615- Twitter Data Mining

Jingrong Cheng

December 10, 2016

Introduction

Human Rights Day is on December 10th, every year, celebrated by the whole world. In 1948, the United Nations General Assembly adopted the Universal Declaration of Human Rights. In 1950, the Assembly passed resolution 423 (V), inviting all States and interested organizations to observe 10 December of each year as Human Rights Day. (12,12, http://www.un.org/en/events/humanrightsday/) This year is 68th anniversary of Human Rights Day and the theme of it is "Stand up for someone's rights today!." The United Nations officially created 2 hashtags -#Standup4HumanRights and #HumanRightsDay. Therefore, it is expectable to see there are many high frequency keywords that are parts of these hashtags on my Twitter searching. The time that I set up for Twitter searching is 10 minutes, and there were more than 12 hundred tweets return back to R. 2271 entries were selected by keyword "HumanRightsDay." Total 1263 Ids are involved in this selection. 1000 tweets with different text are chosen by my Twitter searching. After cleaned the data, it is ready to do some exploration and analysis.

Word Cloud

Figure 1

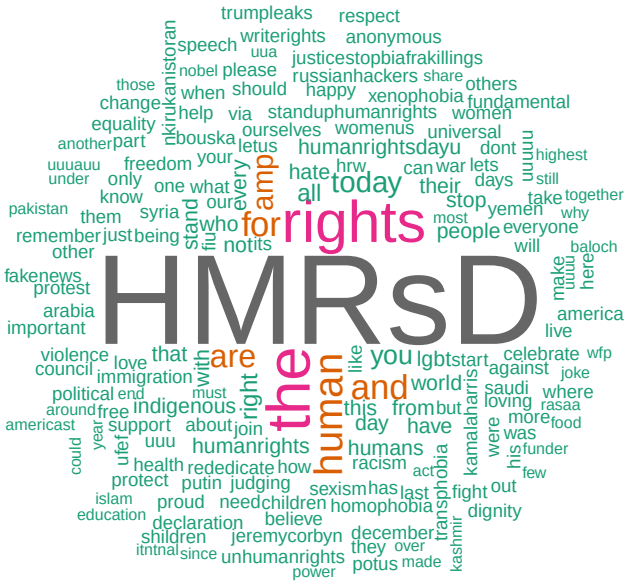


Figure 1 is a word cloud. All the captured terms, which is total 4451, are placed in decreasing order, and only top 180 are chose into the cloud. There was a problem that I met during this process. The most frequent word is "HumanRightsDay" with 2137 frequency, which is exactly what I expected. However, when I created plot, "humanrightsday" is too large to fit in the plot, and I tried several methods to fit the original word into figure, but I couldn't. Therefore, I decided to replace it with "HMRsD." This is a problem that I should inquiry with Professor Haviland, and to learn how to create this figure without changing the context.

Figure 2

Figure 2 is another word cloud. This time I chose the top 400 frequent terms into the cloud. As we could see, "HumanRightsDay" is the most heated term in my Twitter searching, and it has extremely different attention than any other words, according to the size of it in the word cloud

