MA615 Final Project

Danping Liu, Hao Shen, Haoqi Wang, Yuxi Wang

2020/12/14

Initial setting

- The database can be downloand from cowtransfer (https://mssp2020.cowtransfer.com/s/8c75ea77b46944)

```
# load dependency
dbpathT='Covid-tweets-en.db'
dbpathR='Covid-reddit-en.db'
connT=dbConnect(SQLite(),dbpathT)
connR=dbConnect(SQLite(),dbpathR)
```

1.Introduction

As the pandemic continues, we want to find out how do people's thoughts change. What do they care about? Do they have good feelings or bad feelings? What we have done is to build an tool to compare the relationship among keywords trending on twitter, people's sentiment scores and COVID-19 spreading.

Our tool will serve as an exploratory tool, so it will allow users to explore the data by themselves, giving them a general idea of people's considering trends, and helping them to solve their problems, like making the business decision or looking for a research topic.

Firstly, we have got the covid related tweet IDs and detailed tweet information. Due to big dataset, we stored them in a database by SQLite. Secondly, we did simple text mining, in this part, we have done the sentiment score calculation, keyword frequency statistics and reverse geocoding. By using the data and text mining, we draw the interactive plotting and mapping. And we have deploied a shiny app to show this.

Moreover, we get the tweets data to compare with the COVID-spreading through Reddit, which is the social news aggregation, web content rating and discussion website. There is no interference of official account, and we can get dataset easily, which is also the Us-based dataset.

2.API Tools

2.1 Twitter API

Since we only have a standard Twitter developer account which has a limitation of downloading data in the last seven days. So, we chose to download the list of tweet IDs related to COVID-19 from Kaggle, which has a huge ID number and is a json format and then used Twitter APIs to look up the details about these tweets. After data collection and cleaning through SQLite and R, we get the first database which has all the tweets data.

2.2 Bing API

When we draw the map with the number of the tweets, we need to get the geo location. So we reverse geocoding with Bing APIs in order to get the longitude and latitude, which is the location that the person creates the tweet. Then establish another database with geographic information.

The reason why use database 1. We can access to large Tweets data set faster. 2. We can get pre-processing information. 3. SQL for fast frequency calculation than R. 4. the use of SQL is more Convenient to add more data and Convenient for further analysis.

3.Sentiment score calculation

People use Twitter to share their interests and concerns. Word frequency analysis shows what topics they are interested in, and sentiment analysis focuses on how they think about it.

So we decide to measure the sentiment by a sentiment score. What we have done is we break the tweets into words firstly, then we use the sentiment lexicon dictionary to tag each of the word with its sentiment, checking whether it is positive or negative. The sentiment score for each tweet is equal to the number of positive words in this tweet minus number of negative words divided by the total number of sentiment tagged words. And the score is between -1 and 1. -1 is completely negative, 1 is completely positive, and 0 is neutral, or mixed. The sentiment score will be sent back to the database for future analysis, so each tweet has a new variable, sentiment score now in our database.

The overall sentiment score is for a bunch of tweets that contain a keyword or several keywords in a period of time. This process will proceed when we actually draw the line plots on the shiny app. So when the user choose the keywords, we use sql to get the the selected tweets and calculate their overall sentiment score, either daily or monthly. Then we use the daily or monthly overall sentiment score to draw the line plot, and see the trend of people's sentiment for a given keyword, or keywords.

In the same way, we can get another 3 function to get data from two Databases:

Get tweets trend function

Get reddit data function

Get reddit trend function

4.Trendplot function

This function is used to define all our drawing functions.

