CSE587 DIC Lab 1 Part 3

Introduction

This part of the lab includes the gathering of tweets and how we use the Google API to get the latitude and longitude of each tweet based on their geo-location or the potential polygon where it was tweeted from. We will use the rtweet, fiftystater, ggmap and ggplot packages to successfully be able to gather the required tweets and then use ggmap to find the required the geographical latitude and longitude of the tweets that contain the data that can help us find these locations. Then we use ggplot in synchronous with fiftystater to plot the frequency of tweets per state for the 50 states of United States of America.

Process

We gathered tweets using the rtweet package that can gather up to 18000 tweets at a time and you can set an until parameter which can allow us to search tweets for a specific date. Using the right keywords, we run a search to gather all the tweets. Once the process has completed, we saved it to a data frame. Then using a built-in function in rtweet called lat_lng, we find the most accurate latitude and longitude of each tweet. This function uses the (coords coords), (geo coords) or (box coords) in that priority order to find the latitude and longitude depending on which one is available and find those specifications out the of each tweet. If none of those parameters are available then it will set it null (shown as NA on R Studio). Once we have all this information. We want to start data cleaning, because google API will not take empty location coordinates.

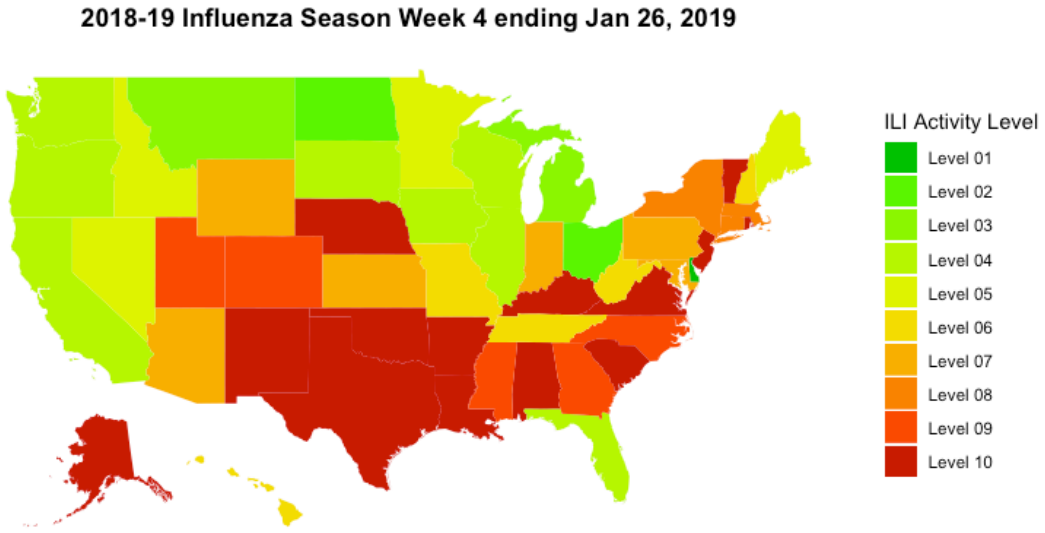
First, we do a simple is.na check and see which tweets actually contain the location coordinates and store those in a different data frame. Once we have this set, we run the revgeocode function on the geo-tagged tweets data frame, to find the potential full address for each tweet. In the return JSON from google API, the state is stored in the same list that has a type of administrative_level_1. Therefore, we have to run a loop search on the JSON to find which list contains that type and then retrieve the state from that list and store it in another column in the data frame. Once we have the states for all the geo-tagged tweets, we now create another data frame that will contain the state names and the count per state. We then load the geo tagged tweets data frame into the state count data frame where it will add the count per state every time it matches the state name in the geo tagged tweets frame.

