Final Project

TWITTER GEOLOCATOIN BASED ON @RBLOGGERS DATA

Haiwei Zhou | MA 615 | Dec 12 2016

Introduction

R-Bloggers.com is a blog aggregator of content contributed by bloggers who write about R (in English) and gives daily news and tutorial about R. The site helps R bloggers and users to connect and follow the "R blogosphere". They also have a twitter @Rbloggers. In this report, I will show the process and results of exploring Rbloggers' twitter data using twitter API and R studio. The main work is datamining from the user's tweets, followers and some relative factors, making analysis and giving geolocation maps. The report contains three parts as following

- Tweets Analysis: Using R to explore the tweets of the user '@Rbloggers'
- Followers Analysis: Focusing on followers' data and geolocation information
- Hashtag Analysis: Searching relative hashtags, plotting maps and making analysis.

In this project, the idea is firstly mining the user's tweets to find the important information such as top topics, high-frequency words and users' sentiment. Then use the high-frequency words as the key word and hashtag to search relative tweets. Meanwhile, the user's friends and followers are also good data to show the network of the user. We can also show this network on the map.

User’s Tweets Analysis

TEXT MINING PROCESS

1. Extract tweets and followers from Twitter website with R studio and twitter package.
2. Clean text using tm package including removing punctuations, numbers, hyperlinks and stop words.
3. Analyze high-frequency words and plot the word cloud.
4. Analyze topics using ‘topicmodels’ package
5. Analyze sentiment using sentiment140 package

EXTRACT TWEETS

After setting up twitter and successfully connected in R studio. I use ‘twitteR’ package to extract the user ‘@Rbloggers’ tweets. The following are the code and some tweets examples. The ‘userTimeline’ function can generate a list of tweets. For further analysis, I use ‘twListToDF’ to generate a dataframe.

```r
#Get users' data of Rbloggers
Rbloggers <- getUser("Rbloggers")
#Use timeline to retrieve users' tweets
tweets <- userTimeline("Rbloggers",n=3000)
head(tweets)
```

```
## [[1]]
## [1] "Rbloggers: I set up a new data analysis blog https://t.co/mY0zj
kfBa4 #rstats #DataScience"
##
## [[2]]
## [1] "Rbloggers: Chaos, bifurcation diagrams and Lyapunov exponents w
ith R (2) https://t.co/JkgI8LIXnv #rstats #DataScience"
##
## [[3]]
## [1] "Rbloggers: Three Shiny Apps to Celebrate the Beauty of Maths ht
tps://t.co/uIJ0Echc5K #rstats #DataScience"
##
## [[4]]
## [1] "Rbloggers: How to weigh a dog with a ruler? (looking for transl
ators) https://t.co/52P0Rj6wXQ #rstats #DataScience"
```

```r
tweets.df <- twListToDF(tweets)
```

CLEAN TEXTS

Use tm package to clean texts and retrieve words and frequency. The first step is to clean tweets we get. Below is an example to show the original results. We randomly choose a number, say 19 and write down the #19 tweets we get.

```
# print tweet #19 and make text fit for slide width
writeLines(strwrap(tweets.df$text[19], 60))

## Basic Tree 1 Exercises https://t.co/VDWS7lWHzC #rstats
## #DataScience
```

Then use the following code to clean tweets by following steps:

(a) Convert letters to lower case
(b) Remove URLs
(c) Remove anything other than English letters
(d) Remove stop words such as "I, is, the, a".
(e) Remove extra space

```
# build a corpus, and specify the source to be character vectors
myCorpus <- Corpus(VectorSource(tweets.df$text))
# convert to lower case
myCorpus <- tm_map(myCorpus, content_transformer(tolower))
# remove URLs
removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeURL))
# remove anything other than English letters or space
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeNumPunct))
# remove stopwords
myStopwords <- c(setdiff(stopwords('english'), c("r", "big")),
                 "use", "see", "used", "via", "amp")
myCorpus <- tm_map(myCorpus, removeWords, myStopwords)
# remove extra whitespace
myCorpus <- tm_map(myCorpus, stripWhitespace)
```

