Hello everyone, I’m Anyu, I’m excited to present my final project here. Since I’m already back in China, I’m not sure about the stability of the network. Sorry for the inconvenience, please let me know if anything goes wrong.

My project mainly contains two parts: the first part is the visualization of covid-19 data, and in the second part, I tried to develop a deep learning model for prediction based on the something new I’ve been learning this semester.

This figure shows the number of countries affected over time, we can see that the coronavirus is affecting most of the countries in the world. It was first reported in China in early December and then discovered by the increasing number of countries worldwide through February and March. By the date of May 2nd, 2020, there are 212 countries reported more than 364 million confirmed cases in total.

Here display two different forms of heatmap of confirmed cases worldwide. The first one is generated by Tableau and the second one is generated by Python. Personally, I prefer the second type because for some countries with smaller number of confirmed cases, using circles is not that obvious to observe and compare. Both of this are designed to be interactive, but I haven’t figured out a way to show it on my slides. We can see that Europe and the United States are hit by the virus most severely currently.

The following are the heatmaps of the United States. The first one is generated by data of each state, and the second one is generated by data of each county: we can see that the district hit by the disease most is the east coast, west coast, and area around the Great Lakes Region. This can be explained by the population density and public transport mobility in these regions. States like Florida has a lot of population mobility during March, considering the spring break. States like California, Illinois, New York had more international transportation and higher population density.

These two figures show the top ten countries with most daily confirmed cases and daily death cases. China and South Korea were the first to experience the outbreak and are now out of the picture due to strict control. From the first figure, we can see European countries’ cases start to increase in late February, and countries like UK, Spain, Italy are showing signs of decreasing the number of new cases. Russia and Canada, whose situation used to be quite stable, have many increasing new cases only since recently, the reason for which is due to either delay of testing or lack of control over international transportation. In the United States, the size of daily increases surpassed all the other countries since the middle of March, and there are still no signs of dropping. It starts to stabilize recently, maybe because the limit of test capacity, or because social distancing policy start to show effect. I noticed that the pattern of increase in the United States has a periodic decrease every five or six days. The reason might be reducing testing ability during weekends, I guess.

From the figure of death cases, we can see that in Europe, France, UK have a turbulent developing trend; Some European countries like Italy, Spain, and seems already passed the highest point of daily increase in death cases. The United States currently has the highest number of death cases worldwide and seems stopped the exponential trend this week, but the trend is quite unstable. From the update of the daily data, I noticed some States do not report data on a daily basis now, maybe that’s one of the reasons why the curve is in a unstable shape.

Since China is regarded as the first country to report the spread of COVID-19, this figure compared the number of confirmed cases in China and the rest of the world. The figure shows that the peak of the pandemic in China happened in February. The lock-down of the central city Wuhan happened on Jan 23𝑡h, and in February, due to the change of diagnosis criteria of confirmed cases, China experienced the most significant number of increases in daily confirmed cases. The number keeps dropping ever after, which indicates the effectiveness of the lock-down policy. On March 14𝑡h, Italy became the first country to have more confirmed cases in China. The world’s cumulative total is still increasing at an exponential rate and not yet reaching its peak. The length of disease spreading period differ from China and other countries, which can be explained by policies: the ’City lock- down’ policy applied by China directly cut the route of transmission at fastest speed, while ’Flatten the curve’ policy implemented by the other countries mainly slow down the process in a limited degree.

In this part, I analyzed the testing abilities of different countries. This figure calculates the cumulative total number of tests, the average number of tests per million people, the mortality rate, and positive rate of each country. The color depth indicates the quantity and level of severity. Due to limited data access, I have 17 countries listed here. We can see that among all these countries, the United States has the most significant cumulative number of tests but have a fewer number of tests per million people. Countries like Bahrain and Austria have a large number per million due to the small number of population. Germany, Italy, and South Korea are also running many tests based on the population scale.

Positive proportion is also an indicator of the sufficiency of total tests. We can see the Philippines, Indonesia, France, United Kingdom, Canada are having high positive rate which is above 40 %, combining the number of tests per million people in these countries, the result indicates the actual number of affected population might be more significant than what it is recorded.