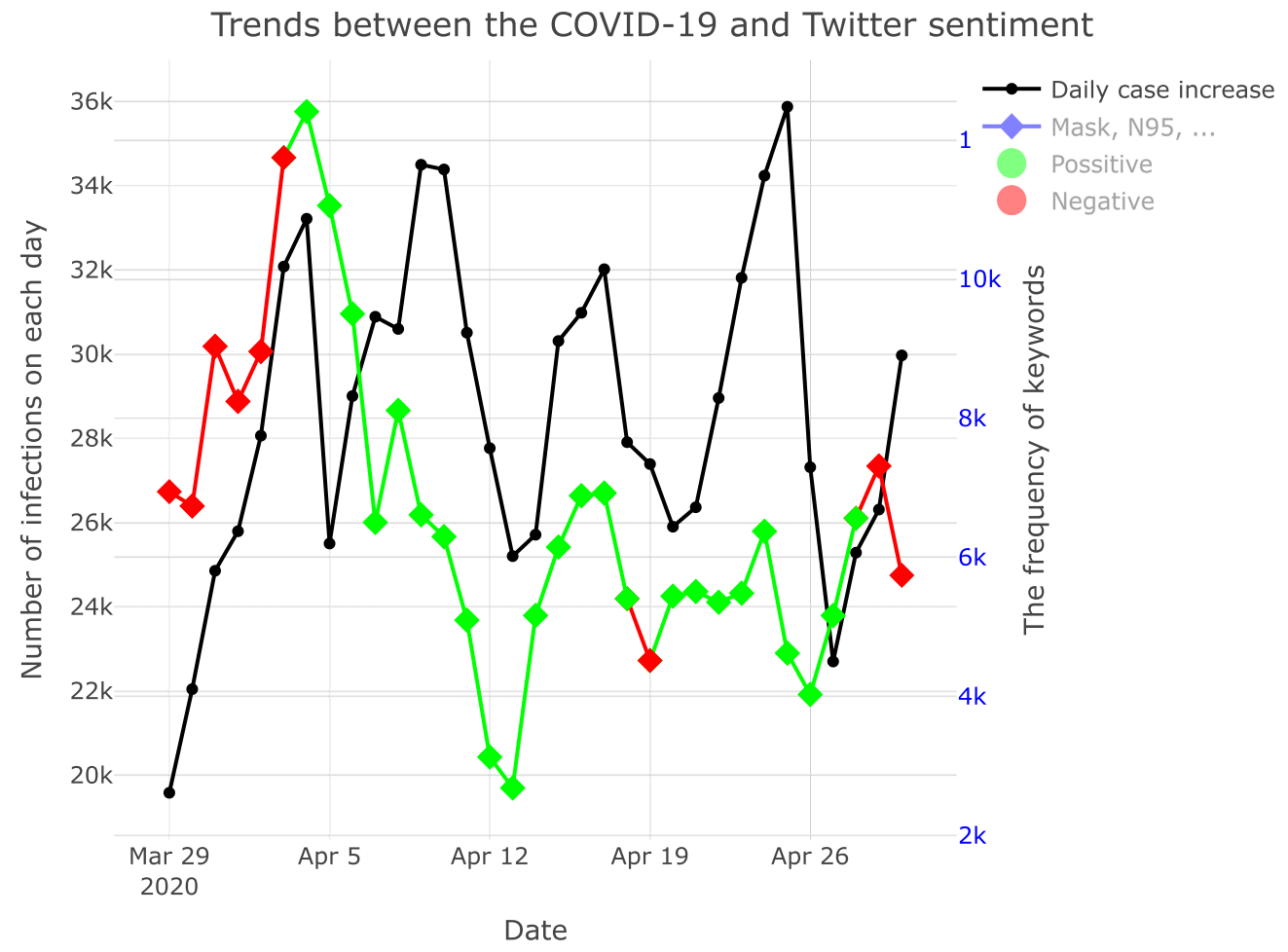
5. Visualization

5.1 Normal Tweets trends

Load Twitter data. And use this data to make plots.

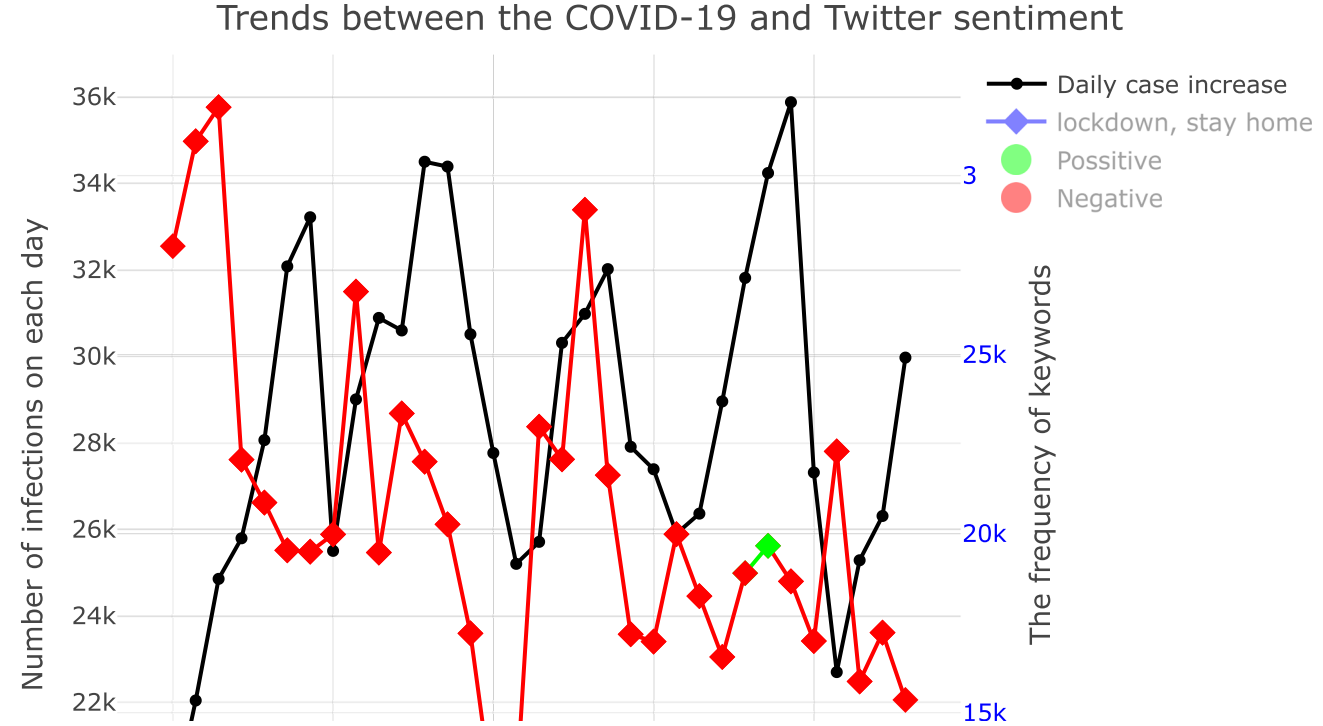
```
# spread data
covid=read.csv('us_covid19_daily.csv')%>%select(date,positiveIncrease)
# a list of groups of keywords
keywords1='Mask#N95#口罩'
keywords2='lockdown#stay home'
keywords1=keywords1%>%str_split('#')%>%.[[1]]
keywords2=keywords2%>%str_split('#')%>%.[[1]]
keyword=list(keywords1,keywords2)
# a list of groups of data
trend1=keyword[[1]]%>%{getTwitterTrend(connT,geoinfo=NULL,keywords=.,period=NULL)}
trend2=keyword[[2]]%>%{getTwitterTrend(connT,geoinfo=NULL,keywords=.,period=NULL)}
trend=list(trend1,trend2)
```

```
trendPlot(covid,keyword[[1]],trend[[1]])
```



