In this project, I found a dataset (fig. 1) about the covid-19 in the US. This dataset records daily COVID-19 cases and deaths in each state from January 22, 2020, till now. The size is 187.99 MB.

After that, I used Python, PySpark, PySpark SQL, and PyEcharts technologies to visualize the data of the new crown epidemic in the United States. I set myself six PySpark SQL questions and solved them all.

Graphical user interface

Description automatically generated

Fig.1

The first step was to clean and convert the dataset from a csv file to a txt file. (fig.2)

Text

Description automatically generated

Fig.2

Then, I used PySpark to cluster the data and created a data frame. I created different data frames by using PySpark SQL to retrieve, filter, and sort data. Because I needed to visualize the data, I saved the data frames in a .json format. (fig.3)

Text

Description automatically generated

Fig.3

At last, I used a powerful and easy-to-use package called “PyEcharts” to visualize. I saved all the results as a .html file so that we can easily review them (screen shoots of codes are below).

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Conclusion:

I think this project is quite challenging and significant to me. I think SQL and big data are inseparable. However, in our lectures, the knowledge of this part is relatively minor. So I hope that when doing this project, I can study spark SQL by myself. More importantly, I used a completely new package for this class. Although this project did not involve very complex models, I am still proud of myself.