**Part 1: Text Processing and Exploratory Data Analysis**

**Introduction**

In the first part of our project, the goal is to understand the data given and process it accordingly to our objective. In this report, it is explained the decisions made for pre-processing the documents, the assumptions made and the information found about the Russia-Ukraine war tweets dataset.

**Documents pre-processing**

On one hand, when reading the tweets from the json file, we decided to keep only the important information of each element. That is, we assumed that for future queries the final output must return only the Tweet | Date | Hashtags | Likes | Retweets | Url for selected documents, and we programmed our code accordingly.

On the other hand, we implemented the funcion *build\_terms()* to remove URLs from the text using a regular expression, convert the text to lowercase, remove punctuation marks, tokenize the terms, remove common stop words, special characters, emojis and reducing words to their root or base forms using the Porter Stemmer algorithm.

In addition, we decided to remove the ‘@’ symbols but keep the username that followed it. This action can be useful to track and analyze specific interactions between users. Furthermore, we decided to remove the character ‘#’ but keep the words starting with it. This choice is made to facilitate the indexing that will be implemented in the following parts of the project and keep information that will be useful for the search engine.

Finally, we changed the date original structure to a pandas datetime object, which will enable more precise temporal analysis and ordering of the data when working with python.

**Documents and Tweets mapping**

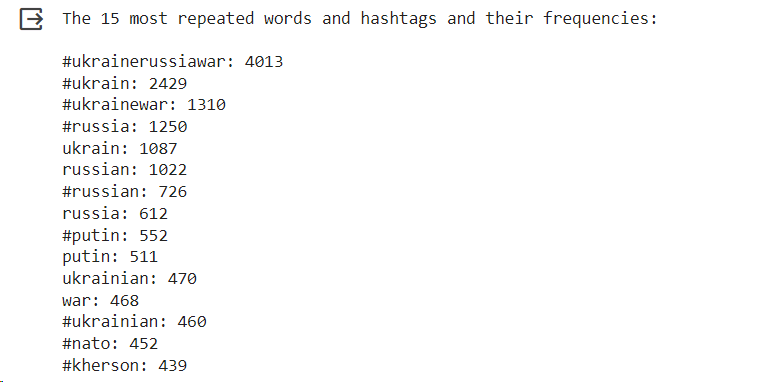
In the notebook we also have implemented the data frame *tweet\_document\_ids\_map* using the provided *Rus\_Ukr\_war\_data\_ids.csv* file. This data structure maps the tweet’s ids with the document ids, keeping all the necessary tweet information and to be considered in the future for the evaluation stage of the project.

**Exploratory Data Analysis**

To explore and gain a deeper understanding of the dataset of tweets related to the Russian-Ukrainian war, which we extracted from the json file, we conducted various types of analysis on the data.

Our initial step involved a date analysis to know when the tweets had been posted. We found that all of them had been tweeted from the 28th of September of 2022 to the 30th of September of the same year. This will help us when looking for more context about the most frequent words and hashtags used.

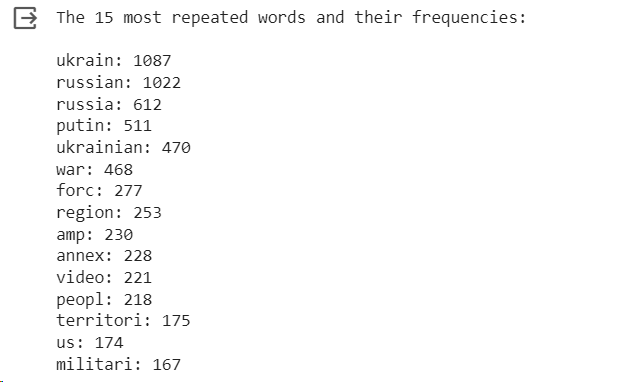
Next, we performed a word count analysis to identify which were the most frequently occurring words and hashtags within the tweets. We obtained the following results:



As expected, the most frequently used term is the hashtag “#ukrainerussiawar”, one of the most popular labels for the war that combines the names of the two countries at the center of the conflict. The terms “#ukrain” ,“#russia”, “russia”, “ukrain” and “#ukrainewar” closely follow behind, which also clearly emphasize the two most mainly countries involved. Additionally, terms like “#putin”, “putin” and “russian” could reflect a focus on the individuals involved in the conflict, specifically on the Russian President Vladimir Putin.

