MSBX 5420

Spring 2020

Team Quandary Peak

Yuxiang Wang, Qingdi Xiao, Wen Huang, Wenbin Yang, Candice Chuang

April 28th, 2020

**Coronavirus Final Project**

**Introduction**

COVID-19 is a strain of coronavirus and has become a global pandemic from March 2020. With COVID-19 becoming a fast-moving virus, people focus on the reasons why the coronavirus is spreading in such a rapid speed and is so difficult to control. To better understand COVID-19 spreading in the past and the future in the United States, the project compared the outbreak among serious regions and the overall situation in the United States. We integrated both statistical aspect (predicting models) and texting mining technique.

**Data Source**

Three datasets have been used in the project:

* 1. Tableau COVID-19 Data Hub

(<https://www.tableau.com/about/blog/2020/3/coronavirus-data-hub-faq>)

The data comes from the dataset maintained and updated by the Johns Hopkins University Center for Systems Science and Engineering. Tableau cleans, reshapes, and makes this data ready for our analysis. This dataset updates by 9:00am EST every day. This data contains Dates and Confirmed Cases and total deaths for each state by providing for all countries over the word. This data also contains the location (Lat and Long).

* 1. Covid-19 in USA

(<https://www.kaggle.com/sudalairajkumar/covid19-in-usa#us_covid19_daily.csv>)

This dataset has details information from 50 US states and the District of Columbia at daily level. This dataset contains date, total case of states, the number of positive and negative cases, the pending cases, the number of recovered and death, and the number of deaths increasing by daily.

* 1. COVID-19 Open Research Dataset Challenge

( <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>)

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 51,000 scholarly articles, including over 40,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease.

**Process**

We uploaded data from local to EMR Cluster, and then stored it to Amazon S3 Bucket, and finally, linked with Leeds EMR Jupyter Notebook Cluster to perform the code.

A picture containing table, device, holding, blue

Description automatically generated

**Tools**

AWS

Amazon Web Services (AWS) is a subsidiary of Amazon that provides on-demand cloud computing platforms and APIs to individuals, companies, and governments, on a metered pay-as-you-go basis. In aggregate, these cloud computing web services provide a set of primitive abstract technical infrastructure and distributed computing building blocks and tools. We used Leeds EMR to save our coding files and data. We also uploaded our data from local to EMR cluster, then transformed the data to S3 Bucket. We can input data directly from S3 bucket to Jupyter File. Then we used Leeds Prod Cluster Jupyter Notebook to create Jupyter Notebook for coding and reviewing.

Spark

Apache Spark is an open-source distributed general-purpose cluster-computing framework. We used Pyspark for data cleaning and Word Cloud. The reason we used spark was because when we were working on a dataframe, we only needed to write codes for one row in spark, and then spark would assign slaves to work on each row of the dataframe instead of writing a for loop. This method optimizes a step-by-step process to a parallel process and will save a lot of time when we run the codes.

Binder

The Binder Project is a software project to package and share interactive, reproducible environments. We used Binder to link our Github Project repo to review our teammates’ codes.

Github

We used Github for merging all parts and displaying our project.

**Text Mining**

1. Topic Model

Topic modeling is an unsupervised machine learning technique that’s capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. We used “tmtoolkit” and “nltk” packages to build our model. Moreover, we combined all text into one dictionary and set up the “stopwords”, “punkt”, “averaged\_perceptron\_tagger”, and “wordnet” for directory. Additionally, we put the texts into document term matrix. Finally, we used “LDA TOPIC MODL” to extract top 16 models over all texts.

1. Word Cloud

The word count is the number of words in a document or passage of text. Our datasets contain over 51000 articles from January 8, 2020 to March 27, 2020. We cleaned data by filtering out news not in 2020, grouping news by date, removing punctuations and unwanted words by a defined function in Spark, and extracting top frequent words via another defined function in Spark. We generated a word cloud for all news in the dataset and several GIFs for daily news word cloud.

**Visualization**

First of all, a heat map of infected people would tell us which areas of the world are infected with Coronavirus. The reddish color indicated the number of infected people in this area. On April 10, the reddish places were southern North America, Western and Southern Europe, Western and Southern Africa, and Eastern Asia, which were densely populated places, so there were also a lot of infected people. The color of the United States was much deeper than other countries in the world.

In the United States, the map of infected people in the United States has changed since the remote class began. The main infected people were concentrated in the East and West coasts of the United States, such as New York, New Jersey, Florida, and California. To output which states had the greater impact on the United States, we made a plot that included 5 Insular areas and fifty states in the United States. Since there were too many infected people in New York, we first normalized this data, and draw the plot, we marked the area where the number of infected people was greater than the average as red, and the green as the contrary. New York and New Jersey had the largest number of infections and the largest influence in the United States, so these two places were also the most dangerous areas up to April 10, 2020, while these states such as Minnesota and Kansas were safer. Also, Colorado was marked as yellow, although the number was below the average, we still had to be careful.

By comparing the number of infected people on the West and East coasts of the United States, there were several extreme values on the East Coast, such as New York City, which made the Y axis of these two plots differ by ten times. After removing the extremes, more cities on the East Coast were with more than 2,000 infected people. So, we can say that the situation on the East Coast was more serious.

**Predictive Modeling**

1. Time Series

We use time series to better understand COVID 19 spreading trends in the past (before 4/10/2020) and future (after 4/10/2020).

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time and Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. When based on time transition, recording each record to a dataframe, the dataset is called “time series dataset”. When analyzing the trends according to time changes, the technique is called “Time Series”. We used ADF/ACF/PACF to get the time series for past 3 months, then use Arima Modeling and SARIMA Modeling to predict. But from the past data, we found the trends were increasing all the time. So, this point might affect the accuracy of our models.

1. Random Forest

Besides the information offered by the U.S CDC, we want to explore if there are any significant factors such as income, gdp, and education level that may potentially affect COVID 19 outbreak. We used Random Forest to the variables which has the most effects of COVID-19.

**Conclusion and Suggestion**

In the first-10-day news (January 8, 2020 to January 19, 2020), the high frequent words were Wuhan, China, Chinese, SARS, and etc. People were focusing on the origin of coronavirus and analyzing the virus explosion in specific areas. However, the last-10-day news (March 16, 2020 to March 27, 2020), the high frequent words became COVID, people, time, new, and etc. The difference means that people’s attention has been transferred from specific areas, such as China or South Korea, to the whole world, because people realized that covid-19 became a global pandemic now. Both governments have done effective actions and people have concerned about the pandemic to slow down globally spreading

Both Confirmed and Death trend keeps increasing and “Economic”, “Education Level”, “Health” issues are more important than population density. We suggest government should take actions based on those low education level states or those states with high risks of potential mental illness.

The United States is in dangerous situation. The East Coast is worse than the West Coast. Although CO is below the average, we still have to be careful. Government has done effective actions and people have concerned about the pandemic.

AWS cluster increases the productivity and has fast innovation, also increases the resource availability and strategic resource usage, really good for team works.

We would like to make streaming data to generate covid-19 data in the future to optimize our project.