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.

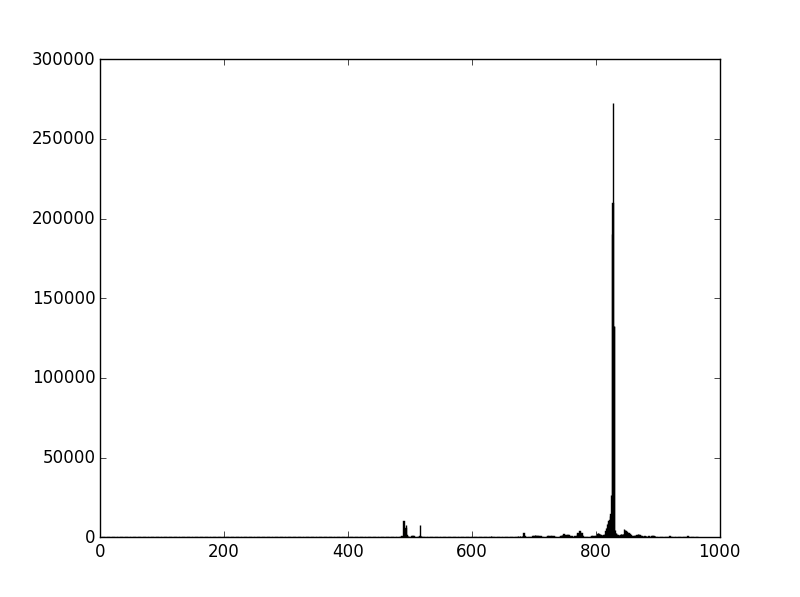
**Analyzing Hashtag Statistics**

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 gohawks.

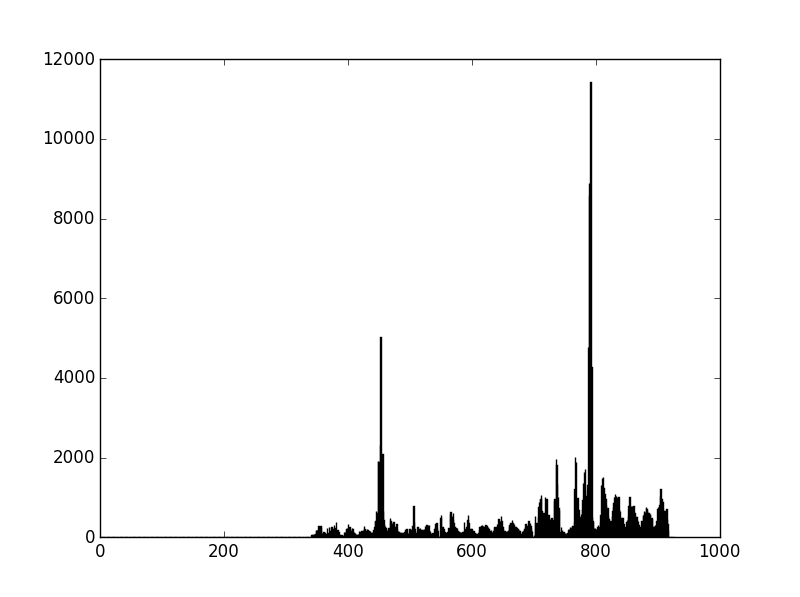
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Hashtag | #superbowl | #nfl | #gopatriots | #gohawks | #patriots | #sb49 |
| Average number of tweets (per hour) | 1399 | 279 | 38 |  |  |  |
| Average number of retweets (per hour) | 3341 | 429 | 53 |  |  |  |
| Average number of followers 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 super bowl 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 ordinary least squares (ols) regression to build a machine learning model from the statsmodels library. We created a matrix of five features for the ols model, split up into one hour time intervals for each data point. For each hashtag we collected the following features to make an ols model:

* Number of tweets: Number of tweets posted in that hour for that hashtag
* Number of retweets: Number of retweets posted in that hour for that hashtag
* Number of followers of users: Total number of followers of users that posted in that hour for that particular hashtag
* Maximum number of followers of users: Maximum number of followers, not a summation, of all users that posted in that hour for that hashtag
* Time of day: Time of day that the hour of features corresponds to. For example, features for 10 AM – 11 AM would number 10 for the time of day feature.

