Ruben Liso

Project 2

Use Case Study - Twitter Data Analysis

Project description :

I used a dataset from kaggle of the tweets related the war between russia and ukraine between the first and the second april of 2022.

Honestly, the dataset was a lot bigger but i split it as much as i could and it is still heavy for calculations.

For each tweet in the dataset we got the following values :

- userid
- Username
- acctdesc
- location
- following
- followers
- totaltweets
- usercreatedts
- tweetid
- original_tweet_userid
- original_tweet_username
- in_reply_to_status_id
- in_reply_to_user_id
- in_reply_to_screen_name
- is_quote_status
- quoted_status_id
- quoted_status_userid
- quoted_status_username
- extractedts

Project development :

First, i set up the Tools and Libraries. Then I extract'ed the data from a Gzip file and stored it in a dataframe.

Next, i cleaned the text thanks to spacy and put it in dictionnary with the tweets' id for keys and the processed text for values. This processed text will be used for several analisys and functions (like the word cloud) in the next parts.

Indeed, i used the already existing data from the dataset that i put in the form of a dataframe to obtain different informations about it. Like the average value of follower of an user, ...

Then, I made a sentiment analysis program using NLTK and TextBlob. It classify the tweets into 3 sentiment categories : neutral, negative, positive and put these value in a dictionnary: sentiment_dict.
Then i made a function to find to find the sentiment of a tweet with its id for argument. This function will be use to filter tweets in the vizualisation part.

For the section 6, I just used Spacy and NLTK to make an Named Entity Recognition (NER) for identifying and classifying entities in tweets and a Part-of-Speech (POS) Tagging for Annotating words with their respective parts of speech.

Finally I made the vizualisation part using Plotly and Dash, this part is consisting of a graph, a wordcloud, a generated csv report and a filtering engine in the database (here the dataframe).

Conclusion and Future Work :

Firstly, I found that there is more neutral tweets than the sum of positive and negative tweets. This is probably due to the fact there is a lot of media or informative account that are covering this war.

Then, I observed that there is more positive messages than negatives ones. This is probably due to the fact the majority of people will use twitter to support the population involve in the war more than to criticize it, even if a non negligible part is doing just that.

I also found a lot of data about the most-used words and hastag, and also about the differents statiscal data. What is found is mostly expected but some interesting fact like that the highest number of retweet of tweets are not usually made by the most followed user can be interesting.

To improve this work, we could switch the tool or method for the preprocessing of the text because some improvment could be made. Maybe using ntlk instead of spacy would have been better.

Also, as the dataset is very extensive, I could have anlyse the relation between each user and tweet with more depth. Like, for example, studying the stats of tweets answered to, or by studying relationship bewteen users by watching if they respond to each other tweets.