Group Member:

Yahui Li 917797914

Bingdao Chen 917781027

Kevin Yu 913705560

Ida He 913701946

Coronavirus Tracking App

Introduction

The purpose of this project is to develop a Shiny App to visualize the coronavirus data. The app is published in https://yhli097.shinyapps.io/STA141BFinalProject/.

It is well known that there is a very big challenge for most countries in the world because of the coronavirus or 2019 novel coronavirus diseases (COVID-19) in recent weeks. The COVID-19 is a new illness that can affect people's lungs in airways. It is caused by a virus called coronavirus. The main symptoms of the 2019 novel coronavirus disease are similar to flu, such as cough, fever and shortness of breath. Due to the coronavirus is new to human this year, there are lots of infected people died at the beginning mainly in the Wuhan, China. Now the coronavirus is spreading exponentially around in the world. The number of new confirmed cases are rapidly increasing daily. Many countries declared a state of emergency. People are definitely care about the coronavirus progression, in particularly its development in the large economy in the world, such as China and the United States. This project aims to provide a comparison of the development of the coronavirus between China and United States through statistics such as the number of new confirmed cases, suspected, cured and dead. This project would visualize these kinds of information with maps and time series plots for people easy to know the newest progress of the coronavirus in the two countries. Also, this project provides a wordcloud of coronavirus related news for people to know the key information every day.

Description of Data

The source of data used in this project mainly includes three parts:

1. The information of situations of coronavirus in United States and China.

2. The geographical information of United States and China.

3.   The news of coronavirus.

The first part of data sets mainly come from the two websites
https://lab.isaaclin.cn/nCoV/ and https://en.wikipedia.org/, the first website provides the
information of situations of coronavirus in China and the second website with the title
'2020_coronavirus_outbreak_in_the_United_States' provides the information of
situations of United States. At the end, the API from first link is unstable and very slow,
we decided to use the CSV version of the same data instead. The csv file can be
downloaded from here https://github.com/BlankerL/DXY-COVID-19-
Data/blob/master/csv/DXYArea.csv. We also obtained time series data collected by
Johns Hopkins University. The link to the data is
https://github.com/CSSEGISandData/COVID-
19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_19-covid-
Confirmed.csv. Both of them including the information of number of people related with
coronavirus which contains the following items:

1). current confirmed count
2). confirmed count
3). suspected Count
4). cured Count
5). dead Count
6). province Name

All of the above items are self-explained, the number of people of the items are the key
information recorded in the two websites for this project. And note that the information is
updated every day. The first website provides APIs , but the second website needs data
scraping methods to extract the items needed for this project.

The second part of data sets are the geographical information of United States and China.
For China, there are geographical information for different provinces and for United
States there are geographical information for different states. Both of the geographical
information are formatted in json data format, the China's geographical information is
provided by the website
https://raw.githubusercontent.com/chemzqm/geomap/master/china-province.geojson and
the United States' geographical information is provided by the website
https://raw.githubusercontent.com/loganpowell/census-
geojson/master/GeoJSON/5m/2018/state.json.

The last part of the data sets is the news of coronavirus in the internet. There are various sources of such news but for this project, the news of coronavirus comes from the yahoo news website https://news.search.yahoo.com. The yahoo news contains lots of news of coronavirus and this kind of information would be mainly used for the wordcloud which gives an overview of key information of situations of coronavirus in the world.

Data Process

There are various data sources. Most of them are in the raw format without cleaning. Before the data can be utilized for the app, we need to clean and tidy the data first. For this project, the application needs to show useful and precise key information about the progress of coronavirus. Therefore, data processing is required. And there are mainly six steps for the data processing:

1). Extract numbers of desired counts (confirmed count of people and etc.) for both China and United States of coronavirus
2). Convert the json format geographical information
3). Merge the geographical information and desired counts
4). Map the coronavirus information into the maps
5). Convert date and transform the coronavirus information into time series
6). Extract words from news of coronavirus

For the first step, because for China, the APIs are provided by the source website, so the processing is much easier. We only need to query from the corresponding API to obtain the desired counts. However, the information for United States is obtained from Wikipedia. Therefore, the data processing is much more complicated. First, the data scraping is used to download the information table of the coronavirus and then with the table, extract the desired counts of coronavirus information for different states.

For the second step, as the geographical information for China and United States are stored in the json format, data processing is performed by converting the json format into a data frame for each province or state.

For the third steps, the counts information and geographical information are merged by province or state in order to display the boundaries of the province or state properly on maps.

For the fourth step, after merging the geographical information and counts information, the details of the different counts information(current confirmed count, confirmed count,

suspected count, cured count, dead count) were binded to the corresponding province or state in the maps so that the full counts information could be correctly displayed when people clicked some province or state in the maps.

For the fifth step, as besides the maps, the time series of each item of counts information are also important to people, so the past dates were processed and bind to the corresponding information and finally converted into time series. Thus, the time series plots for each item could be correctly displayed.

For the last step, as the wordcloud is much easier to show for key information of news, the words are needed to be extracted from the news, these words are the words which are shown most times in news.

User Guide

The interface of the shiny application of COVID19 UI PLATFORM is divided into two main panels - China and United States. The layout of the two main panels are similar. Each panel contains 3 main sub-panels, corresponding to maps, time series plots and summary information. Also, there is a panel under more information tab. It shows wordcloud of news for coronavirus.

To use the application, for US or China, the operating parts are similarly. Using China as an example, click the China tab and then people could select the province interested by choosing the province from the available options. Once the province is chosen, then click the China tab panel, the map, time series plot and summary information would be displayed for that province. And people could click the map to show the details of the information or directly check the time series plots and the summary information tab panels. It is the same for US, people could choose the interested state and click it then the map, time series plot and summary information would be displayed for that state in the United States panel.

At last, people might interest in the news of coronavirus, then people could click more information panel and select news interested. Also, the application provides two options for wordcloud of the news. The first option is the minimum frequency and the maximum number of words by sliding the sliders due to people's choices then the corresponding wordcloud could be shown in the Word Cloud panel.

To be noticed that it might take some extra time for the user to read the data and run the app. Thanks for the patience.