COVID-19 Real-time Tracker

Chao Zhou, M.S.

Nima Zahadat, Ph.D.

The George Washington University

Table of Contents

Abstract

COVID-19 is one of the most devastating disease in human history due to its high transmissibility, wide affected age group, and its lethality to human life. The virus has already affected the whole world, leading to an ongoing pandemic. The purpose of this project was to have a real-time COVID-19 China cases tracker that employed dynamic data mining, data processing, data persistence, data visualization technique so that people can have more understanding on the effect of COVID-19 virus using real-time information access regarding the pandemic data from Chinese CDC and other government agency. This Tracker focused on web crawling from Tencent and Baidu official webpage, then data processing and storing in a relational SQL database for interactively manipulating and cleaning data. Finally, different statistical models and machine learning techniques are used to try to see the future trend of COVID-19, and they are demonstrated via data visualization depending on any modern web browser.

Keywords: data mining, data scraping, flask, MySQL

COVID-19 Real-time Tracker

COVID-19 pandemic is one of the most disaster disease in human history. It is still an ongoing pandemic worldwide that lasted over almost a year and a half. Symptoms of COVID-19 are variable, however, often include cough, headache, short of breath, even loss of smell and taste. These symptoms may happen in a time spanning from one day to fourteen days after exposing to the virus. Many people who get infected were not even aware of that they already caught the virus. Luckily, most people only have mild symptoms after getting it, however for elderly people, the symptoms are often very severe and fatal. However, in some countries, there are still many people do not believe the deadly effects caused by COVID-19. The aim of this project is to produce an unbiased data visualization in order to reveal the more factual side of this pandemic in China. In this article, COVID data from China's government agency and other third-party company is used together to form a more diverse dataset. At the same time, this project is also carried out with the core principle of data science field and demonstrate a complete life cycle of a data science project. In the beginning, dirty data, also known as raw data, is gathered and processed through a data pipeline. Then, exploratory data analysis is done upon it for building a more specific question on the raw data. Finally, data is ingested and stored in a database for later use and the whole life cycle is repeated as necessary. Although the big picture is vague right now, it will become clearer in later section. Each section header will tell a big theme that will be discussed. There are mainly 7 sections in this article after this introduction.

LITERATURE REVIEW

A useful COVID-19 real-time tracker must employ several principles to make sure the information displayed is as accurate as possible. Data visualization is an increasing popular

method for communicating data effectively, at the same time, many scientists does not have the chance to be trained systematically to make good use of the technology (Stephen 2020). In order to make good unbiased visualization, we need to understand the main learning media in forms of human senses. According to Stephen, visual learning is one of the most dominant way of understanding information which has been rapidly reformed by newly updated technology support. Many journals now even require graphical illustration so that later the article may be advertised using those figures. Due to the increasing popularity of visualization of information, there rises many challenges for doing neutral data visualization with the least bias. Data visualization is powerful and also very dangerous if it is used in a malicious way. According to Elisabeth, different people have very different interpretation and expectations even on the same visualization (Elisabeth 2016). The visualization system should not standardize the interpretations of data to only one direction, instead, it should incorporate ambiguity and uncertainty in the data, so the graph won't persuade its audience into single certain interpretation of data. In this project, these principles are always used throughout the whole production process to produce an unbiased and neutral yet very beautiful and interactive data visualization tool.

RESEARCH METHODOLOGY

Python 3.7 is the main programming language used in this project. Many of the common packages were used during this process. The discussion for those packages will be explained in detail in later paragraph. Due to the ongoing pandemic, there are also newly generated data every single day. In order to display and analyze data interactively, data scraping from other news site is necessary instead of having an up-to-date cleaned data. After data being scraped from website, the data is stored in a MySQL database, then it gets preprocessed for Exploratory Data Analysis later.

Pie Chart, Line Chart and data visualization on a map are also utilized in this process. Finally, data analysis was used to produce a conclusion rely on the previous procedure. The whole life cycle is repeated if necessary.

TECHNOLOGY INVOLVED

As mentioned above, Python 3.7 is the main programming language being used in the whole process. This section will expand on the purpose of each package involved in the process so that the work can be easily replicated if needed. A 3-step process is utilized in the data processing which are data ingestion, data processing, and data analysis and visualization. In the data ingestion period, webpage data is first identified and get scraped from Tencent and Baidu using beautifulsoup, urlib package, then the data is processed using standard library like pandas, NumPy and statlib. Now the data is ready for processing and storage using pysql package. Finally, pandas, echart and flask library are used to build a webpage front-end and back-end.

