EE239AS Project 4

Popularity Prediction on Twitter

Winter 2016

Team members:

Rebecca Correia (204587944)

Salil Kanetkar (704557096)

Vedant Patil (104590942)

Introduction

Twitter is one of the most popular ways for people to share information about a large variety of trending topics. Nowadays, more and more people are engaging themselves in this ever-growing online social networking service. Twitter basically allows its users to post and read short messages called "tweets" that have a maximum limit of 140 characters. Not surprisingly so, Twitter was named as the "SMS of the Internet" in 2013.

The way Twitter functions is pretty straightforward – Users can post "tweets" about any topic of their interest that can be subscribed to by other users called "followers". If a follower chooses to share the tweet, the activity is called as "retweeting". Users are also allowed to mark any tweet that they like as their "favorite". One of the most interesting concepts that was introduced by Twitter, however, was that of a "hashtag". The tweets that are posted by users could be grouped together by either topic or type simply by using some catchy words or phrases prefixed with a "#" symbol, also called as a hashtag. Additionally, Twitter also enables users to call for other users' attention towards a tweet or activity by typing in a "@" symbol. This function is called as "mentioning a user".

Owing to the rich information that tweets and user activities on Twitter could provide, Twitter has proved to be a great platform for performing a variety of social network analysis right from predicting how the audience would react to a particular event in the future to predicting the outcome of elections. One such analysis involves studying the popularity of hashtags. There are several factors that trigger a hashtag to become popular, such as occurrence of a major event, tweets by a famous person or even something controversial.

The aim of this project is to make use of the underlying structure of Twitter to make predictions about the popularity of a certain given hashtag. We have been given Twitter data grouped according to six popular hashtags related to the Super Bowl event held in 2015 collected over a stipulated time period. Given the trends for the tweets belonging to the different hashtags over a period of time, our task is to predict the amount of popularity that each of the hashtag will attain in the future.

Problem 1

In this part, our task is to explore the training tweet data and calculate some important statistics for each hashtag such as the average number of tweets per hour, average number of followers of users posting the tweets and average number of retweets.

The first calculation that we intended to make was that of the time elapsed for any given hashtag. For this purpose, we initialized the start_time and the end_time to some extreme values against which we could calculate the total number of hours elapsed easily. We parsed each line, i.e. an individual tweet, contained within a hashtag file as a JSON object called as 'individual_tweet_data'. By using the documentation of Twitter API as a reference, we then extracted the information related to the attributes of our interest from the tweet data as follows:

retweets = individual_tweet_data["metrics"]["citations"]["total"]

user_id = individual_tweet_data["tweet"]["user"]["id"]

followers = individual_tweet_data["authors"]["followers"]

date = individual_tweet_data["firstpost_date"]

Once we computed the appropriate desired counts from the attributes given above, we calculated the statistics using the following formulae:

Average number of tweets per hour = $\frac{Total\ no.\ of\ tweets}{Total\ no.\ of\ hours\ elapsed}$

Average number of followers of users = $\frac{Total\ no.\ of\ followers}{Total\ no.\ of\ unique\ users}$

Average number of retweets = $\frac{Total\ no.\ of\ retweets}{Total\ no.\ of\ tweets}$

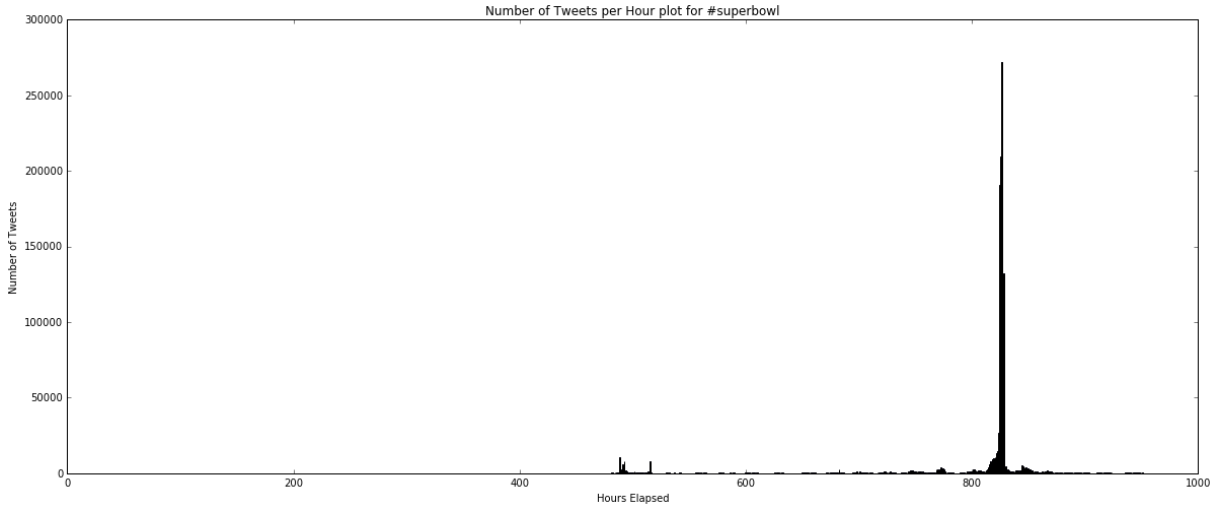
This process was repeated for all the six hashtags that were provided to us, thereby yielding the following results:

Statistics:

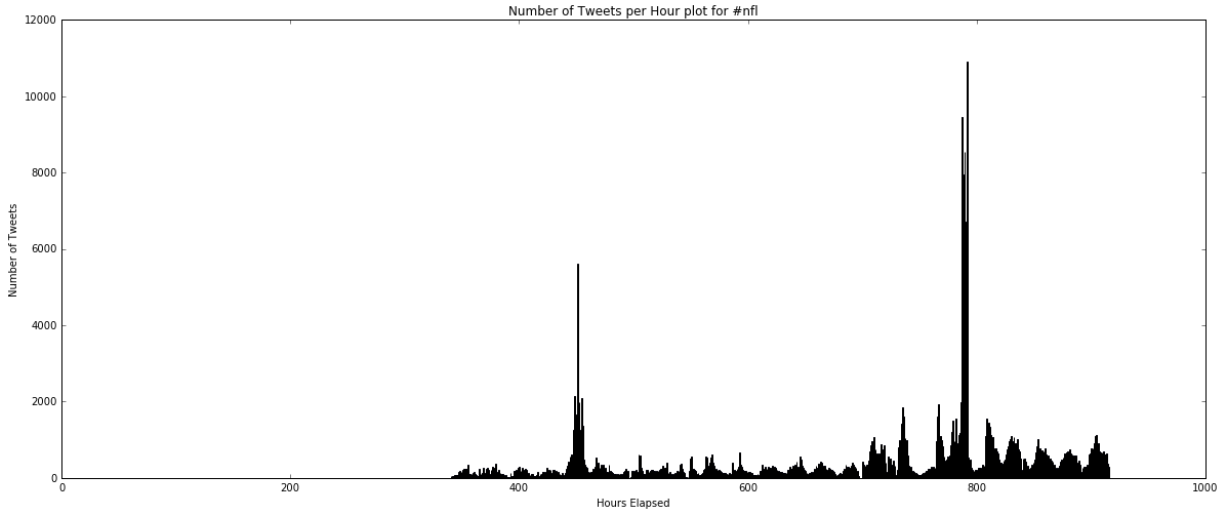
Hashtag	Average no. of tweets/hour	Average no. of followers of users	Average no. of retweets
#gohawks	193.556	1544.969	2.014
#gopatriots	38.407	1298.824	1.400
#nfl	279.421	4289.746	1.538
#patriots	499.197	1650.321	1.782
#sb49	1420.878	2235.164	2.511
#superbowl	1400.588	3591.604	2.388

