Investigating COVID-19 Spread

CSE564: Project Proposal

Proposed By - Harshit (112687784), Siva (112957009)

Background

The 2019–20 coronavirus outbreak is an ongoing pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak was identified in Wuhan, China, in December 2019, declared to be a Public Health Emergency of International Concern on 30 January 2020 and recognized as a pandemic by the World Health Organization on 11 March 2020. As of 17 April 2020, more than 2.21 million cases of COVID-19 have been reported in 210 countries and territories, resulting in more than 151,000 deaths. More than 565,000 people have recovered.

Medical materials and other goods shortages caused by the 2019–20 coronavirus pandemic is one of the major issues of the pandemic. On the medical side, shortages of personal protective equipment such as medical masks, gloves, face shields, gear, sanitizing products, are also joined by a potential shortage of more advanced devices such as hospital beds, ICU beds, oxygen therapy, mechanical ventilation, and ECMO devices. So, we would like to take this opportunity to visualize the effect of COVID-19 at a global level as well as in the United States. We would want to observe certain trends on COVID-19 as it is influenced by various political, economic and social factors. At last, we would also like to visualize the problems and shortages that we are facing during these testing times so that these visualizations can help governments and other institutions understand what prospects to work on in order to avoid any sort of similar pandemics.

Datasets Description

There are a lot of time series as well as statistics data available on COVID. We particularly use datasets from the below sources for our project.

1. NY Times COVID Dataset: This data is adopted from https://github.com/nytimes/covid-19-data. We have three datasets - U.S. Data, U.S. State-Level Data, U.S. County-Level Data. We plan to combine and use these datasets to visualize the effect of COVID-19 in the United States. These datasets have data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time. The daily number of cases and deaths can be found in these files.

2. John Hopkins COVID-19 Cases Data: This data is available at https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases. Here we have data related to the COVID-19 at a global level. There are three datasets:
    - time_series_covid19_confirmed_global.csv
    - time_series_covid19_deaths_global.csv
    - Time_series_covid19_recovered_global.csv

   Fields available in these datasets include Province/State, Country/Region, Last Update, Confirmed, Suspected, Recovered, Deaths.

3. Health Infrastructure Data from World Bank: This data is available at http://wdi.worldbank.org/table/2.12#. This describes various health spending per capita by Country, as well as doctors, nurses and midwives, and specialist surgical staff per capita. This dataset, when combined with COVID-19 data, can be used to answer questions about the preparedness of a certain country to deal with the epidemic.

Problem Statement

Currently, the target for each government and institution is to mitigate the level of damage due to this pandemic. As experts say this is just the first wave, many more waves are yet to come. So, proper planning needs to be done. To adopt some future measures there has to be existential inference suggesting what factors are valuable for avoiding damages in the future. Thus, the problem statement is to visualize infection rates with different factors. There can be a myriad of factors governing it. For every geopolitical region, although the total number of infections saw a similar curve, they had different rates. These may depend on economy size, population, economy type, etc. Seeing these trends, different institutions need to act differently against the pandemic.

One of the important factors is the health infrastructure of different regions. The hospital capacity, number of medical professionals, etc decide the adversity during the pandemics. But these trends cannot be understood to be positively affecting every region, there are many more parameters along with these, as we have seen many countries with good health infrastructure also getting severely affected. Thus these health factors need to be investigated to better understand the threat and take future measures.

We can have a number of other factors that can affect the infections. These can be standard of living, population density, number of vehicles in a region, etc. We will try to investigate how these factors are correlated. The trends such as rate of growth (daily/weekly) might also play a good factor regarding how strict the social distancing rules need to be. Qualitatively all these factors affect in some way or another. But quantitative planned measures need to be taken to balance the health and economy of any region. Thus the problem is to analyze various trends to help institutions take proactive future measures.

Approach

Our approach is to design a dashboard that will help scientists with trend analysis. We have different visualizations to show weekly/daily trends for different regions. The dashboard will have different views, each corresponding to different categories (health, economy, census, etc).

- Firstly we will try **data fusion** to come with insights of interests for the problem statement.
- We will visualize weekly infections/death rate on responsive **maps** (and/or a **mosaic plot** to visualize most affected regions) and use **line charts** to show daily analysis. These will help understand how much the pandemic affected a particular region. The datasets above contain daily death/recovery/infection rates corresponding to each state/country.
- We will have **bubble plots** to show country-wise/state-wise trends in terms of infections and population.
- We will also have **parallel coordinates** visualizations to investigate the correlation between variables. These will be for adversely affected countries. Additionally we will run **PCA** on all available investigating variables to show **scatter plot matrix** for 2 PC variables and infection rate.
- The dashboard will have **heat maps** to show weekly infection count for some important and densely populated cities around the world.
- To investigate the causation of various socio-economic factors on the pandemic, the dashboard will have visualization showing correlations using **scatter plots** for different regions. Below factors will be observed:
    - Population density
    - Health infrastructure
        - Number of hospital beds
        - Number of health professionals
    - Economy
        - GDP
        - Economy Type (IT, Agriculture, etc)
    - Environment conditions (weather, pollution rate, etc.)

Technologies

The proposed system will have a python backend on the Flask server, which will respond to the request from the client. The backend will be responsible for all the data processing tasks. The client-side frontend is planned to be designed with a bootstrap template with visualizations in SVG elements designed with D3.js. The library chosen will also help to provide an interactive webpage to the scientists/users.

- Backend
    - Flask server
    - Numpy, Pandas, Scikit
- Frontend
    - HTML, CSS [Bootstrap]
    - Javascript (AJAX, D3, jQuery)
- Code Management
    - Github - https://github.com/harshit13/covid19-viz

References

- NY Times - https://github.com/nytimes/covid-19-data
- JHU - https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases
- World Bank -
    - https://data.worldbank.org/indicator/sh.med.phys.zs
    - http://wdi.worldbank.org/table/2.12#
- WHO -
    - https://www.who.int/data/gho
    - https://www.who.int/data/gho/data/indicators/indicator-details/GHO/hospital-beds-(per-10-000-population)

Contact

- Harshit - hharshit@cs.stonybrook.edu
- Siva Boppudi - sboppudi@cs.stonybrook.edu