Frequency Bar plot

Figure 3

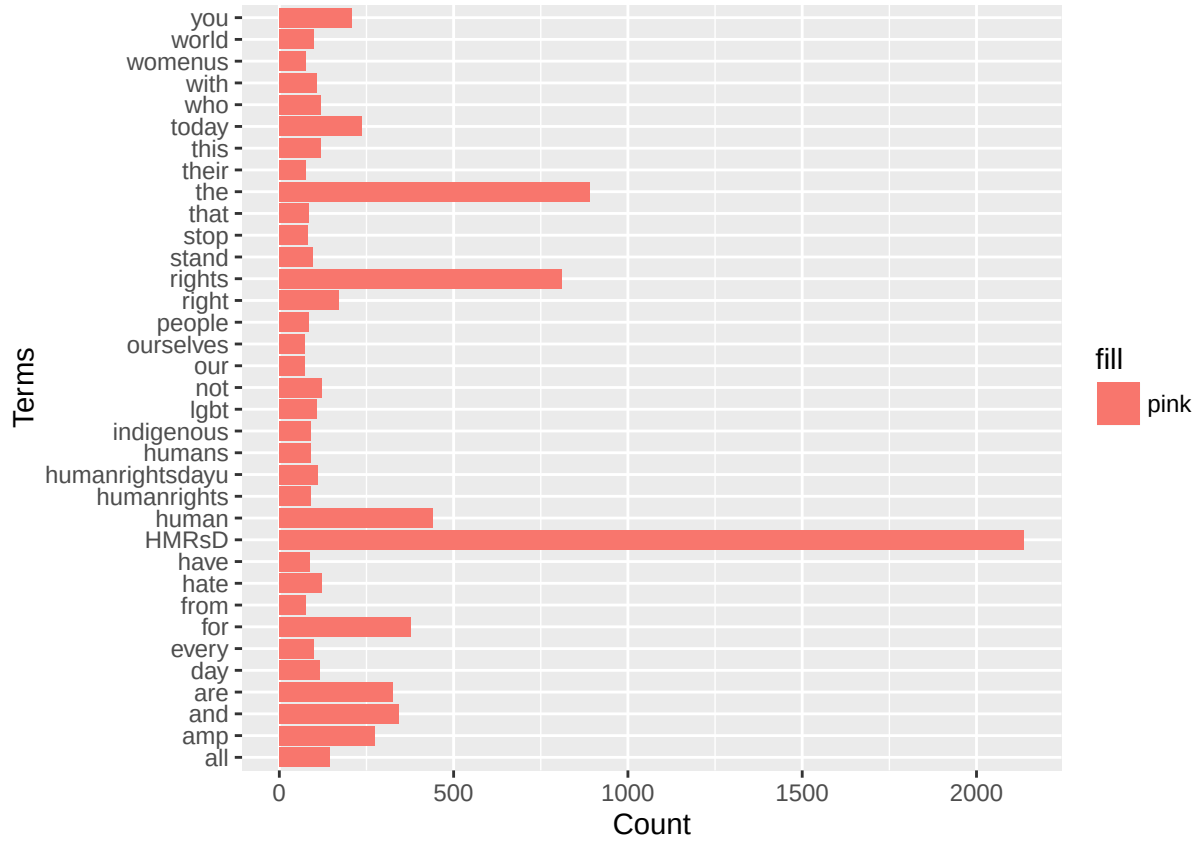
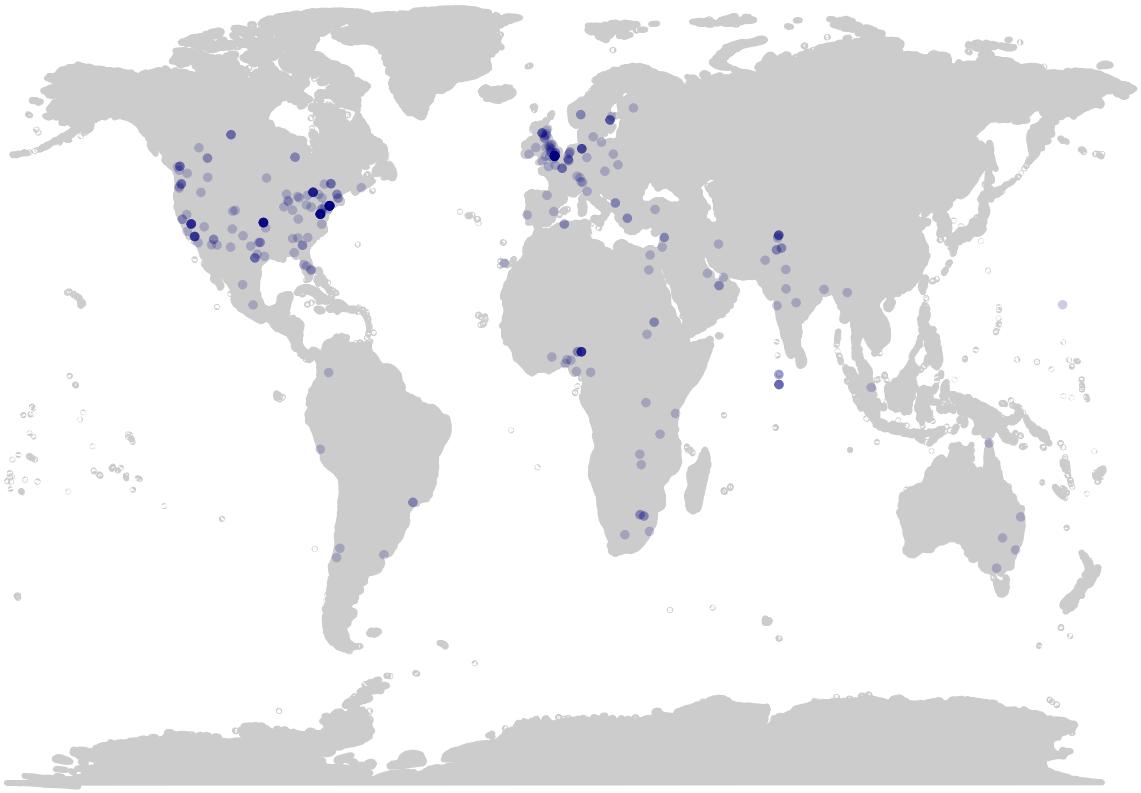


Figure 3 is the frequency bar plot. Top 35 frequent terms are selected to create the barplots. HMRsD stands for "HumanRightsDay", which is the most frequent word in my Twitter searching. "The" is the second frequent word, and "rights" is third as well as "human" is the fourth. It is reasonable to see those words having high frequency since those are campaign related keywords.

Mapping Plot

Figure 4



Since I searched tweets around the world within 10 minutes, I am able to create a world mapping plot for the tweets that have been tweeted at that period of time. In the figure, all the location points are presented in color navy. As we can see, tweets are distributed mostly at United States and Europe, and some of them distributed in Africa. It is interesting to see that there is no twitter posted in China, since Twitter and Facebook are not able to access in China. Increasing debates and topics about human right are appearing around the United States and Europe recently due to various reasons. I hope see more and more people to care this topic in the future.

