Joel Duran, Èric Dalmases and Marc Aguilar

Report | Project Part 1

The project has been published on GitHub at the following URL and is available within this specific tag:

URL: https://github.com/ericdalmases/IRWA.git
Tag: IRWA-2023-part-1
(https://github.com/ericdalmases/IRWA/releases/tag/IRWA-2023-part-1)

Data Processing

We started the first deliverable of the project by defining a class to represent each individual Tweet. This will be useful in order to keep our code clean and easily modify the fields each tweet contains if needed. We have added some extra fields such as the user_id or the username since it can become useful for the development of the project.

The Tweet class includes a method named fromJson which facilitates the creation of tweet objects from JSON data. This method parses aforementioned fields from the provided JSON data and instantiates a Tweet object.

When a new tweet instance is created the _processTweetText method performs processing of the tweet content. In order to do so we applied the techniques we saw in the first lab with some little modifications that we have decided given the context of the data.

First of all we turn the tweet text to lowercase. Then, we thought that leaving both mentions and hashtags would be useful when doing the data analysis, as they normally either describe the general topic of the tweet or tag an important person that was involved in the conflict. Therefore, we keep both of them and we remove the URLs since they are not relevant. Although we delete almost all icons and symbols, we leave the most relevant flags (such as 🇺🇸 or 🇪🇺) which can be very impactful when performing queries due to the conflict context. To finish with the processing we tokenize the text by splitting by blank spaces, we remove the stop words and we perform stemming.

Then we basically load all the tweets from the JSON file to its corresponding Tweet instance and we store them all in a list of tweets.

After finishing this process and evaluating the results we noticed that there were words such as “dictator” or “dictation” that became shortened after the stemming part to “dictat”. Although this can lead to confusion in the data analysis process, it will not affect when performing the search since both queries and tweet contents will go through the same preprocessing pipeline.

Data Analysis

First we decided to extract some basic statistics after turning the list of tweets into a dataframe and building the vocabulary dictionary. We observed that there are **4000 tweets** and **8684 unique tokens**.

Most frequent terms

To start with the analysis, we considered that knowing which are the most repeated words could be useful to better understand the data. The best way to visualize those words are the word clouds, so we decided to do different plots depending on some conditions.

First we wanted to know which were the most frequent words before performing any kind of processing. As it can be observed in *Figure 1*, the most frequent terms are "https", "t", "co" which basically are the ones used in urls. This was a clear indication that the text processing phase would be crucial.

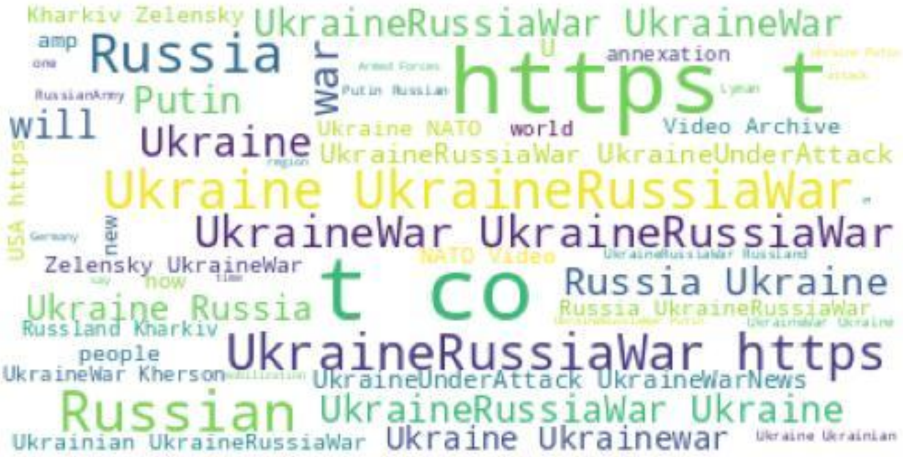

Figure 1

In *Figure 2* we can observe a total different word cloud with the most frequent terms after performing the processing. Here we can observe the terms "fascist", "russia", "wrong", "dictat", "putin" as we commented above or even mentions such as "@melsimmons". We can observe how the majority of the terms are related to the conflict itself. There appear some hashtags but with less frequency, so we decided to do a word cloud only with hashtags.

