests) with geopolitical regions (India) and institutional entities (e.g., government bodies).

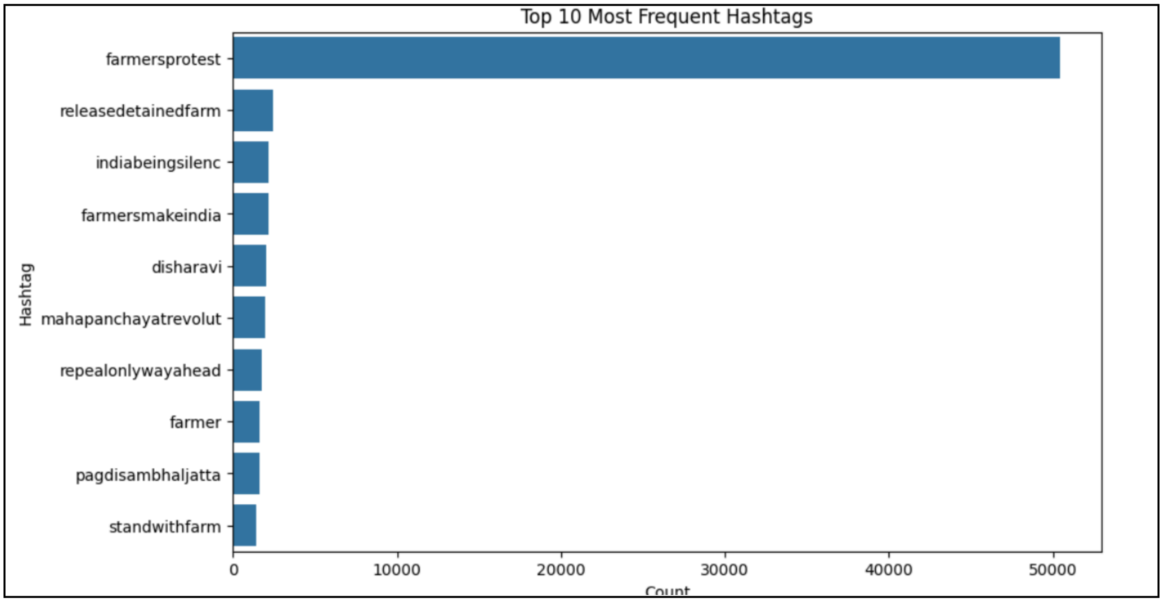
Hashtag Count Distribution: A histogram showing how many hashtags are used per tweet.



We observe that the most common number of hashtags per tweet is 3, followed by 2. This suggests that users (on average) do not use so many hashtags during their protests on Twitter.

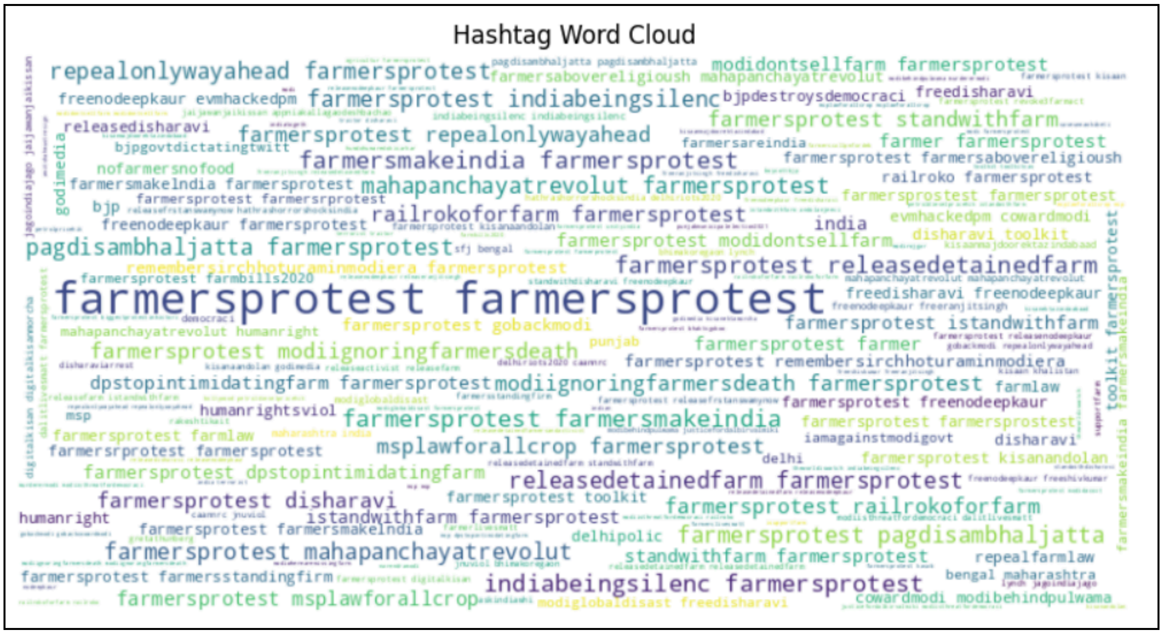
Now is time to perform some analysis on Hashtags:

Most Frequent Hashtags: A bar chart showing the top 10 most frequently used hashtags in the dataset.



It makes sense that the top hashtag is farmersprotest, as we are analysing the farmers protest in India in 2021. We need to take into account that hashtags are a union of words, so that it's more difficult to find occurrences / frequencies of the same hashtag, so that we found a more sparse list.

Hashtag Word Cloud: A word cloud visualizing the most frequently used hashtags.



The above word cloud shows us how sparse are the word frequencies when talking about hashtags. If we compare it with the first word cloud, we can see that more words are shown, and with less frequency each. Again, hashtags are less likely to be written in the same way, so that frequency decreases.

4. Conclusion

Through this pre-processing and exploratory analysis, we have cleaned the dataset and identified key patterns. The cleaned dataset is now ready for future tasks, such as building an inverted index and developing the recommender system. The EDA results give us a clearer understanding of the tweet content and user engagement, which will guide us in the next phases of the project.

Github Link: https://github.com/AlbertGarrell/IRWA-2024-u198736-u198740-u199896

Github Tag: https://github.com/AlbertGarrell/IRWA-2024-u198736-u198740-u199896/releases/tag/IRWA-2024-part-1