Figure 5

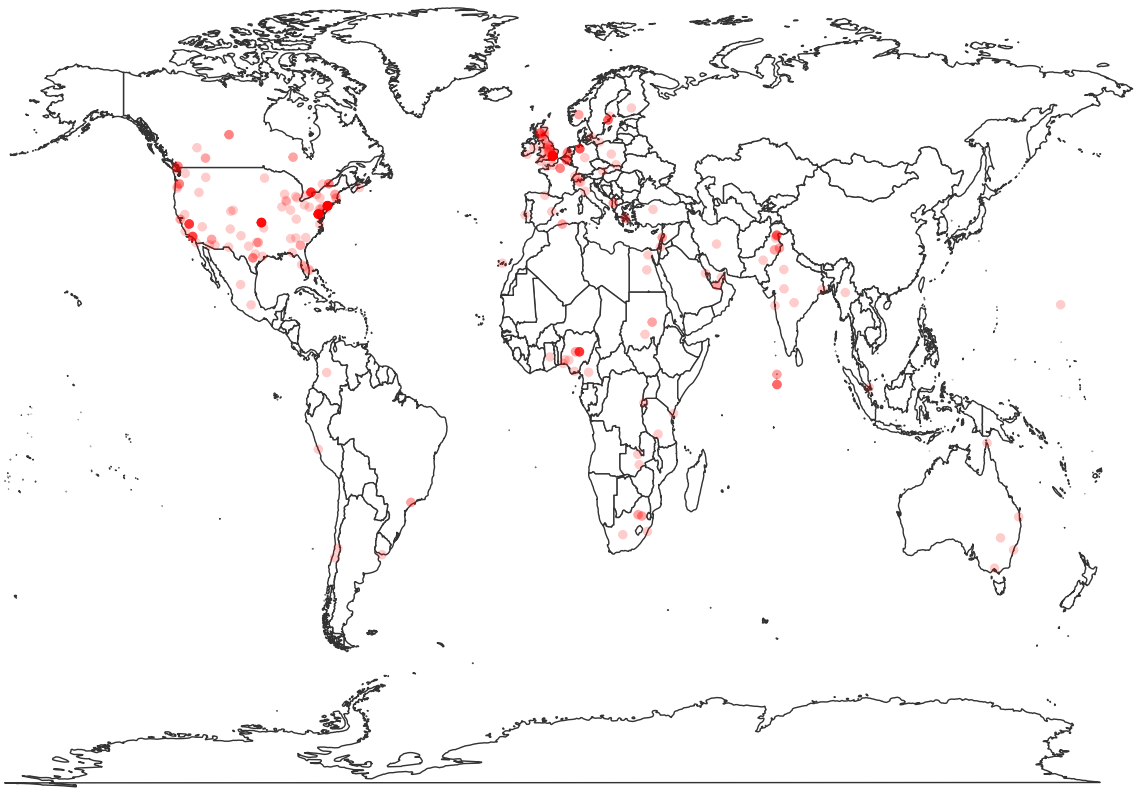


Figure 5. There is another presentation for world mapping plot. All the locations are pointed in red color. From the clear country border, it is clear to see the location of most of the twitters posted from United States and Europe on the internet.

Figure 6

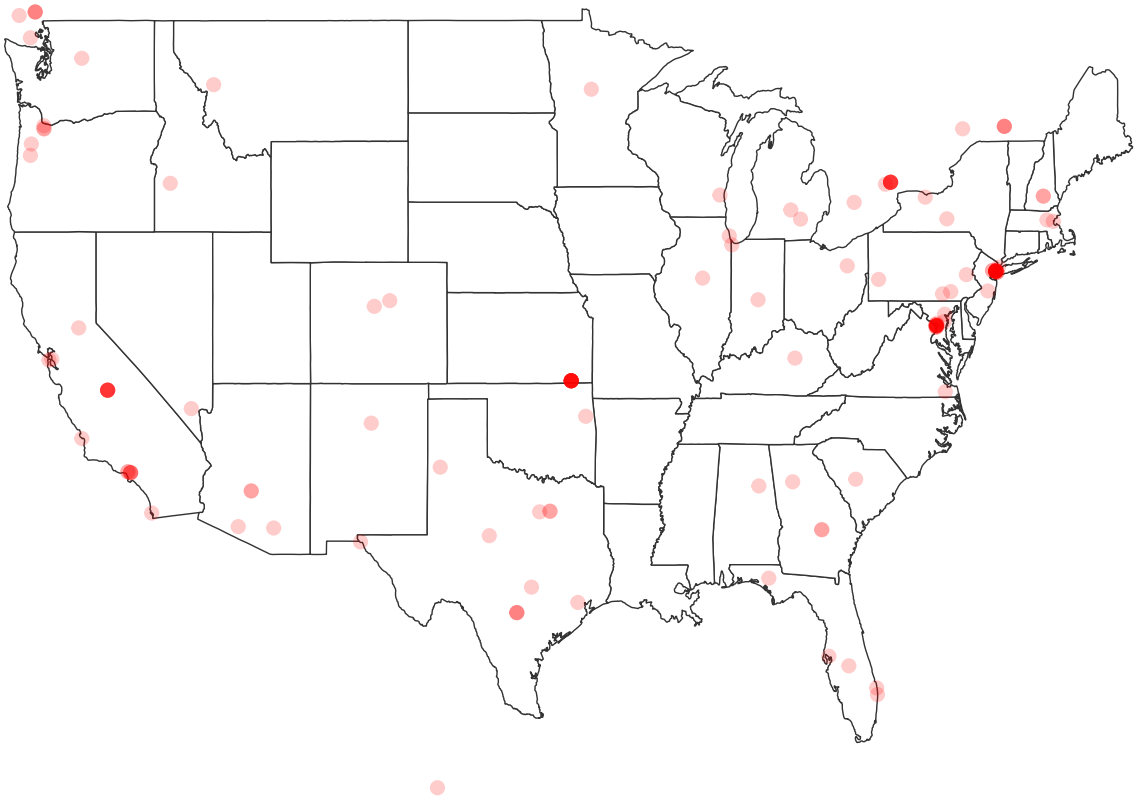


Figure 6 presents the twitter location in United States, since US is one of the area that has most of the points. We could tell from the figure that users from east and west coast are more active about Human Rights Day.

