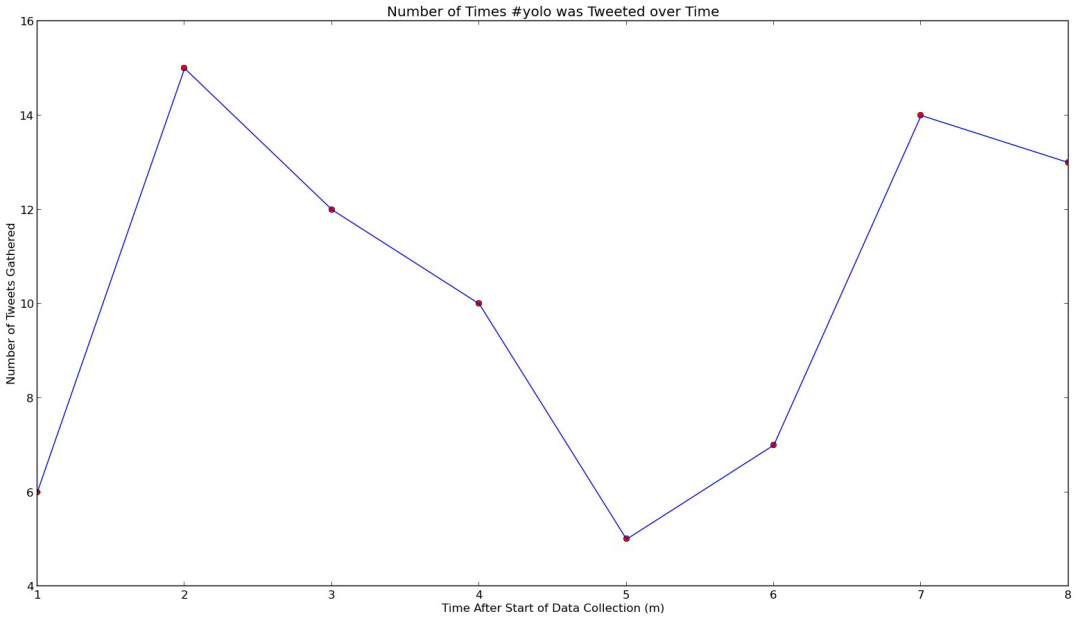
Anders Johnson
Software Design
9/29/14

Data Mining Project Write-Up:

Project Overview:
The goal of my project was to plot the frequency at which certain hashtags were used in tweets over a specific time period. I used a Twitter API to live stream tweets and save them into a text document with a corresponding time stamp in order to find the frequency of the hashtag usage per minute.

Implementation:
My code is separated into two scripts, “streaming_tweets.py” and “my_data.py.” The first script that needs to be used is “streaming_tweets.py.” Its main purpose was to use the pattern library to live stream tweets that contained #yolo, #smh, and #ftw and store them into separate text files for each hashtag with a time stamp. I encountered a lot of trouble with maxing out the Twitter API and could only run data collection of one hashtag at a time, so I gathered tweets that contained #yolo for 8 minutes and stored it in the file “yoloTester.txt”. When I attempted to gather data for the other two hashtags, my API was maxed out and giving me an error saying that I had sent Twitter my license key too many times. This made it impossible for me to collect data for #smh and #ftw, but I left them in my code anyways.

The next script that should be run is “my_data.py.” First, I put all of my data from “yoloTester.txt” into a dictionary. However, I wasn't able to organize the dictionary, so I had to hard code the data to plot it in the correct order, which was something that I truly wanted to avoid. The script then used pyplot from the matplotlib library to generate a line plot of the data, which can be found below.



Reflection:

Overall, I had mixed feelings about this project. I really learned a lot from it, and enjoyed mining data from Twitter. I also developed better skills at searching the web for relevant information to help me with Python, rather than just staring at my keyboard not knowing what to do. However, I found the Twitter API extremely frustrating to work with. Even though my code would work at one point of the day, it might not necessarily work at another point. This is the main reason why I only have one line plotted on my graph as opposed to three. I could not get data from Twitter, even though the code was perfectly fine since it was an exact replica of the working code that I had used just minutes prior. Because of the difficulties with the API, I did not get as far in the project as I would have hoped and did not enjoy it as much as I wanted to.

I felt that my project was initially scoped well, but my final output was not what I had hoped. I wanted to have real data for three hashtags over a longer time period than five minutes. However, because of difficulties with the Twitter API, I had to settle for what I had. Looking back, if I had started data mining earlier, I may not have run into as many issues because I would have collected data over a a more spaced out time period. In the end, I am really pleased that I was able to plot something, but if I had the time I would definitely be interested in doing future work on my topic and diving deeper into it.