Take the plot of input ‘mask’ and ‘N95’ as an example. We can see that the black line represents the number of people infected with the new coronavirus that day. The other line represents the word frequency of mask and N95. In addition, green means positive of this keyword’s sentiment in one day, and red means negative of this keyword’s sentiment in this day.

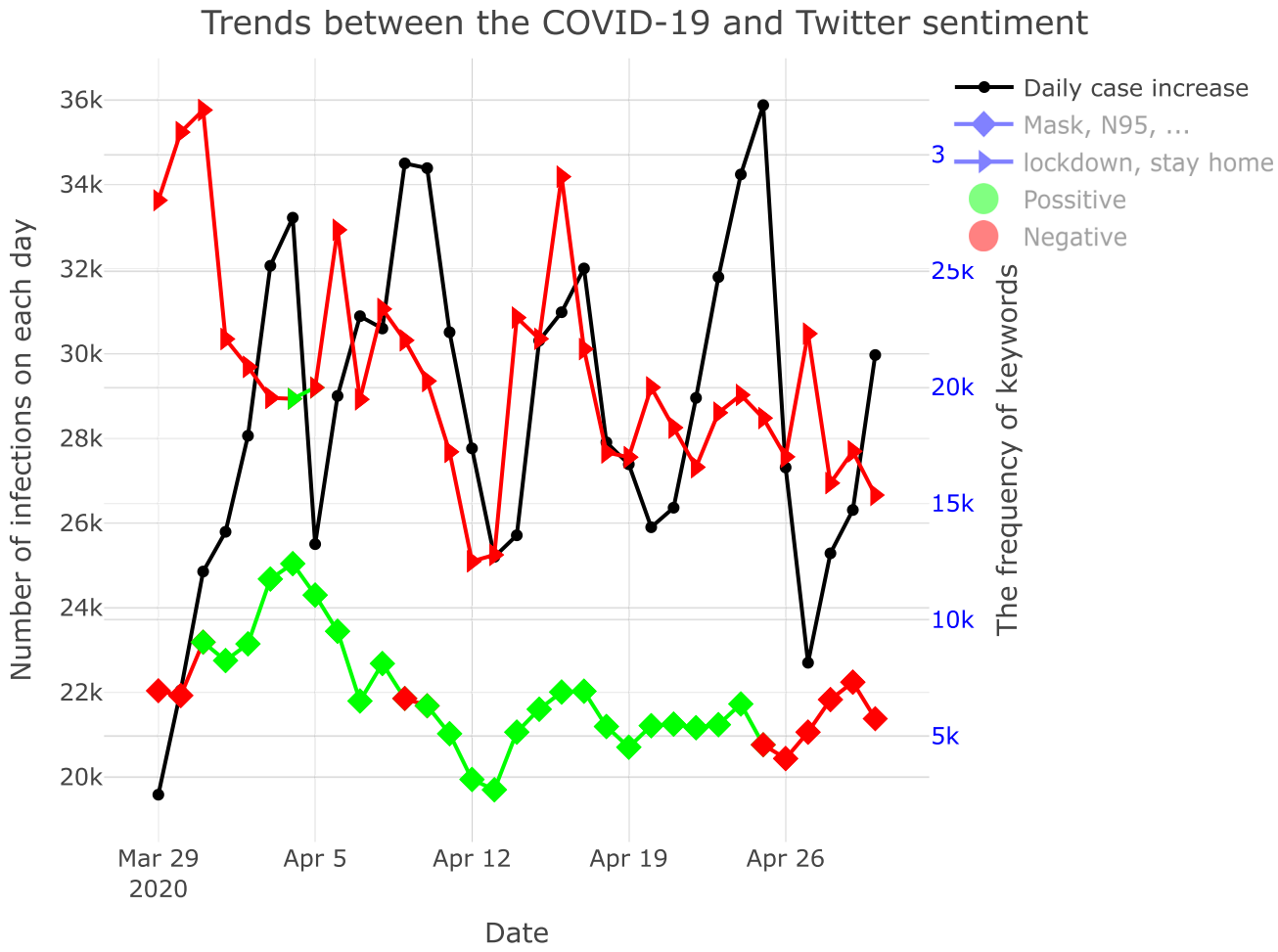
```
trendPlot(covid,keyword[[2]],trend[[2]])
```



