Project: Criminal Activities of Wagner Group

Name: Shahid Attar

Advisor: Dr. Dustin White

Subject: ECON 8320 Tools for Data Analysis

Problem Statement:

Creation of a Data set which points out the activities of Wagner Group of mercenaries at a global level. The source can be Twitter, Google News or any other source of reliable information.

Overview:

I have developed a program which uses the **Tweepy Library in Python** to download the data from Twitter and return a data frame which is downloaded in the CSV format after the execution. All the details such as Text/Description, URL, Location, Date, Time etc. are extracted from the text of the tweet. I will be explaining the methodology, challenges and the future scope of the project in the following sections.

Methodology:

At first, when I was thinking about choosing the data source, I thought the most suitable option would be a news source like Google News for getting the required data, but when I checked I found out that we could only extract a mere 100 records from it each day that too from an API which is not maintained by Google (Google News API has been deprecated in May 2011). On the other hand, Twitter was very flexible with the number of records and with a rate limit of 450 requests per 15 minutes [1].

After selecting Twitter as the source of my data I subscribed for the free version of the API which was only allowing 50000 tweets per month which would limit the extent to which I could experiment with the code and extraction. Therefore, I applied for the elevated access and the limit was raised to 2.5 million tweets per month.

The process is as follows:

1.) Search for suitable methods from Tweepy documentation [2] which can help in searching the tweets.
2.) Explore options to build the query [3].
3.) Extract small data (10 tweets) to check if all the parameters are present in the data.
4.) If some parameters are not present, then check the developer portal if it is possible to get those parameters [3].
5.) Extract the data in raw format.

6.) Create multiple functions to extract the required information from the data using Spacy and Core Python. Store this data individually in different lists.
7.) Cleaning the data using basic python and creating a data frame using Pandas.
8.) Export the data frame to .csv format.

Challenges in the project:

Extraction of data using the API was normal but **the construction of query** to exclude retweets and navigating to different pages using Pagination was something new. As I did not have any idea about constructing a query, adding required parameters like 'referenced tweets', 'entities', 'geo' took a lot of time as I had to go through all the documentation of Tweepy and some of the Twitter Developer Portal.

Once I was able to extract all the information, the next thing was using Spacy to extract all the information from the text of the Tweet. The drawback with Spacy (like other ML tools) is that it does not always identify all the entities correctly.

At first, I planned to create a list of criminal activities and match the activities in the tweets with those in the list and keep those tweets only. But I think that this would be helpful when we perform a total database search instead of just the recent tweet search which only gives us access to the tweets of the last 7 days.

Also, because Twitter is not explicitly a news platform, there are a lot of **irrelevant tweets** present in the feed which deter the quality of our data. Also, there are tweets which have videos and images as their information, but the text of the tweet is not detailed enough to be reflected on our data frame.

Results:

We were able to break down 1000 tweets by using Twitter API with Python and libraries like Spacy and Pandas we were able to create a rough data set and export the same to csv. The dataset has a lot of missing information mainly because of Twitter being the source and the error of a Machine Learning Library like Spacy.

Future Scope:

If I get a chance to do the project in the future, I would use the Academic Research Access for this project which would provide me with the ability to search the whole database and a large capacity of tweets (10 Million per month) to work with.

I would also search articles which are not in English and more in the proximity of Russia and Allies of Russia which would help me get something useful as most of the English news media are already in the radar of the Intelligence agencies.

I would analyze the positive tweets related to the Wagner Group and create a Graph Network of such people to check who they follow the most. This can help in analyzing the person of interest.

At last, I would consider checking the news agencies of the places where Twitter is banned [4] and the countries are the ones supporting the Russian Invasion.

REFERENCES:

[1] https://developer.twitter.com/en/docs/twitter-api/rate-limits

[2] https://docs.tweepy.org/en/stable/client.html

[3] https://docs.tweepy.org/en/stable/expansions_and_fields.html

[4] https://www.statista.com/chart/26601/countries-blocking-or-severely-restricting-access-to-twitter/