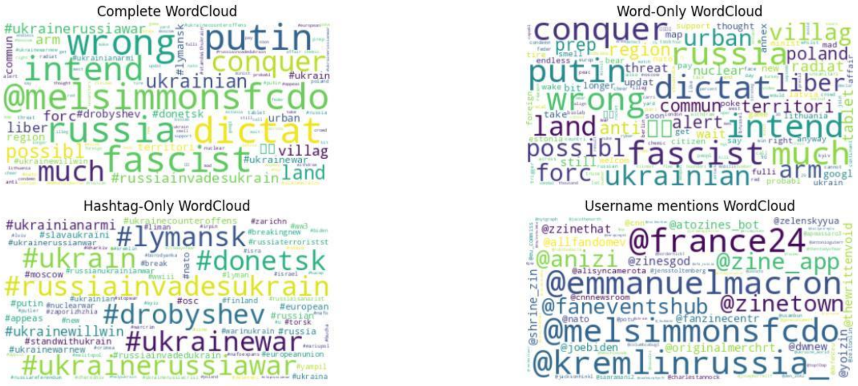

Figure 2 (Top-Left), Figure 3 (Bottom-Left), Figure 4 (Bottom-Right), Figure 5 (Top-Right)

As it can be observed in *Figure 3,* the most frequent hashtags are *"russiainvadesukrain", "lymansk", "donetsk", "ukrainewar"* and *"ukrainerussiawar".* All the ones which can be read are related with the conflict itself. *Figure 4* illustrates the most frequently mentioned users in the tweets. Among them, notable figures such as *"melsimmonsfcdo," "emmanuelmacron," "kremlinrussia,"* and *"france24"* stand out, with the first three being politicians and the last being a news channel.

The distribution of our data

To see how our data was distributed we started by plotting the number of words the tweets contained, Figure 6. As it can be observed, it follows very closely a normal distribution with an average number of words per tweet of 25 words.

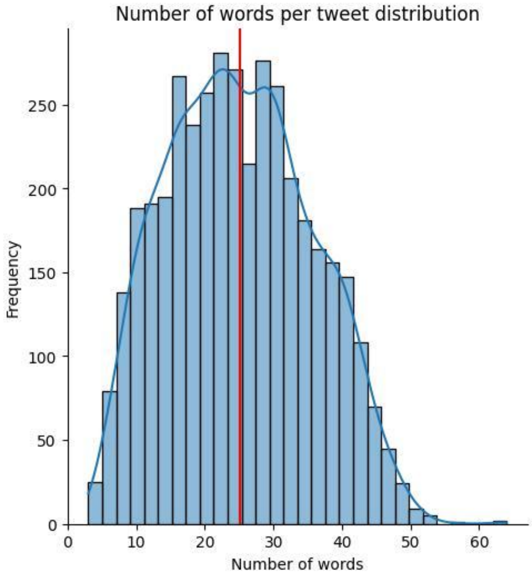

Figure 6

Then we plotted the distribution of the number of tokens, Figure 7. This is similar to the previous one since it also follows a normal distribution but with a slightly lower average of 18 tokens per tweet. We can conclude that our processing pipeline has successfully cleaned our tweets.