Simply Statistic Analysis

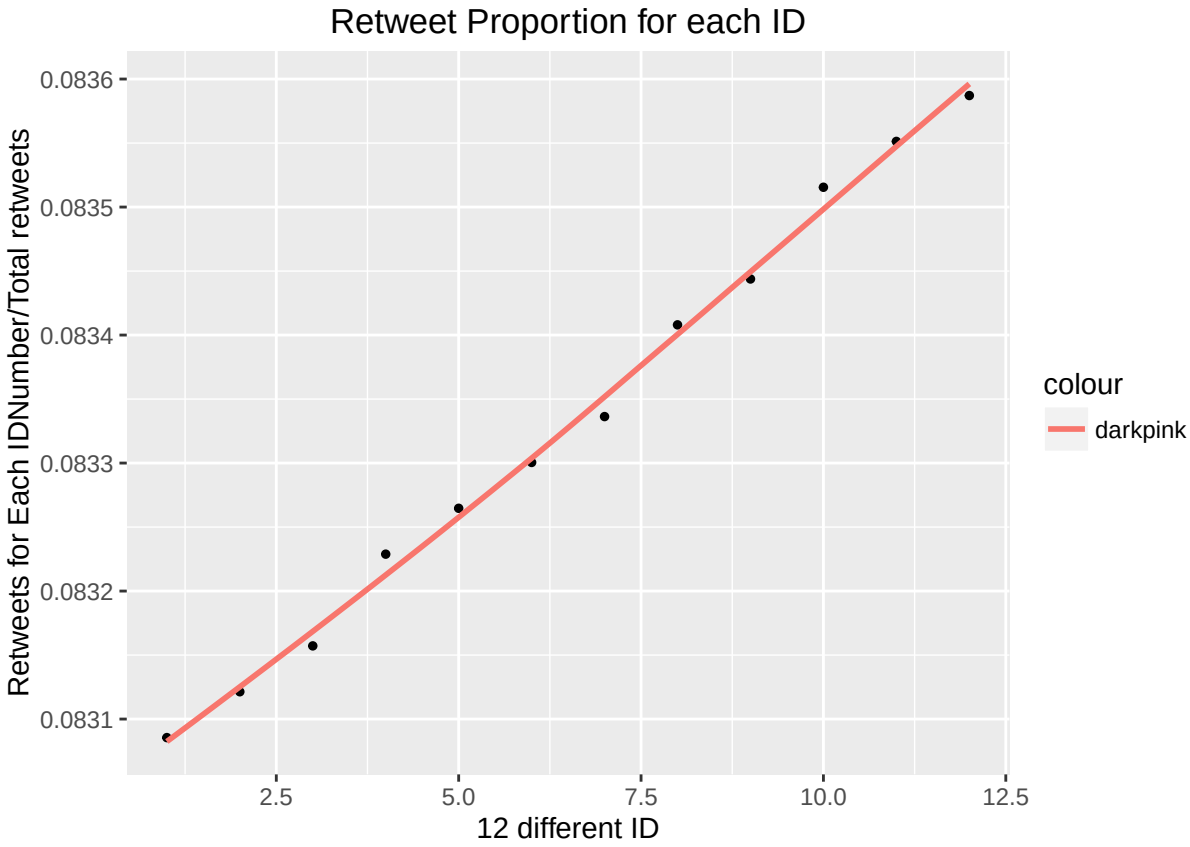
Tweets with top 4 total retweets are selected to do the following analysis. Retweet proportion for each ID and connection between followers and number of retweets will be analyzed for each of the selected twitter.

No.1

number of retweets: 27911

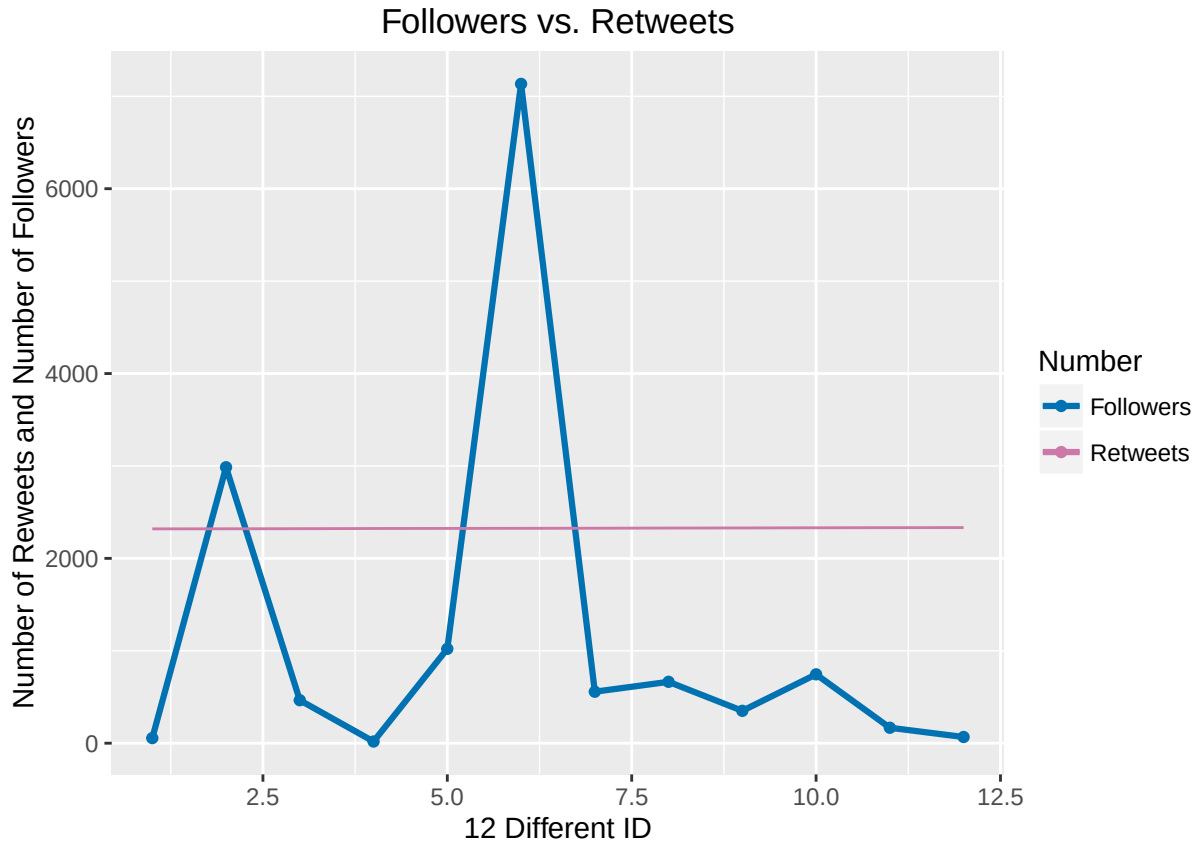
Twitter text: RT @JamesMelville: Humans should help other humans. Where they come from is irrelevant. #RefugeesWelcome #HumanRightsDay https://t.co/dqbr0

Figure 7



For this twitter, 12 different users are captured by my search. The plot shows the retweet proportion for each user, which presents how the retweet number for each user contributes to the whole publicity for this twitter. From the figure, each of the user retweets about the same amount as each other. There is no super large or small amount of retweets among these users, which is interesting. Each of them retweets around 8.3% of the total retweets number.

