Part 1: Text Processing and Exploratory Data Analysis

Abril Masgrau Turró, Jorge Martínez Rosa and Enric Riba Sáez
Information Retrieval & Web Analytics
University Pompeu Fabra

0) Introduction

This project is about Ukraine-Russian war tweets during 2022. With python3 programming language and some libraries we pre-process a text with tweets to a structured processing data with tweet information. This data will be analyzed to find interesting results.

1) Pre-process the documents

At the beginning, we imported all the libraries required for the task. Then, we've loaded the two documents needed (in the variable "path", you have to add your own link where you have these documents).

In the function called *build_terms()*, we've preprocessed the text applying some criteria. First of all, transforming to lowercase and removing the dots so .war can be interpreted as war. Then, text is split into words (tokenization). We remove stop words using ntlk and stem terms to make it easier to search for words. Finally, we prune the hashtag from the words in the tweet text. Then, we return the processed tweet in a structured dataframe.

Hashtags (#) are removed because they are stored in the column name "Hashtags" and it's not necessary to keep them.

2) Exploratory Data Analysis

After pre-processing the documents, we analyze our dataset to find relevant information about the Ukraine-Russian War.

2.1 Number of words

First of all, we count how many words all the tweets have in all the documents. After that, we search for the most common number of words and plot their distribution. According to the shape, it seems to follow a Poisson distribution.

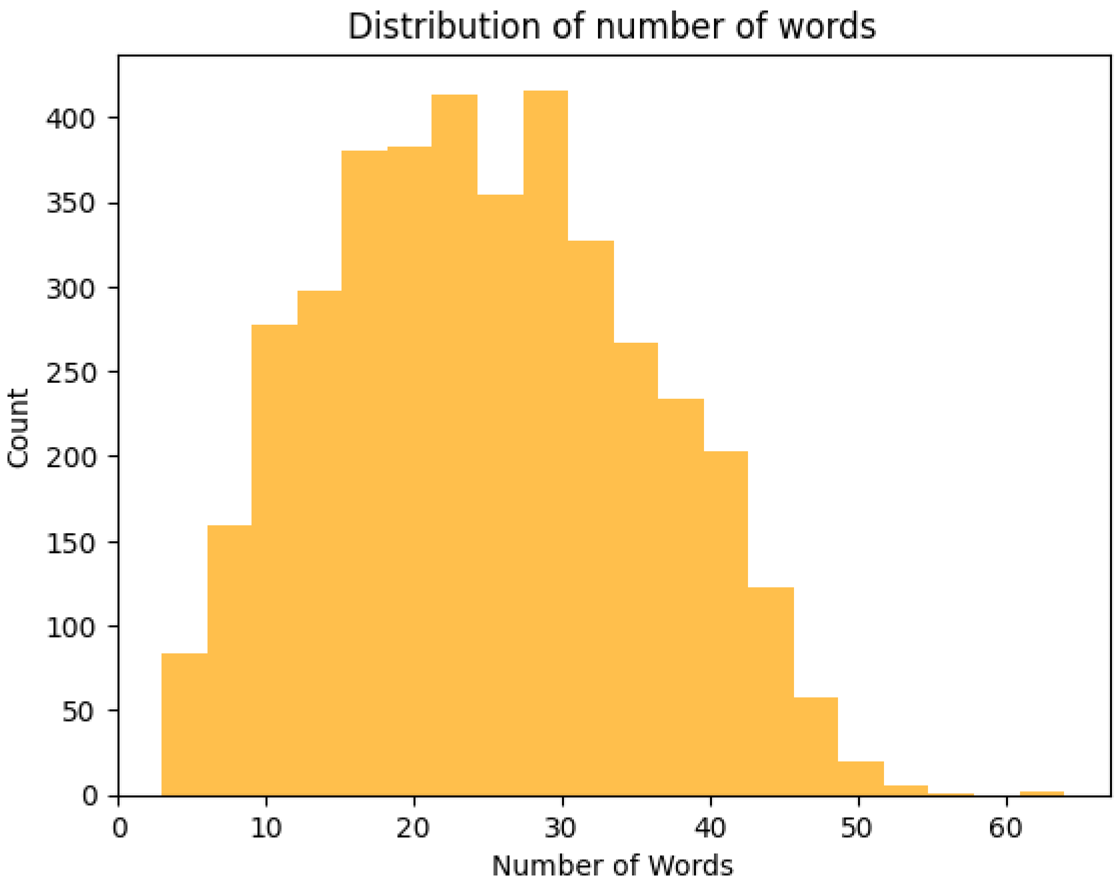
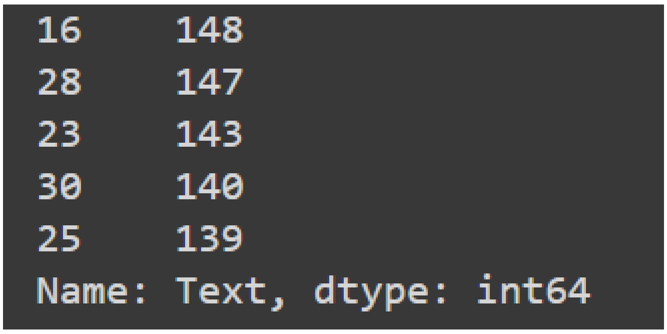