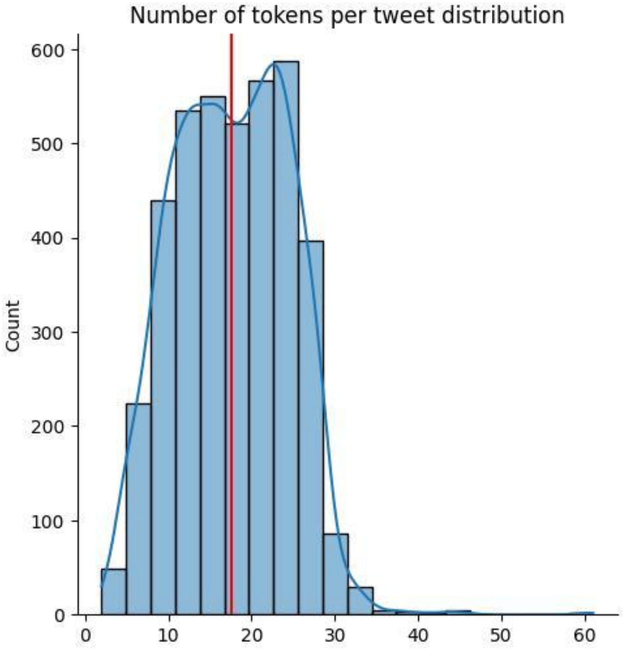

Figure 7

Subsequently, we plotted the distribution of the number of characters, Figure 9. While this plot doesn't provide highly informative insights, we can observe that there is a strange peak around 280 characters. This could be caused by the maximum character limit of tweets, which was set at 280 characters.

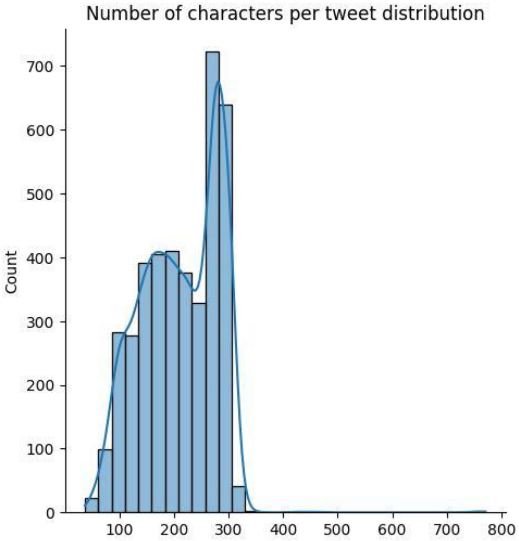

Figure 8

Finally, we wanted to explore the distribution of the number of hashtags per tweet, Figure 9. We can observe that it more or less follows a Poisson distribution with its peak around 5 hashtags per tweet.

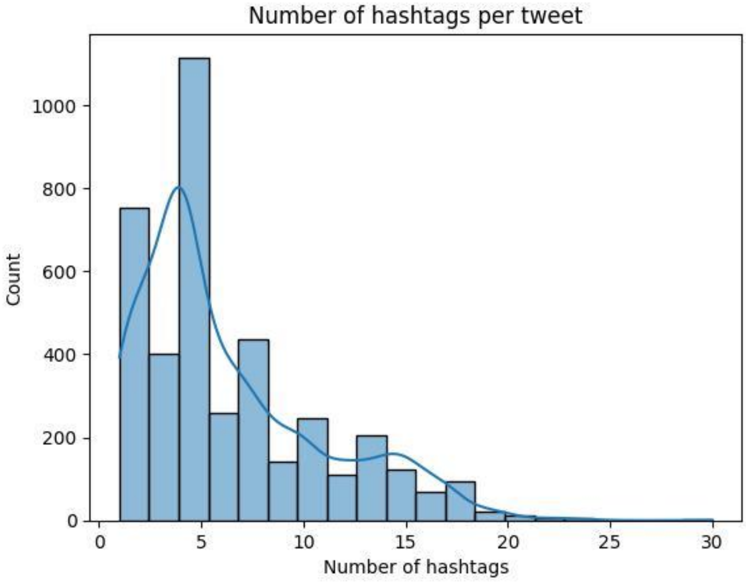

Figure 9

Correlations of our data

After exploring some distributions of our data, we wanted to know if there existed any kind of correlation between some variables. In Figure 10, we plotted the correlation matrix with the variables *id, user_id, likes and retweets*. As expected, user_id and id do not have any correlation with any of the variables but, likes and retweets are very correlated which means that if one of the variables change, the other will change at an approximate constant rate.