While, terms like “nato” could suggest that there are discussions about Ukraine's international alliances and/or its possible intervention in the war. And, the term “#kherson”, which is an Ukrainian province and city, reveals a focus on a specific area of the ongoing conflict. We did some research and we found that on the 30th of September of 2022, which is one of the dates when the tweets were posted, Russia claimed to have annexed the Kherson oblast/province.

Additionally, we redid the word count analysis to only identify the most frequently used word (we excluded the hashtags), to get a better understanding of the discussions revolving around the conflict. We obtained the following result:



The most frequently repeated words, including “ukraine”, “russian”, “russia”, and “putin”, represent the main actors in the ongoing conflict. While other terms like “force”, “region”, “territory”, “military” and “annex” provide important contextual clues, such as the use of force and the annexation of Ukrainian regions to Russia, like Kherson.

Also, the frequent use of the word “video” could indicate to us that users are sharing and discussing videos related to the war. This might include citizen-shot footage of the events or news reports.

We also found that, on average, tweets are quite lengthy, with an average word count of 17.46 words and an average character count of 129.26. Additionally, the corpus contains a wide vocabulary with 6506 unique words.

Additionally, we ranked the most liked posts and we obtained the following results:

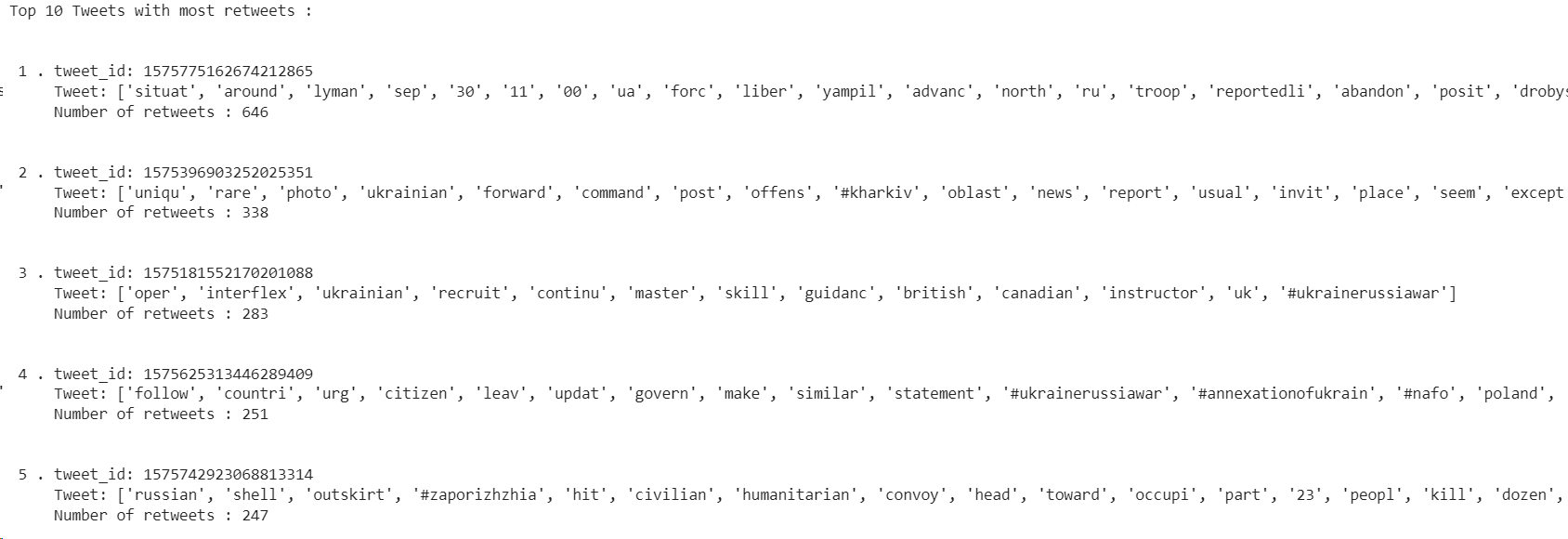


We can observe that the most liked tweet is likely, considering the word list, a news report about the situation around Lyman and the Ukrainian forces' advance after the liberation of Drobysheva on the 30th of September of 2022.

The second tweet featured a rare photo from a Ukrainian command post in Kharkiv while the third tweet discussed the ongoing training of Ukrainian recruits with the guidance of British and Canadian instructors, underlining the international support for Ukraine.