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, using the hashtag #nfl, is illustrated below.

OLS Regression Results

==============================================================================

Dep. Variable: y R-squared: 0.717

Model: OLS Adj. R-squared: 0.715

Method: Least Squares F-statistic: 465.3

Date: Fri, 20 Mar 2015 Prob (F-statistic): 5.43e-249

Time: 06:02:09 Log-Likelihood: -6820.0

No. Observations: 926 AIC: 1.365e+04

Df Residuals: 920 BIC: 1.368e+04

Df Model: 5

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

-------------------------------------------------------------------------------------------------------------------------------

const 86.7198 24.896 3.483 0.001 37.860 135.579

x1 0.3037 0.111 2.743 0.006 0.086 0.521

x2 -0.0037 0.055 -0.067 0.946 -0.112 0.104

x3 0.0001 1.59e-05 8.236 0.000 0.000 0.000

x4 -0.0001 2.19e-05 -6.409 0.000 -0.000 -9.75e-05

x5 -1.0519 1.826 -0.576 0.565 -4.635 2.531

==============================================================================

Omnibus: 986.013 Durbin-Watson: 1.983

Prob(Omnibus): 0.000 Jarque-Bera (JB): 451093.615

Skew: 4.278 Prob(JB): 0.00

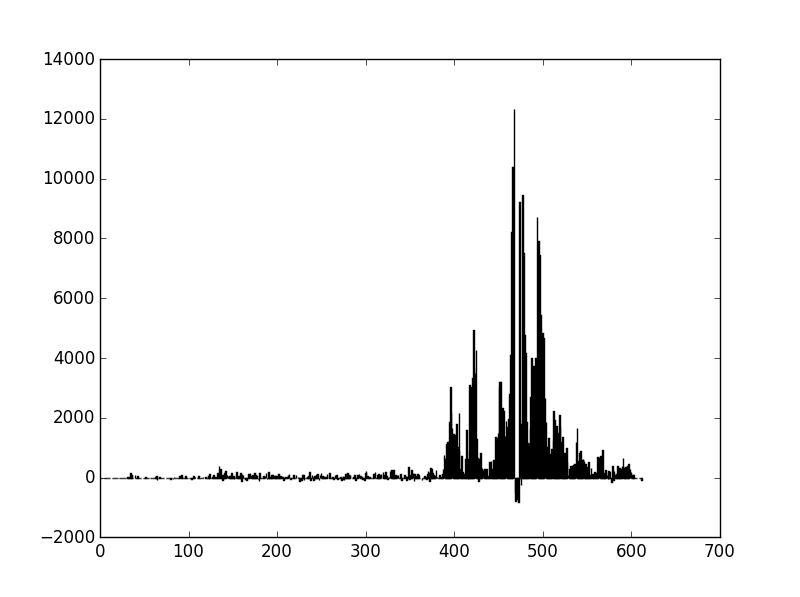
Kurtosis: 110.788 Cond. No. 7.77e+06

==============================================================================

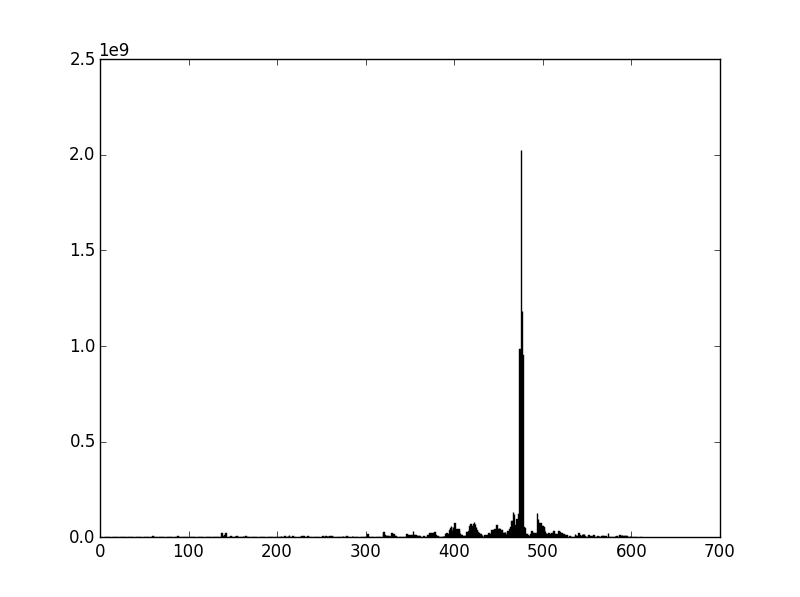
**An Alternative Regression Model**

Some of the alternative features we looked at from the tweets were the fields were the ‘impression’ and ‘acceleration’ of a tweet. These features correlated better with the number of so we decided that these would yield better results. The histograms of these features shown below illustrate the correlation with the number of tweets. In the alternative model, we did not use the number of followers of users and the max number of followers.