DATA ANALYSIS

Tencent and Baidu actually used data from Chinese CDC for their website. The data is constantly generating new data samples every day. In order to capture the change persistent in time, data scraping and storing into a database is necessary. Since the data is highly tabular, relational database is used instead of a non-relational one. Because the data is already very structed and relational database deals with this kind of data very well. It is very fast to do query in a SQL database. In the beginning, there are multiple datasets in different formats. Those formats vary from csv file to h5 file. They are aggregated using python and Jupyter notebook and in the end form a single final dataset can be used for further exploration. In this final dataset, there are over

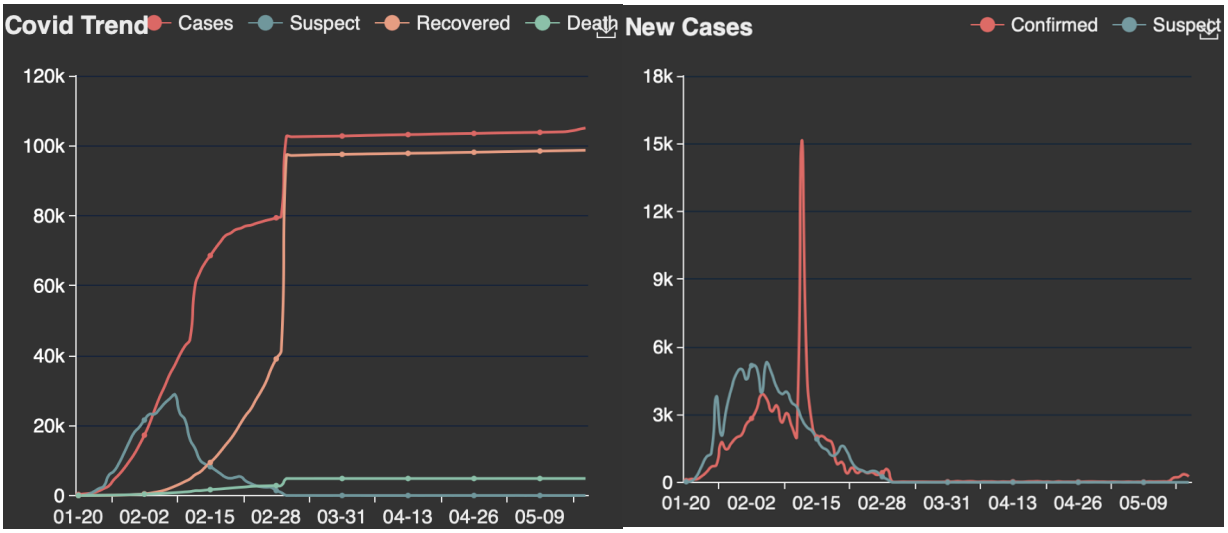
20 factors describing many different aspects of COVID pandemic in different cities and provinces in China. There are daily case confirmed, daily case recovered, daily death, their location and daily uncertain cases. In the appendix, there will be illustration to make things clear. The data from CDC provides a foundation for this paper, so that those datapoints can combine to tell a story.

Many data visualization technique is applied in the front-end of the html page. For example, in the line chart, several different dimensions of the data appear in the same chart, however, you can pick the one or multiple dimensions you would like to explore to enforce the neutrality of the data visualization. Color consistency is also carefully considered throughout the whole project theme. The line chart and COVID is also interactive. It encourages its user to interact with the transformed information. There is also a word cloud made out of daily news buzz words related to COVID. In the middle of the demonstration page, total cases, total recovered, total cured case number are displayed so that anyone visits the page can see those number in the first place. The code is publicly available at https://github.com/Nateczhou/Covid_Realtime_Tracker.

KEY FINDINGS

The EDA process provides many interesting insights into pandemic in China. On the top left corner, we can clearly see number of COVID cases were increasing so rapidly at the beginning of pandemic simply because the country needs some reaction time to deal with the suddenly

appeared virus. However, when time passed February, the total number of cases increased so little



that completely flatten out the later line chart. This is due to the really restricted rule in China, China's policy on stopping traffic between any cities works as a result derived from this visualization. In the middle of the page, there is also an interactive map in the middle, each province got colored with different brightness individually according to its number of cases, the darker it gets, more Covid cases it has. We can clearly see which province have the need to be immediately treated, we can also observe the epicenter of this pandemic. By using these visualizations, we can always see the real-time case number to have a grasp on the current state of the pandemic in certain, which could help not only government, but also individual citizen has a more accurate sense of the status of the pandemic in present.

RECOMMENDATION

The realization of the real-time COVID-19 dataset pattern from data visualization could be improved by adding a more intuitive user interface and organizing the chart in a much more logically smoother way so that people who use this system can identify key information from the vast amount of data without too much struggle. Another creative way to add more complexity to

the future improvement of this project is to add natural language processing (NLP) to extract vast information from social media, different sources of news to produce a more interesting project. This project can also be extended vertically which is going beyond one country.

CONCLUSION

COVID-19 indeed is a widespread and fast spread virus. If those visualization could be used when the COVID just started, people may be able to treat this virus more seriously in time, so that more lives could be saved, and less lives need to sacrifice unnecessarily.

By looking at the visualization in the EDA process, more insights get discovered so more insights can be dug into by using the previous EDA outcome. This iterative process provides a more scientific way for exploring data which raw number on a paper cannot provide. Therefore, deeper question would be asked after the first iteration of data science work. When abnormality in a line chart has been seen, more question could be asked. For example, why is the confirmed case line completely flatten out after around March, why is there a huge spike in February? What's the story behind it? During this iterative data science life cycle, more and more interesting insight can be dug out, and more underneath fact will float above the water surface.