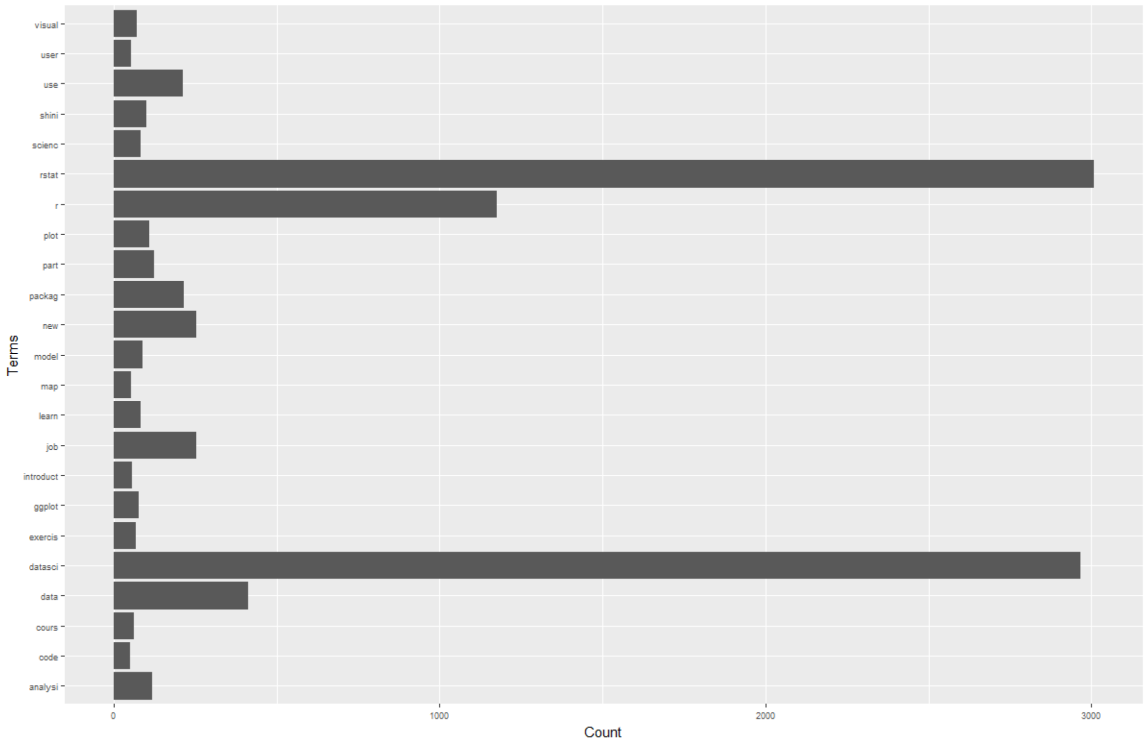
We can show the result to compare with the words without cleaning.

```
writeLines(strwrap(myCorpus[[19]]$content, 60))

## basic tree exercis rstat datasci
```

After cleaning tweets, we can do some basic analysis.

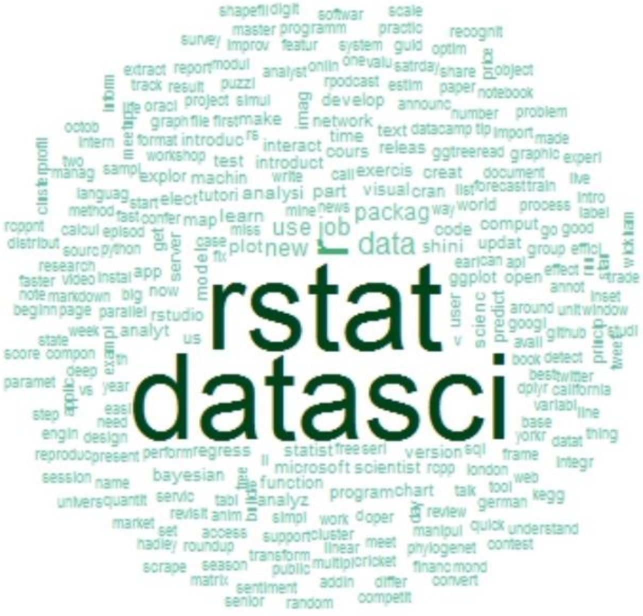
WORD FREQUENCY

The word frequency analysis includes three parts:



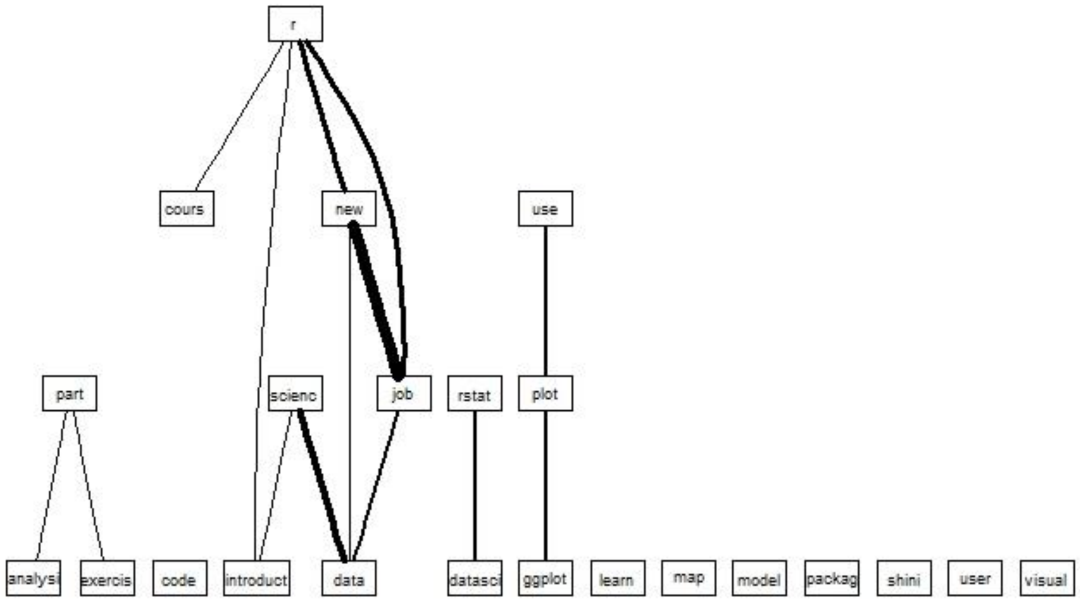
Graph 1: Word Frequency

I use “wordcloud” package to generate a word cloud based on the frequency of words in the tweets and it is shown as graph 2. We can find that the two largest words, rstat and datasci have highest frequency. I will use the hashtag #rstats to do further analysis.



Graph 2: Word Cloud

We can also plot the network of the words based on their association. The thicker the line is, the more connection the two words have.



Graph 3: Words network

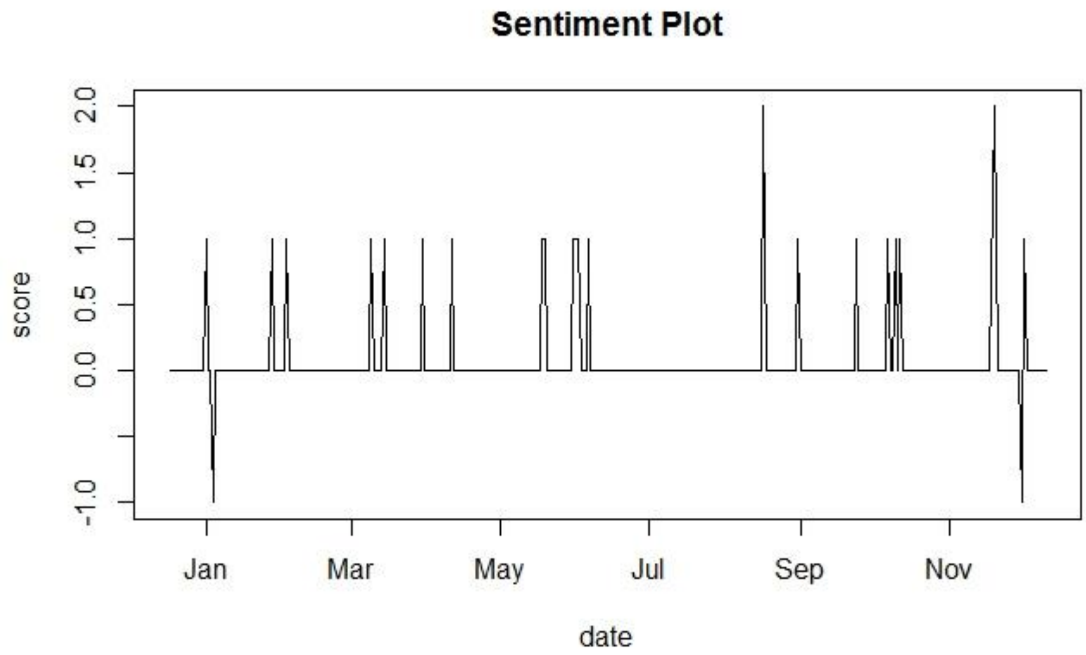
SENTIMENT ANALYSIS

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral. We use the package 'sentiment' to make some basic analysis of user's 3000 tweets.

