Estellers, Oriol - 242142
Fuentes, Raimon - 242176
Ribas, Pol - 241620
Github link and TAG: IRWA-2023-part-1

Part 1: Text Processing and Exploratory Data Analysis

After setting up the environment for processing and analyzing a document corpus consisting on a set of tweets related to the Russo-Ukrainian War, we have started by implementing the following features:

- **Lowercasing:** All characters of every word have been turned into lowercase, which will result in a higher precision when performing searches and retrievals.

- **Removement of punctuation marks:** We have removed all punctuation marks, except for the #, @ and ‘, as the two firsts give relevant information about the hashtags and mentions, respectively, which will be important for further analysis. On the other hand, the third punctuation mark is present in contractions, which are very common. In other words, their elimination would lead to having a wide range of unmeaningful words, such as *weve*, *cant*, *Im*,... among others.

- **Tokenization:** By tokenizing, we have splitted the tweets into a list of terms. Instead of using the function “split”, like we did in the first practice session, we have decided to use a function (TweetTokenizer(...)) that is specifically designed to handle unconventional grammar such as abbreviations, hashtags and emoticons. We have done so because in social media, non-standard language constructions are usually common, so working with a library that is specifically designed with this purpose will give us more coherent results.

- **Emoji elimination:** Using the emoji library, we have removed all emojis in the tweets to obtain a standardized corpus. People don’t use emojis when searching for documents, so it’s senseless not to remove them when processing the tweets.

- **Elimination of stop words:** We have removed words that carry very little useful information due their highly usage. That is, words such as *a*, *the*, *is*, *are*,... By doing this, we can focus on the more significant words and phrases. Moreover, by reducing the amount of words we analyze, we improve efficiency and reduce dimensionality, as there are less computations and less resources required.

- **Stemming:** On the same line, we have reduced words to their root or base form by stemming them. We obtain less variations, so we "standardize" the vocabulary. Additionally, we also obtain a more precise search engine, as with only a word we can obtain more results, as it will take into consideration its different forms.

- **Number transformation:** We have transformed all numerical characters to cardinal words, such as substituting a 2 by the word *two*, so tweets only contain alphabetic characters (except for the ones that we haven't removed voluntarily). It will enhance the results given by the search given.

All these transformations will be applied to the dataset and, afterwards, the results will be stored in a new column of the dataframe called *normalized*.

For the exploratory data analysis section, we have decided to study the following aspects:

- **Average number of words per tweet:** In order to compute the average number of words per tweet, we have counted how many words each tweet has in the normalized version, as we want to focus on the relevant words that give us information. Using the numpy library, we conclude that, on average, each tweet is composed of nearly eighteen words, with a standard deviation of seven point six words. With this information, the indexing structures and storage mechanisms can be optimized, as we won't be working with a very long corpus.

- **Top-5 hashtags:** The five most used hashtags are {'#ukrainerussiawar': 4.012 times, '#ukrain': 2.163 times, '#ukrainewar': 1.326 times, '#russia': 1.232 times, '#russian': 724 times}. They are all generic, we can't extract any specific information about the writer's point of view. We have gone over all normalized tweets searching for the tweets and storing their number of occurrences in a dictionary.

➢ **Top-50 most used words with Word Cloud:** The word cloud that we obtain when plotting the top-50 most used words among the entire (normalized) vocabulary are:

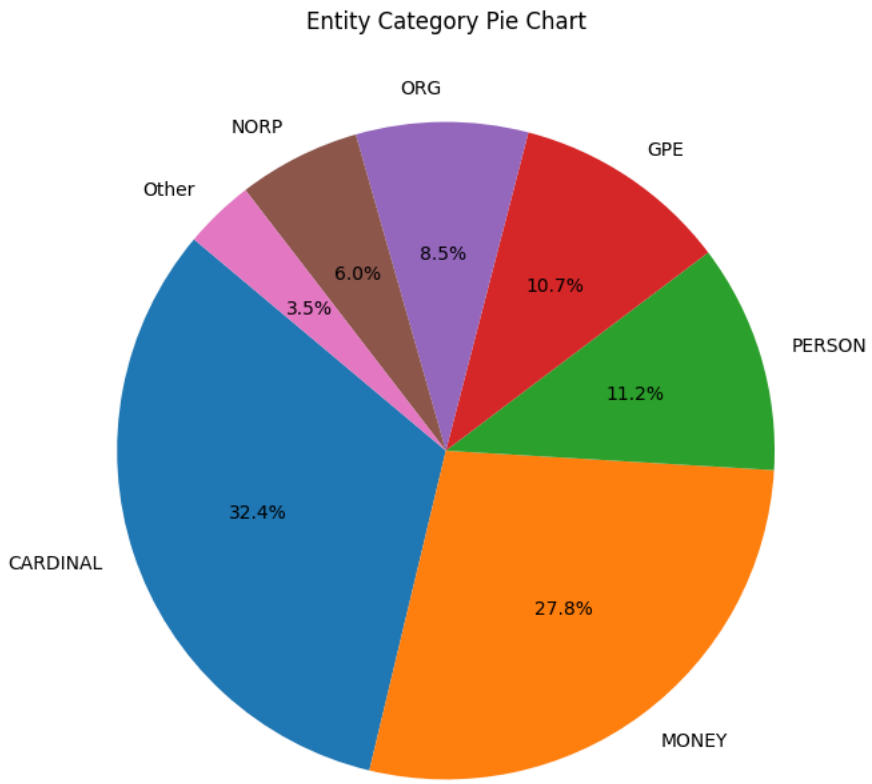
At first glance, we see that most words are either “russia”, “ukraine”, “war”, the combination of these three or political leaders' names, such as Zelenski or Putin. Other relevant words used related to the war, there aren’t any words that can’t be linked to the theme, meaning that we have successfully removed stop words.