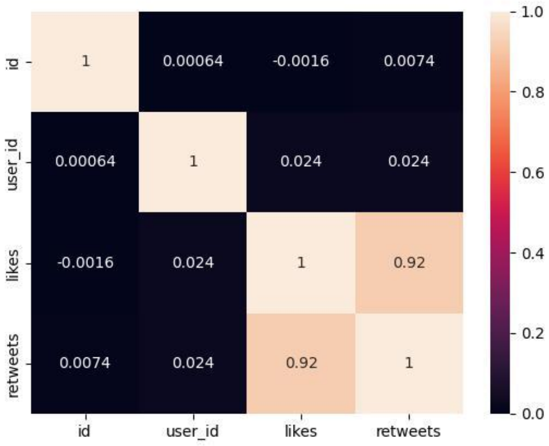

Figure 10

In Figure 11, we can take a closer look into this correlation. The line represents the linear relationship between both variables, while the blue shading denotes the extent of variability.

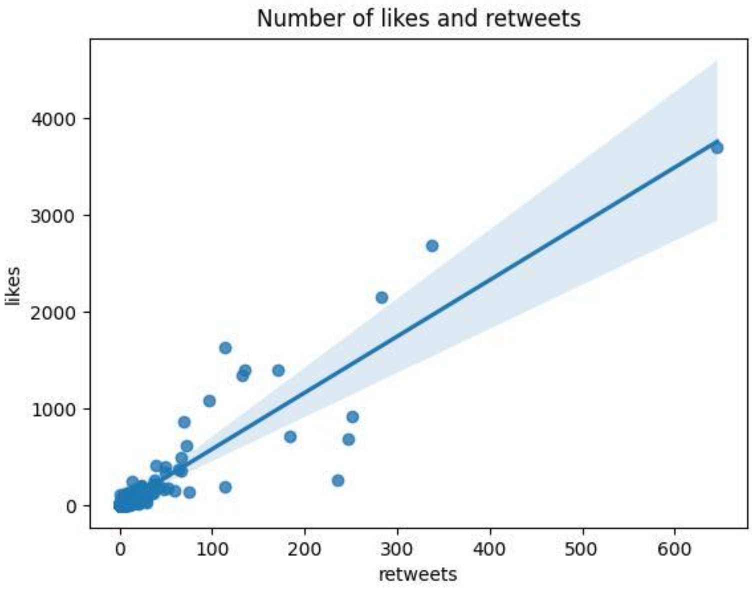

Figure 11

General Analytics

In Figure 12, we wanted to know the users who tweeted the most. It can be seen that it follows a Geometric distribution. Which means that there are few users that tweet a lot and the rest follows a constant number of tweets posted.