Figure 8



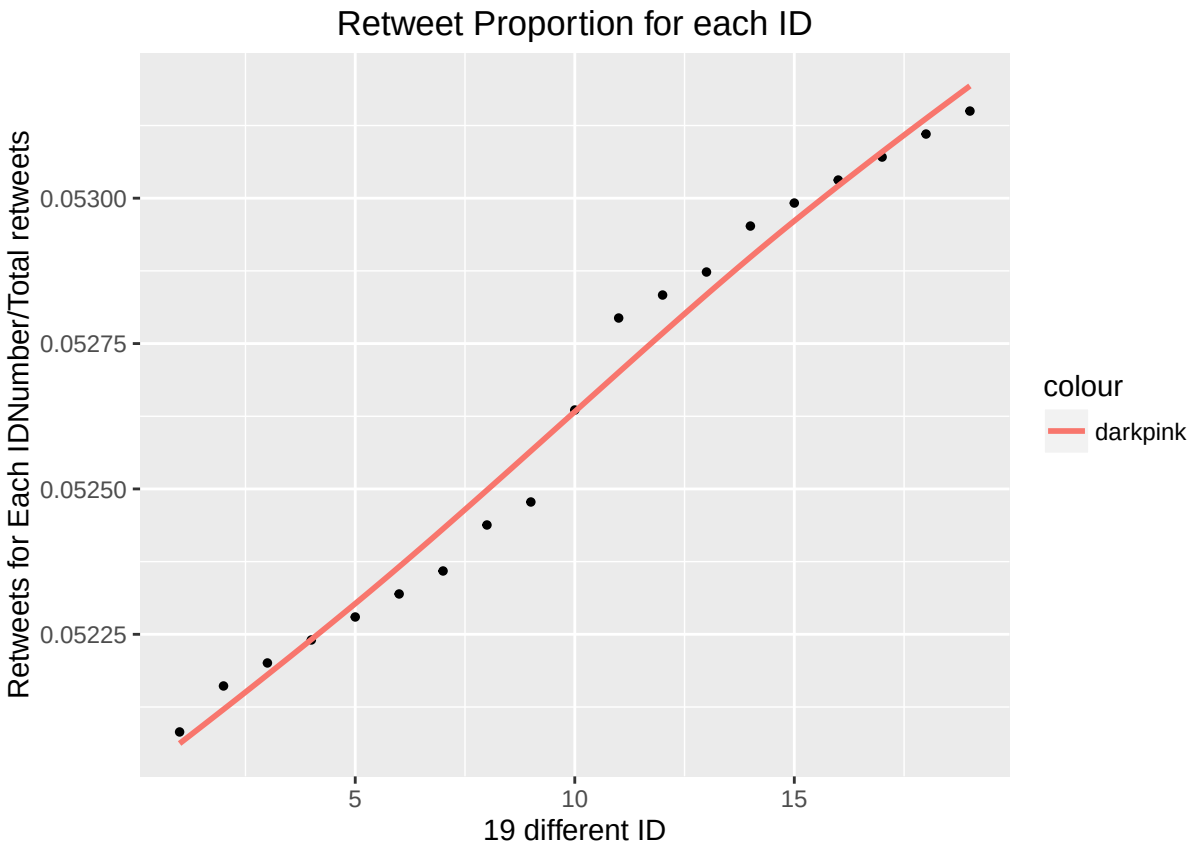
In this figure, line in pink color represents the number of retweets of each user for this twitter, and the blue line demonstrate the number of followers that each user has at the data mining time period.