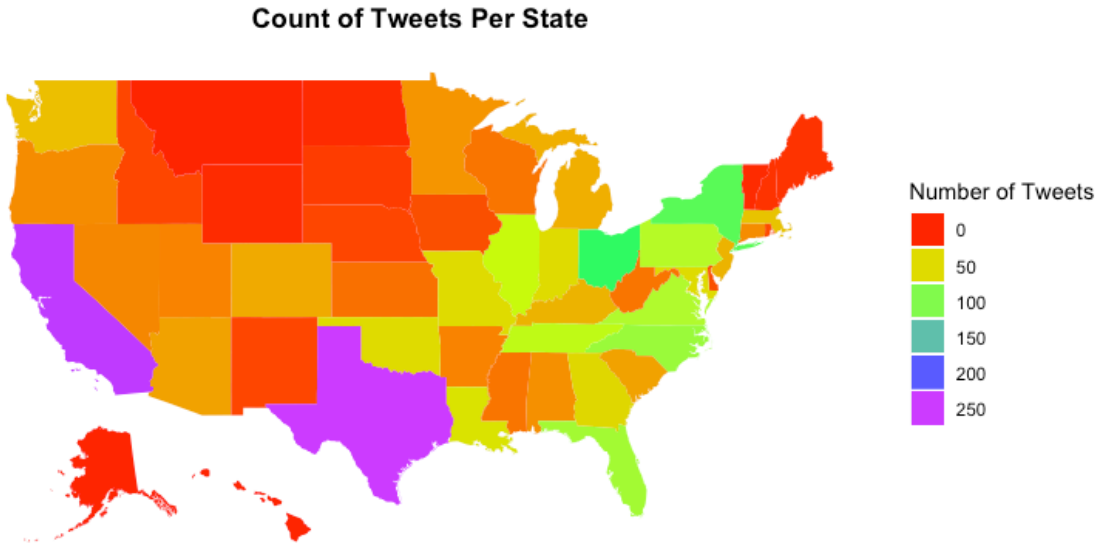
Once we have this data, we know use ggplot with fiftystater package to plot the US map. In addition, geom_map and coord_map is using to accurately plot the boundary lines in each state and then we plot the state count data frame into this map. Then, using the scale fill attributes we setup a gradient scale that can color each state based on the count of tweets per state.

Once we have the geo tagged tweets, we want to find out how many tweets has specific keywords. “Flu” is one of the main keywords, therefore we use a grepl function which will search the text field for the keyword and pull all the relevant tweets and we save it in new data frame. We do the same with “pneumonia” as the second keyword since it the next diagnoisis after flu.

Analysis

The map below is the CDC Heatmap that show the activity for flu like illness in each state, up to Week 4 of 2019. We can see that Texas and its neighboring regions have a high-level activity compared to some other states. In addition, we can see some states have very low activity which is propagated into their tweet counts, because there is no problem, there is no discussion.