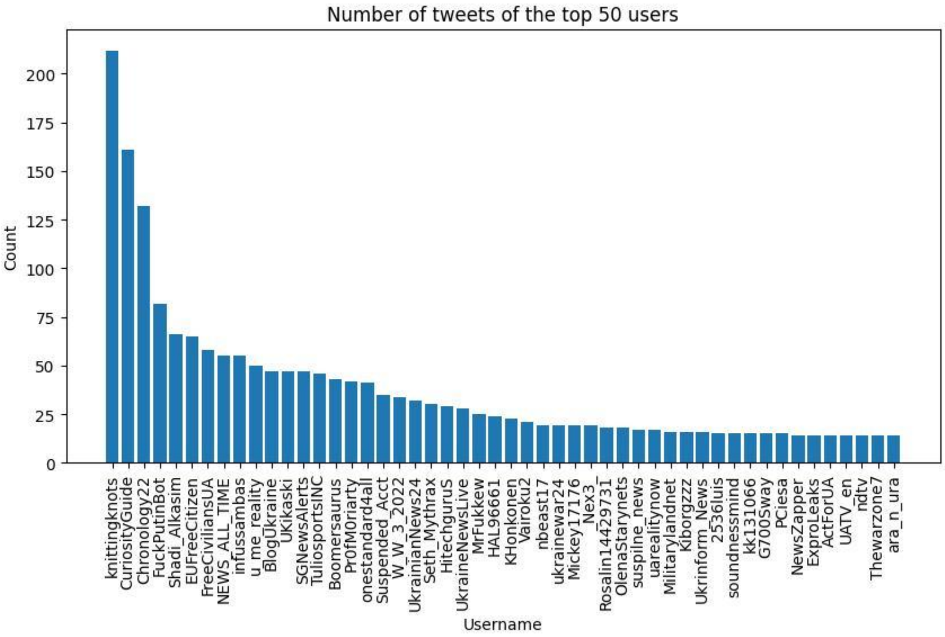

Figure 12

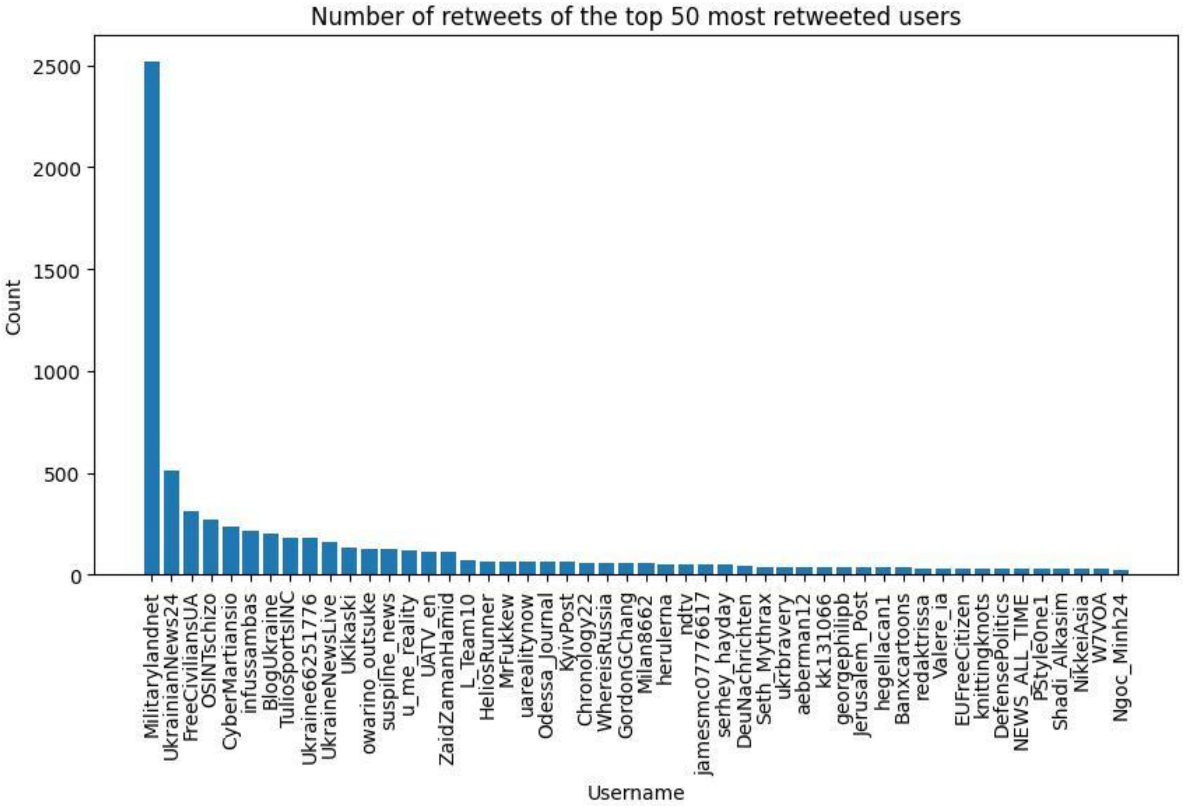

Figure 13