EE239AS - Project Report 2
Crawling and Data Collection on The Web
Winter 2015

Ayush Minocha (104410979)
Nitish Mehta (404434455)
Jessica Abraham (304432126)
Khushboo Singhi (904435315)

February 23, 2015

Twitter is a popular platform where users share opinions and thoughts on the latest events. It is a hot resource for gathering public opinion on various topics and news. Information extracted through tweets could be potentially useful on a commercial basis in devising marketing strategies. This project is based on crawling Twitter to extract useful data, and making useful observations about the public opinion related to an event by plotting tweets on a graph based on temporal properties and frequency.

In order to enable this, there are various APIs which help in crawling Twitter data. We have used the Topsy API in this project, which provides the /content/tweets resource to extract the top tweets for a list of hashtags. Search parameters could additionally be used. Topsy API sends requests using the HTTP GET method and returns the results in json format which are ranked based on the sort method. We have used this API extensively in our project. Topsy also provides several other Content APIs like Bulk Tweets, Photos, Citations and other Metrics and Insights APIs.

We chose the 2015 Super Bowl event, and crawled Twitter using the Topsy API to collect a subset of tweets related to this event, which were posted during a particular time period. The tweets were extracted based on hashtags and timestamps. We plotted graphs based on the temporal variance and frequency of tweets. By observing these graphs, properties like popularity of the event over time can be determined. We also extracted the retweet patterns for the tweets in the selected slot, which further augments information about popularity of events. We found a high correlation between the two most popular hashtags in the given set. The implementation details of each part are given below.

1 Part 1

We have selected Slot 8, which has a timeframe of 18:50:00 to 20:00:00 on 2/1/2015. The hashtags in our slot are: '#SuperBowl', '#NFL', '#DeflateGate', '#DeflatedBalls', '#SNL' and '#Colts'.

We retrieved the tweets for the hashtags '#NFL' and '#SuperBowl' using the Topsy API and stored the top 5 tweets in top_tweets.txt.

The top 5 tweets for '#SuperBowl' are as follows:

1. Tweet Text: Pete Carroll, MVP. For the Patriots. #SuperBowl
   User: Piers Morgan
   Posting Date: 2015-02-01 19:08:03

2. Tweet Text: Relax. My man's got this. #RussellWilson #SuperBowl
   User: Piers Morgan
   Posting Date: 2015-02-01 18:50:32

3. Tweet Text: PHOTOS New England Patriots celebrate fourth #SuperBowl title –>`http://t.co/0Yp4Wu0rNB`
   User: Yahoo Sports
   Posting Date: 2015-02-01 19:28:05

4. Tweet Text: 'World Champions!'....

   Oh, pur-lease.

   #SuperBowl
   User: Piers Morgan
   Posting Date: 2015-02-01 19:30:40

5. Tweet Text: No #Beastmode???? I'm so confused right now. So confused. #SuperBowl `http://t.co/y54ZZHxoQQ`
   User: Scooter Braun
   Posting Date: 2015-02-01 19:40:53

The top 5 tweets for '#NFL' are as follows:

1. Tweet Text: It's a true shame that the #NFL put so much $ into an anti violence campaign only for the #SuperBowl to end in violence!
   User: Kelly Osbourne
   Posting Date: 2015-02-01 19:11:05

2. Tweet Text: #NFL #SuperBowl Seahawks 24 - 28 Patriots | FINAL `http://t.co/cypLIPPI9L` `http://t.co/IDYVWJ0pG4`
   User: Meridiano
   Posting Date: 2015-02-01 19:08:48

3. Tweet Text: #OtrosDeportes | ¡Go Patriots! Los nuevos campeones de la #NFL `http://t.co/su5LmWihxv` `http://t.co/wP6r4YENMG` User: Meridiano
   Posting Date: 2015-02-01 19:40:21

4. Tweet Text: Wow!#NFL#SuperBowl
   User: Caron Butler
   Posting Date: 2015-02-01 19:03:24

5. Tweet Text: ¡Increíble intercepción! Los Pats serán campeones de la #NFL
   User: San Cadilla
   Posting Date: 2015-02-01 19:00:54

2 Part 2

In the second part of the project we extracted all the tweets containing one or more hashtags given in our slot and saved it to tweets.txt. The Topsy API can return a maximum of 500 tweets at a time. After analyzing the number of tweets returned for different time intervals (k=1, 5, 10, 20 seconds) we concluded that we should divide the total time slot into periods of 10 seconds. If the number of tweets returned by the API was 500 then we further divided the

time slot into two 5 second intervals and resent the request. Furthermore if the API returned 500 tweets for the 5 second interval we divided the interval into 1 second intervals and resent the request.

However we did not encounter any case in our slot where the one second division was required. This can be observed from the results_number for every pair of start_date and end_date for every query_string in search_log.txt.

3 Part 3

We aggregated the results obtained in the previous part to get the total number of tweets for each hashtag in our timeframe. Figure 1 shows the total number of tweets versus the hashtags.

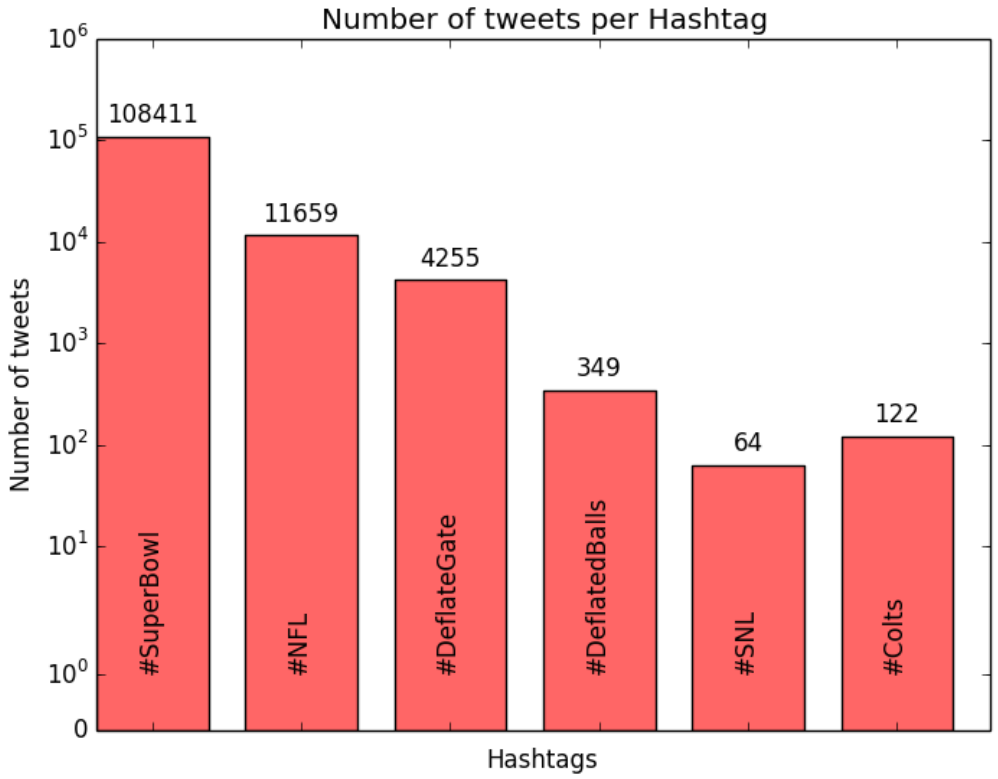

Figure 1: Bar Graph for total number of tweets for each Hashtag

From the bar graph it is evident that the most popular hashtag was '#SuperBowl' with 108411 tweets in our slot while the least popular hashtag was '#SNL' with 64 tweets in our slot.

We then used the hashtag '#SuperBowl' and pulled out the tweets for each second in the time frame, and thus found the count of tweets for every second. We used this data to plot the distribution of the number of tweets per second in the given interval that contained the hashtag '#SuperBowl'. The plot is given in the Figure 2.

According to the plot there was a spike in the number of tweets per second at around 19:05:00 which could mean that something interesting like a touchdown happened in the match during that time. Another possibility could be that some interesting tweet led to a lot of discussion at that time.

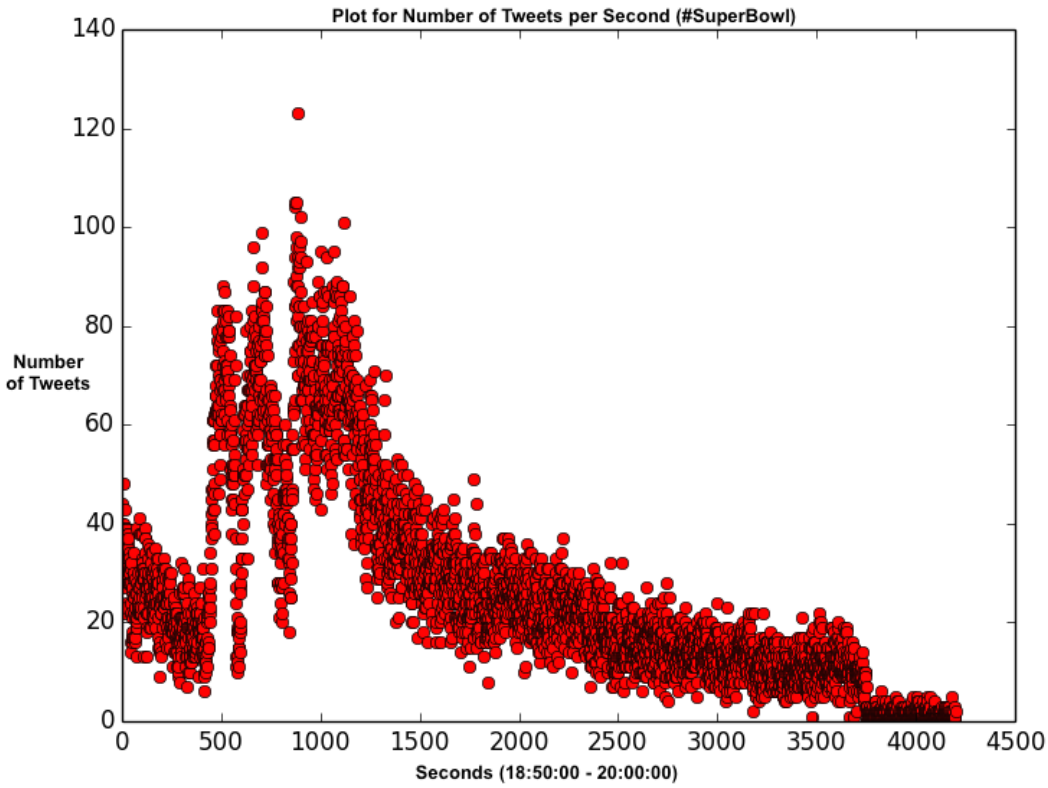

Figure 2: Plot for number of tweets per second for '#SuperBowl'

4 Part 4

Every tweet object has a unique id associated with it. So if two tweets are the same then they share the same id. Therefore to find the unique tweets we just take all the tweets with unique id's. When we query the tweets using the Topsy API, we set the optional parameter of 'include_metric' to be 1. This gives all the metrics related to the tweet like the citations (retweets), replies, peak, impressions, momentum etc. To find how many times the tweet has been retweeted we use the total citation metric, in the metric field of the tweet object. So we now have the number of times a unique tweet has been retweeted.

We use this and make a python dictionary where the keys are 'k' which is the number of retweets and the values are the corresponding number of tweets that have been retweeted 'k' number of times. On plotting these values as seen in Figure 3 we can see that most of the values are very small and in the range of 0 to 10, where as only for smaller values of 'k' we have large number of tweets. So to get a better view of the plot, we reduce the range of the y-axis from (0,5000) to (-5,50). This new plot is shown in the Figure 4. This distribution looks like a plot of $y = \frac{1}{x}$, and thus we cannot fit a line to this plot. If we try to fit a line to this plot, we get a line $y = -0.032 \times x + 66.7$ because most of the points are close to zero. This can be seen in the Figure 5.

For the plot in the log-log scale, we took the log of all the k's and the number of tweets corresponding to each k. This plot looks a little linear as seen in Figure 6. This is also explainable from the fact that the graph on linear scale looked like $y = \frac{1}{x}$ and so in the log-log scale it should look like $log(y) = -log(x)$ which is equivalent to $y = -x$. Thus we get an almost linear graph in the log-log scale. On fitting a linear model to the distribution we get Figure 7, which is a good approximation.

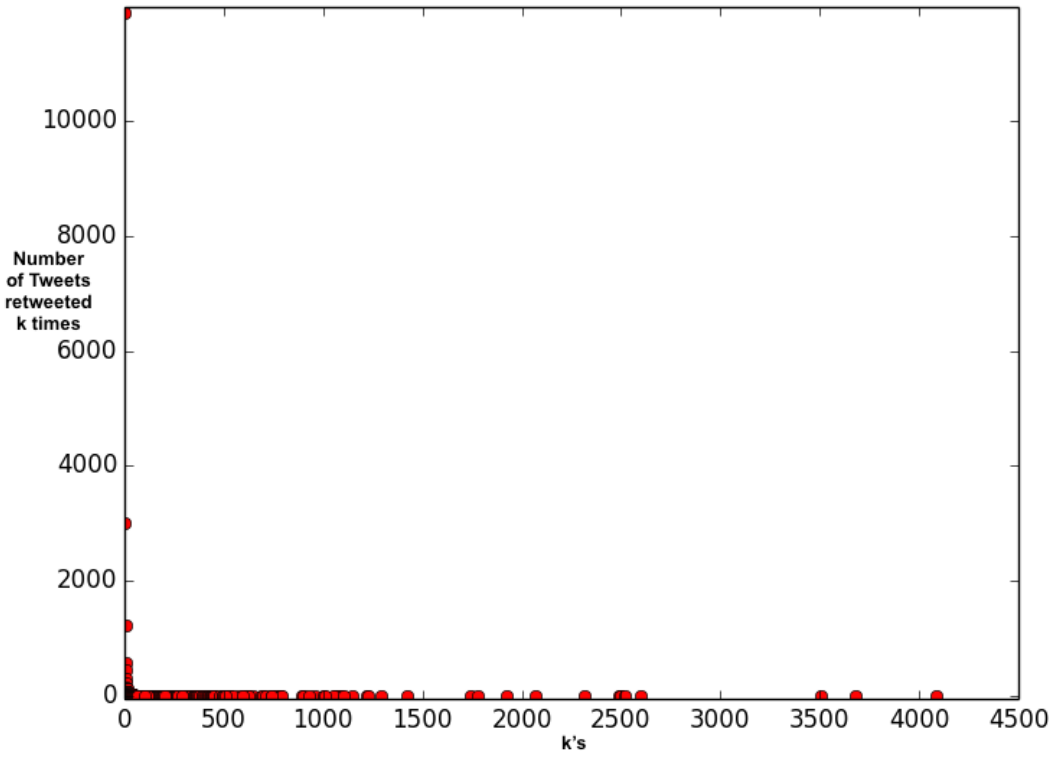

Figure 3: Plot for number of tweets retweeted k times on a linear scale

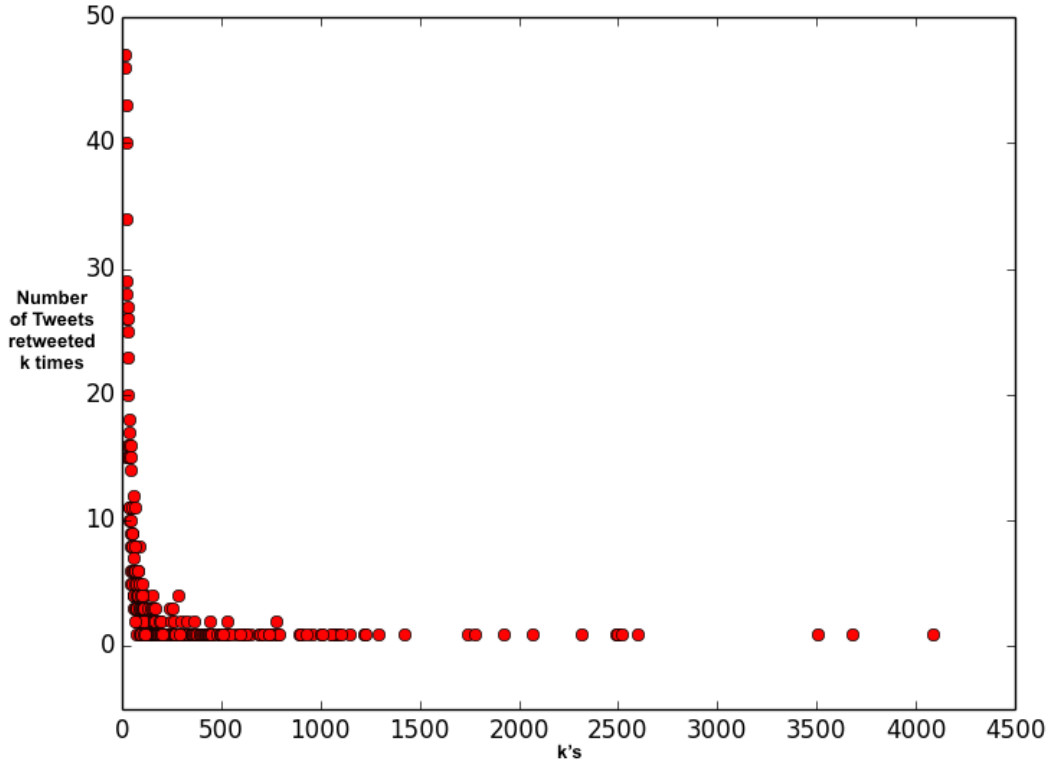

Figure 4: Plot for number of tweets retweeted k times on a linear scale with y-limits adjusted

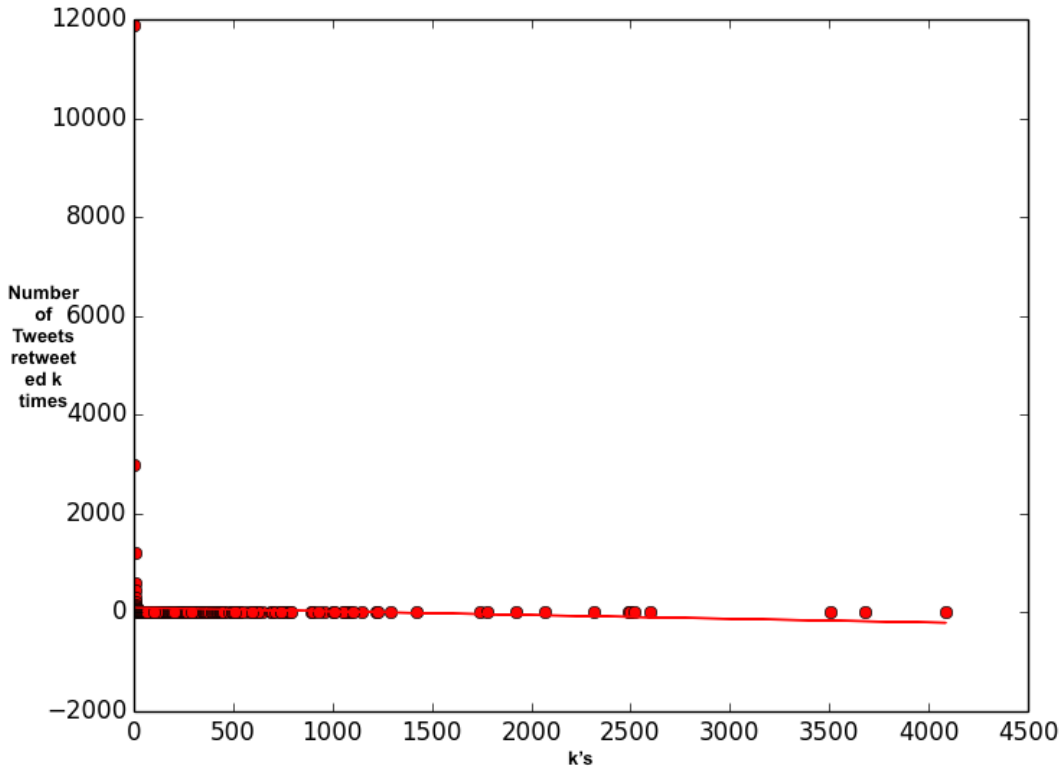

Figure 5: Plot for number of tweets retweeted k times on a linear scale with a line fit to it

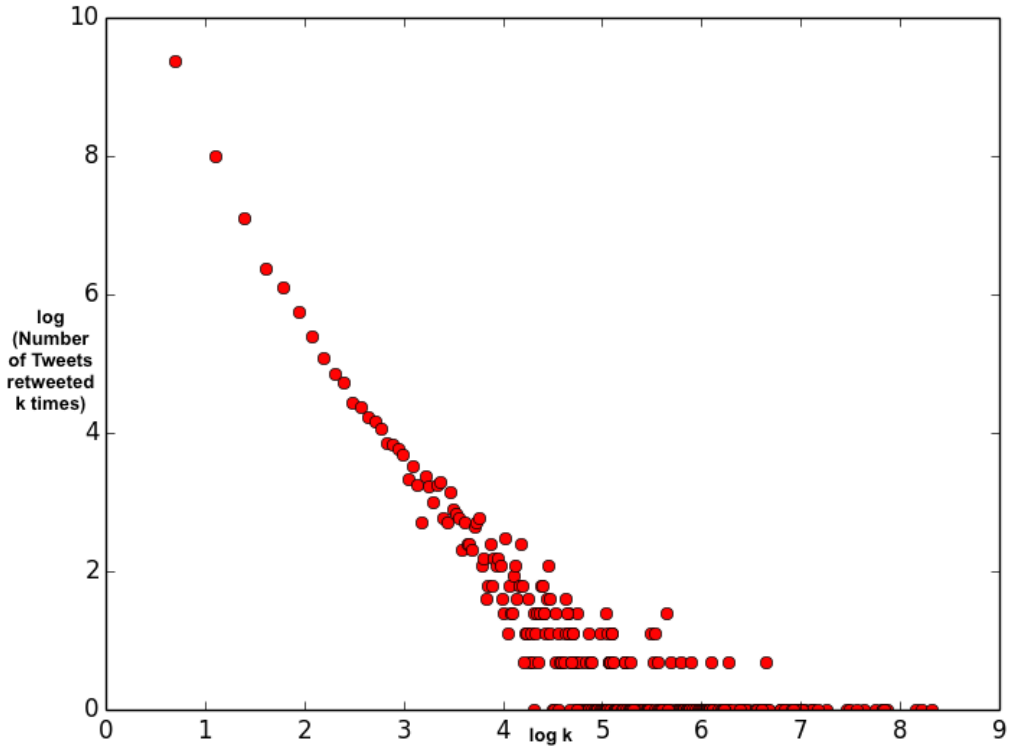

Figure 6: Plot for number of tweets retweeted k times on a log-log scale

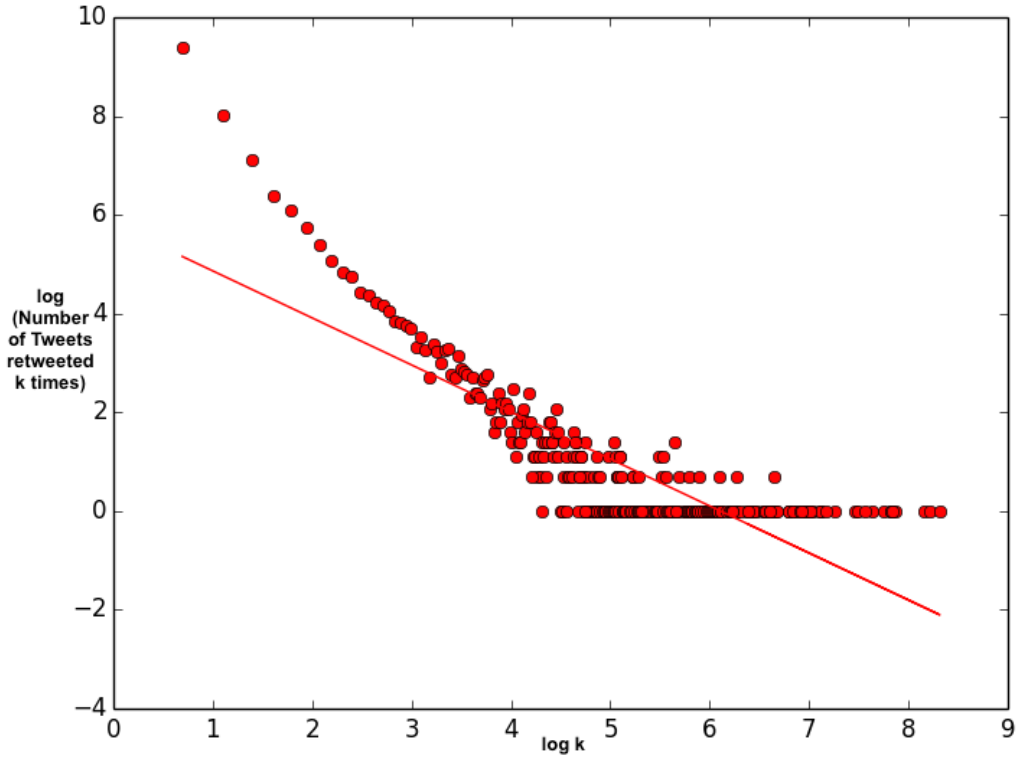

Figure 7: Plot for number of tweets retweeted k times on a log-log scale with a line fit to it

5 Part 5

The graph in Figure 8 shows the correlation between the first and second most popular hashtags which are '#SuperBowl' and '#NFL' respectively in our time slot. '#SuperBowl' has 108411 tweets and '#NFL' has 11659 tweets in total during this time period.

For each of the two hashtags, we found the number of tweets per second. We then used these pairs of values of number of tweets at time 't' for '#SuperBowl' and '#NFL' to plot the graph. Thus each point in the graph represents the number of tweets of '#SuperBowl' and number of tweets of '#NFL' at some time 't'.

In Figure 8, the x-axis represents the number of tweets containing the hashtag '#SuperBowl' per second while the y-axis represents the number of tweets containing '#NFL' per second. It is quite evident that the scale of x is much greater than scale of y as the number of tweets of '#SuperBowl' exceed '#NFL' by about 10 times.

By observing the graph, we can conclude that a linear model can fit the distribution well, and that we can draw a line to approximate the number of tweets for the first and second most popular hashtags. This can be justified by the fact that the tweets containing '#SuperBowl' have a high probability of being accompanied by '#NFL' as SuperBowl is the final game of the national football league. But the inverse is not necessarily true. Hence, as the number of tweets containing '#SuperBowl' increase, the number of tweets containing '#NFL' also increase.

6 Part 6

The file 'tweets.txt' contains all the tweets recovered in Part 2 of this project.

We have used the 'ast' library of python to convert this text file to a dictionary, and then retrieve all the required field values from the dictionary. We used the ast.literal_eval() helper

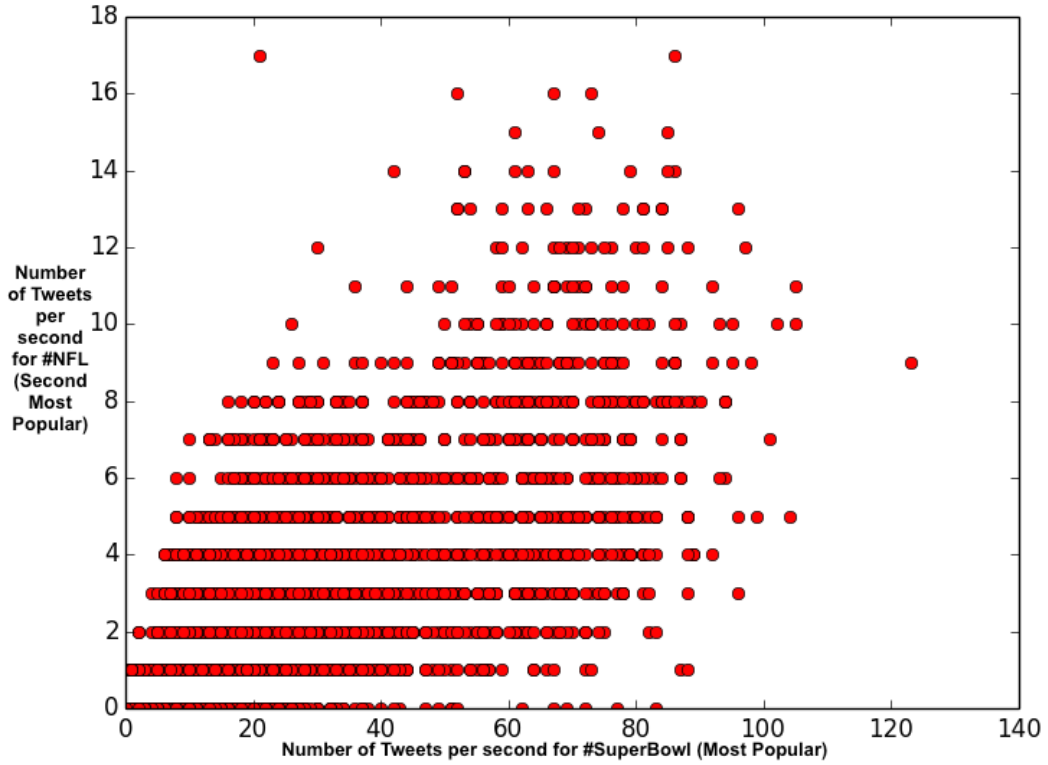

Figure 8: Plot for time rate of most popular hashtag tweets vs time rate of second most popular hashtag tweets

function to evaluate the Unicode encoded tweets.

The tweets post date is in the UNIX format, which needed to be converted back to the local time. This was accomplished by using the datetime library in python. The datetime.fromtimestamp() function converts the UNIX timestamp to a naive datetime object that denotes the local time.

So in this part we iterated through each line of the tweets.txt file and stored it in a dictionary and then displayed the posting date, the tweet text, number of retweets and the user who posted the tweet as specified in the problem statement.

Below are the first five tweets from the file

1. Posting Date: 2015-02-01 18:50:03
   Tweet Text: El Extra es bueno, la ventaja es de nuevo de los Pats. @Patriots 28-24 @Seahawks #SuperBowl `http://t.co/W0DcPGe2Hq` `http://t.co/3CN0iBkD5j`
   Number of Retweets: 29
   User: MedioTiempo

2. Posting Date: 2015-02-01 18:50:04
   Tweet Text: It's all or nothing now...Russell and 2min to go to win. #SuperBowl
   Number of Retweets: 103
   User: David Boreanaz

3. Posting Date: 2015-02-01 18:50:09
   Tweet Text: Russell Wilson is brilliant. It's not over until the plus-sized lady sings #SuperBowl

Number of Retweets: 3
User: Shoq

4. Posting Date: 2015-02-01 18:50:04
   Tweet Text: TOUCHDOWWWWWWNNNNNNNNNNNN!!!!! BRADY BABY! #SUPERBOWL
   Number of Retweets: 2
   User: Marcel Middleton

5. Posting Date: 2015-02-01 18:50:05
   Tweet Text: So this #SuperBowl has come down to one question: Quien es mas Montana?
   Number of Retweets: 12
   User: Brian Murphy

7 Conclusion

To conclude, online social platforms like Twitter which house public opinion on latest events or varied topics like politics and sports serve as vital resources in gauging public views on these events. The sheer volume of data on these social platforms makes it a challenging data mining and machine learning task. As observed through this project, the number of tweets posted in one second of time on a single topic is extremely high. By plotting the collected data on various graphs, we can draw certain inferences about it. This data can also be used in further analysis like opinion mining, sentiment analysis and sarcasm detection using complex machine learning techniques. These systems are now gaining popularity and provide useful information to the commercial industry.