Histogram:

"Number of tweets in hour" over time for #superbowl



"Number of tweets in hour" over time for #nfl



As we can see from the above histograms, there were burst occurences during both, Superbowl as well as NFL. Looking at the hourly trends of the superbowl event, we can observe that the number of tweets shot up massively at around the 830th hour. On the other hand, NFL displayed more burstiness with respect to its number of tweets per hour with the highest number of tweets being posted at around the 800th hour.

Problem 2

In this part, our task is to make use of 5 features, namely, number of tweets, total number of retweets, sum of the number of followers of the users posting the hashtag, maximum number of followers of the users posting the hashtag and the time of the day, in order to fit a linear regression model that makes predictions for the next hour, given the previous hour data.

For any given hashtag, in order to generate the training set, we first extract each of the above features from the tweet data on an hourly basis i.e. considering a 1-hour window. Keeping the hourly window into consideration, we have calculated our features in such a way that any timing between 'n' hours and 'n+1' hours gets considered as the 'nth' hour. In saying so, each time a new timing is encountered, we first set the features for the corresponding hour to a default value of 0, and increment it appropriately each time it appears again in the dataset. Once we have the hourly features in hand, we create a predictors-labels matrix such that for each hourwise features, the label indicates the tweet count for the next hour. This predictors-labels matrix is then used to fit a linear regression model using OLS.

Observations

Accuracy

We have used R-Squared as a parameter for accuracy. It is a measure of how well your model has fit the data. A value of '1' indicates a perfect fit. Although it should not be '1'. This indicates that your model is over fitting the data.

T-test

T-test is basically used for assessing how statistically significant a particular explanatory variable is. It is important to find such significant explanatory variables in order to fit the regression model most efficiently. It is calculated by dividing the estimated coefficients of the parameters used to fit the model by their standard error. Higher the value of the t-statistic, more is the likelihood that the parameter has an actual value that is different from zero. This means that we would want to consider the features with the highest t-test values to be significant.

P-value

Given a null hypothesis for the probability distribution of the data, the p-value can be said to be the probability that the outcome would be atleast as extreme or probably even more extreme than the outcome that was observed. As the p-value goes on decreasing, the evidence against this null hypothesis keeps increasing. This means that we would want to consider the features with the lowest p-values to be most significant.

The results obtained after fitting the OLS regression model can be seen as follows, arranged according to the 6 hashtags:

X1: Maximum followers count

X2: Number of retweets

X3: Time

X4: Tweet Count

X5: Follower Count

1. **#gohawks**

OLS Regression Results

Dep. Variable:	y	R-squared:	0.488
Model:	OLS	Adj. R-squared:	0.486
Method:	Least Squares	F-statistic:	184.4
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	7.69e-138
Time:	19:49:37	Log-Likelihood:	-7811.3
No. Observations:	972	AIC:	1.563e+04
Df Residuals:	966	BIC:	1.566e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	66.6970	46.808	1.425	0.155	-25.159	158.553
x1	0.0004	8.16e-05	4.478	0.000	0.000	0.001
x2	-0.1657	0.043	-3.823	0.000	-0.251	-0.081
x3	1.9088	3.490	0.547	0.585	-4.939	8.757
x4	0.5770	0.122	4.747	0.000	0.338	0.816
x5	-0.0006	0.000	-4.709	0.000	-0.001	-0.000

Omnibus:	1840.941	Durbin-Watson:	2.337
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4354376.854
Skew:	13.180	Prob(JB):	0.00
Kurtosis:	329.834	Cond. No.	3.17e+06

The ***R-Squared*** value is only ***0.488*** which means that the model did not fit very well. From the p and t values we can see that the significant features are ***Maximum follower count*** and ***Tweet Count***.

2. **#gopatriots**

OLS Regression Results

Dep. Variable:	y	R-squared:	0.664
Model:	OLS	Adj. R-squared:	0.662
Method:	Least Squares	F-statistic:	267.6
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	1.19e-157
Time:	19:49:43	Log-Likelihood:	-4447.7
No. Observations:	683	AIC:	8907.
Df Residuals:	677	BIC:	8935.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	8.0471	12.254	0.657	0.512	-16.013	32.107
x1	0.0011	0.000	5.354	0.000	0.001	0.002
x2	0.4125	0.260	1.586	0.113	-0.098	0.923
x3	0.1565	0.910	0.172	0.864	-1.630	1.943
x4	-0.5872	0.239	-2.453	0.014	-1.057	-0.117
x5	-0.0012	0.000	-6.286	0.000	-0.002	-0.001

Omnibus:	793.266	Durbin-Watson:	2.106
Prob(Omnibus):	0.000	Jarque-Bera (JB):	450270.294
Skew:	4.813	Prob(JB):	0.00
Kurtosis:	128.417	Cond. No.	6.45e+05

The ***R-Squared*** value is ***0.664*** which means that the model fit decently well. From the p and t values we can see that the significant features are ***Maximum follower*** count and ***Retweet Count***.

3. #nfl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.604
Model:	OLS	Adj. R-squared:	0.602
Method:	Least Squares	F-statistic:	280.9
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	2.36e-182
Time:	19:50:39	Log-Likelihood:	-6992.8
No. Observations:	926	AIC:	1.400e+04
Df Residuals:	920	BIC:	1.403e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0%	Conf. Int.]
const	61.7117	29.839	2.068	0.039	3.151	120.272
x1	-0.0001	2.5e-05	-5.697	0.000	-0.000	-9.34e-05
x2	-0.1778	0.065	-2.716	0.007	-0.306	-0.049
x3	-1.2138	2.198	-0.552	0.581	-5.528	3.100
x4	1.3405	0.110	12.216	0.000	1.125	1.556
x5	0.0002	3.38e-05	5.811	0.000	0.000	0.000

Omnibus:	1045.607	Durbin-Watson:	2.159
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1262900.553
Skew:	4.465	Prob(JB):	0.00
Kurtosis:	183.699	Cond. No.	5.43e+06

The ***R-Squared*** value is only ***0.604*** which means that the model did well. From the p and t values we can see that the significant features are ***Tweet count*** and ***Follower Count***.

