Jingzhi Yu (UID: 604514516), Haiman Duan (UID: 404406184), Mengnan Jia(UID: 804186335)
February 21, 2015

Report for Project 2 of EE 239AS

Part 1

The crawling results for the top 5 tweets of the hashtag #SuperBowlXLIX is saved as top_tweets in the fold.

The results are shown as below:

1） "Agradezco a @PepsiCo y a su CEO, la destacada Indra Nooyi, por su invitaci \xf3n al #SuperBowlXLIX. Bien x @Patriots! http://t.co/OAK4aR6MgE " Post by: FelipeCalderon Posting time: 2015-02-01 19:03:57

2) Captan noche del #SuperBowlXLIX desde el espacio http://t.co/eamJJb5xy7 http://t.co/qfVDrk8XbC Post by: El_Universal_Mx Posting time:2015-02-01 18:51:06

3) Katy Perry's halftime show: What's the verdict? http://t.co/CYiSumOi2f via @grinsli #SuperBowl #SuperBowlXLIX #SB49 http://t.co/8ESlGsXG3v Post by: CNN Posting time: 2015-02-01 19:04:02

4) Russell Wilson threw the game away. #SuperBowlXLIX https://t.co/ZrGr2V8lKN Post by: ComplexMag Posting time: 2015-02-01 19:06:13

5) When your coach calls the worst play in the history of the universe. #SuperBowlXLIX http://t.co/RbZwGJmOmC Post by: Travon Posting time: 2015-02-01 19:05:46

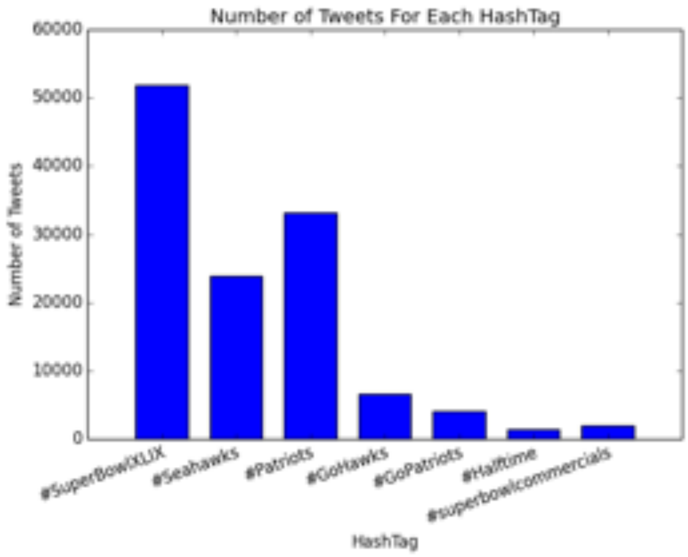
Part 2

The crawling results are stored in tweets.txt and search_log.txt in the fold.

Part 3

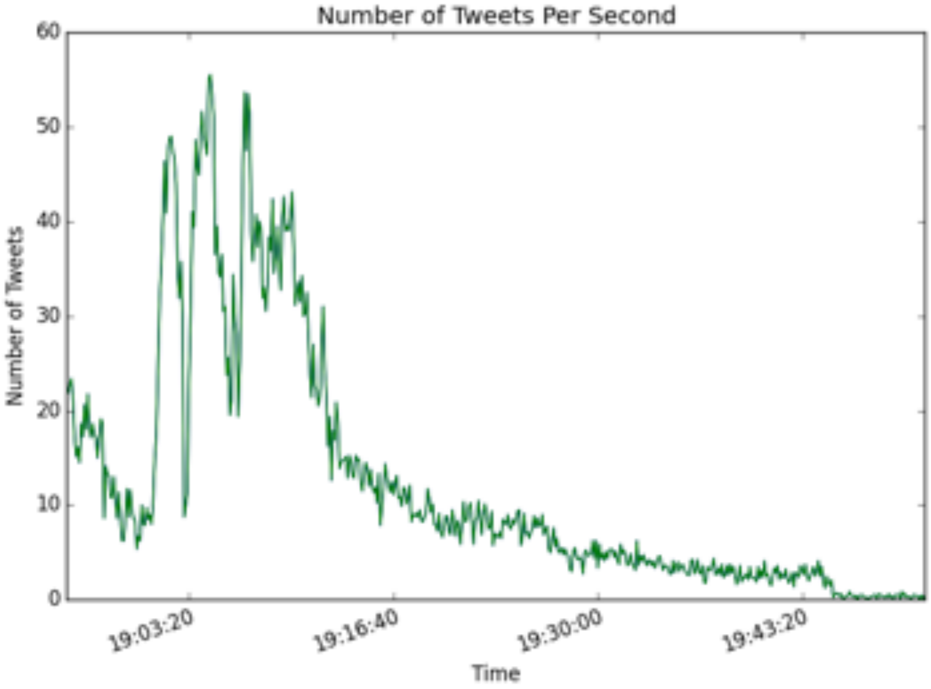
1) A histogram that shows the number of tweets for each hashtag is shown below as Figure 3.1.