Positive	Neutral	Negative
25	**2971**	**4**

Table 1: Sentiment results

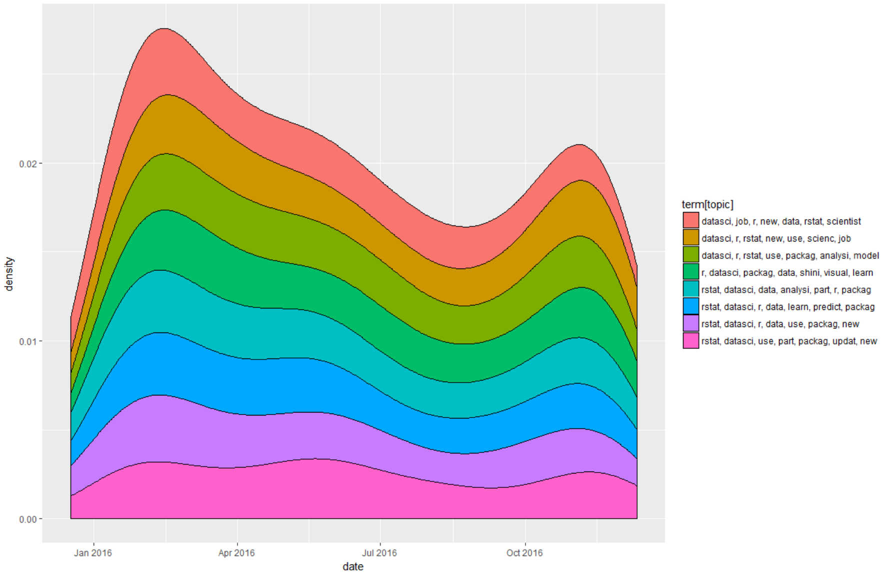
We can also show the sentiment change by time. Now assign 1 to positive, 0 to neural and -1 to negative. Plot the score response to date.



Graph 4: Sentiment Plot

TOPIC ANALYSIS

The last step I do for the user's tweets is the topic analysis, which is used to find the most popular topics based on the words we get. Using the package 'topicmodels' to get the plot below. According to the graph, we can find that the most popular topic is "data science, job, r, stat, scientist, new and data". It is highly accorded with what R-bloggers do every day.



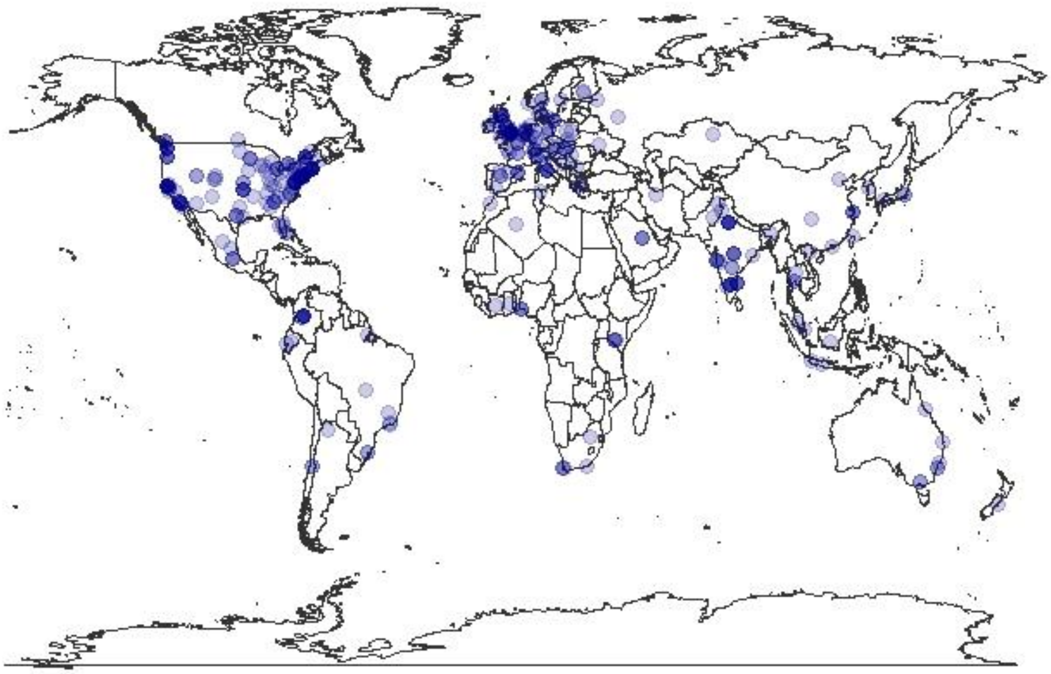
Graph 5: Topic models

Followers Analysis

Since R-bloggers is a twitter who shares r news and tutorials every day, it is interesting to find its followers who are interested in R and has high possibility to work as a statistician. Thus The next part is extract followers' data to do some analysis and draw the maps.