No.2

number of retweets: 25287

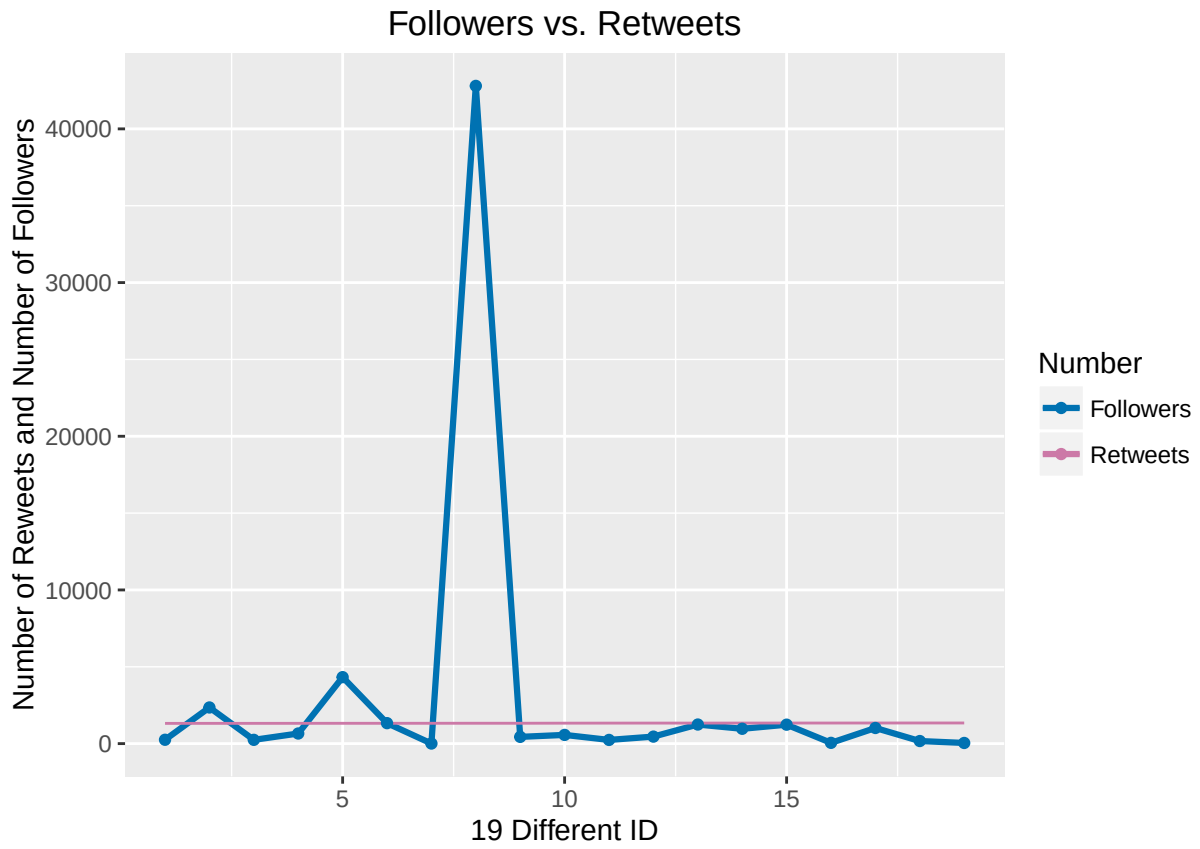
Twitter text: RT @WHO: Today is #HumanRightsDay. The highest attainable standard of health is a fundamental right of every human being

Figure 9



Second top twitter has 19 users involved. After no.1 twitter's analysis, it is not surprised to see that the retweet proportion for each user is almost the same to each other. Each user contributes about 5.2% of the total retweets.

Figure 10



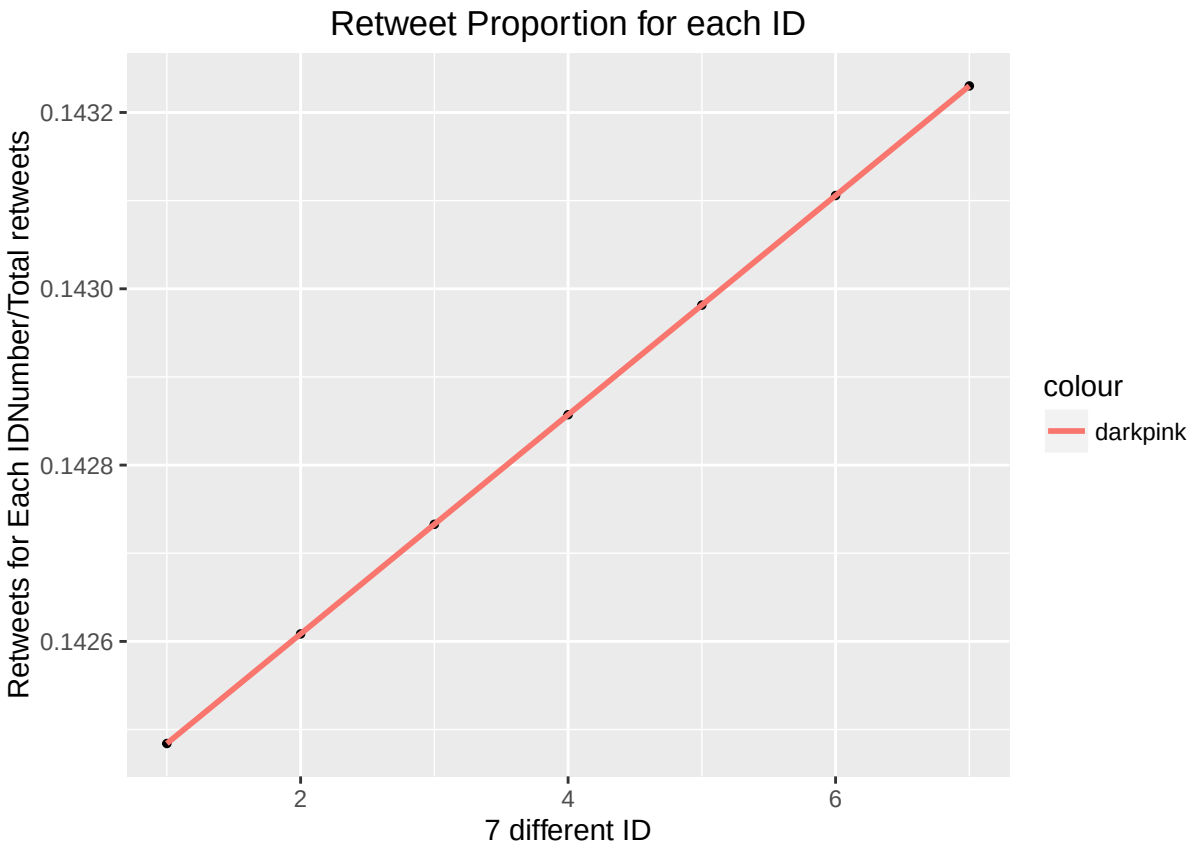
In this figure 10, line in pink color represents the number of retweets for each user, and the blue line indicates the number of followers that each user has at the data mining time period. The result didn't show the pattern that I would expect either. The retweets number doesn't correlate with the number of follower.

