EE239: Crawling and Data Collection on The Web

Amogh Param (UID 704434779)

Karishma Bhatia (UID 304432391)

Naren Nagarajappa (UID 004414529)

Introduction:

In this project we implemented a crawler for Twitter to mine information for the most popular hashtags on Feb 1st 2015 from 15:00:00 to 16:00:00 PST (slot 24). There are multiple ways of mining Twitter for data in this project we use the hashtag and timestamp as the main means to categorize data collection. Using the Twitter API we were able to query the data and get a JSON file which we could further analyze to breakdown the most popular tweets. With this information, we can analyze the public opinion based on tweet trends and their semantic correlation.

Part 1:

We reserved slot number 24 for the project, which consists of the Hashtags '#SuperBowlXLIX', '#Seahawks', '#Patriots', '#GoHawks', '#GoPatriots', '#Halftime' and '#superbowlcommercials', between the time slots 15:00:00 and 16:00:00 on 2/1/2015. For the top_tweets, we chose '#SuperBowlXLIX' as the Hashtag. The search was to find top 5 tweets with the chosen hashtag and output of the search is stored in the 'top_tweets.txt' file. The text field, the user posting the tweet and the posting date for each of the top 5 tweets is listed here:

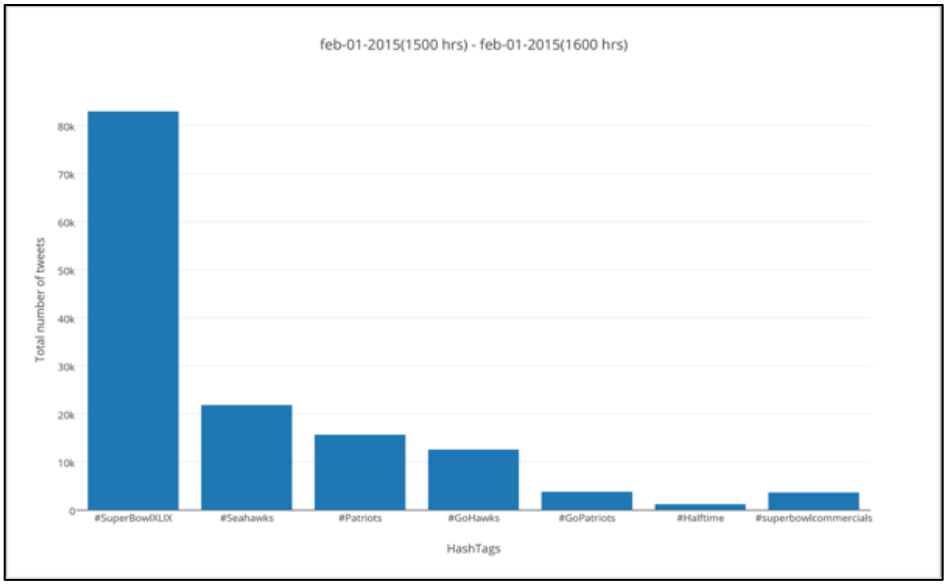
Text: "Are you ready for #SuperBowlXLIX? Here are 5 things to know about the big game: http://t.co/t0xa0URZ04 #SuperBowl http://t.co/OZWCnaYoer" **User :** CNN **Post Date :** 2015-02-01 15:07:08
Text: "Agradezco a @PepsiCo @pepsi y a su CEO, la destacada Indra Nooyi, por su invitación al #SuperBowlXLIX" **User :** Felipe Calderón **Date :** 2015-02-01 15:07:36
Text : "#NowPlaying @SNFonNBC #SuperBowlXLIX #SuperBowl" **User :** Jaime Camil **Date :** 2015-02-01 15:04:25
Text : "Let's do this @Seahawks #SuperBowlXLIX #Seahawks #GoHawks http://t.co/5aMTeZ6hFE" **User :** Courtney Sixx **Date :** 2015-02-01 15:25:21
Text : "ARE YOU READY FOR SOME FOOOOOTBAAAAAALLLL?! #SuperBowlXLIX http://t.co/ldppffISty" **User :** E! Online **Date :** 2015-02-01 15:30:10