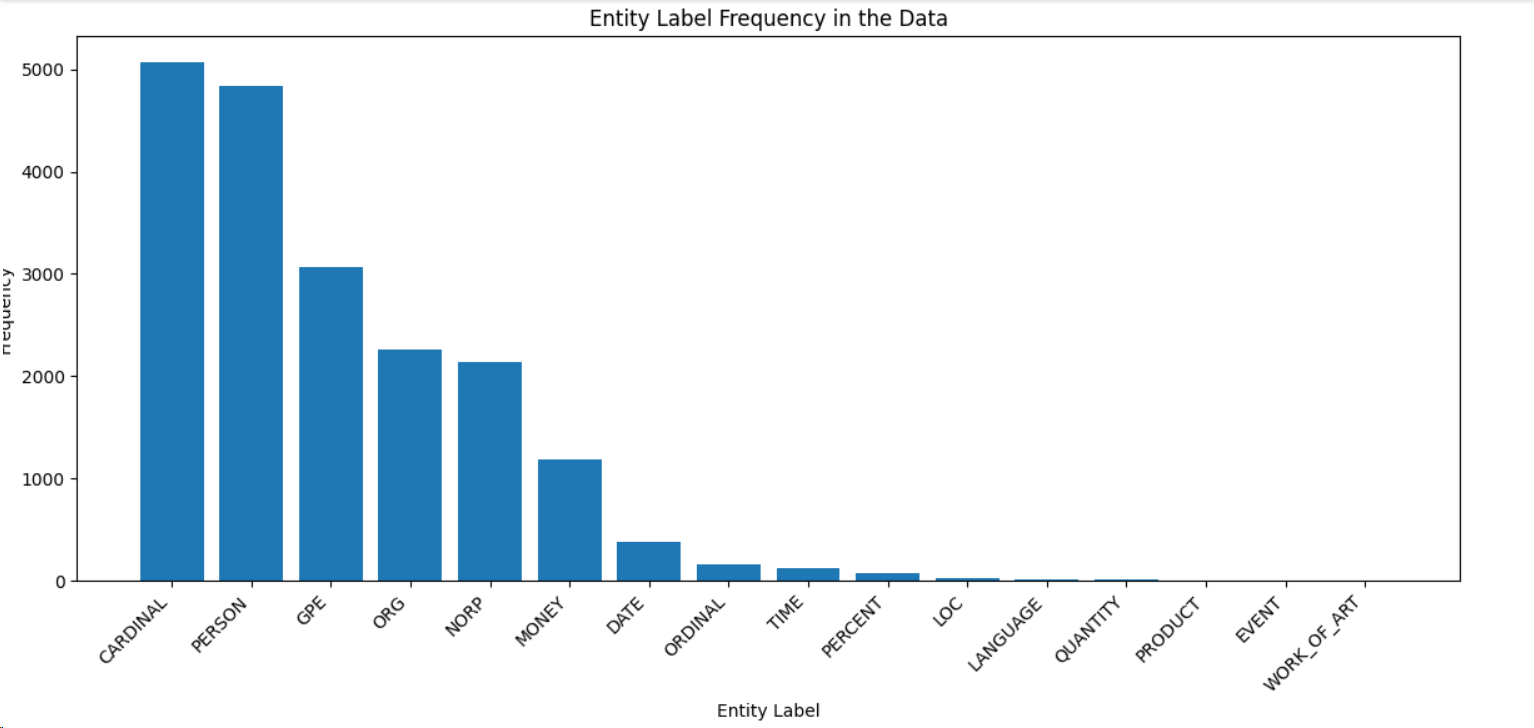
In conclusion, the tweets in the top 10 list reflect various aspects of the conflict, from military operations and strategies to international involvement and humanitarian concerns.

Then, we also ranked the most retweeted posts and we obtained the following results:



We can observe that the ranking very closely resembles the previous one, as it includes in the top 3 the same tweets as before. Also, the content of the tweets is very similar as they also reflect various aspects of the war, from military operations to international involvement and humanitarian concerns.

At last, we conducted an entity recognition analysis of the tweets:



We found 1540 unique entities and that the most prominent entity labels detected include individual names (PERSON), geopolitical entities (GPE), and numerical values (CARDINAL). This could reflect that users are discussing people, numbers, regions or territories involved in the conflict.

Additionally, mentions of organizations (ORG) and references to nationalities, religious, or political groups (NORP) tells us that the discussions about the war are quite complex and involve different groups and organizations.

**GitHub link**

[**https://github.com/PatriciaGaray/IRWA-2023-u242781-u192945-u190355**](https://github.com/PatriciaGaray/IRWA-2023-u242781-u192945-u190355)

TAG: IRWA Project Part 1