We have used the WordCloud library, as it already has the implementation of the graphic we are looking for.

➢ **Entity recognition:** We wanted to see if we could extract any conclusions by looking at the entities of each word. To do that we have downloaded an english language model which has information about the entities of each word in the language. After that we have searched for the entity of each word of our tweets, and we got the following result:

We can see that the most frequent entity is cardinal, which makes sense since we are talking about different events happening through time, which also explains why there are so many dates. We can also observe that the second most important entity is money, which is obvious because a conflict like this has a lot of economical consequences. The entity person also appears in this list as well as GPE, since a lot of tweets are about the people and locations involved in the conflict. The entities

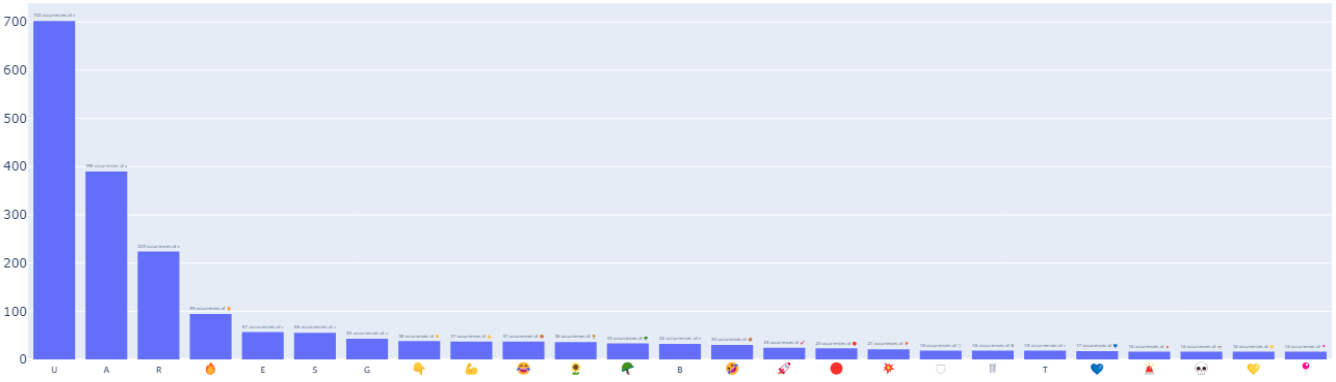
ORG and NORP are also in the list since there are a lot of organizations, religions and political parties involved in or affected by the conflict.



➢ **Vocabulary size:** Studying the vocabulary size can also be a good way to explore the data. To do so, what we have done is iterate through all the words in each tweet and store in a dictionary all the unique words appearing. The results obtained are that the number of distinct words in the tweets' dataset is 13581 words. Since we know that the dataset is formed by 4000 tweets, we can compute that the average number of unique words per tweet is 3.4. We can relate this information to the one obtained before when we calculated the average number of words per tweet analysis, which was 17.9. Therefore, when comparing these two we observe that the percentage of unique words used per tweet is 18.97%. Since the percentage of unique words per tweet seems very low, we deduce the tweets usually had similar words with each other or at least words that someone had used before. This appears to be reasonable if we take into account that the tweets were all related to the same subject, which was the conflict of Ukraine and Russia, and therefore people usually talked about the same things and that is why a lot of words will be repeated.

For this analysis we have used the normalized tweets because if we had used the original ones, we would have probably obtained even a lower percentage of unique words per tweet, since a lot of stop words would have been taken into account.

➢ **Histogram of emojis used:** We have decided to plot a descending histogram of the emojis used. A relevant insight is that out of the top-7 most used emojis, 6 (including the first three) are letters.



To generate this histogram, what we've done is, using regex, search in every tweet (non-processed version, as in the processed the emojis are deleted) all the emojis it contains. If the tweet contains any emoji, it's added to another array, which will be converted into a unique text. Afterwards, we look for all emojis in the text, store them and count the number of occurrences for each of them. Finally, we plot the top-25 using a dynamic histogram.

We wanted to see if they were related in some way to the conflict. We can observe that the most used ones are letters, which are not related specifically to the war since they can be used in all kinds of tweets. However we can see that emojis like helmets and explosions are very used, which makes sense because they are war-related.

- **Top-5 most mentioned countries:** To achieve that we first defined a list with all of the countries in the world. After that, we iterated through every unnormalized tweet and every word in the dataframe and checked which countries were mentioned, and how many times. We did not use the normalized text since the list of countries that we use to identify them has capital letters and the normalized tweet doesn't.

  Since we have a dataset of tweets about a political conflict, we thought it would be interesting to see which countries were mentioned more frequently. The result was the following: {'Ukraine': 537, 'Russia': 323, 'Germany': 25, 'Belarus': 22, 'Finland': 13}.

  Ukraine and Russia are obviously there because they are the "protagonists" of the conflict. Germany may be there since it may be one of the most affected big countries economically by the war. Belarus may be mentioned a lot because it is a bordering country with both Ukraine and Russia, and it also has a good relationship with the latter. And finally Finland is probably in a lot of tweets since they wanted to enter the OTAN and Russia was against that.

- **Top-5 most active users:** To do that we have first identified the user of each tweet by looking for the '@' character. Then we have counted the number of tweets by each user and output the top 5 with the most tweets.

  We wanted to see which users posted more tweets about the Russia-Ukraine conflict. These 5 accounts are the following: {'@zelenskyyua': 44, '@youtub': 40, '@potu': 31, '@eucommiss': 22, '@mfarussia': 22}. We can see that some of the users have names related to the conflict like *eucommiss* or *mfarussia*.