Take the plot of input ‘lockdown’ and ‘stay home’ as an example. We can see that the black line represents the number of people infected with the new coronavirus that day. The other line represents the word frequency of lockdown and stay home. In addition, green means positive of this keyword’s sentiment in one day, and red means negative of this keyword’s sentiment in this day.

```
trendsPlot(covid,keyword,trend)
```

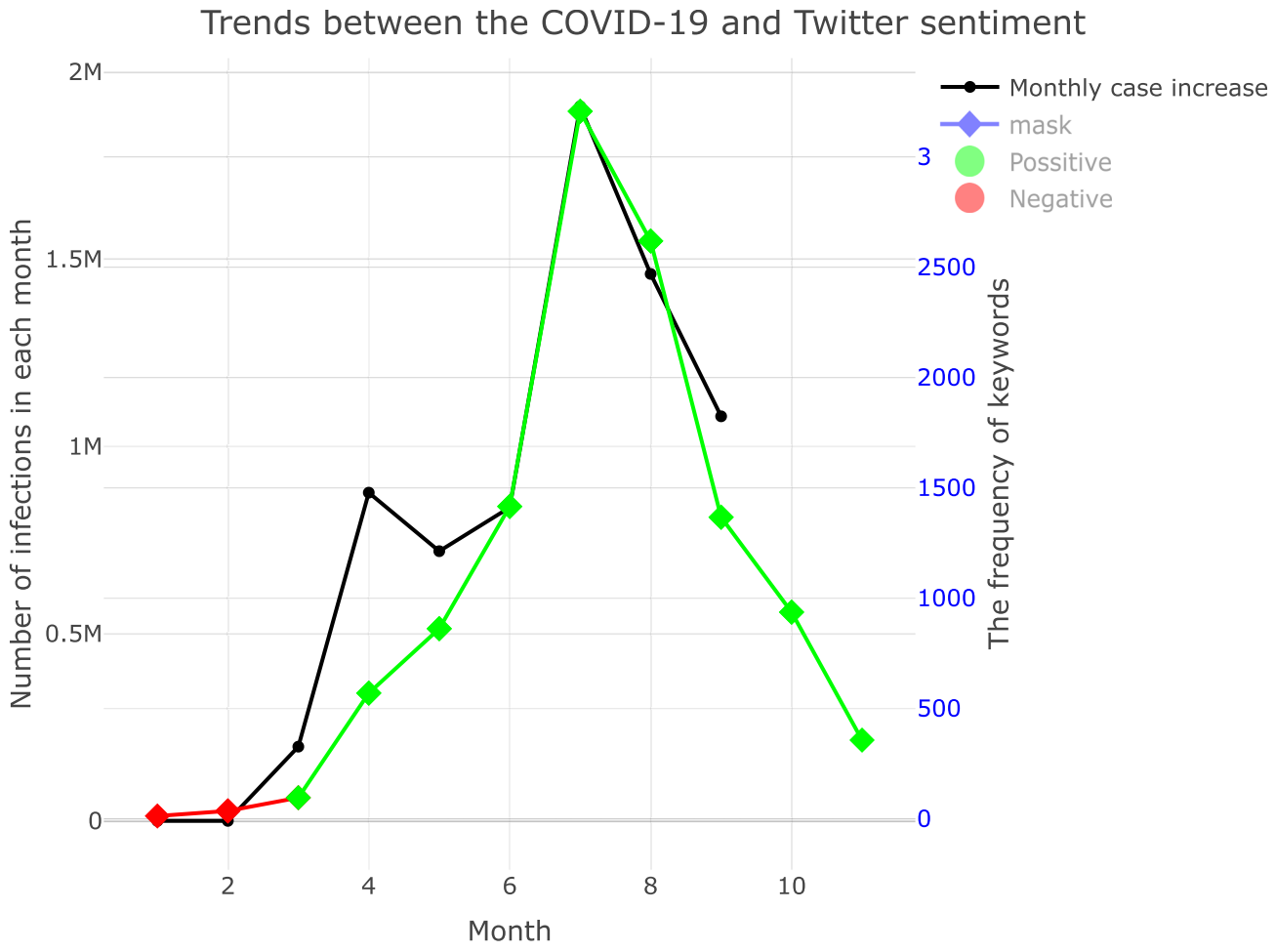
```
## Warning: Can't display both discrete & non-discrete data on same axis
```



From this plot, we can see there are three lines here. The dark line means the number of infected people every day. The diamond line means the word frequency of the keyword one which is ‘mask’. And the triangle line means the word frequency of the keyword one which is ‘mask’. The second and third lines mixed two colors. Green means positive of this keyword’s sentiment in one day, and red means negative of this keyword’s sentiment in this day.