In this figure, I generated a plot of the world’s mortality rate and mortality rate in each continent. After a drop in February, the mortality rate of the disease keeps increasing, mainly due to the shortage of healthcare resource and the spread of disease in countries which do not have robust public health systems. The mortality rate in Europe is the highest and is still increasing, which partly can be explained by the structure of the population. The outbreak in nursing homes in Washington State can explain the first peak of North America. From the figure, we can see that the mortality rate in South America keeps increasing, which is mainly due to a lack of medical resources. In Africa the mortality rate starts dropping, partly due to the high temperature.

The next part is developing the deep learning model for prediction.

On the left is the structure of the network, and the right part are the input features. There are two kinds of input features. Temporal inputs are time series data. we used a Multi-Stacked Fully Connected Bidirectional LSTM(Long Short Term Memory) model. LSTM is known as developed to capture the high dependency in time series. And bidirectional networks are proved better than the traditional LSTM in many fields capturing time-series information. (By stacking LSTM layers, potential temporal correlation can be considered. ) Demographic inputs are not related to time, so it goes directly into dense layers. The model output are number of confirmed cases and fatalities on the 14th day.

（Dense Layer：fully-connected layers, function as classifier.

* In the data preprocessing part, based on the data provided by JHU, I added the demographic features including population distribution by country and age; smokers percentage by country, and indicator data of quarantine, restrictions, and schools.

Here I generated the correlation matrix plots, the deeper the color is, the higher the correlation between the two variables is. Red means a negative correlation, while blue means a positive correlation. These correlation matrices give us strong evidence that the features and data we used to predict confirmed cases and fatalities are meaningful and significant.

The correlation among Confirmed cases, fatalities, and percentage of the population with ages larger than 60 years old is shown in figure.1. This plot shows that number seems to be not related to the portion of the population with ages larger than 85 years old. It may because the total number of old people that are older than 85 years old is tiny. Thus, I did not consider the percentage of the population with ages larger than 85 years old in the model.

The correlation among Confirmed cases, fatalities, and percentage of the population with ages 0-59 years old is shown in Fig.2. We can see that the confirmed cases and the fatalities are almost not correlated with the population with ages from 30 to 40 years old. They are negatively related to the percentage of people ages from 0 to 30 years old. It intuitively makes sense since young people have a stronger immune system.

The correlation among Confirmed cases, fatalities, and other features introduced above are shown in Fig.3. From the correlation matrix, we can conclude that the confirmed cases and the deaths are positively related to the population and density of each country. A country/region with more population has more chances to have a more infected population and more fatalities. Confirmed cases and fatalities are negatively related to quarantine policy applied for the area. If a school open or not, restrictions used for the region and the number of hospital beds of each country.

To evaluate the accuracy of the model, we use rooted mean squared error.

Next is the result of the model. The first fig shows the loss function of training and testing. The second one shows the results of the prediction of the number of confirmed cases for the 14𝑡h day. We see that the prediction of the daily confirmed cases in the test data set is precise except for a very steep peak around the 40th sample. It may because some countries/regions change their test policy or improve their testing ability that day. We may consider the steep peak as an outlier in the data set. If we have some more information about the policymaking of each country/region, we could create some new features that could improve the prediction results.

Fig.18 shows the results of the prediction of the number of fatalities for the 14𝑡h day. We see that prediction results are quite precise when we predict the fatalities on the 14𝑡h day.

We can see that prediction of the fatalities is much more precise than confirmed cases in the 14𝑡h day. It is intuitively reasonable since a country/region can change its policy of testing the infected population or improving their testing ability within a few days. However, the percentage of patients in serious conditions would not change too much. The ordinary patients would not die immediately. Thus, the variation in the number of fatalities is not very big.

Now the following is some of my thoughts. Whether deep learning is suitable for predicting disease. Using deep learning for disease prediction does not consider the features of the disease. Compared with traditional epidemiology models, like SIR, SEIR, it ignores features of disease itself, like R0, transition rates, recovery rate etc, thus it can only be applied for real-time update, and still require more features for future trend prediction. It might be useful for the decision-makers of each country to adjust their decisions and policy based on real time prediction.

Another question is about the difference of policy applied in China and some other countries. In China, we all know the ‘lock down policy’, which shows immediate effectiveness, almost functions like: stop the curve, but almost everything shut down in the country: like factories, transportations… In some other countries, the ‘flatten the curve policy’ preserves some of the basic functions of the society, which helps people kind of live a normal life, but actually stretched the spreading period of the disease. I think which policy is better actually depends on different countries, with different economic situation.