Figure 1. Plot Distribution of Number of Words in a tweet



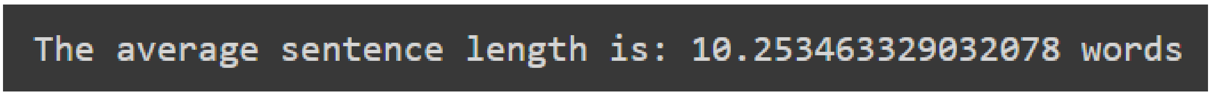
```
16     148
28     147
23     143
30     140
25     139
Name: Text, dtype: int64
```

Figure 2. Display about most common number of words in a tweet

As we can see, the most common number of words in a tweet is between 15-30 words. Which indicates that people use more than 50 characters, tweets in these documents are frequently short phrases. Also, 16 is the number of words most repeated in the document.

2.2 Average sentence length

We discover the average sentence length that it’s 10 words approximately for a tweet. That means that short sentences are in fact the most common tweets.

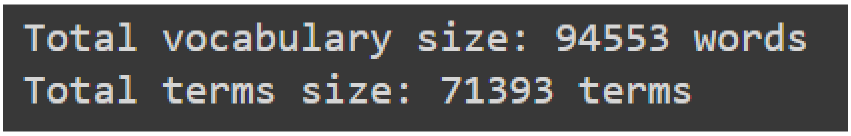


```
The average sentence length is: 10.253463329032078 words
```

Figure 3. Display about average sentence length in a tweet

2.3 Original Text and Processed Text

In the documents, there are 94.553 unique words in total and 71.393 unique terms in the processed text, two big amounts that are close to each other.

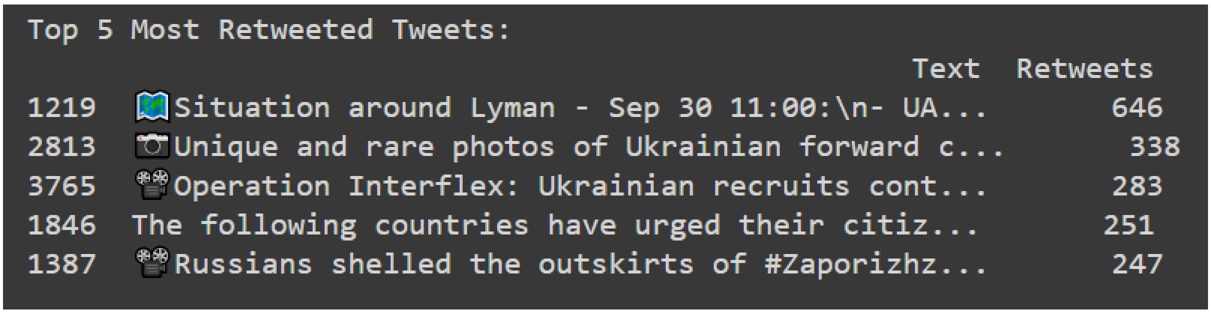


```
Total vocabulary size: 94553 words
Total terms size: 71393 terms
```

Figure 4. Display about original text and processed text

2.4 Most Retweeted Tweets

We can see that the top 5 most retweeted tweets are some related to breaking news, images and videos of the war situation, having several hundreds of retweets.

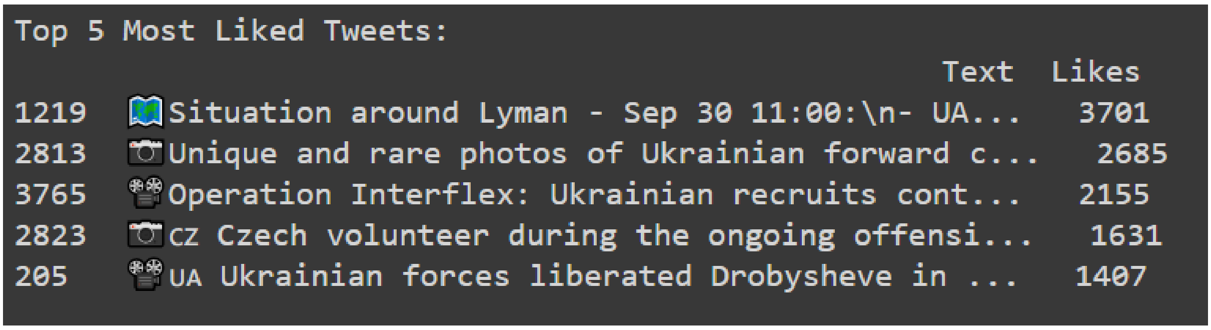


```
Top 5 Most Retweeted Tweets:
                                                    Text  Retweets
1219  🗺Situation around Lyman - Sep 30 11:00:\n- UA...       646
2813  📷Unique and rare photos of Ukrainian forward c...       338
3765  📽Operation Interflex: Ukrainian recruits cont...       283
1846  The following countries have urged their citiz...       251
1387  📽Russians shelled the outskirts of #Zaporizhz...       247
```

Figure 5. Display about the most retweeted tweets

2.5 Most Liked Tweets

We can see that the top 5 most liked tweets are some related to breaking news, images and videos of the war situation, having several thousands of likes. Furthermore, the top 3 most liked tweets are the same as the top 3 most retweeted tweets, even with the same rating order.

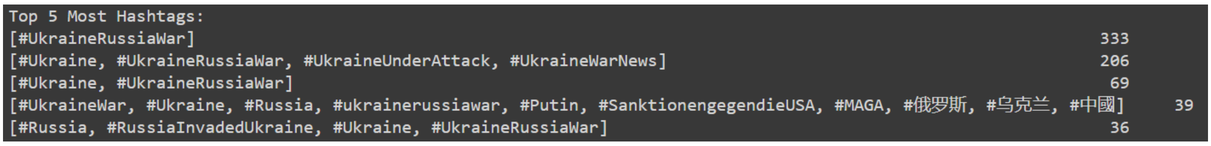


```
Top 5 Most Liked Tweets:
                                                    Text  Likes
1219  🗺Situation around Lyman - Sep 30 11:00:\n- UA...   3701
2813  📷Unique and rare photos of Ukrainian forward c...   2685
3765  📽Operation Interflex: Ukrainian recruits cont...   2155
2823  📷CZ Czech volunteer during the ongoing offensi...   1631
205   📽UA Ukrainian forces liberated Drobysheve in ...   1407
```

Figure 6. Display about the most liked tweets

2.6 Most used Hashtags in a Tweet

We can see that the top 5 most used hashtags directly mention the Ukraine-Russian War with at least one of these three words or others related. In fact, #UkraineRussiaWar is in each of the five most used hashtags. Moreover, there's a significant difference between the number of uses of the two most used hashtags and the rest.

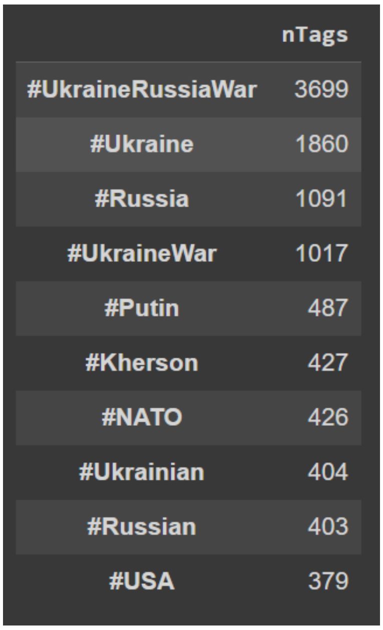


```
Top 5 Most Hashtags:
[#UkraineRussiaWar]                                                                                     333
[#Ukraine, #UkraineRussiaWar, #UkraineUnderAttack, #UkraineWarNews]                                     206
[#Ukraine, #UkraineRussiaWar]                                                                            69
[#UkraineWar, #Ukraine, #Russia, #ukrainerussiawar, #Putin, #SanktionengegendieUSA, #MAGA, #俄罗斯, #乌克兰, #中國]     39
[#Russia, #RussiaInvadedUkraine, #Ukraine, #UkraineRussiaWar]                                            36
```

Figure 7. Display about the most used hashtags in a tweet

2.7 Most used Hashtag in a Tweet

As it was expected, all terms strictly related to the war are the most used hashtags in these datasets. Moreover, #UkraineRussiaWar is the most used hashtag in a tweet with a huge difference, as before. In addition, some words or hashtags that appeared in the previous case (figure 7) also appear here, but others don't.



	nTags
#UkraineRussiaWar	3699
#Ukraine	1860
#Russia	1091
#UkraineWar	1017
#Putin	487
#Kherson	427
#NATO	426
#Ukrainian	404
#Russian	403
#USA	379

Figure 8. Display about the most used hashtag in a tweet

2.8 Word Cloud

Finally, we map the most used words as a word cloud where the bigger ones are the most frequent. As we observe, there's a small difference between the most used words and the most used hashtags, but all of them are closely related to the Ukraine-Russia War and mention Ukraine or Russia.

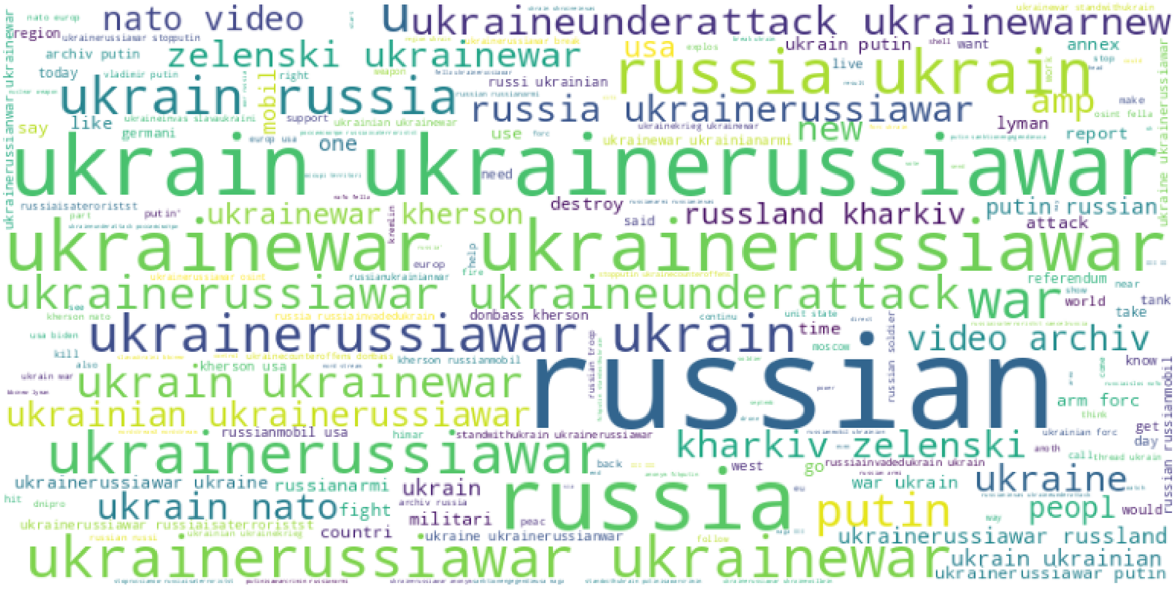


Figure 9. Word cloud of the most used words

3) Conclusions

As a conclusion, in this document we can find that most of the tweets are related to Ukraine and Russian as it's expected because they are the spotlight of the war. Nevertheless, NATO and the USA are quite popular because of their relationship with this war and their political issues with Russia. Putin and Zelenski, politicians, are the most popular names in this dataset.

4) Link

https://github.com/JorgeMRPOO/IRWA-2023-part-1