5.2 Geo Tweets trends

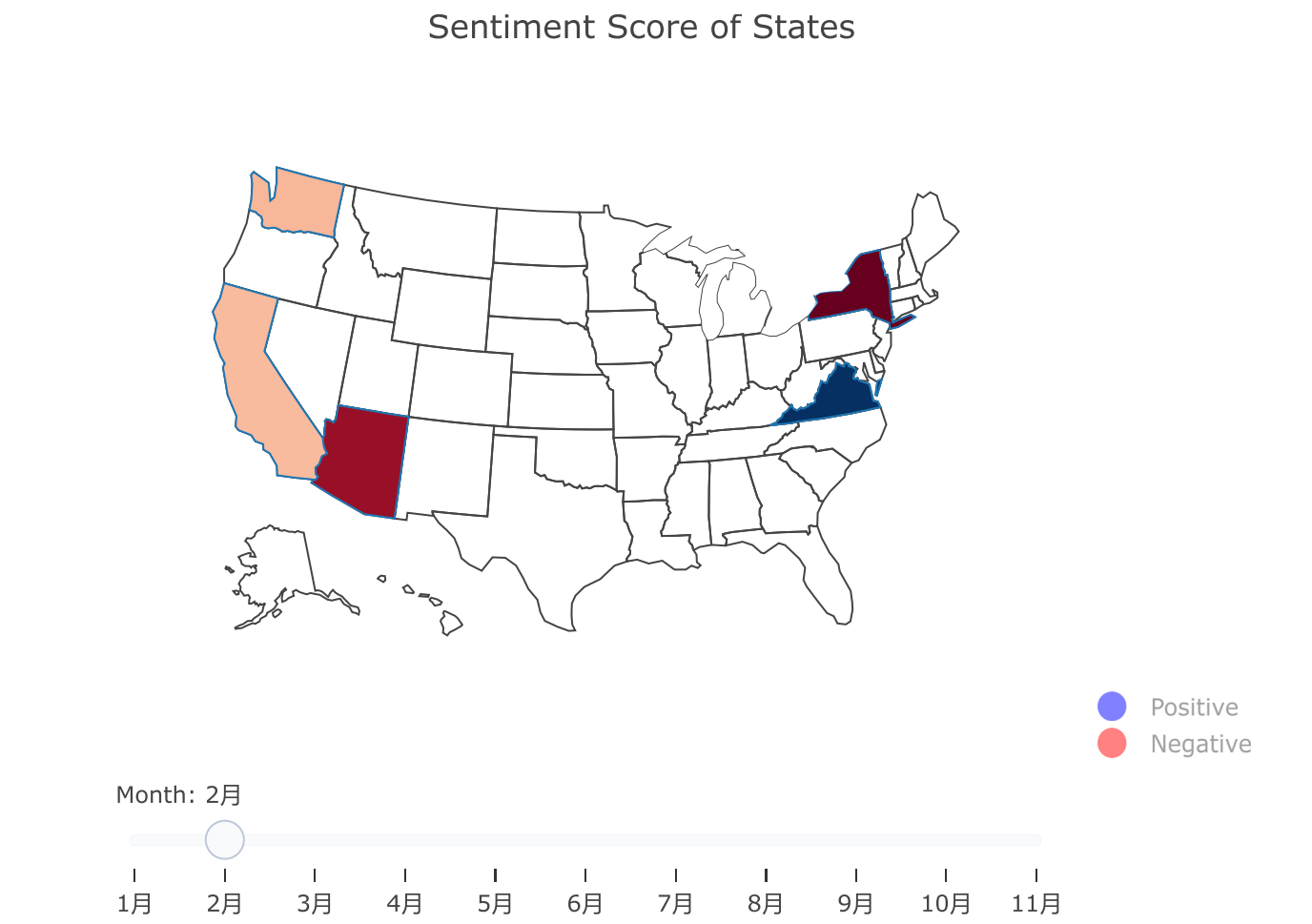
```
# spread data
covid=read.csv('us_states_covid19_daily.csv')%>%
    select(date,positiveIncrease,state)%>%
    mutate(month=month(ymd(date)))%>%
    {aggregate(positiveIncrease~month,.,sum)}
# a group of keywords
keyword='Mask#N95#口罩'
keyword=keyword%>%str_split('#')%>%.[[1]]
# a group of data
trend=keyword%>%
    {getTwitterTrend(connT,geoinfo='country',trend='month',keywords=.,period=NULL)}%>%
    filter(country=='United States')%>%
    mutate(month=as.integer(month))%>%
    select(-country)
# example
geoTrendPlot(covid,'mask',trend)
```



In this plot, we can see the monthly frequency of mask-related words and the number of virus infections. The purpose of drawing this picture is to see the trend of virus infection and the trend of word frequency changes over a long period of time. And after understanding the situation of the epidemic situation and keywords each month, it will help to understand the mapping.