No.3

number of retweets: 8043

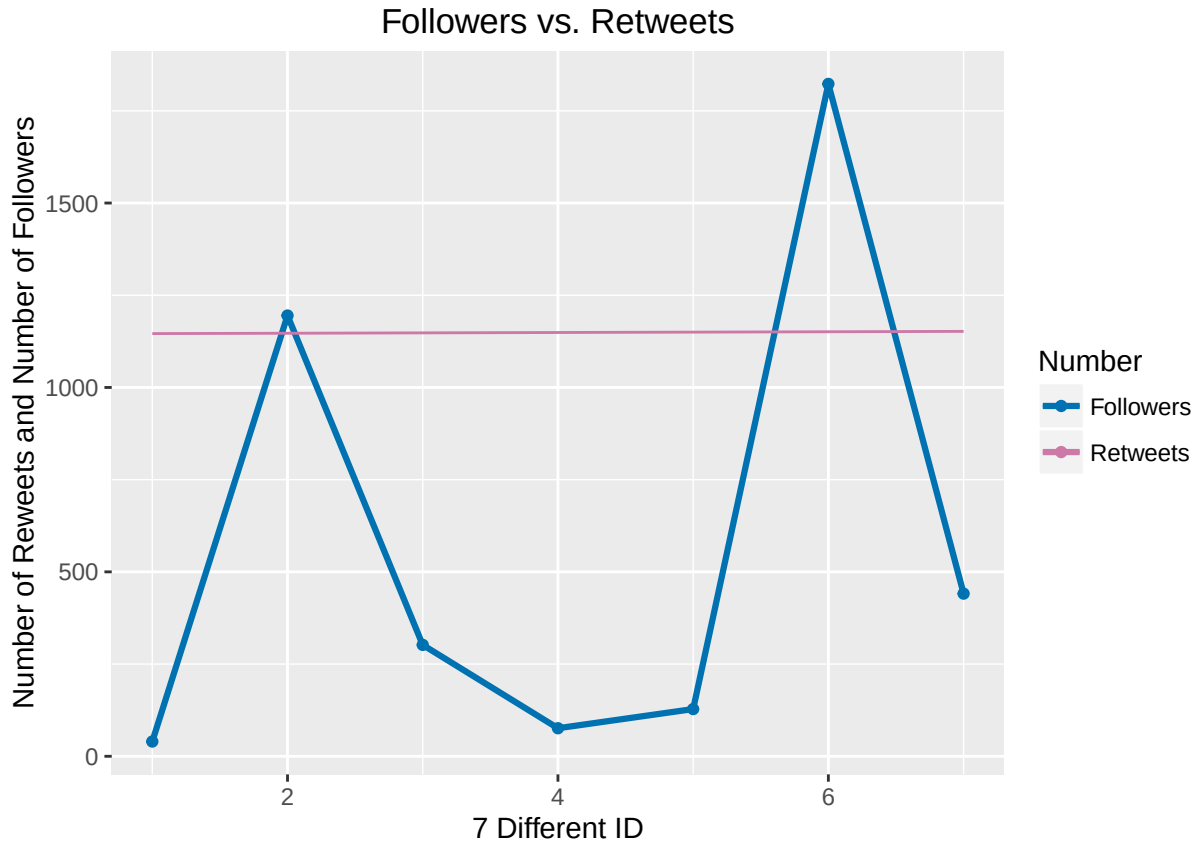
Twitter text: RT @UN_Women: Today is #HumanRightsDay. In 1995, @HillaryClinton delivered this powerful speech. RT if you agree! https://t.co/oRUy3M3fgu

Figure 11



Third top twitter was distributed by 7 users at the time of Twitter searching. The retweet proportion for each ID is about the same. Each user contributes about 14.3% of the total retweets.

Figure 12



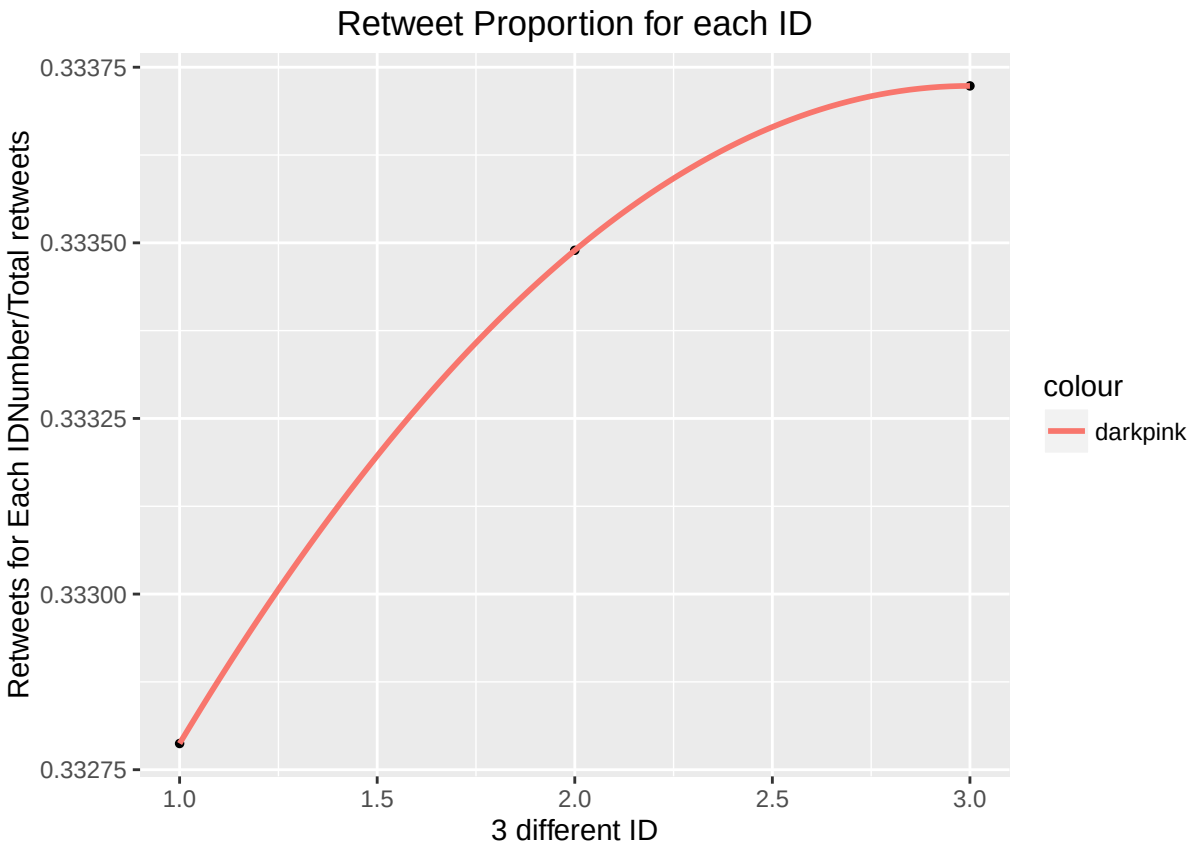
In this figure 12, line in pink color represents the number of retweets for each user, and the blue line indicates the number of followers that each user has at the data mining time period. The retweets number for each user doesn't correlate to number of followers for each user, which is not what I expected. I may need further analysis on this particular part of research.

