**Twitter Data Analysis - Milestone 1**

As the title says, this project will aim on collecting analytic data from one of the largest social network called Twitter. To do so we will be using Twitter API, which is something like a wrapper for python. First of all one need to create a twitter developer account and obtain the credentials. We need to visit this site[[1]](#footnote-1) and fill up some forms in order to have access to creating new app in Twitter API. While creating an app I have used the name “OTH - Twitter Data Analysis“. All of the tasks require working Twitter API. In my case I have chosen this library import tweepy. All the fetched data will be stored in data frames which are available in the library import pandas. Also for the work with credentials I needed to include this library import json. It allows me to create and read JSON file on a disk. Last library I need is import csv, because I will not only show results in python, but I will save all the results in to the disk as well.

The assignment consists of four tasks in total. Before digging deeper into each of the task there is some work to be done. At first we need to fill our valid API credentials inside the program 1.1 Creentials and run it. It creates a JSON file twitter\_credentials.json in the same folder as the python file is. In this file there are stored all the keys and tokens. Before running each of the sub-program which will take care of those tasks, user need to also run an initializing program 1.2 Initialize that will import all the necessary libraries, load credentials of Twitter API and also ask user about twitter event he want to search. In each of the sub-program I will be fetching tweets which are newer than the date 01.01.2020 and whose total number will not be larger than 1 000 in order to do not exceed the limit of query on Twitter API and also to be sure my query will not be running too long. For getting data form Twitter I will be using tweepy, precisely I will put my queries inside the function tweepy.Cursor(). In order to not exceed pagination limits I need to apply this parameter to the API wait\_on\_rate\_limit=True. I will be fetching data (running the query) in the beginning of each sub-program.

In the first task 2.1 Derive the sentiment we need to decide on the sentiment of each tweet, which has the given event in it. For that I will use the library from textblob import TextBlob. It gives us module that is able to handle this task on its own. At first we access text of the tweets in for loop by tweet.author.name. Then we need to remove all the retweets and URLs in text of all tweets because it can distort the results. We store text of tweets into a data frame. After that we just simply use TextBlob to validate text on a scale -1 to 1 for each text. After that we will show the texts and its sentiment values. We can also show the results graphically using a histogram. For that we need another library import matplotlib.pyplot.

In the second task 2.2 Top 10 hash tags and users we need to come up with top 10 hashtags and users in given event based on number of tweets in our dataset. Firstly I get hashtags and authors of tweets using for loop and accessing it by tweet.entities.get('hashtags') and tweet.author.name. Then I will use 2 python dictionaries. In the first one I will use hashtags as keys and I will iterate through dictionaries to count occurrence. And in the second one I will use users as keys and basically do the same process but with authors/users. In the end I will put data into a data frames and I will print the results.

In the third task 2.3 Followers of a given user we are supposed to show all the followers of given user in our dataset. In the beginning we will print table of all available users/authors in our dataset and ask the user for which one of them he wants the results. Then we will use following tweepy’s API api.followers\_ids to access list of followers. Then we simply put result into data frame or array and print the results in python.

In the fourth task 2.4 Tweets and profiles of all followers we need to find all tweets and profiles of all follower of given user even outside our data set. We firstly ask user about which account will be searched and we make a similar query to fetch all the followers of given user api.followers\_ids. The query will differ from the previous one, because it will include all the users not only the one that are using event hashtag. After then we add it to data frame and print it as well. According to coming up with all the tweets of the given user we will use our API again api.user\_timeline. That will give us all the tweets that we can now store into data frame and print.

I have decided to run a query in the start of each sub-program because I think that running only one query in the beginning and storing all the informations into JSON or CSV would be too slow considering higher limit of parameters which I have described in the first paragraph. I assume that writing too much data on disk is quite slow operation in general. In order to have access to result outside of Jupiter I will export results to CSV files. In the task about sentiment of tweets I will save histogram to the disk as PNG as well.

1. Twitter Developer Account: <https://developer.twitter.com/en/apps> [↑](#footnote-ref-1)