5.3 Geo Tweets map

```
# spread data
covid=read.csv('us_states_covid19_daily.csv')%>%
    select(date,positiveIncrease,state)%>%
    mutate(month=month(ymd(date)))%>%
    {aggregate(positiveIncrease~state+month,.,sum)}
# a group of keywords
keyword='Mask#N95#口罩'
keyword=keyword%>%str_split('#')%>%.[[1]]
# a group of data
trend=keyword%>%
    {getTwitterTrend(connT,geoinfo='state',trend='month',keywords=.,period=NULL)}%>%
    filter(country=='United States')%>%
    mutate(month=as.integer(month))
# example
geoTrendMap(covid,trend)
```



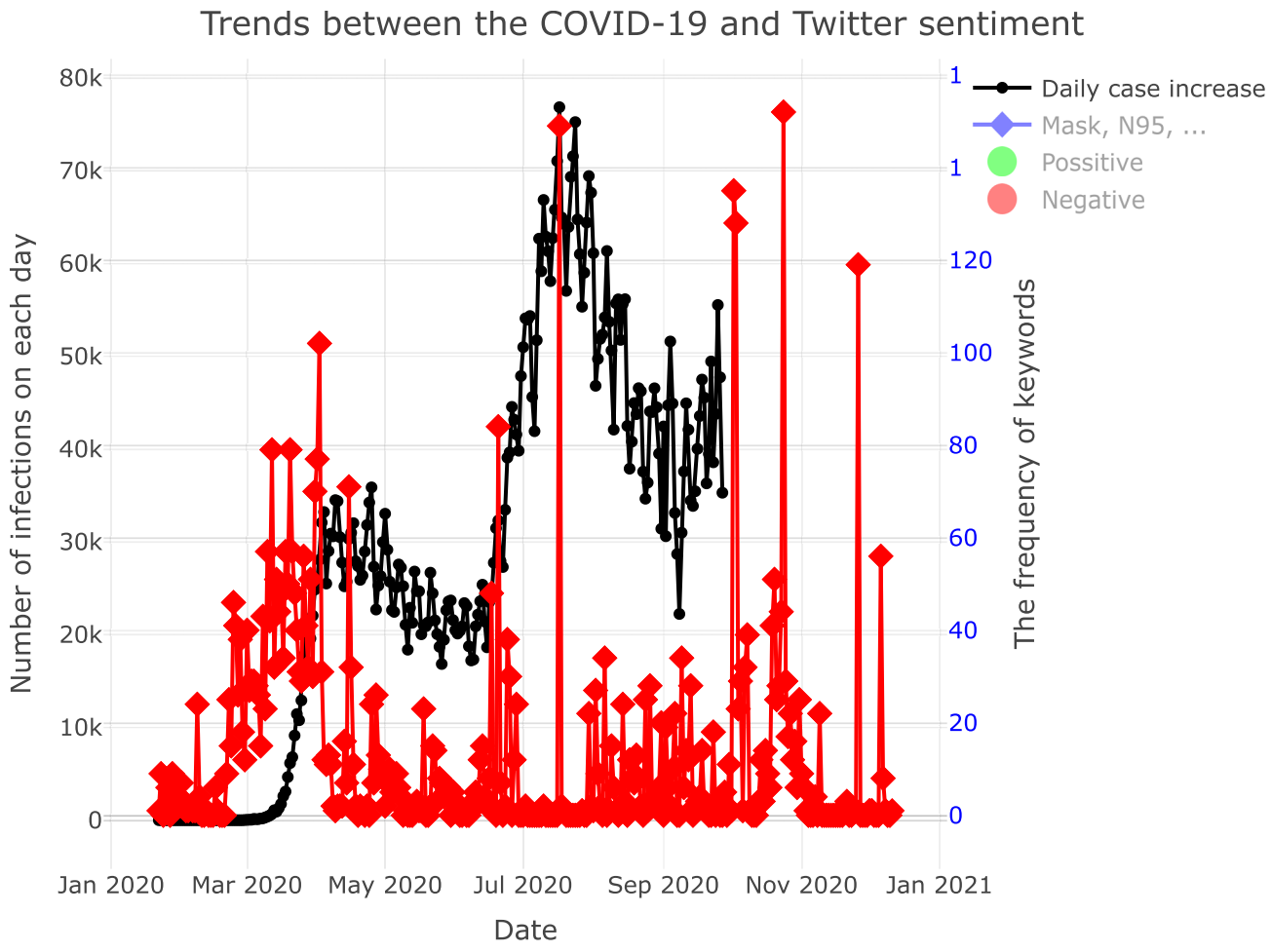
We conducted a plot of tweets each month, and also get some geographic information of these tweets in our dataset from January to November. We can slide the time bar to get the information of each month. When the mouse is placed in the area of this state, we can see the name of this state and the number of tweets collected in this state. Also, we can see the sentiment score in each state. For the color scale, dark red means the overall sentiment of this state's tweets is very negative. Sky blue means that the overall sentiment of the state's tweets is positive. The second information is that the number of infections and deaths in a specific day.