To conclude, this COVID real-time tracker obviously can be improved much further, however, data ethics should be always remembered in people's mind when producing a potentially high impacted project. Therefore, the insight from massive data is not only accurate, neutral but also more ethical.

BIOGRAPHY

Chao Zhou is a graduate student in the Data Science Program at The George Washington University. His interest includes machine learning, artificial intelligence, deep learning, data visualization. He really likes to learn state-of-the-art technology related to computer.

References

Crystal L., Tanya Y., Gabrielle I, Graham M. J., Arvind S. (2021, March 1). The Data Visualizations Behind COVID-19 Skepticism. http://vis.csail.mit.edu/covid-story/

Emanuele G., David A. (2020). COVID-19 Data Hub. https://joss.theoj.org/papers/10.21105/joss.02376.pdf

Stephen R. M. (2020). Principles of Effective Data Visualization. https://doi.org/10.1016/j.patter.2020.100141

Elisabeth K. D. (2016). Deceptive visualizations and user bias: a case for personalization and ambiguity in PI visualizations. https://doi.org/10.1145/2968219.2968326

Thirumalaisamy P. V., Christian G. M. (2020). The COVID-19 epidemic. Trop Med Int Health. 2020 Mar; 25(3): 278-280

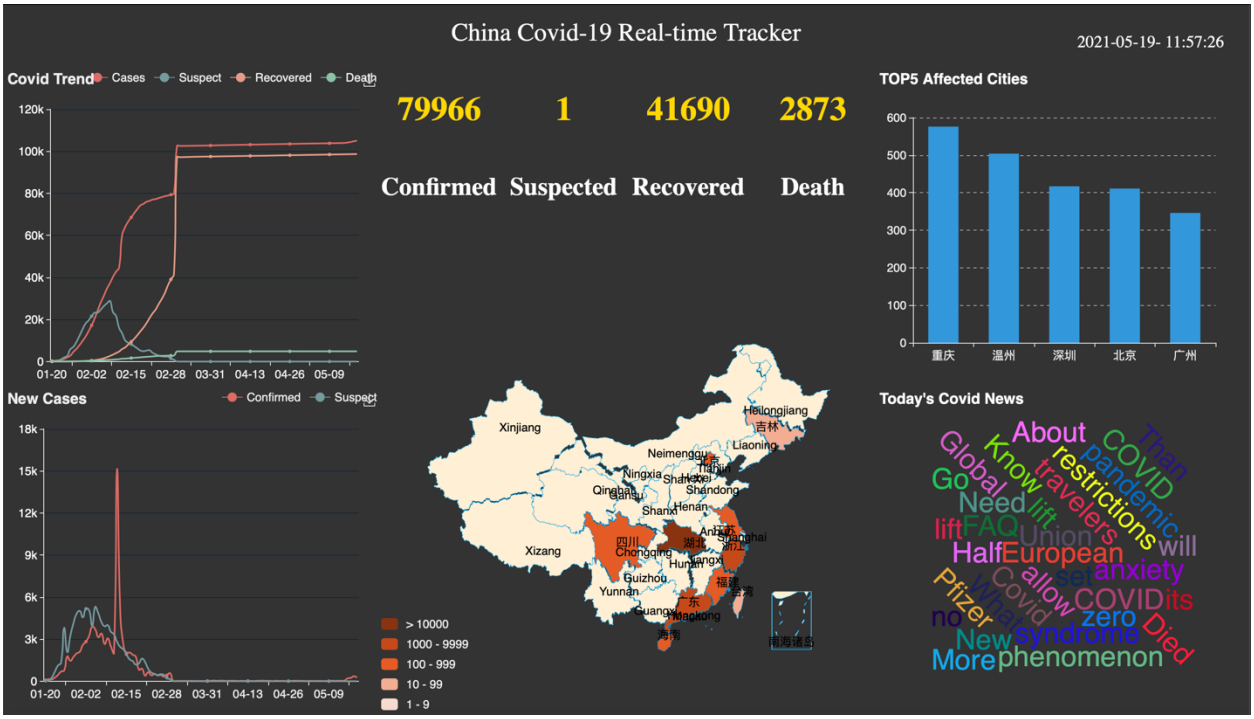
Baidu. (2021). Echarts. https://echarts.apache.org/en/index.html

Flask. (2021). Flask Guide. https://flask.palletsprojects.com/en/1.1.x/

Heng et.al. (2020). Coronavirus disease 2019 (COVID-19): current status and future perspectives. PubMed

CDC. (2021). COVID-19 Databases and Journals. https://www.cdc.gov/library/researchguides/2019novelcoronavirus/databasesjournals.html

Figures


China Covid-19 Real-time Tracker
2021-05-19- 11:57:26
Covid Trend
Cases
Suspect
Recovered
Death
79966
1
41690
2873
Confirmed
Suspected
Recovered
Death
TOP5 Affected Cities
重庆
温州
深圳
北京
广州
New Cases
Confirmed
Suspect
Today's Covid News
> 10000
1000 - 9999
100 - 999
10 - 99
1 - 9