Detecting COVID-19 through Twitter

Candidate Numbers: 118,110, 113

Introduction

We created a system that visualizes active Covid-19 cases with Covid-19 twitter related activity, in an attempt to see if there is a relation between the two. To do this we required large amounts of data from different sources, and using big data technologies is therefore the only option to tackle such a task. We chose to do this because we were not able to find anyone who had visualized this previously.

In order to see if there is any relation between the two elements, we chose three different periods for further investigation. The selected periods reflect on different Covid-19-situations in the world. The chosen periods are:

- Period 1: 21-24 March. In March Covid-19 was declared by the WHO as a global pandemic.
- Period 2: 5-8 July. In this period there were lower cases of the virus in some parts of Europe, and many countries in Europe allowed travel to other parts of the world, without having to quarantine.
- Period 3: 13-16 October. In this period, there was a new wave of Corona-virus cases in many parts of the World.

After gathering data from these three periods we created a web application to draw the activity on a map to see if the areas with high/low Corona cases matched areas with high/low Twitter activity about the Corona virus.

Big data technologies were important in our project in order to handle the large amount of data used. We used big data technologies to collect, process and extract information from large amounts of social media and confirmed covid-19 data.

Technologies used

The first of which being a dataset from the John Hopkins University-github where they had aggregated multiple data sources from governments around the world to track data about Covid-19. This dataset was complete with a lot of features, but we ended up using the gps coordinates, the region name and the amount of active Covid cases per location. The other dataset was a dataset collected by Rabindra Lamsal and uploaded at IEEE DataPort containing geotagged-tweets related to Covid-19 using about 90 keywords. The keywords used can be found on the upstream link to the dataset.The technologies used included the Twitter API using DocNow's Hydrator to hydrate tweets, Python to clean and preprocess the datasets, and HTML,CSS, and JavaScript to create the web application to visualize the data. More specifically we used the LeafletJS map API with OpenStreetMap and MapBox to create the map we drew the data on. We chose these technologies as they were technologies we were comfortable with, but we did consider using the Google Maps API at first for the map. We were recommended by fellow students to try the LeafletJS API instead because of its ease of use. During the data preprocessing and cleaning phase, we used pandas library in python to let us convert the datasets from csv and jsonl format to dataframes. The dataframes were then manipulated and outputted again as a clean dataset in csv format.

Technologies we used for this project and their purpose are presented in the table in the appendix.

Technologies not used

We considered using other technologies during this project. Mainly related to how we would use and store the data. If we were to collect our own data we would need to stream this data using Apache Spark, and use MongoDB or Apache Hadoop for storing the data. This would allow us to continuously stream current data, but because of limitations like the amount of tweets we could collect each month from using the Twitter API we chose to use already collected data . With our current access level to the Twitter API we could only collect 500000 tweets a month, so it was more convenient and "safer" (in that we could not run out of tweets per month) to use daily updated open source datasets available online. We still needed to use our Tweet limit on the Twitter API to hydrate the Tweets using Hydrator.

Flow Chart see appendix

Challenges

Twitter Dataset: The main challenge with the dataset chosen was that not many tweets had gps coordinates, which resulted in difficulties in mapping all tweets to our visualization. The data we used was limited to only English speakers as the tweets collected from IEEE DataPort only searches for english keywords.

Abstracting: There are a lot of other factors that come into play when people decide what to tweet about. Assuming that if a country suddenly tweets a lot more about Covid is due since there are many more infections, is probably abstracting too much. Maybe they are talking about it in a way of "there is a lot of Covid in that other country," or maybe the news is talking about it, hence many people tweet about it. Due to these limitations in the data, it can be difficult to see a clear correlation, since the data is not that representative to the population globally.

Covid-19 data set: Covid infection varies a lot from country to country and so does the average number of tweets from that country. The regions related to the active Covid-19 cases are also very varied. This is most likely due to the data that John Hopkins University has been given from different governmental agencies is limited. Some agencies have supplied reports on a country level while others have been on the level of an entire country. If one looks at the difference in reporting from Japan and France you can see the data in the dataset varies widely.