5.4 Normal Reddit trends

Load Reddit data. And use Reddit data to draw plots.

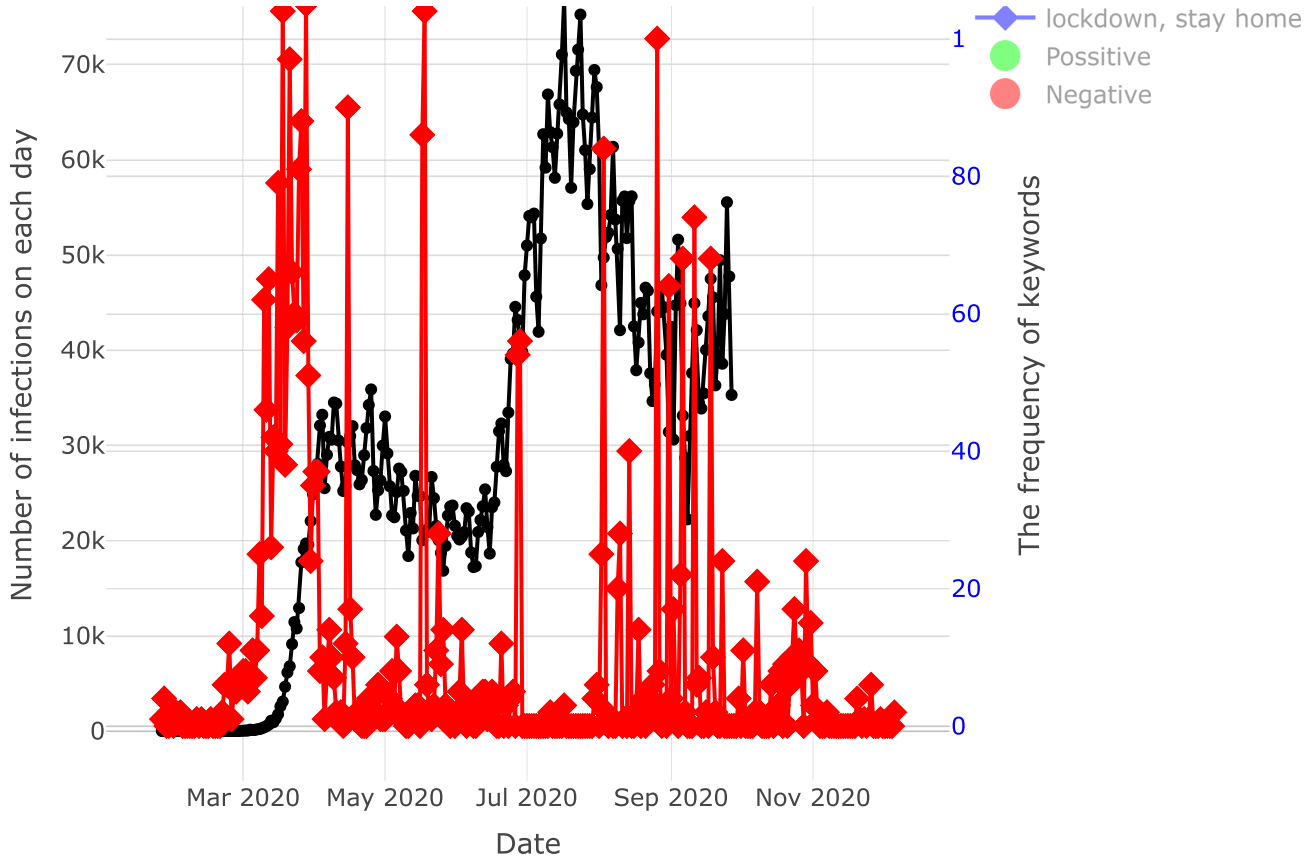
```
# spread data
covid=read.csv('us_covid19_daily.csv')%>%select(date,positiveIncrease)
# a list of groups of keywords
keywords1='Mask#N95#口罩'
keywords2='lockdown#stay home'
keywords1=keywords1%>%str_split('#')%>%.[[1]]
keywords2=keywords2%>%str_split('#')%>%.[[1]]
keyword=list(keywords1,keywords2)
# a list of groups of data
trend1=keyword[[1]]%>%{getRedditTrend(connR,keywords=.,period=NULL)}
trend2=keyword[[2]]%>%{getRedditTrend(connR,keywords=.,period=NULL)}
trend=list(trend1,trend2)
# examples
```

```
trendPlot(covid,keyword[[1]],trend[[1]])
```



Take the plot of input 'mask' and 'N95' as an example. We can see that the black line represents the number of people infected with the new coronavirus that day. The other line represents the word frequency of mask and N95. In addition, green means positive of this keyword's sentiment in one day, and red means negative. In the plot, you can see that there are only red lines, which means that the sentiment of the tweets related to the mask is negative every day. But what needs to be emphasized here is that the amount of Reddit data we use is small. This may be related to the number of users of the software.