FOLLOWERS' GEOLOCATION

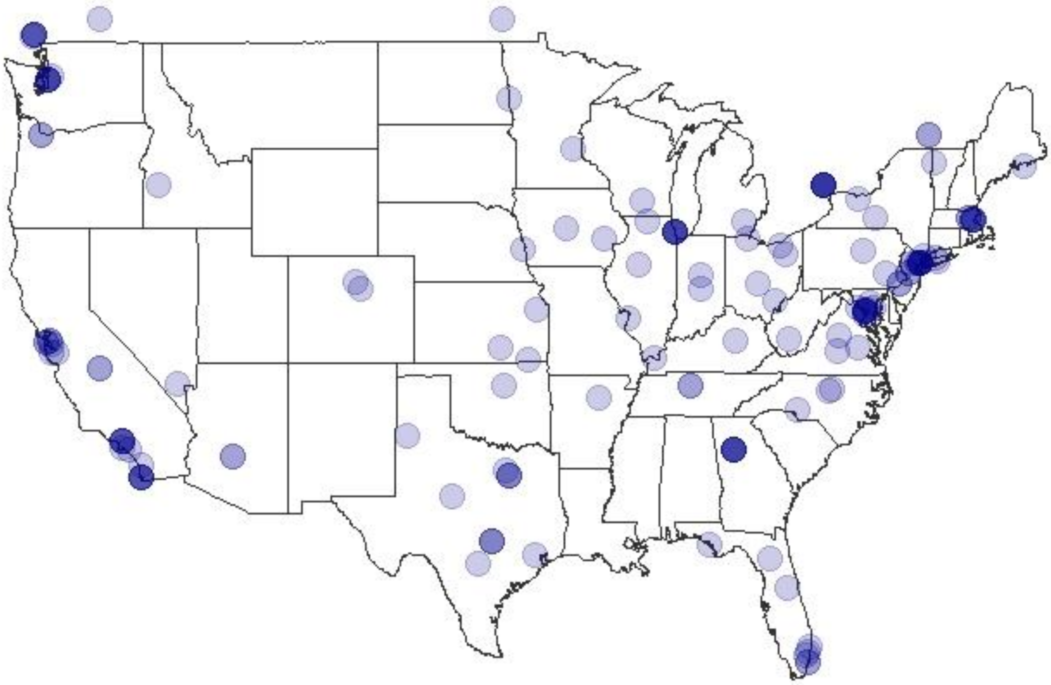
Using twitteR package, I downloaded data on every follower of the Rbloggers. I then geocoded these users' locations (as self-reported in their bios) using the Google Map, doing so via a modified version of the geocode. With this information, we then generated maps of each active user's dispersion and measured their activity levels in the country.



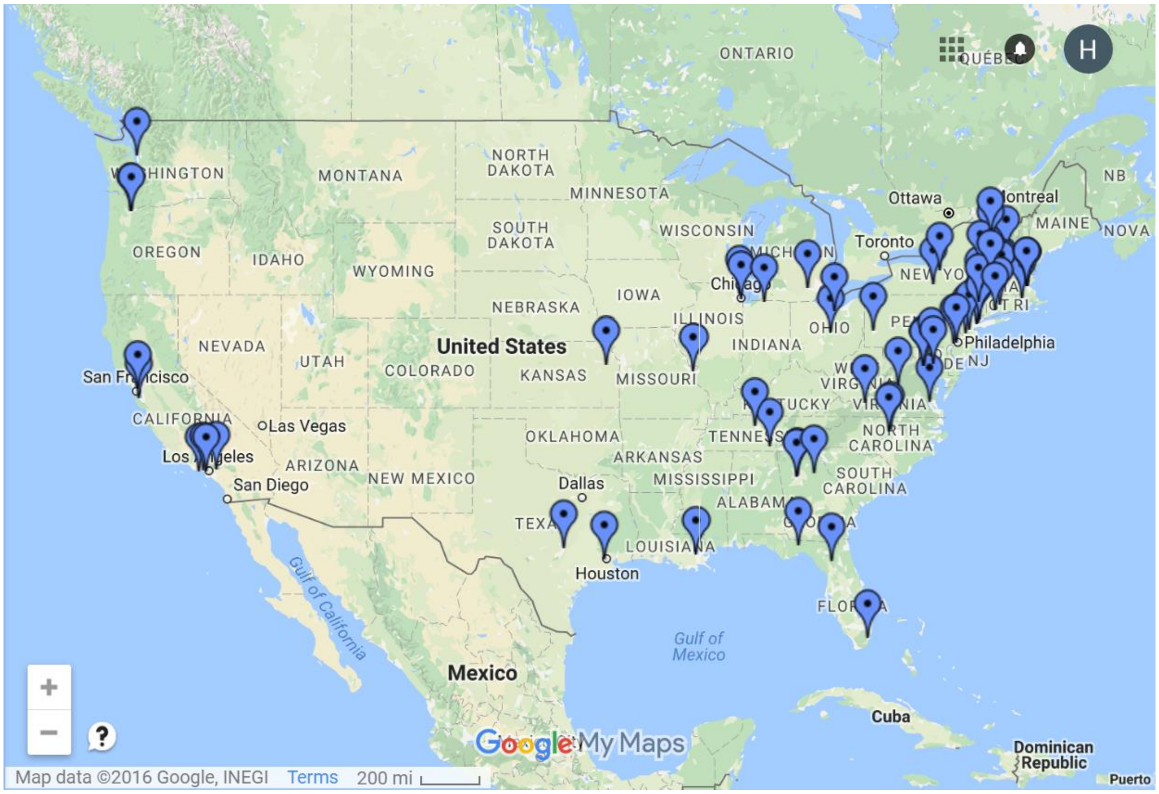
Graph 6: Followers' location in world map