4. #patriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.716
Model:	OLS	Adj. R-squared:	0.715
Method:	Least Squares	F-statistic:	491.1
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	2.86e-263
Time:	19:52:18	Log-Likelihood:	-8753.1
No. Observations:	980	AIC:	1.752e+04
Df Residuals:	974	BIC:	1.755e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0%	Conf. Int.]
const	136.4441	114.603	1.191	0.234	-88.453	361.341
x1	0.0003	4.28e-05	7.779	0.000	0.000	0.000
x2	-0.9485	0.073	-13.020	0.000	-1.091	-0.806
x3	-1.4734	8.492	-0.173	0.862	-18.138	15.192
x4	1.7833	0.079	22.488	0.000	1.628	1.939
x5	-0.0002	8.94e-05	-2.750	0.006	-0.000	-7.04e-05

Omnibus:	1873.390	Durbin-Watson:	1.696
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4047796.756
Skew:	13.529	Prob(JB):	0.00
Kurtosis:	316.684	Cond. No.	9.74e+06

The ***R-Squared*** value is ***0.604*** which means that the model did well. From the p and t values we can see that the significant features are ***Tweet count*** and ***Follower Count***.

5. **#sb49**

OLS Regression Results

Dep. Variable:	y	R-squared:	0.821
Model:	OLS	Adj. R-squared:	0.819
Method:	Least Squares	F-statistic:	527.8
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	2.37e-212
Time:	19:55:13	Log-Likelihood:	-5692.9
No. Observations:	582	AIC:	1.140e+04
Df Residuals:	576	BIC:	1.142e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	162.3844	350.147	0.464	0.643	-525.336	850.105
x1	0.0002	2.96e-05	7.410	0.000	0.000	0.000
x2	-0.3676	0.043	-8.464	0.000	-0.453	-0.282
x3	-16.3977	26.012	-0.630	0.529	-67.488	34.693
x4	1.1411	0.052	21.885	0.000	1.039	1.244
x5	-0.0003	6.91e-05	-4.103	0.000	-0.000	-0.000

Omnibus:	1160.867	Durbin-Watson:	1.726
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2239973.135
Skew:	14.031	Prob(JB):	0.00
Kurtosis:	305.626	Cond. No.	6.19e+07

The ***R-Squared*** value is ***0.821*** which means that the model did very well. From the p and t values we can see that the significant features are ***Maximum Follower Count*** and ***Tweet Count***.

6. **#superbowl**

OLS Regression Results

Dep. Variable:	y	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.741
Method:	Least Squares	F-statistic:	551.8
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	6.28e-279
Time:	19:59:44	Log-Likelihood:	-9909.3
No. Observations:	963	AIC:	1.983e+04
Df Residuals:	957	BIC:	1.986e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	274.4117	457.086	0.600	0.548	-622.594	1171.417
x1	-0.0004	2.58e-05	-13.823	0.000	-0.000	-0.000
x2	0.0247	0.126	0.196	0.845	-0.222	0.272
x3	-12.0723	33.391	-0.362	0.718	-77.600	53.455
x4	1.6753	0.258	6.489	0.000	1.169	2.182
x5	0.0013	0.000	9.861	0.000	0.001	0.002

Omnibus:	1887.423	Durbin-Watson:	1.699
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5780466.256
Skew:	14.126	Prob(JB):	0.00
Kurtosis:	381.501	Cond. No.	9.08e+07

The ***R-Squared*** value is ***0.741*** which means that the model did well. From the p and t values we can see that the significant features are ***Tweet count*** and ***Follower Count***.

Problem 3

In this problem, our job is introduce new features of our own and use them along with the features given in the previous problem to fit the regression model.

According to us some of the features that could affect the popularity of tweets are:

1. Author count – the total number of unique authors (or users) who posted a tweet within a given hourly window.
2. Mention count – the total number of users who were mentioned in the tweets that were posted within a given hourly window.
3. URL count – the total number of tweets containing a link of a picture, a song, a video, or just some general news.
4. Total listed count – the total number of public lists that the users who have posted tweets within a given hourly window are a member of.
5. Maximum listed count – the maximum value for listed count that is attained within a given hourly window.
6. Favorites count – the total number of times the tweets appearing within a given hourly window have been "liked" by the users.
7. Maximum favorites count – the maximum value for favorites count that is attained within a given hourly window.
8. Sum of the ranking score – the total amount of influence that the tweets within a given hourly window have on the audience.
9. Total number of verified users – the total number of verified users that post tweets within a given hourly window.

Now, we use the above 9 features along with the 5 features that we were given and fit the OLS regression model in the same way as we did in the previous problem.

We followed this approach in finding the most significant features in all the files. We trained the model with all the 14 features. We found the significant features from these 14 and trained a new model.

The 14 parameters are as follows:
X1: Ranking score of the tweet
X2: Retweet Count
X3: Tweet Count
X4: Maximum lists a user is part of
X5: Total verified users
X6: Total user count
X7: Total favorites a tweet has received
X8: Total URL mentions
X9: Maximum favorite count
X10: Followers count
X11: Total user mentions
X12: Time
X13: Total lists a user is a part of
X14: Maximum followers
From the previous model training we identified the best features and trained new models.
Too many parameters can lead to over fitting. So we selected the best features for each hashtag.

Observations:

1) #gohawks

OLS Regression Results

Dep. Variable:	y	R-squared:	0.669
Model:	OLS	Adj. R-squared:	0.664
Method:	Least Squares	F-statistic:	138.0
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	4.33e-218
Time:	16:53:38	Log-Likelihood:	-7600.1
No. Observations:	972	AIC:	1.523e+04
Df Residuals:	957	BIC:	1.530e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-5.1195	38.941	-0.131	0.895	-81.540	71.301
x1	8.1197	0.870	9.338	0.000	6.413	9.826
x2	-0.2187	0.052	-4.174	0.000	-0.321	-0.116
x3	-40.1807	4.364	-9.207	0.000	-48.745	-31.616
x4	-0.0662	0.017	-3.891	0.000	-0.100	-0.033
x5	-28.0786	14.256	-1.970	0.049	-56.056	-0.101
x6	3.5365	0.752	4.704	0.000	2.061	5.012
x7	-0.3706	0.519	-0.714	0.476	-1.390	0.648
x8	3.7796	1.184	3.193	0.001	1.457	6.102
x9	0.4525	0.524	0.864	0.388	-0.576	1.481
x10	-0.0008	0.000	-8.311	0.000	-0.001	-0.001
x11	1.8306	0.426	4.296	0.000	0.994	2.667
x12	0.8453	2.850	0.297	0.767	-4.747	6.437
x13	0.0690	0.012	5.761	0.000	0.046	0.093
x14	0.0007	0.000	5.020	0.000	0.000	0.001

Since we had a lot of parameters on which we trained the data, we had to filter out the best parameters. The highlighted ones are the top three parameters. We used the p and t values to determine the significant ones.