* Total Impression: Volume of tweet mentions over time for list of keywords.
* Total Acceleration: Sum of acceleration. Acceleration of tweet ranges from -100 to 100.



*Figure* : Total Accelerations for #superbowl



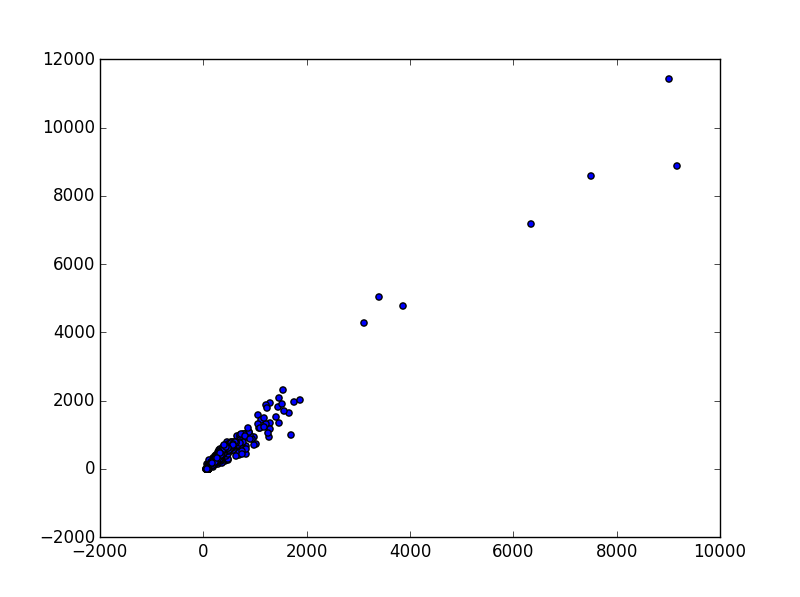
*Figure :* Total Impressions for #superbowl

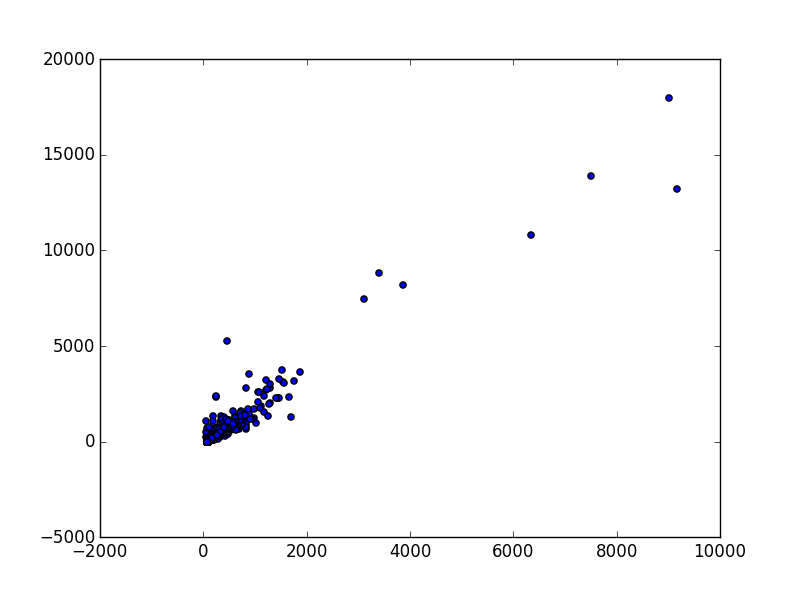
The OLS regression results for #nfl are printed below. By comparing them to the results from the original feature set, we see that there is not a significant improvement in the results. Because of that we decided to continue the next sections with the original five feature sets.