Combination of the two dataset: The combination of the datasets in general worked poorly, because the twitter data and Covid-19 data did not go well together. The Covid-19 data is clustered from specific countries/places since the data is collected from government/countries. While the tweet-data in general is geo-tagged to specific places.

Future work

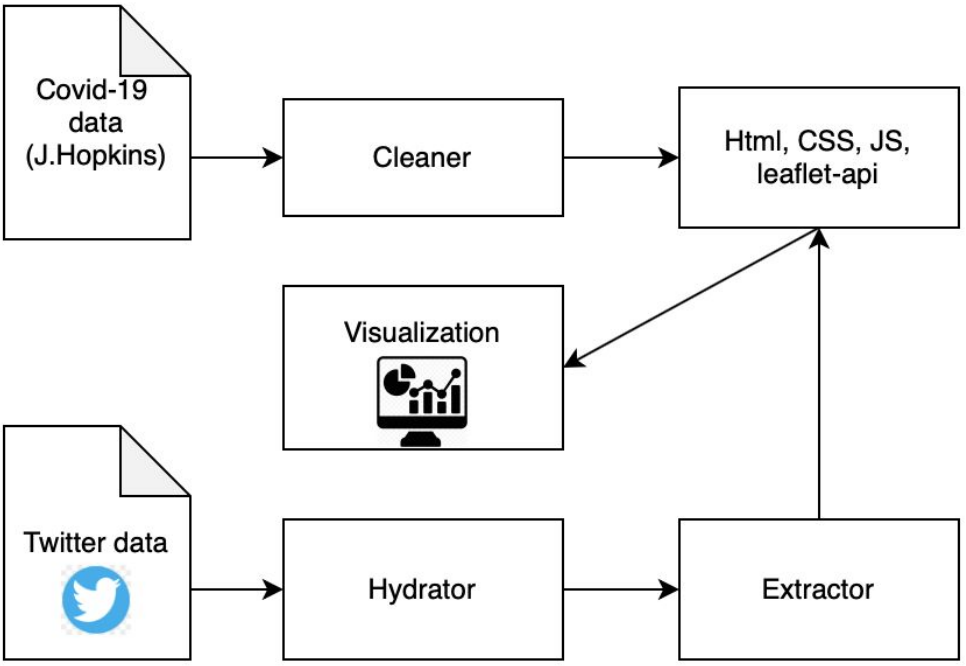
A future goal of ours could be to see if we streamed in continuously data from twitter that had gps coordinates. Then do a sentiment analysis where we split them into categories, negative and positive and tried to see if there is a correlation between an increase in negative or positive tweets about corona. Then compare that with what we currently have.

Correlation in the data

In our project we wanted to see if there is any correlation between the twitter activity related to Covid-19 and actual Covid-19 cases. Due to the challenges we had with the limitation in our dataset, such as the data is mainly english and not representable for the population and that the combination of the chosen dataset did not work well together, it was difficult to see if there was any relation. However, based on our visualized map, we can see that the areas with a high confirmed Covid-19 cases and twitter activity varies and there is no clear correlation.

Appendix:

Technologies	**Used for**
VS Code (Code OSS - branch)	IDE for working with HTML, CSS, and Javascript
Jupyter Notebook (Anaconda branch)	IDE for working with Python
Twitter API	Hydrating the twitter dataset
HTML	Marking up the web application
CSS	Styling the web application
JavaScript	Adding functionality to the web application
LeafletJS (JavaScript API)	Adding map functionality to JavaScript for visualization of the data
Python (Using pandas library)	Working with datasets as dataframes to do cleaning and preprocessing
GitHub	Version control
GitKraken	Git GUI for pushing project to github.
Hydrator (by DocNow)	Hydrating tweets from tweet ID.


Covid-19
data
(J.Hopkins)
Cleaner
Html, CSS, JS,
leaflet-api
Visualization
Twitter data
Hydrator
Extractor