X1: Ranking score of the tweet

X13: Total lists a user is a part of

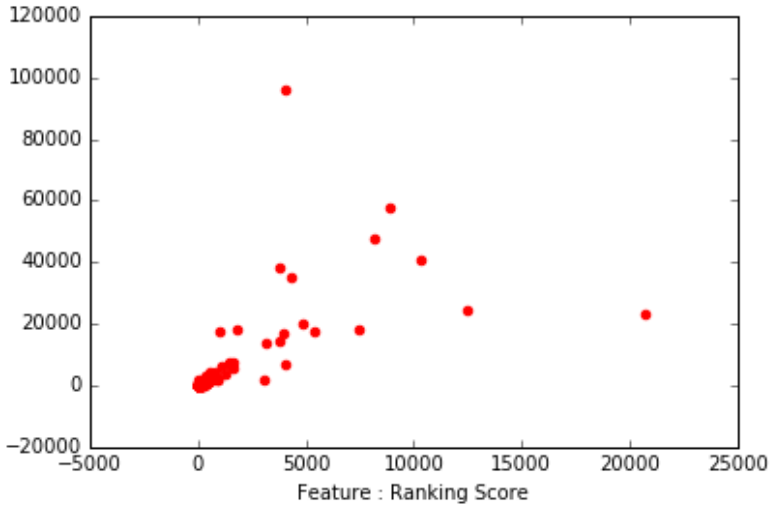
X14: Maximum followers

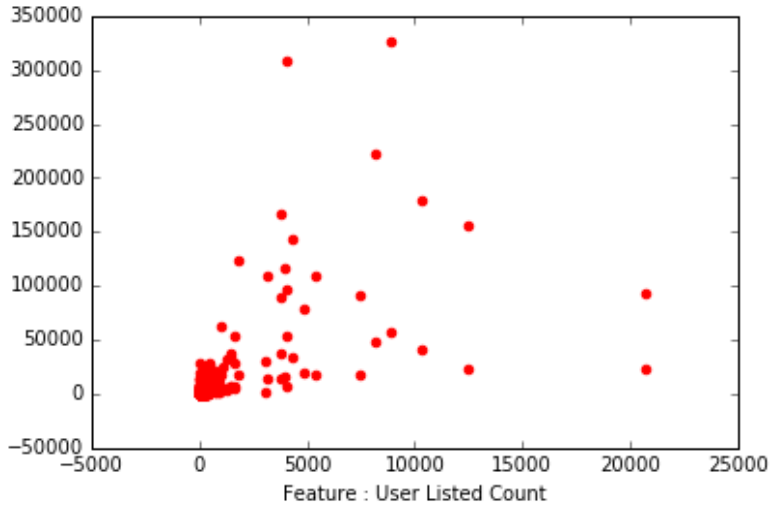
X6: Total user count

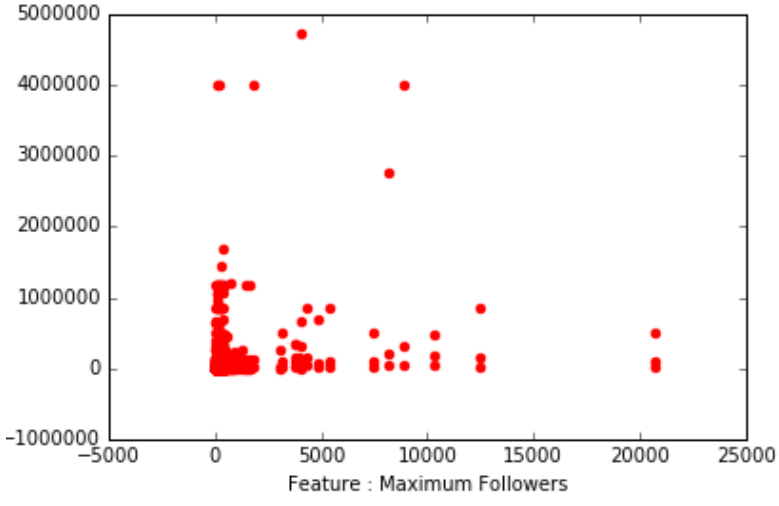
X11: Total user mentions

We also trained another model using these five features, the R-Squared value dropped to 0.578. This was definitely expected as we reduced the number of features.

Top three features scatter plots:







2) #gopatriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.868
Model:	OLS	Adj. R-squared:	0.865
Method:	Least Squares	F-statistic:	313.5
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	2.10e-282
Time:	16:53:44	Log-Likelihood:	-4128.9
No. Observations:	683	AIC:	8288.
Df Residuals:	668	BIC:	8356.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-8.0085	7.810	-1.025	0.306	-23.343	7.326
x1	3.5906	0.396	9.073	0.000	2.814	4.368
x2	-0.9003	0.190	-4.726	0.000	-1.274	-0.526
x3	-13.6052	2.076	-6.554	0.000	-17.681	-9.529
x4	-0.0293	0.033	-0.889	0.374	-0.094	0.035
x5	3.8290	18.433	0.208	0.836	-32.365	40.023
x6	-2.9499	0.626	-4.713	0.000	-4.179	-1.721
x7	-9.9348	3.274	-3.034	0.003	-16.364	-3.506
x8	12.5442	0.955	13.130	0.000	10.668	14.420
x9	5.2093	7.291	0.715	0.475	-9.106	19.525
x10	-0.0020	0.000	-7.097	0.000	-0.003	-0.001
x11	4.7811	0.423	11.299	0.000	3.950	5.612
x12	-0.2513	0.578	-0.435	0.664	-1.385	0.883
x13	0.0348	0.028	1.254	0.210	-0.020	0.089
x14	0.0019	0.000	6.835	0.000	0.001	0.002

From the 14 features shown in the model above, we chose the five significant ones. They are as shown below:

X1: Ranking score of the tweet

X8: Total URL mentions

X11: Total user mentions

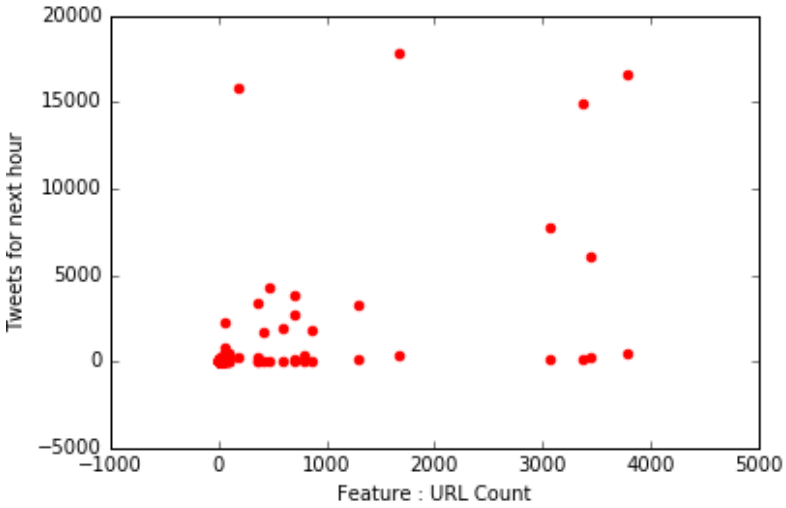
X13: Total lists a user is a part of

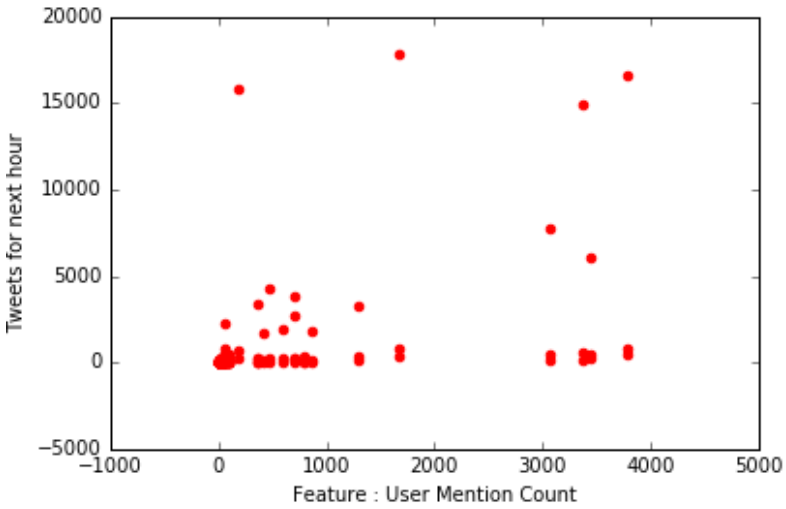
X14: Maximum followers

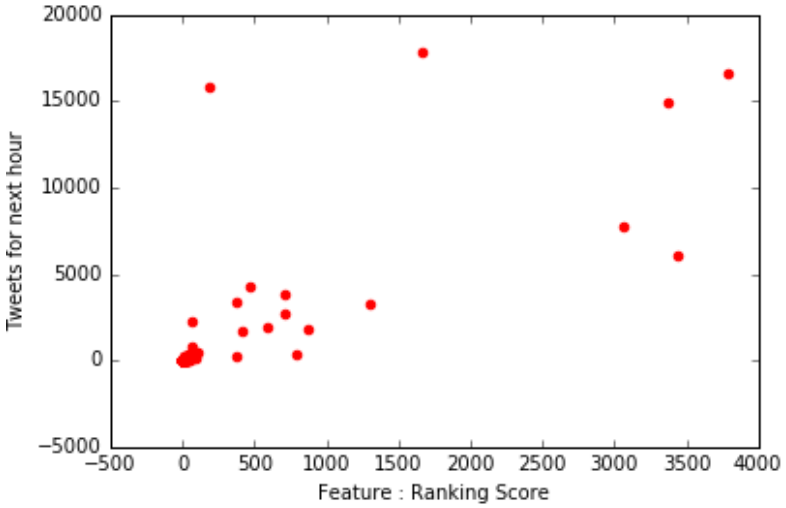
The ones which are highlighted are the top three features for #gopatriots.

We again trained a model with these new five features. The R-Squared value dropped to 0.731. This goes in sync with what we observed for #gohawks. With the reduction in number of features, the R-Squared value decreases.

Top three features scatter plots:







3) #nfl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.761
Model:	OLS	Adj. R-squared:	0.757
Method:	Least Squares	F-statistic:	207.1
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	2.31e-271
Time:	16:54:41	Log-Likelihood:	-6759.5
No. Observations:	926	AIC:	1.355e+04
Df Residuals:	911	BIC:	1.362e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	41.1437	24.480	1.681	0.093	-6.900	89.187
x1	-1.4545	0.271	-5.372	0.000	-1.986	-0.923
x2	-0.0310	0.057	-0.540	0.589	-0.143	0.082
x3	8.2194	1.299	6.326	0.000	5.670	10.769
x4	-0.0345	0.007	-5.094	0.000	-0.048	-0.021
x5	-4.9049	8.615	-0.569	0.569	-21.812	12.002
x6	-1.3839	0.299	-4.635	0.000	-1.970	-0.798
x7	-1.2582	0.811	-1.551	0.121	-2.851	0.334
x8	-0.8485	0.161	-5.283	0.000	-1.164	-0.533
x9	-0.9311	0.999	-0.932	0.351	-2.891	1.029
x10	-0.0004	4.07e-05	-9.467	0.000	-0.000	-0.000
x11	1.1429	0.481	2.375	0.018	0.198	2.087
x12	-1.3769	1.727	-0.797	0.426	-4.766	2.013
x13	0.0470	0.006	8.278	0.000	0.036	0.058
x14	0.0004	4.44e-05	8.251	0.000	0.000	0.000

The best features are mentioned below. The highlighted ones are the most significant ones.