At first, I generate a world map plot to show the followers' distribution. Since the Rbloggers is an account about R programing and statistics learning, the main followers are supposed to be relative academicians and engineers. Thus the result seems good since we can find that the most compact points are in the US and Europe where the higher education and data analysis are matured product and has the largest popular of R users.

We can also find that both R and twitter originate from the US, thus the next step is to find the geolocation of the followers in the US.



Graph 7: Followers' geolocation in US

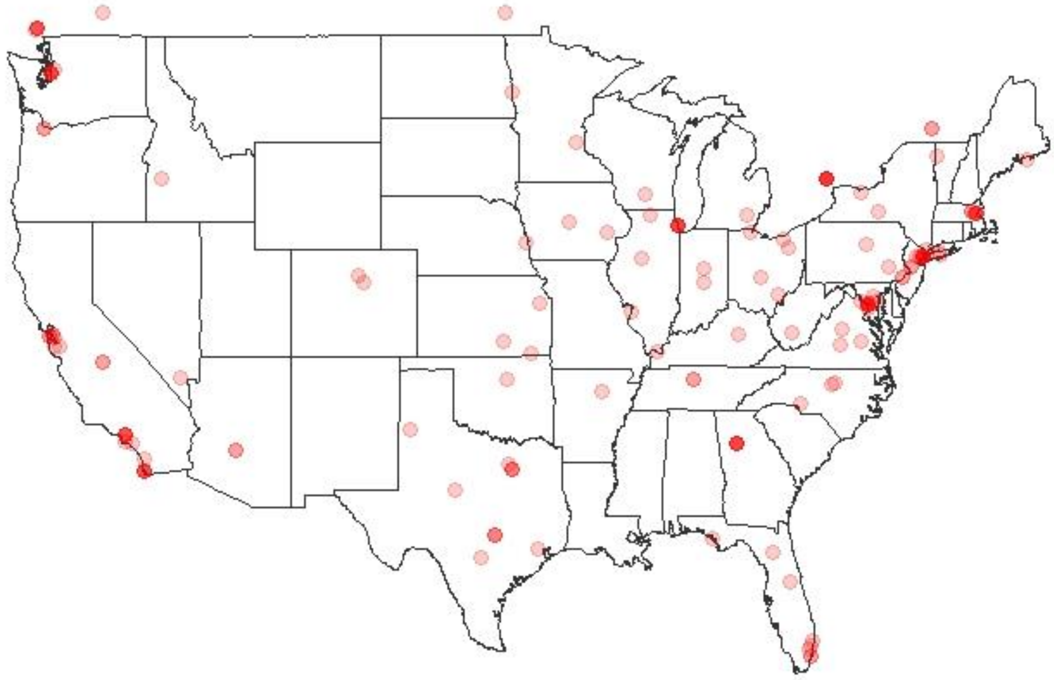


Graph 8: Maps of US Universities

After plotting the geolocation of the followers in the US, I found a picture called “Maps of US Universities” as comparison. We can find the points on the two maps seem to has strong relationship. It is reasonable since a lot of R-bloggers followers are professors and students.