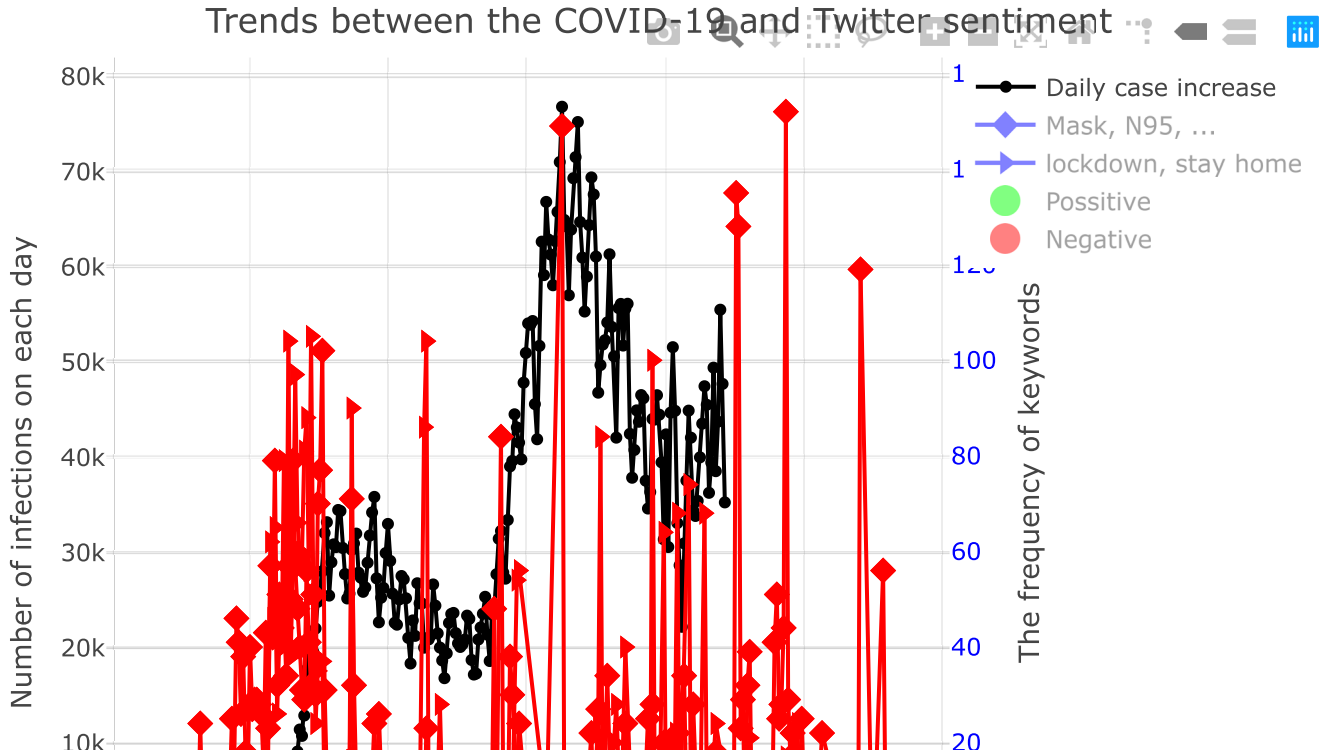
```
trendPlot(covid,keyword[[2]],trend[[2]])
```



Take the plot of input ‘lockdown’ and ‘stay home’ as an example. We can see that the black line represents the number of people infected with the new coronavirus that day. The other line represents the word frequency of lockdown and stay home. In addition, red means negative of this keyword’s sentiment in this day. In the plot, you can see that there are only red lines, which means that the sentiment of the tweets related to the mask is negative every day. What needs to be emphasized here is that the amount of Reddit data we use is small. This may be related to the number of users of the software. This is why the volatility in the graph is so large.

```
trendsPlot(covid,keyword,trend)
```

```
## Warning: Can't display both discrete & non-discrete data on same axis
```



From this plot, we can see there are three lines here. The dark line means the number of infected people every day. The diamond line means the word frequency of the keyword one which is ‘mask’. And the triangle line means the word frequency of the keyword one which is ‘mask’. The second and third lines mixed two colors. Red means negative of this keyword’s sentiment in this day. The volatility in the graph is large. What needs to be emphasized here is that the amount of Reddit data we use is small. This may be related to the number of users of the software.

```
##           used (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells 1444415 77.2    2656211 141.9  2656211 141.9
## Vcells 3004801 23.0    8396560  64.1  8396560  64.1
```