X3: Tweet Count

X13: Total lists a user is a part of

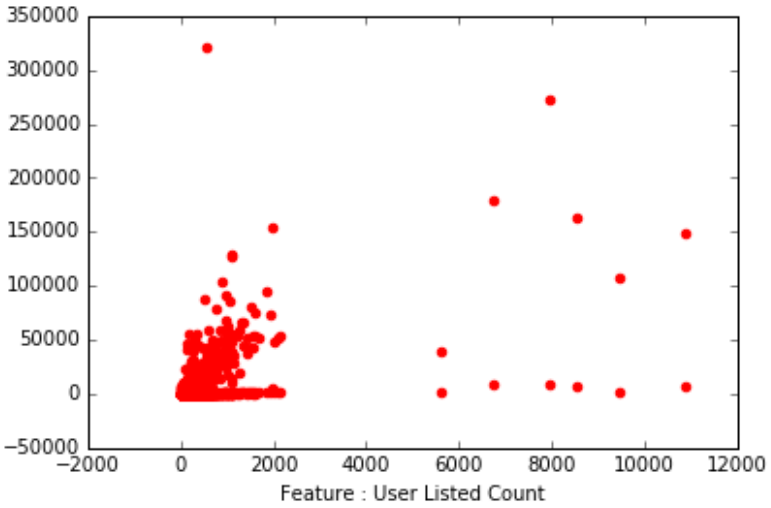
X14: Maximum followers

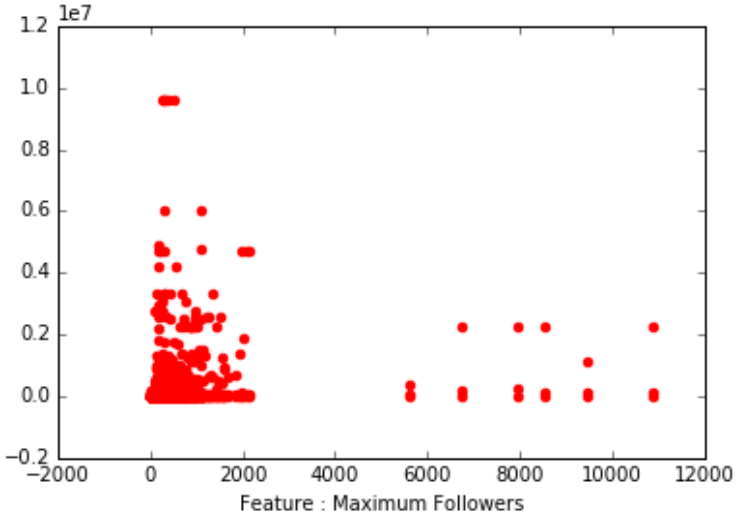
X11: Total user mentions

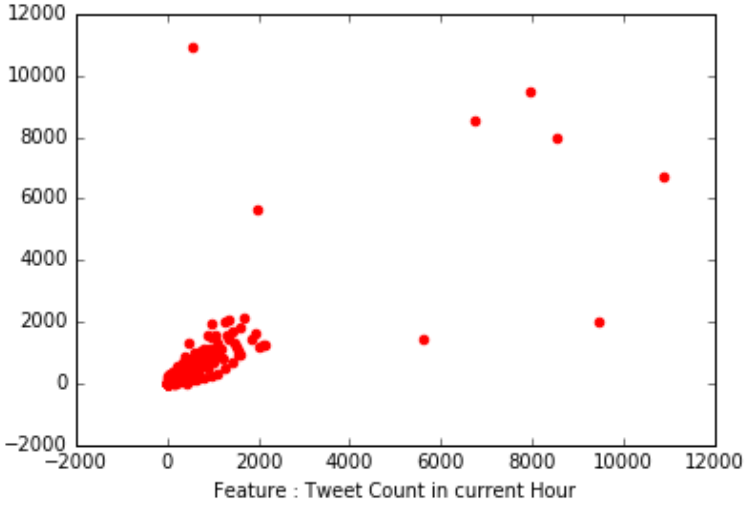
As before, we used the p and t values to compare the best features.

We also trained a new model with these four features. The R-Squared dropped to 0.598.

Top three features scatter plot:







4) #patriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.783
Model:	OLS	Adj. R-squared:	0.780
Method:	Least Squares	F-statistic:	248.9
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	2.34e-308
Time:	16:56:20	Log-Likelihood:	-8621.0
No. Observations:	980	AIC:	1.727e+04
Df Residuals:	965	BIC:	1.735e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-68.5071	104.429	-0.656	0.512	-273.442	136.428
x1	5.5770	0.761	7.330	0.000	4.084	7.070
x2	-0.2615	0.129	-2.024	0.043	-0.515	-0.008
x3	-26.8776	3.742	-7.183	0.000	-34.220	-19.535
x4	0.0274	0.025	1.095	0.274	-0.022	0.077
x5	85.0327	23.409	3.633	0.000	39.095	130.971
x6	1.1836	0.913	1.297	0.195	-0.607	2.974
x7	-0.5275	0.409	-1.290	0.197	-1.330	0.275
x8	1.7192	1.024	1.678	0.094	-0.291	3.730
x9	0.5702	0.461	1.237	0.216	-0.334	1.475
x10	0.0006	0.000	5.728	0.000	0.000	0.001
x11	1.1124	0.454	2.452	0.014	0.222	2.003
x12	1.0192	7.510	0.136	0.892	-13.719	15.758
x13	-0.0361	0.016	-2.315	0.021	-0.067	-0.006
x14	-0.0007	0.000	-5.300	0.000	-0.001	-0.000

Following is a list of the features we thought to perform well for the model. The ones which are highlighted are the top three ones.

X1: Ranking score of the tweet

X5: Total verified users

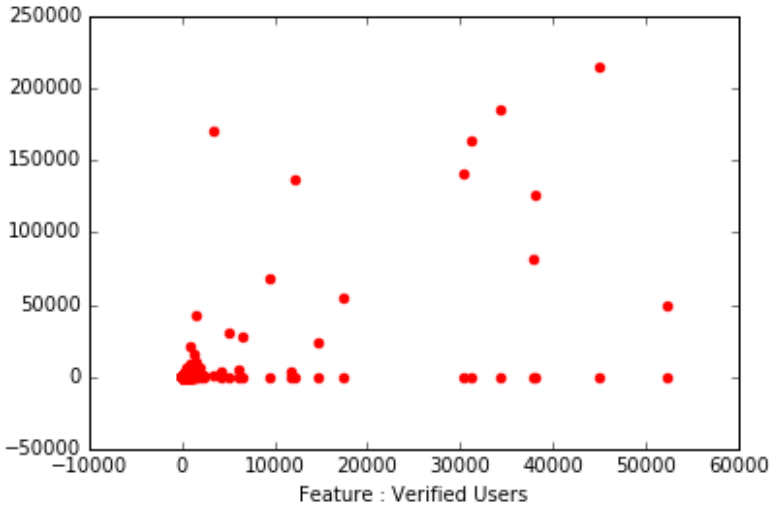
X10: Followers count

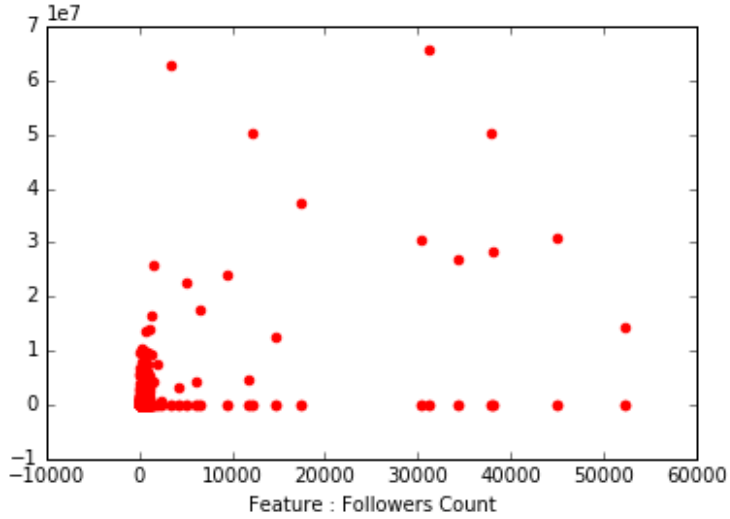
X6: Total user count

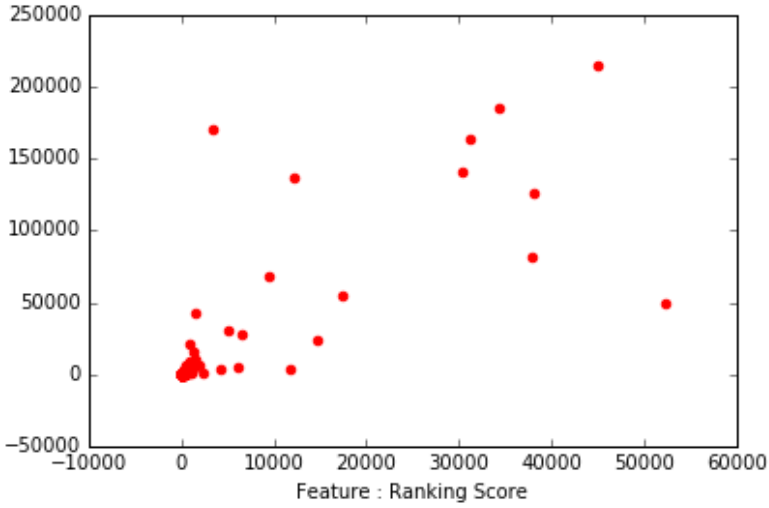
X11: Total user mentions

The trend that we observed in the previous cases continued for #patriots as well. On training a new model we understood that the R-Squared dropped to 0.753. However this decrement was a lot less as compared to the previous hashtags.