Part 2:

In this section we wanted to focus on finding the tweets for #SuperBowlXLIX, #Seahawks, #Patriots, #GoHawks, #GoPatriots, #Halftime, and #superbowlcommercials. To do this we created a similar file as in Part 1. Since the maximum amount of tweets is 500 per query we decreased our time span to intervals of 10 seconds to find all the tweets for our selected time interval of 1 hours on Feb 1st. We stored all the tweets for the hashtags in the *tweets.txt* file and stored all the log information for each query of 10 seconds to *search_log.txt*. Looking through this file it can be observed that there are significantly more tweets for #SuperBowlXLIX than for any of the other hashtags. A potential explanation for this can be inferred by the generality of the hashtag. This means that the more general a hashtag (i.e., #SuperBowlXLIX is more general than #Patriots -- it can be looked at as #SuperBowlXLIX is a superset of #Patriots) the more likely it is to be attached to all specific hashtags about the same general event. We will examine this trend further in Part 3.

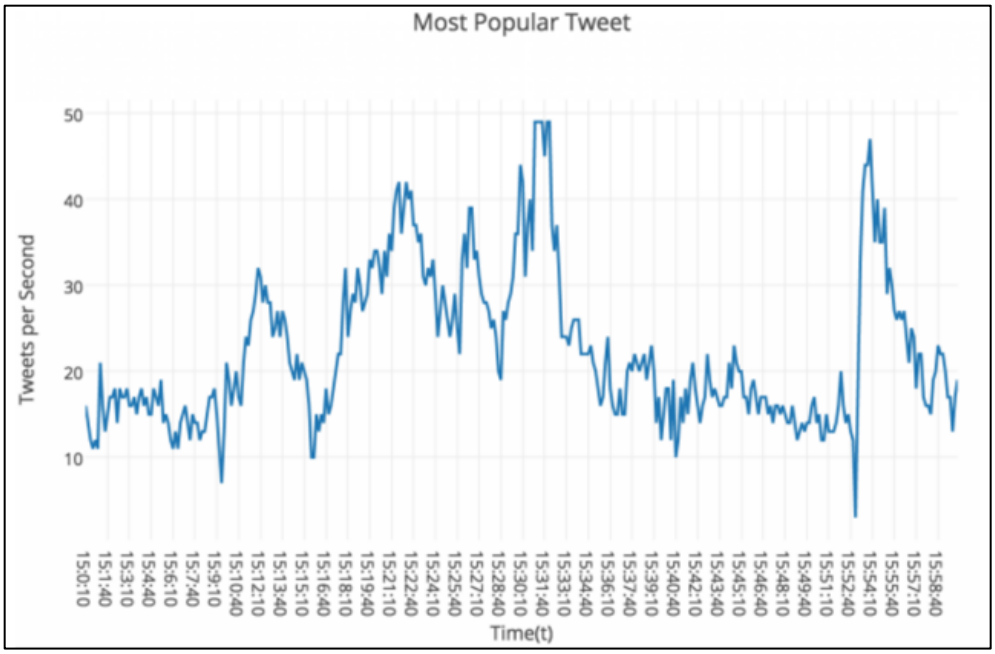
Part 3:

In this section we took the data from Part 2 and created a bar graph to analyze the overall trend of the hashtags mined.



This graph shows that the most popular hashtag was for #SuperBowlXLIX - in fact, it was more than triple the amount of the next most popular hashtag. Given this information we can draw some assumptions about public sentiment during the Super Bowl. For example, we can hypothesize that tweets from less general categories (i.e, specific team tweets or specific show segment tweets) will have both the less general hashtag (#Seahawks or #Patriots) and the more general hashtag (#SuperBowlXLIX), so the most general hashtag.