Figure 3.1



2) In our statistical results, the most popular hashtag is #SuperBowlXLIX. The number of related tweets that were tweeted each second in the given time period was plotted as below in Figure 3.2. It can be shown that the number of tweets reach several peaks during 19:00 and 19:16.

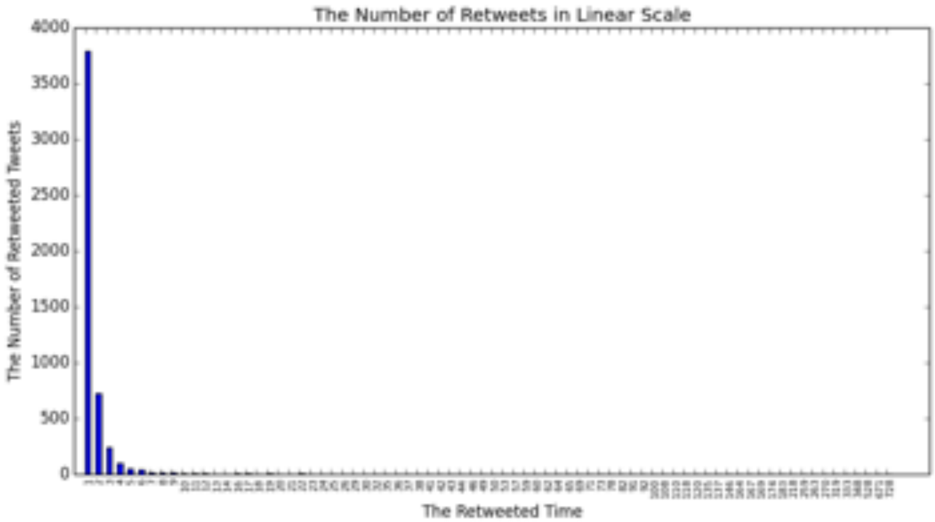
Figure 3.2



Part 4

1) We believed it is a better way to plot the number of tweets that are retweeted k times over the number of retweets k (k=1, 2, 3, …) in linear scale as a histogram, as Figure 4.1 shows.

Figure 4.1



However, as it is obvious, due to the enormous difference between those favorite tweets and those less popular ones as well as the scale of number of retweets times, the plot doesn't look very indicative.

2) Then we plotted the data into a log-log scale, as Figure 4.2 shows. And it is much better for an observation.

Part (5)

As what is required in the project instruction, we plotted all the data with the tweeting rate of #SuperBowlXLIX, which is the most popular one over #Patriots, the second popular one, as what Figure 5.1 has been shown. However, there is apparently not a wise choice to draw a conclusion of their correlation. They are mostly scattered points, while most of them crowded in the zero point and its adjacent area.

So besides drawing the scatter plot, we plotted another picture. We calculated the ratio of the tweeting rate of #SuperBowlXLIX over that of #Patriots. And plot them vs. time, just as Figure 5.2 shows.

