Part 1: Text Processing and Exploratory Data Analysis

1. Pre-process

In this section of the practice, we have preprocessed the document of tweets discussing the Ukraine-Russia war. To do this, we have performed the processes as instructed in the prompt, and then we have added 5 more processes to ensure that the tweets are as desired. These additional 5 preprocessing steps are as follows:

- If there is an extra line added to any tweet, it is removed. In other words, we eliminate line breaks to keep everything on a single line.
- We remove double spaces, if any, due to typing errors.
- We removed leading and trailing spaces before and after each tweet, in case any users had added them.
- We eliminated links to web pages. Why? After careful consideration and observing how we obtain the tweets once they have been converted to lowercase and punctuation marks have been removed, we noticed that it is not possible to access these links. Therefore, they do not provide us with any valuable information.
- We modified the date format to make it much easier to work with. This way, in the next section, we will be able to analyze the data based on the dates.

Additionally, another point to mention is that we have decided not to remove the # and @ symbols from the tweets. We believe this is valuable information to retain, as it allows us to clearly identify the hashtags present in the tweet and the usernames. These users will also be useful for analyzing the data we have.

2. Exploratory Data Analysis

To carry out this section, we have followed the examples provided in the prompt. To do so, we have structured our analysis based on the column we are going to examine.

- Tweet: We wanted to analyze which words are the most common in the set of all tweets. To do this, we have created a word cloud. In it, you can see that the most frequently repeated words often include the words "Russia" and/or "Ukraine." Actually, the analysis that surprised us the most was that there are more words that appear in the cloud about “Ukraine” than “Russia”.
  Furthermore, with the usernames we retained earlier, we have analyzed which ones are mentioned the most, displaying the top 10 most mentioned users. With this analysis, we saw that the president of Ukraine was the most mentioned user, which is kind of surprising because we don’t find in the top 10 the president of Russia. Additionally, we got amazed by the fact that the 2nd top mentioned user was Youtube.
- Date: For this variable, what we wanted to determine was whether there were any tweets from before the start of the war or from the same day. In our study,

we observed that there are none. So, all the tweets given were posterior to this date.

- Hashtags: Another question we were interested in was whether more *Hashtags* were associated with more *Likes* and/or *Retweets*. To investigate this, we believed that the best way to determine this result was by creating a correlation matrix. In this matrix, we can see that there is very little correlation between *Hashtags* and the other two variables; they are not correlated. However, there is a positive correlation between *Likes* and *Retweets*. The conclusion of this analysis is that the more *Likes* you get, the more *Retweets* you will get. Similarly, having more *Hashtags* doesn't imply that you get more *Likes* or *Retweets*.
- Likes: For this variable, first and foremost, we obtained a description of it, including its mean, median, maximum, and so on. Next, we created a histogram to visualize the distribution it follows. However, we found that this was not the best idea since the maximum value was 3701 likes, making it difficult to observe the distribution properly. Therefore, we created a box plot where we could appreciate the outliers and the information described in the initial study. Finally, we analyzed the top 10 tweets with the highest number of *Likes*.
- Retweets: The studies we used for this variable are the same as the previous one, with the difference that here we also examined another case. For the identical studies, we encountered a similar problem with the histogram because the maximum number of *Retweets* was also far from the mean. Therefore, the distribution is not very clear. However, from the data, we can see that the majority of tweets have 0 *Retweets*.

  Regarding the top 10 tweets with the highest number of *Retweets*, we wanted to determine if they were the same as in the previous variable. In this analysis, as the correlation matrix had anticipated, we found that there is a set of tweets that make up the top 10 for both *Likes* and *Retweets*.

  The additional study involved checking whether tweets containing the word "rt" or "retweet" had a higher number of *Retweets* or not. We did it in order to check whether there was a significant impact on writing things like “rt if you agree” or “retweet if you agree”, that sort of things, on the number of *Retweets* or not. In this analysis, we observed that this strategy had no impact for the tweets observed/given, and the number of *Retweets* for all cases is very low.

As part of our analysis, we observed that having outliers or very high values for the variables *Retweets* and *Likes* can be a challenge, as they represent rare cases within this dataset. In future experiments, it might be beneficial to assign a smaller weight to these outliers or consider removing them to avoid skewing the overall analysis.