Designed report by Abdelrahman Rezk

COVID-19-Arabic-Tweets-Dataset

We have collected more than 3, 000, 000 tweets from twitter API, besides cleaning these tweets and we have make some analysis to get what is behind these tweets.

- Eng: Ayman Mahgoub
- Researcher at electronics research institute
- E-mail: Ayman Mhgb

- Eng: Abdelrahman Rezk
- Teaching Assistant at Arab Open University & NLP Engineer
- E-mail: Abdelrahman Rezk

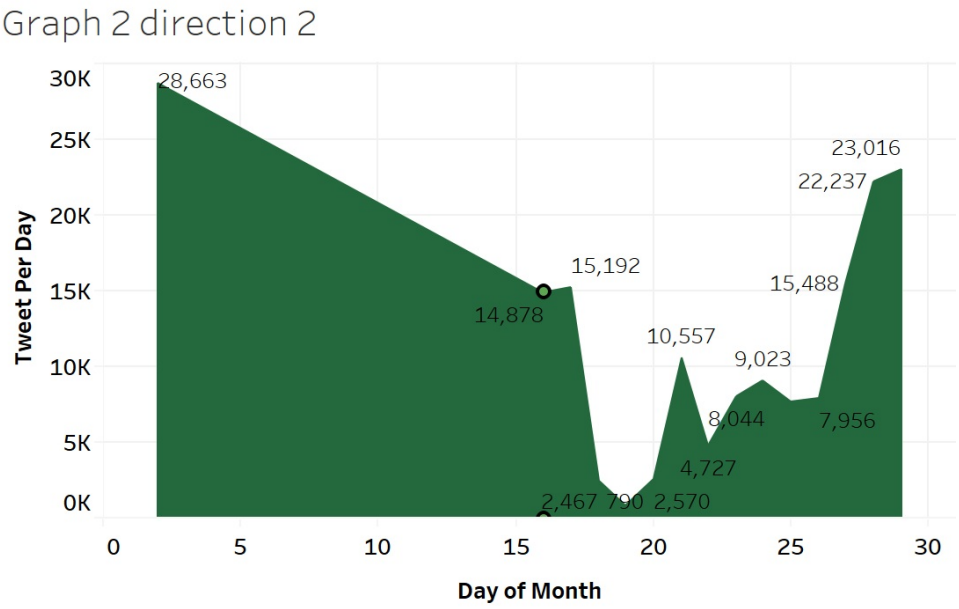
Corresponding Code Files

Files included in direction:

- config_files
- Tableau for graphs

Files Name:

- Analysis.py



Words frequency

Once we have collect and clean our dataset of some rubbish, we start the process of analysis these tweets, and one of these ways to analysis is to count the uniqe words to know which most of frequenct words that describe most of tweets.

Designing the function that get the frequency of each word after of that we have graphed these words using Tableau.

We have impletent some of the function in this Analysis and others we have used from previous work:

- words_frequency
- drop_rows
- analysis_pipline
- count_tweets_file

words_frequency

Using the sklearn library and its method of countvectorizer which behave to count for each word in our data how this word frequency in our dataset, but based on our cleaning we have passed the list of all words of one direction which by we have introduce before its contain more than one file and for each file we have set of tweets, then from each file get all tweets, then append each of these tweets to one list, after that split all of tweets to words.

After that we have initialize Dataframe that contain each word meet their frequency in our dataset.

drop_rows

After the spliting process we have some of the words are numbers and some are in another language which not Arabic language so we have used some of the Analysis to handle like these cases and drop all of the words that are not Arabic words but it also have some of passed words during this process because of some complicated cases, but most of Non Arabic words are dropped.

analysis_pipline

Based on the idea of pipeline all of work in on way we have designed one function to throw all of the function will be used for this analysis like what we have discuess above others was from previous work and have their Docstring to get intuation about their work.

Tehe other function used is:

- read_direction_analysis >> and you can find in the direction_and_file_handleing notebook.

count_tweets_file

Because our data have different direction which represent different month that the tweets comes in, so we have get per file how much the tweets are, then we have a file for each direction represent these counts, after that we have graphed these counts Vs their days.

Some Snapshots