Figure 4.2

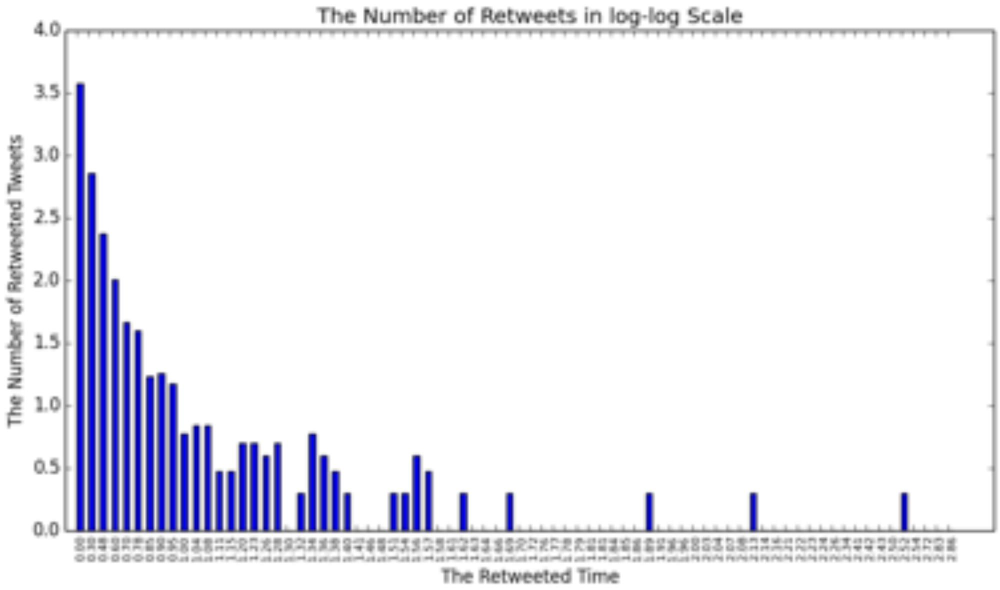


Figure 5.1

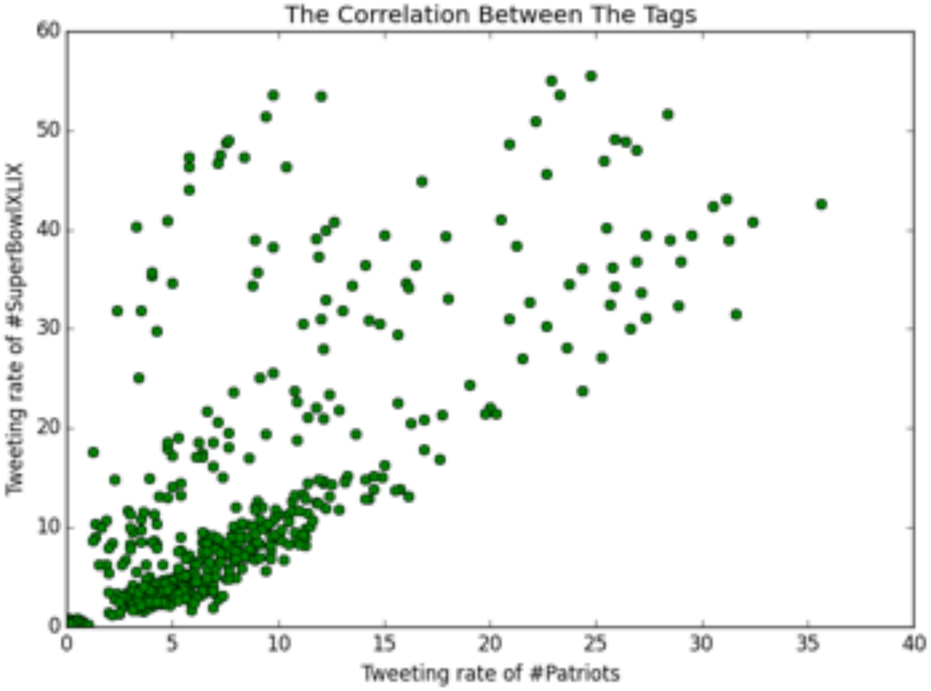
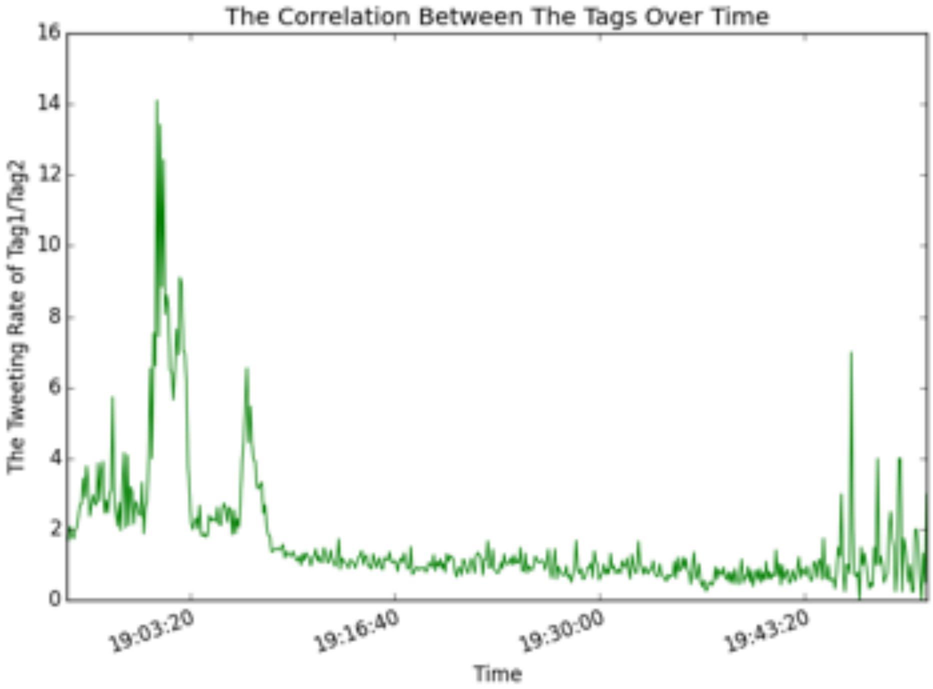


Figure 5.2



And since this plot shows some drastic oscillation as time elapses, it hardly to draw a fast conclusion on the correlation between the tweets related to the two tags

Part (6)

Please see attached file named proj2_6.py for the program.