No.4

number of retweets: 4273

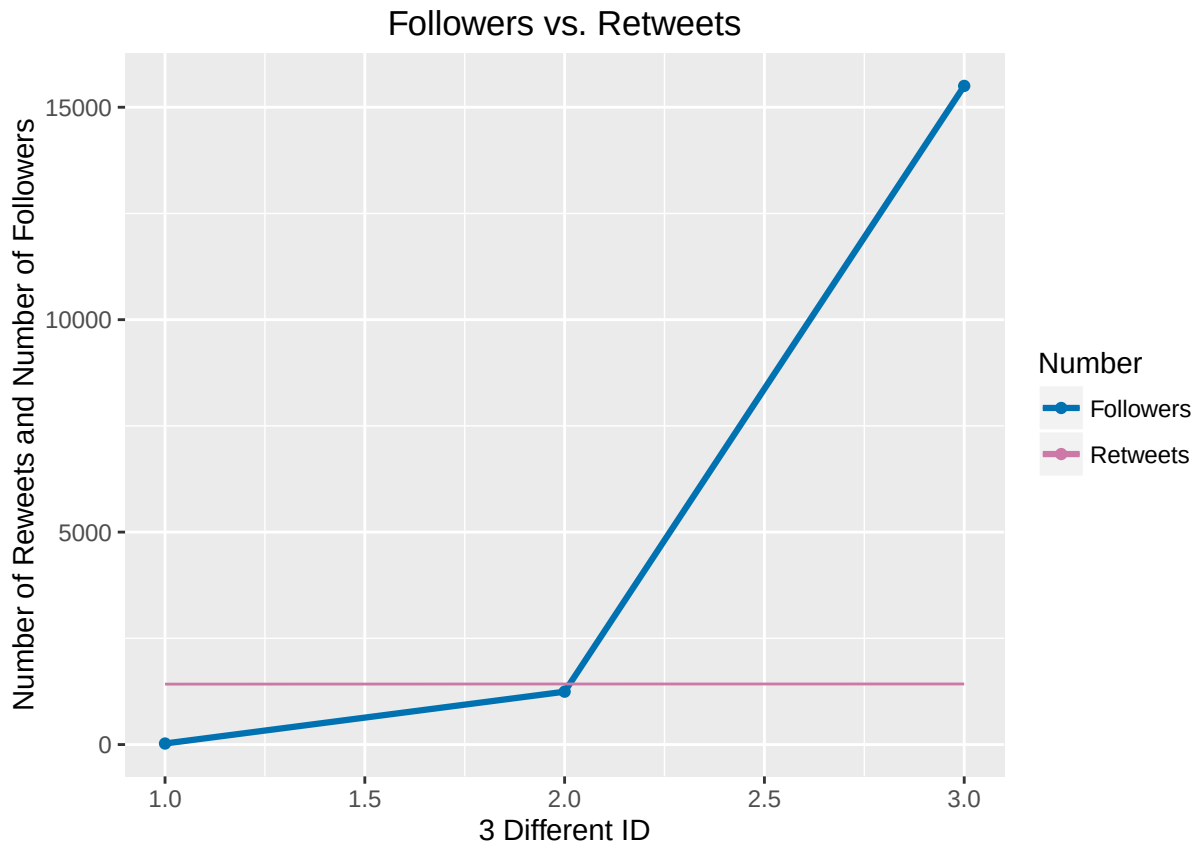
Twitter text: RT @UNHumanRights: Dec 10 is #HumanRightsDay. Let's #StandUp4HumanRights - for greater freedoms, stronger respect & more compassion https:/

Figure 13



Fourth top twitter was distributed by 3 users at the time of Twitter searching. The retweet proportion for each ID is about the same. Retweets for Each user's twitter contributes about 33.3% of the total retweets.

Figure 14



In this figure 14, line in pink color represents the number of retweets for each user, and the blue line indicates the number of followers that each user has at the data mining time period. The retweets number for each user doesn't correlate to number of followers for each user, which is not what I expected.

Conclusion:

Twitter data mining is a great experience that letting us connect with the real social world by R programming language. In this report, there are certain results that don't reflect to my expectation, which is an interesting point that would lead me to do more exploration in social media environment. Assumptions may not be proved by the research. Also, learning how to use Shiny to create report and present figures properly is one of my next directions. I plan to polish this report during winter break.