We then took the most popular hashtag, #SuperBowlXLIX, and created a graph showing the tweets per time interval to analyze information about the progression of tweets.

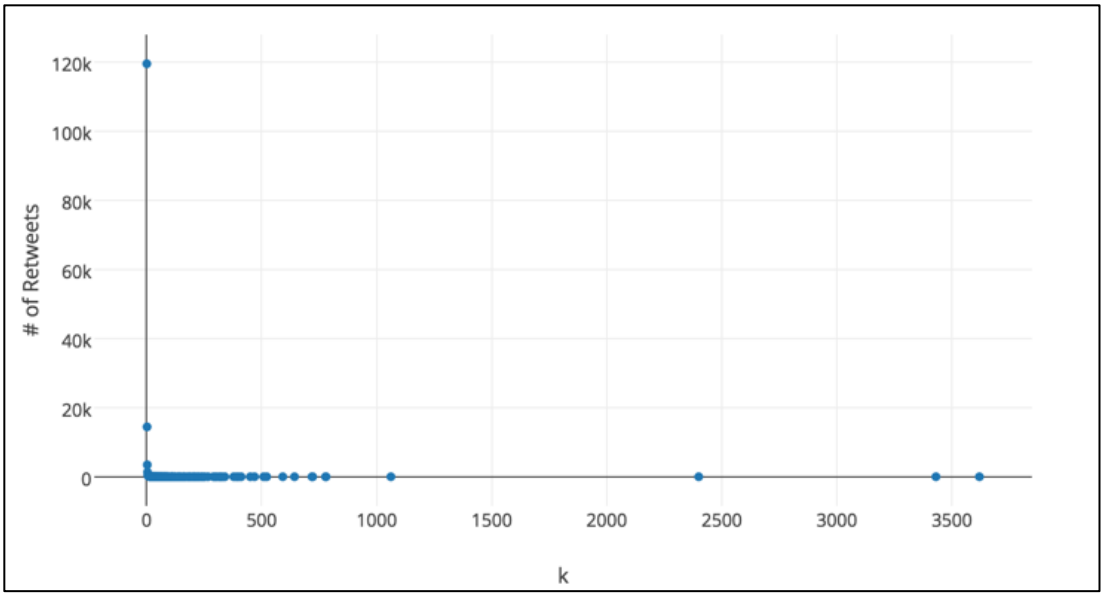


The kickoff for SuperBowlXLIX was at 6:30pm Mountain Time - since our data is in PST we will use 5:30pm (17:30 PST) as the kickoff time. With this information we can see that our tweets were generated before the game started at 2:00pm local time in Arizona (3:00pm PST). To try and understand the public sentiment additional information is needed on what events were happening at this time. NBC broadcasted a slew of interviews as a pre-show rally on national TV from noon to 6:00pm. Additionally, fans at the stadium would be participating in pre-game tailgates ensuing excitement and fans at home would be preparing for super bowl screening parties. With this information we can hypothesize that the hashtag spikes at the 30min and 1 hour blocks can be attributed to NBC interviews that end at the 30min and 1 hour markers, since all the other per-game events are continuous and do not specifically correlate to the times of the spikes.

Additionally this added information can be used to hypothesize why the #SuperBowlXLIX is the most popular at the time interval. This can be hypothetically attributed to fans that are very excited about the game in general and not necessarily tweeting about the actual team/events at the game since it has not started.