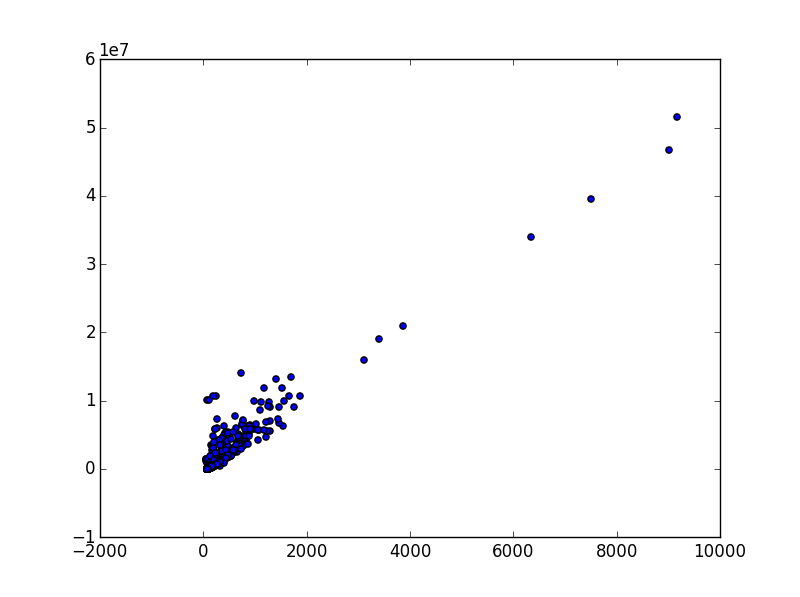
|  |
| --- |
| OLS Regression Results  ==============================================================================  Dep. Variable: y R-squared: 0.575  Model: OLS Adj. R-squared: 0.572  Method: Least Squares F-statistic: 165.3  Date: Fri, 20 Mar 2015 Prob (F-statistic): 6.56e-111  Time: 21:16:04 Log-Likelihood: -4775.6  No. Observations: 616 AIC: 9563.  Df Residuals: 610 BIC: 9590.  Df Model: 5  ==============================================================================  coef std err t P>|t| [95.0% Conf. Int.]  ------------------------------------------------------------------------------  const 127.4792 46.517 2.740 0.006 36.126 218.833  x1 0.7360 0.159 4.626 0.000 0.424 1.048  x2 -0.2563 0.080 -3.211 0.001 -0.413 -0.100  x3 9.89e-05 2.35e-05 4.203 0.000 5.27e-05 0.000  x4 -8.161e-05 3.36e-05 -2.426 0.016 -0.000 -1.56e-05  x5 -0.2852 3.349 -0.085 0.932 -6.862 6.292  ==============================================================================  Omnibus: 719.201 Durbin-Watson: 2.324  Prob(Omnibus): 0.000 Jarque-Bera (JB): 408387.471  Skew: 4.793 Prob(JB): 0.00  Kurtosis: 128.775 Cond. No. 9.91e+06  ============================================================================== |

When looking at the regression results we saw that with these features, there was not a significant improvement in the R-squared error.

**Analyzing Features Chosen**







**Building Three Different Models**

We built three different models to match the bursty nature of the data, as the pre and post burst time periods would exhibit unique combinations of features leading to unique regression models. The first model uses the weeks and times leading up to the day of the Superbowl. The second model uses the time period of the majority of the day of the Superbowl, with the last model being built from data of tweets that were posted or happened after the Superbowl.

* Before Feb 1, 8:00 AM – The first model, corresponding to before the Superbowl
* Between Feb 1, 8:00 AM – 8: 00 PM – The second model corresponding to tweets on the day of the Superbowl
* After Feb 1, 8:00 PM – The third and final model built from tweets after the Superbowl

**Cross Validation**

When doing cross validation, we split the data into ten parts. We had training indices which consisted of 9/10th of the data and the test indices were 1/10th of the data. We ran ten iterations in which we built a regression model of the test data and then predicted the data for the test indices. To test the accuracy of the model we calculated the mean squared errors to compare the predicted values versus actual values of the test data. The indices were selected at random at each iteration and there were no overlaps of indices in the test data between iterations. This ensured that the data was thoroughly tested.

**Testing Data Results** – Predict the number of tweets in the next hour: