Guillem Gauchia Torres - u186410 - 240215
Àlex Herrero Díaz - u186402 - 240799
Adrià Julià Parada - u188319 - 242195
Date: 21/10/2023

Part 1: Text Processing and Exploratory Data Analysis

This project has been carried out by implementing in a notebook the necessary functions and code to process a database of tweets about the current war between Ukraine and Russia.

Lab1_Code.ipynb is the Notebook used to carry out the objectives of the practice and to obtain results on the analysis of the data. The code files are located at the following Github Repository:
https://github.com/AlexHerreroDiaz/IRWA-2023/tree/main/IRWA-2023-part-1

1) Pre-process documents

To perform the preprocessing part of the documents, in this case the tweets dataset, we are asked to obtain the following information from the tweets:

- Tweet
- Date
- Hashtags
- Likes
- Retweets
- Url

We also believe it is important to extract the following two fields:

- Doc_id
- Tokens

To store the information of each tweet in the database we have chosen to create a dictionary containing the previously mentioned fields, for this we make use of the ***create_dictionary*** function that we have implemented.

Therefore, to carry out the processing of the text field in a token list, we have taken into account the following procedures implemented in the ***build_terms*** function:

1. We remove the URLs from the texts
2. We execute the ***split_hashtag_words*** function to treat hashtags so that if there is a hashtag containing multiple words together it splits them accordingly (e.g. if we have #RussianUkrainianWar we get Russian Ukrainian War).
3. We switch to lowercase and remove punctuation, emojis, symbols, numbers, and strings beginning with '#' and '@'.
4. Once the text of the tweet is formatted, we divide it into tokens using the stemmer to obtain only the root of each word.

Finally, we map for each element of the dictionary, using the ***get_data_ids*** function, so that using the file ***Rus_Ukr_war_data_ids***.csv we assign to each tweet according to its id to which document it belongs.

In this way we will obtain a dictionary with the following elements as shown in the following example:

EXAMPLE TWEET
Tweet: Putin Suffers Most Humiliating Ukraine Defeat Yet Around Key City of Lyman in Donetsk https://t.co/UE1xogD6GT #Ukraine #UkraineRussiaWar #UkraineUnderAttack #UkraineWarNews
Id: 1575839192541798401
Date: Fri Sep 30 13:25:15 +0000 2022
Hashtags: ['Ukraine', 'UkraineRussiaWar', 'UkraineUnderAttack', 'UkraineWarNews']
Likes: 3
Retweets: 1
Url: https://twitter.com/twitter_username/status/1575839192541798401
Doc: doc_556
Tokens: ['putin', 'suffer', 'humili', 'ukrain', 'defeat', 'yet', 'around', 'key', 'citi', 'lyman', 'donetsk', 'ukrain', 'ukrain', 'russia', 'war', 'ukrain', 'attack', 'ukrain', 'war', 'news']

The time it took to create the dictionary containing 4000 items was 5.15s.

2) Exploratory Data Analysis

Once the dictionary is created, we can make a study of the data and try to show outstanding characteristics to better understand them. After the process performed to create the dictionary, we obtain the following data:

From the **4000** *tweets* that have been processed we found that the vocabulary obtained has a size of **7189** stemmed words, it is not a very large number considering the amount of tweets, and this is due to two possible reasons:

1. By obtaining the root of the words, many variants of each word are reduced.
2. Since this is a set of tweets dealing with the topic of the Ukraine-Russia war many of the words will be similar.

Among all the *tweets* we have seen interesting to observe what is the usual average number of tokens obtained from the publications, the result obtained was about **21.33** tokens per tweet on average.