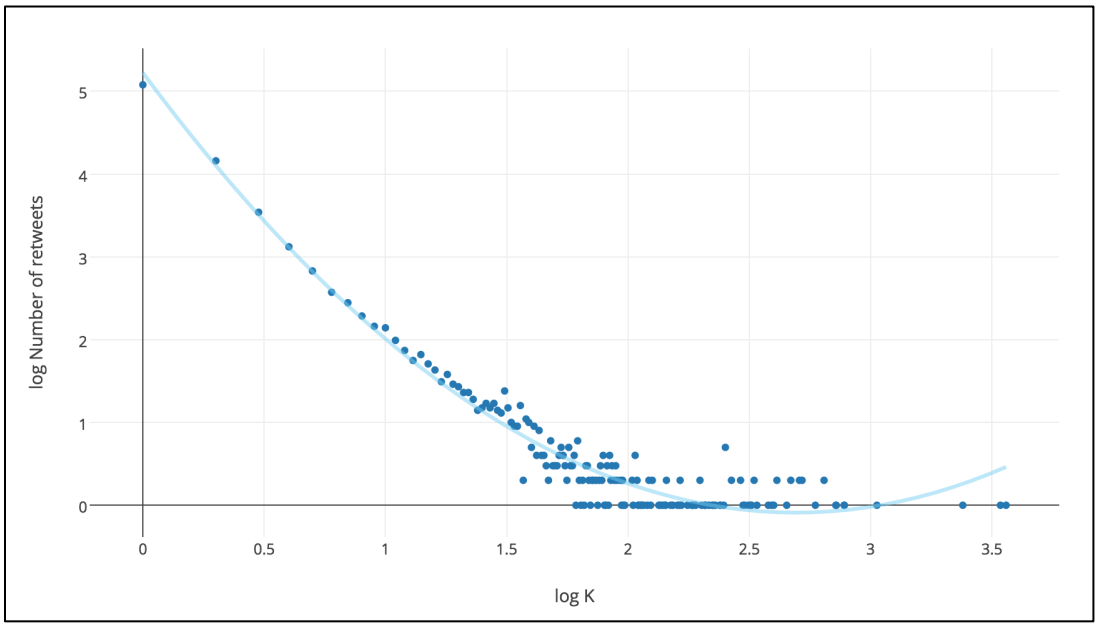
Part 4:

In this section we print out the number of tweets per retweet amount (k). We plot the results in the below graph:



This graph shows that there are only a few tweets that were retweeted more than 50 times. Whereas, a large number of tweets were retweeted under 15 times. This could be explained by the social influence and following that celebrities have on social media. The TOPSY API seems to provide incorrect values for the retweet count accessed by the index path: `tweetObject["tweet"]["retweet_count"]`. Therefore, we use the following index path instead, to get a better estimate of the actual retweets: `tweetObject["metrics"]["citations"]["total"]`

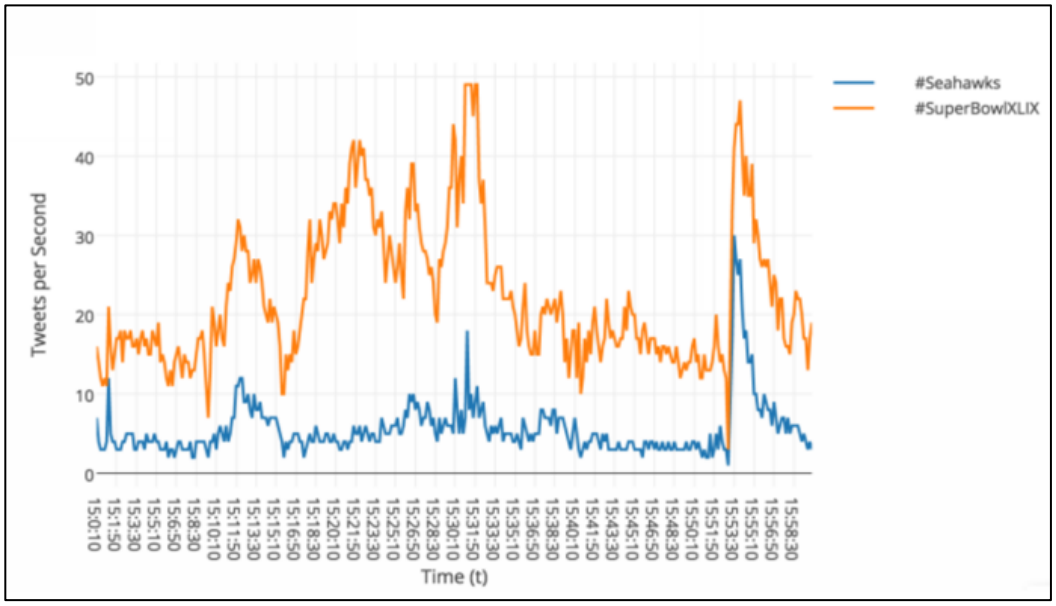
Additionally we graphed the above metrics in the log-scale, which can be seen here:

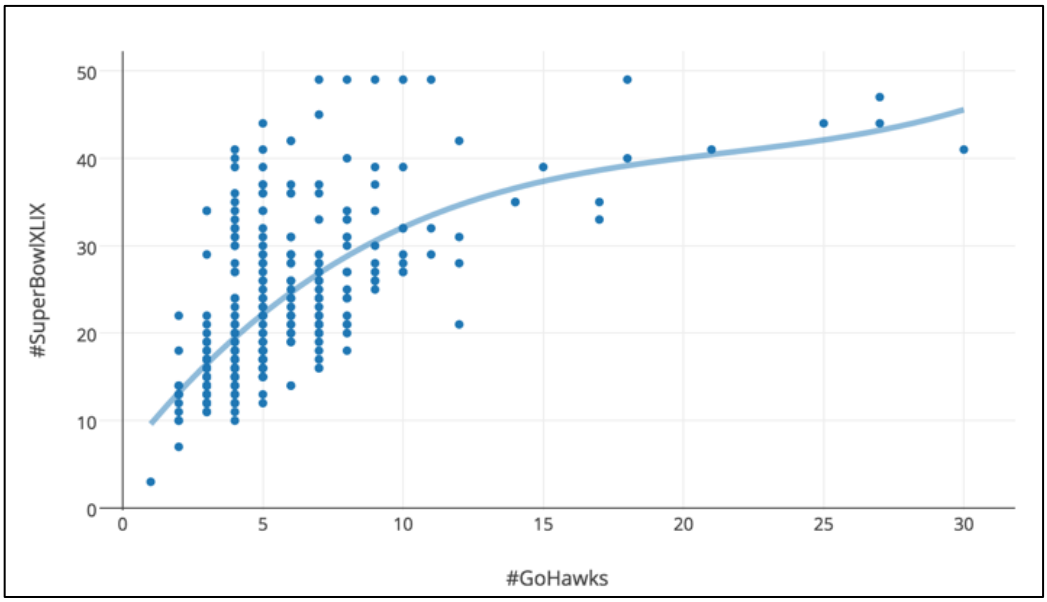


With this graph we can fit a quadratic function as shown in light blue above.

Part 5:

In this part we wanted to analyze the trends of the most popular and second most popular hashtags. To do this we took the data from part 2 and plotted both tweets per time interval against the time interval.





Analyzing this graph should be easy to see that there is a correlation between the progression of tweets between #SuperBowlXLIX and #Seahawks. When there is a spike for #SuperBowlXLIX there is also a spike for #Seahawks that follows a similar pattern to the #SuperBowlXLIX tweets. Interestingly there is a spike at exactly 15:53 for both hashtags, which can be potentially attributed to an external factor of that time.

In the second graph we fit a 3rd degree polynomial function to show the fit and the correlation of the hashtags.

Part 6:

In this part we make a parser that can return output of with a tweet's post date, text, number of retweets, and the user posting it.

Concluding Notes:

For each of the above parts, there is a separate Python script in the `src/` folder. In order to test any of the parts, please run the corresponding script file.