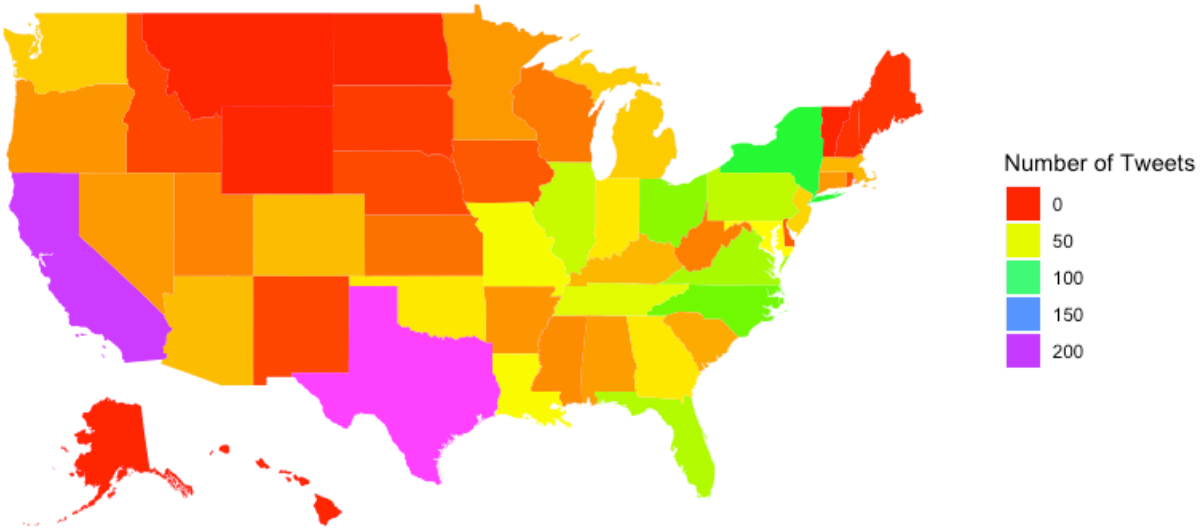




The above maps are the total counts per state. When we compare this with the CDC heatmap which is based on the number of outpatients that were admitted across the country due to flu-like illness, we can see that where the activity level for flu was high is where the number of tweets were high as well for that period. For example, Texas has one the highest states with tweets and even though its neighboring states have lower amount and this could be from many of those residents living in and near the Texas border or otherwise have a lower tweet frequency than other states.

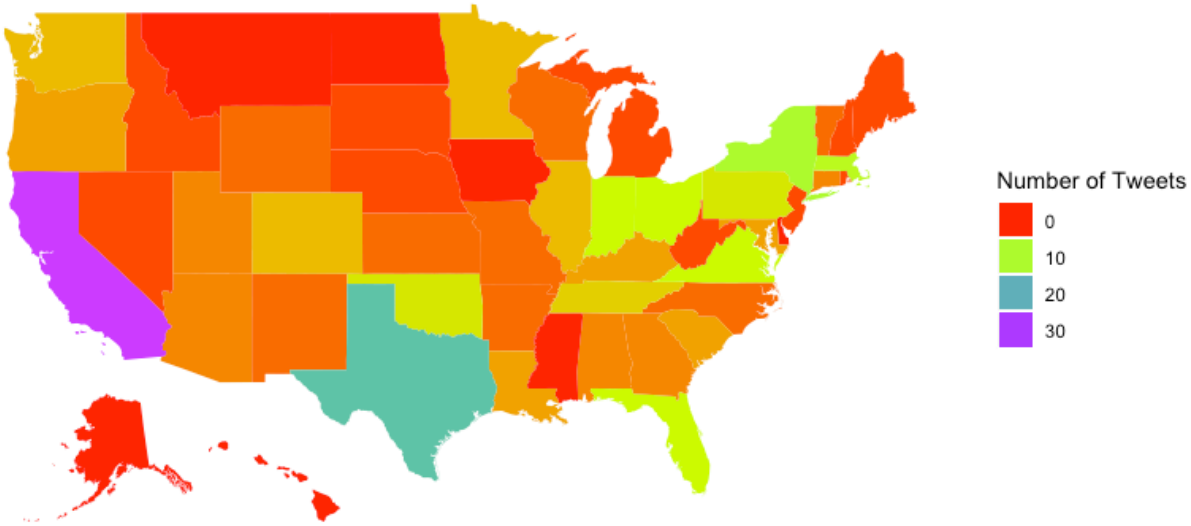
In a pool of about 24,000 tweets, we were able to get 2085 USA based tweets, which is only 8% of the tweet base. This heatmap almost reflects the CDC Heatmap, as it shows that the state with high activity level is the one with high tweet count.

Count of Tweets Per State with 'flu' Keyword



The above image shows the heatmap where “flu” is the keyword and similar to the previous image we can see that Texas is one of the key states with highest tweets. This is because of the intensity of flu in Texas causing people to talk about it. In addition, with “flu” being a keyword, about 86% of our tweets had this word at 1800.

Count of Tweets Per State with 'pneumonia' Keyword



The above image shows the heatmap where “pneumonia” is the keyword. Pneumonia is important because at this point of flu you are visiting the hospital and getting medicines. Therefore, this data is very close the heatmap. This is a serious condition and therefore many people might not want to tweet about it, however it is important to know how serious the flu is getting in some states. We were able to gather 224 tweets from the geo-tagged tweets, which is about 10%.

Conclusion

To conclude, we learn that where there is a problem or a disease spread like flu, there is enormous activity in that area. In addition, we can recognize that where is not flu problem, there is not much discussion online about it. On the other hand, we learnt how to draw multiple heatmaps with different data and examine the results and check that they were as expected. Please find the Shiny App for this at https://rakshitmtannupri.shinyapps.io/App-1/