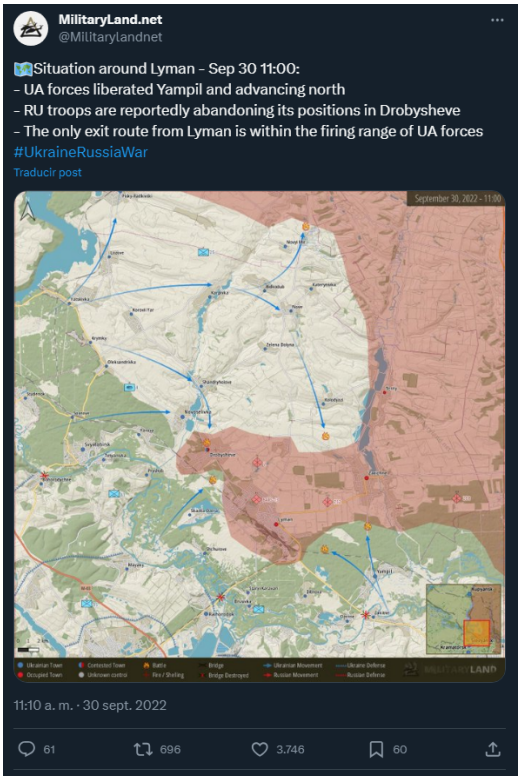
On the other hand, we have the words in their base form from which we have searched which words are the most frequent in the database:

Top stemmed 10 words with more occurrences in the data:	
ukrain	3989
russia	3947
war	3928
russian	1590
putin	1126
ukrainian	1051
armi	552
nato	543
kherson	501
state	496

Obviously from the results we can see that the top ten most frequent words are very much related to the context of the war between Ukraine and Russia as expected.

Top 10 most liked Tweets:

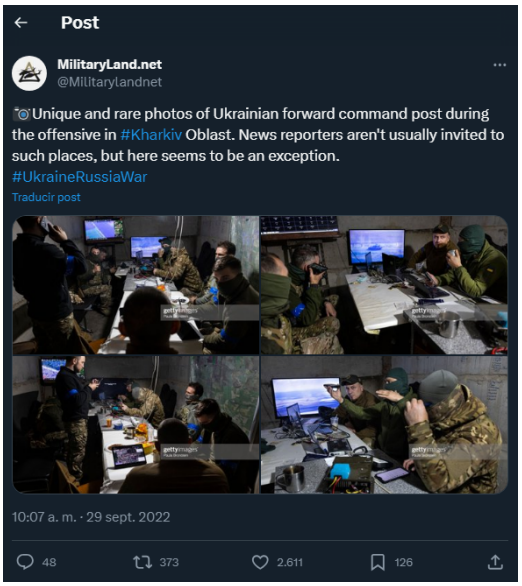
1. Tweet: doc_1220 has 3701 likes
https://twitter.com/Militarylandnet/status/1575775162674212865?lang=ca
2. Tweet: doc_2814 has 2685 likes
https://twitter.com/Militarylandnet/status/1575396903252025351
3. Tweet: doc_3766 has 2155 likes
https://twitter.com/Militarylandnet/status/1575181552170201088
4. Tweet: doc_2824 has 1631 likes
5. Tweet: doc_206 has 1407 likes
6. Tweet: doc_2119 has 1407 likes
7. Tweet: doc_3802 has 1348 likes
8. Tweet: doc_451 has 1083 likes
9. Tweet: doc_1847 has 923 likes
10. Tweet: doc_1245 has 868 likes



The image that appears next to the ten *tweets* with the most *likes* belongs to the *tweet* with the most *likes* in the dataset. We can see a variation in the *tweets* due to subsequent interactions made by Twitter users after the day the information was extracted from that *tweet* in the dataset.

Top 10 most liked Retweets:

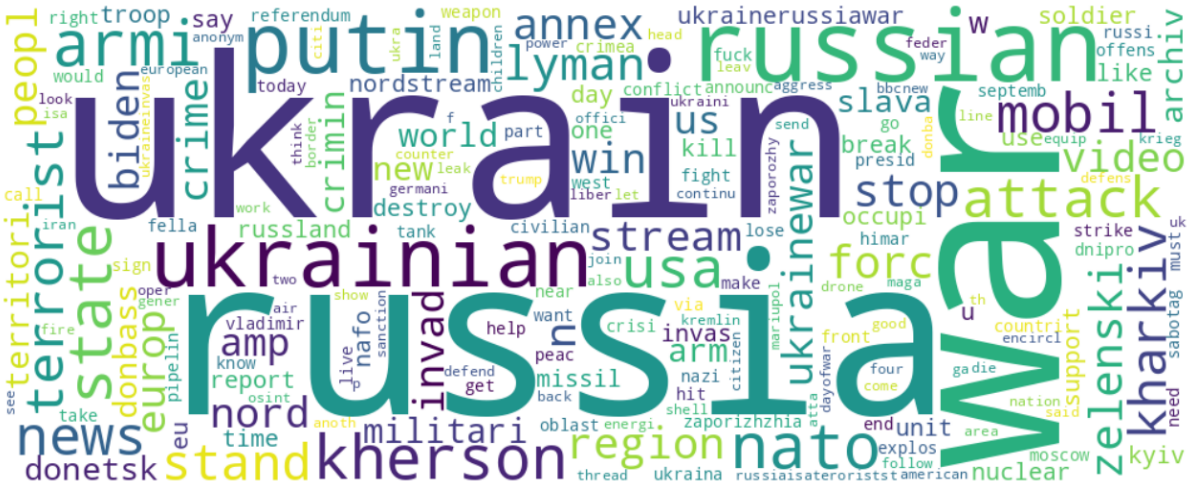
1. Tweet: doc_1220 has 646 retweets
https://twitter.com/Militarylandnet/status/1575775162674212865?lang=ca
2. Tweet: doc_2814 has 338 retweets
3. Tweet: doc_3766 has 283 retweets
https://twitter.com/Militarylandnet/status/1575181552170201088
4. Tweet: doc_1847 has 251 retweets
5. Tweet: doc_1388 has 247 retweets
6. Tweet: doc_1210 has 236 retweets
7. Tweet: doc_1533 has 184 retweets
8. Tweet: doc_206 has 171 retweets
9. Tweet: doc_2119 has 136 retweets
10. Tweet: doc_3802 has 133 retweets



The image next to the ten most *retweeted tweets* belongs to the second most *retweeted tweet* in the dataset. We can see a variation in the *tweets* due to subsequent interactions made by Twitter users after the day the information was extracted from that tweet in the dataset.

We highlight that the first three *tweets* with more likes and more *retweets* are the same.

To all this, we have created a WordCloud with the tokens obtained from the database and the result obtained is as follows:



Here we can see that the most prominent words are ***'ukrain'***, ***'russia'*** and ***'war'*** being data on tweets about the war between Ukraine and Russia is an expected result, among the following words we can see names of political leaders such as Putin, Zelenski and even Biden, Ukrainian cities or regions such as Kherson or Donbass and words that represent acts related to the war.
As curiosities we can see that the cities and regions that are most mentioned in the tweets are part of Ukraine because this war between Russia and Ukraine is being a Russian invasion into the country of Ukraine and all the battle fronts are developed in Ukrainian territory.

Finally, using the model called NLP spaCy, we have extracted the entities and also what type they are. Here there are the top ten of most frequent entities and the most types of entity:

Top 10 most frequent entities		**Top 10 most frequent types of entity**	
Entity: russia	6623	Type: GPE	7616
Entity: russian	2205	Type: PERSON	3418
Entity: armi	572	Type: ORP	2697
Entity: putin	448	Type: NORP	2539
Entity: nato	333	Type: DATE	333
Entity: kherson	152	Type: CARDINAL	289
Entity: one	6	Type: LANGUAGE	107
Entity: slava	91	Type: EVENT	95
Entity: moscow	90	Type: LOC	68
Entity: nord	89	Type: ORDINAL	59

As we can see, the most frequent entities are mostly related to Russia as its president, its capital and its own name. Moreover, the most frequent city is not the capital of Russia or Ukraine, it's Kherson, which is a city located in a key point of this war.
Furthermore, the most frequent type of entity by far is the Geo-Political Entity. This can be interpreted as lots of Geo-Political Entities are involved directly or indirectly in this war.