Top three features scatter plots:







5) #sb49

OLS Regression Results

Dep. Variable:	y	R-squared:	0.872
Model:	OLS	Adj. R-squared:	0.869
Method:	Least Squares	F-statistic:	276.1
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	2.33e-242
Time:	16:59:11	Log-Likelihood:	-5594.9
No. Observations:	582	AIC:	1.122e+04
Df Residuals:	567	BIC:	1.129e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	86.7395	309.418	0.280	0.779	-521.006	694.485
x1	5.4273	1.855	2.926	0.004	1.784	9.071
x2	0.6144	0.114	5.382	0.000	0.390	0.839
x3	-30.4870	8.955	-3.404	0.001	-48.076	-12.898
x4	-0.0719	0.018	-4.047	0.000	-0.107	-0.037
x5	-126.5027	26.485	-4.776	0.000	-178.524	-74.481
x6	1.9504	0.992	1.967	0.050	0.003	3.898
x7	-0.2062	0.119	-1.731	0.084	-0.440	0.028
x8	-4.2190	1.248	-3.381	0.001	-6.670	-1.768
x9	-0.6107	0.471	-1.298	0.195	-1.535	0.314
x10	-5.859e-05	9.31e-05	-0.629	0.529	-0.000	0.000
x11	5.2693	0.493	10.684	0.000	4.301	6.238
x12	-1.0491	22.294	-0.047	0.962	-44.838	42.739
x13	0.0836	0.014	5.818	0.000	0.055	0.112
x14	-0.0002	0.000	-2.268	0.024	-0.000	-3.14e-05

The list for best set of features and the top three ones is shown below:

X2: Retweet Count

X11: Total user mentions

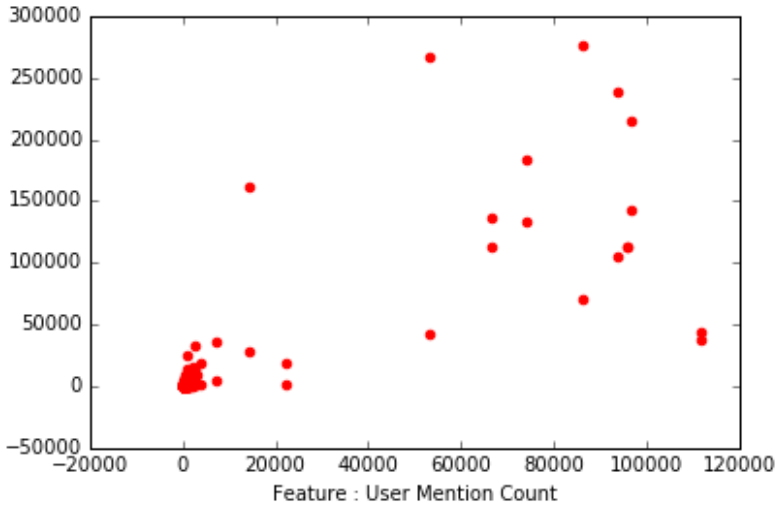
X13: Total lists a user is a part of

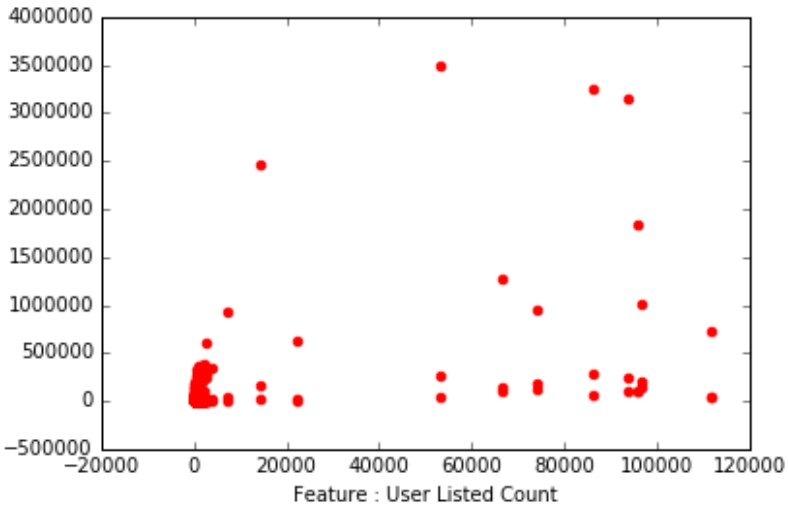
X1: Ranking score of the tweet

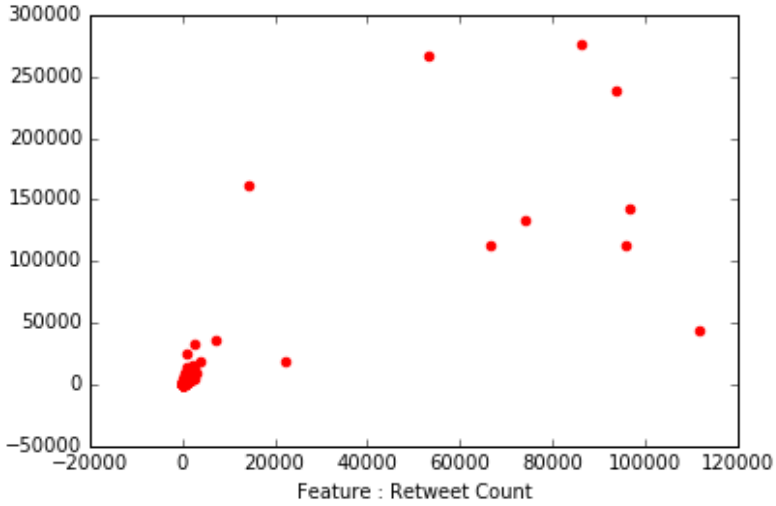
X6: Total user count

The model with the new set of features gave us an R-Squared of 0.843. This trend was observed in case of #patriots as well where the R-Squared dropped by a very small margin.

