Popularity Prediction on Twitter

Introduction

Predicting popularity of an event or subject is useful problem with numerous applications in fields such as social network analysis. Twitter is an example of a platform that provides a public discussion service that allows us to research such problems. Twitter, as a company, is very interested in these phenomenon and is able to see trending hashtags and oftentimes, their popularity is a reflection of popularity of public sentiment on that topic with implications in the real world. We have seen multiple examples of this, as recent as this year with #Blacklivesmatter leading to protests across the United States or a few years ago with the Arab Spring protests across the middle east that were organized through twitter and social media platforms.

We will attempt to use Twitter to predict popularity of topics through hashtags. We will focus on Superbowl 49 that happened on February 1st, 2015. We want to focus on hashtags that will have a high volume of tweets to allow for proper analysis of a bursting fashion, in which the number of tweets will peak and then decrease. We quantify the popularity of a topic as the number of tweets that are posted using that particular hashtag. Using multiple features of tweets from multiple time frames, we will attempt to build a regression model that can predict number of tweets for consecutive time frames.

Analyzing Hashtag Statistics

We looked at 6 hashtags for the purposes of this study: #superbowl, #nfl, #gopatriots, #gohawks, #patriots, #sb49. We divided the tweets into timeframes of one hour. The average number of tweets per hour, follower of users posting the tweets, and number of retweets are listed below for each of the six hashtags in table 1. The #superbowl hashtag had the highest average number of tweets at around 1400 tweets per hour, followed by #nfl hashtag with approximately 279 tweets per hour. There is a dramatic drop off in the other hashtags compared with #superbowl. This is indicative that the Superbowl is a much more important social event rather than just sports, as there are many people posting about the superbowl event whereas there are less tweets about the actual sport side of the event with hashtags such as nfl, gopatriots or gohaws.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Hashtag | #superbowl | #nfl | #gopatriots | #gohawks | #patriots | #sb49 |
| Average number of tweets (per hour) | 1399 | 279 | 38 |  |  |  |
| Average number of retweets (per hour) | 3341 | 429 | 53 |  |  |  |
| Average number of followers of users posting tweets (per hour) | 3341 | 429 | 53 |  |  |  |

Table 1: Average number of tweets per hour, follower of users posting the tweets per hour, and number of retweets per hour are listed for each hashtag.

The histograms for each of the hashtags are depicted in figures 1-3 below. From the plots we see that users were extremely active around the time of the super bowl, posting a high number of tweets. We see a peak in the number of tweets at one point during the superbowl game. The hashtag #superbowl saw a staggering 272,322 tweets between 5-6 PM on February 2, 2015 which was during the game.

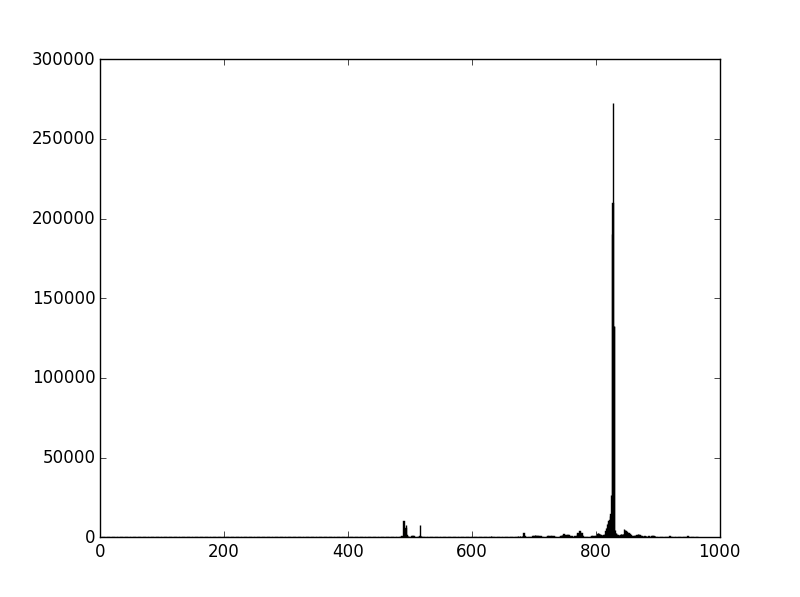


Figure 1: Histogram of number of tweets in hour for #superbowl

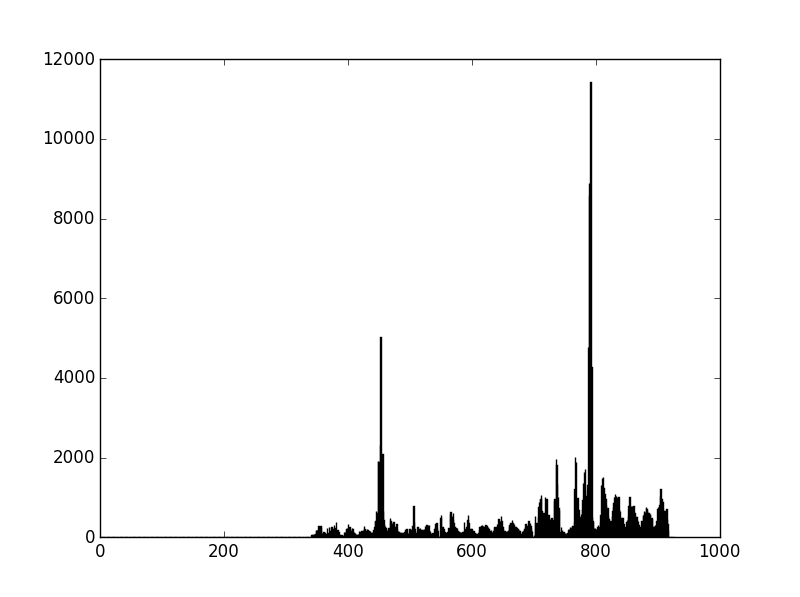


Figure 2: Histogram of number of tweets in hour for #nfl

Building A Linear Regression Model

For the regression analysis we used the ordinary least squares model from the statsmodels library. We created a matrix of the five features specified and passed that into the model. For the predicted value, we set yi = num\_tweetsi+1. For the last item in y, we just set it to yn-1. We pass that into the OLS algorithm to generate the model. An example of our regression result is illustrated below.