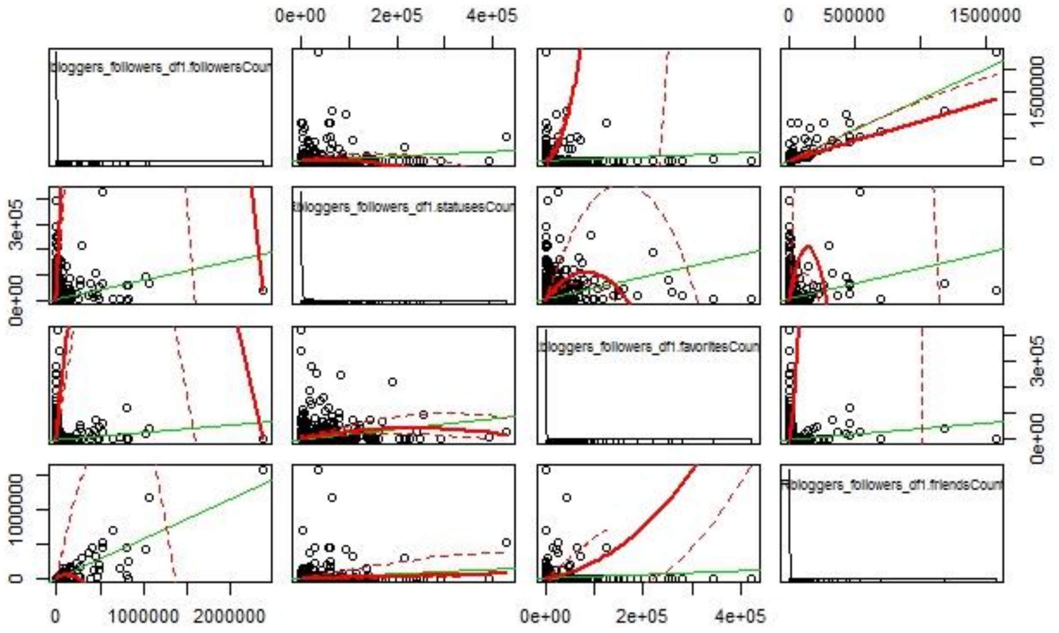
ACTIVE USERS

Since there are over 40k followers of R-bloggers, I want to find how many active users, to determine the followers’ quality of this user. I also give the geolocation maps of the active users to find is the distribution different from the all followers’ map



Graph 9: Active Users Geolocation in US

After giving the plot, I tried to find the influence factors of the activity of the user. Thus I collect the data from the followers’ data set, fitting a linear model of the users’ followers count, friends count, favorite counts and the response: their status count which shows their activity.



Graph 10: Scatterplot Matrix of the Four Variables

According to Graph 10, we can find that there is weak relation between the four factors except friends' count and followers' count. It is obvious that the more friends you have, the more followers you will have.

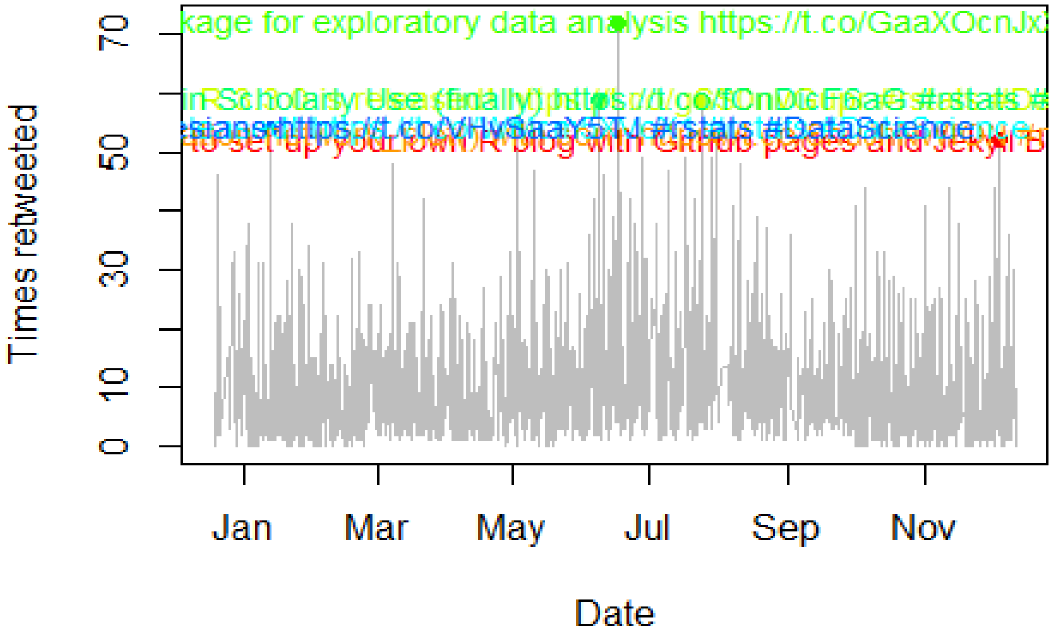
TOP RETWEETED TWEETS

Another interesting data is that we can find the top retweeted tweets from R-bloggers. it responses the followers' interests, thus I extract the tweets which were retweeted more than 50 times.

```
> tweets.df$text[selected]

[1] "How to set up your own R blog with Github pages and Jekyll Bootstr
ap https://t.co/eGnx3a0Bg7 #rstats #DataScience"

[2] "Correlation network_plot() with corrr https://t.co/8huJhNWfON #rst
ats #DataScience"

[3] "R 3.3.0 is released! https://t.co/gSS0nvGupr #rstats #DataScience"


[4] "Introducing xda: R package for exploratory data analysis https://
t.co/GaaXOcnJxX #rstats #DataScience"

[5] "R Passes SAS in Scholarly Use (finally) https://t.co/fCnDicF6aG #r
stats #DataScience"
```

```
[6] "R 3.3.0 is released! https://t.co/WMcm5XVoX1 #rstats #DataScience"


[7] "R Users Will Now Inevitably Become Bayesians https://t.co/VHvSaaY5
TJ #rstats #DataScience"
```

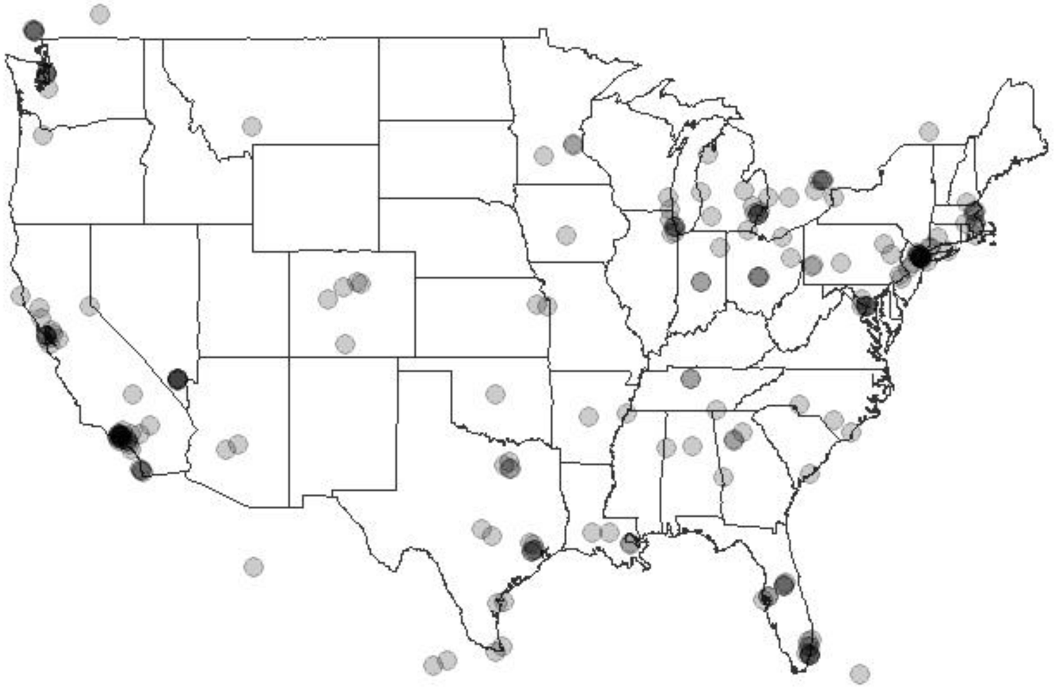
The graph 11 shows how the times of the R-bloggers' tweets been retweeted based on time and I point out the most popular 7 tweets shown above.



Graph 11: Top retweeted tweets

Hashtags Analysis

The last step is using the high-frequency words we found to do the hashtag searching. Thus we choose "#DataScience" and "#rstats". I set the r to search for 10 minutes and finally get a data set which contains over two thousand observations. Similar to the work before, I plot the geolocation of the hashtags to find where are the people who are interested in data science and r. The results below is similar to the graph of R-bloggers' followers map, Thus it's reasonable.



Graph 11: Hashtag searching results locations

Conclusion

This project focuses on the twitter user ‘R-bloggers’. The main work is data mining and cleaning data we get. To explore the geolocation data and try to plot it is very interesting and challenging. I found that the location of R-bloggers’ followers are similar to the people who are interested in data science and also similar to the universities in US. The location is also similar to the companies which need to do data analysis such as Bay Area and New York.

During the project, I found a lot of interesting topics and tools to help me finish the project. I also use some statistical knowledge to make some analysis on the data I get. Since time is limited, I only do the works above and due to the slow processing of R studio, it’s hard to extract large data. To do large data mining and analysis, some other programming language such as python is necessary.

Finally, I would express my gratitude to all those helped me in this class during the whole semester, especially to Professor Haviland for the well-organized lectures and useful suggestions after class.