Top three features scatter plots:







6) #superbowl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.914
Model:	OLS	Adj. R-squared:	0.912
Method:	Least Squares	F-statistic:	717.5
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	0.00
Time:	17:04:08	Log-Likelihood:	-9382.5
No. Observations:	963	AIC:	1.880e+04
Df Residuals:	948	BIC:	1.887e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-18.5130	270.784	-0.068	0.946	-549.919	512.892
x1	-0.9847	1.335	-0.738	0.461	-3.604	1.635
x2	1.0810	0.119	9.072	0.000	0.847	1.315
x3	0.7341	6.589	0.111	0.911	-12.197	13.665
x4	-0.0468	0.017	-2.771	0.006	-0.080	-0.014
x5	16.8871	29.855	0.566	0.572	-41.703	75.477
x6	3.3122	0.862	3.844	0.000	1.621	5.003
x7	-4.8663	0.303	-16.059	0.000	-5.461	-4.272
x8	-6.6089	1.325	-4.989	0.000	-9.208	-4.009
x9	6.6447	0.659	10.087	0.000	5.352	7.937
x10	-0.0005	4.67e-05	-9.659	0.000	-0.001	-0.000
x11	1.1973	0.667	1.795	0.073	-0.111	2.506
x12	5.1149	19.549	0.262	0.794	-33.249	43.478
x13	0.0568	0.009	6.462	0.000	0.040	0.074
x14	0.0004	0.000	2.809	0.005	0.000	0.001

We identified out the significant features using their p and t values from the list above. They are as shown below:

X2: Retweet Count

X9: Maximum favorite count

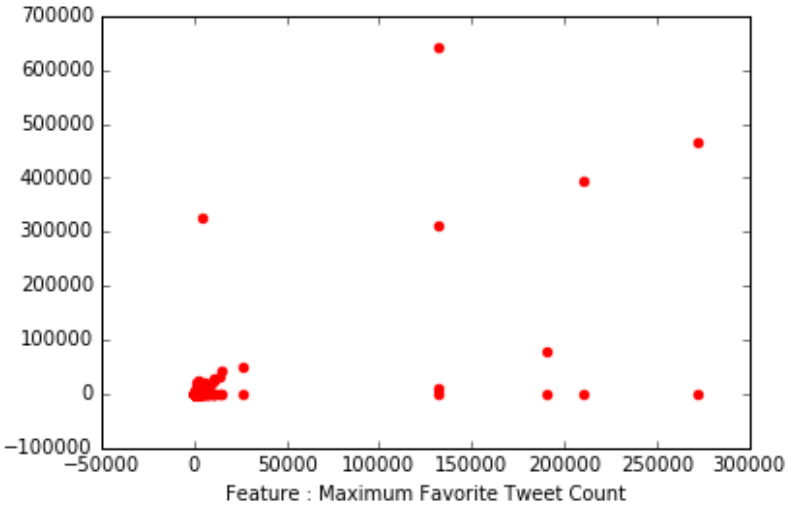
X13: Total lists a user is a part of

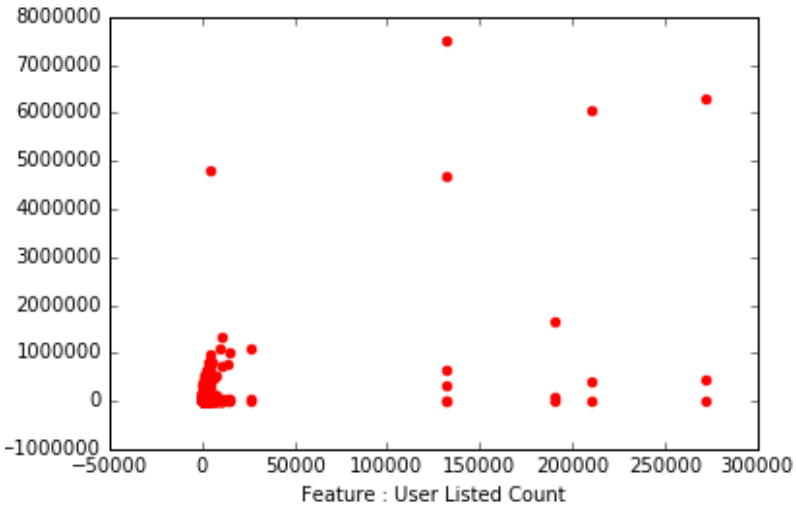
X6: Total user count

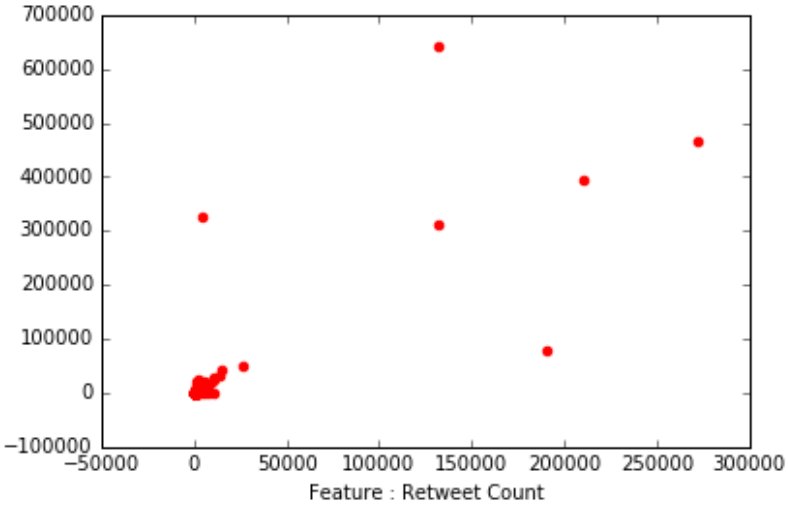
X14: Maximum followers

The R-Squared value for a new model which was trained using only these five features was 0.765 which was much lower than what we had obtained with all the features. In general, reduction in features led to lower R-Squared values. One reason could be that the model was over fitting with so many features, and we reduced the number of features, the over fitting reduced.

Top three features scatter plots:







Problem 4

Part-1

Now, in this part, we utilize the 14 features obtained from the previous parts of the project organized in the form of (features, predictant) pairs for each window. This feature data is split into 10 parts in such a way that 90% of the data is used for fitting the model, while the remaining 10% of the data is used as the testing data. This process is repeated 10 times, i.e., we perform 10-fold cross validation on the feature data for each of the hashtag.

In order to validate how well our model is performing, we calculate the prediction error given by $|N_{predicted} - N_{real}|$ for each fold, and then take the average over the 10 folds.

Observations:

Hashtag	Average prediction error
#gohawks	201.985
#gopatriots	46.838
#nfl	209.403
#patriots	570.734
#sb49	1296.939
#superbowl	1446.295

Part-2

Since we know the Super Bowl's date and time, we created different regression models for different periods of time. First, when the hashtags haven't become very active, second, their active period, and third, after they pass their high-activity time.
The time slots are as shown below:
1. Before Feb. 1, 8:00 a.m.
2. Between Feb. 1, 8:00 a.m. and 8:00 p.m.
3. After Feb. 1, 8:00 p.m.

Observations:

Hashtag	Period 1	Period 2	Period 3
#gohawks	195.330	3586.817	3618.150
#gopatriots	15.0497	5931.725	3.3970
#nfl	129.859	8306.454	323.191
#patriots	193.289	644817.015	119.640
#sb49	95.050	98712.984	233.679
#superbowl	235.638	249011.408	459.229

The error seems to be a disaster for period 2. The reason is that we had only twelve training points. And it is very difficult to give good predictions. A solution could be to use **sliding windows** to increase the number of data points.

Problem 5

In question 5, our task was to test the models we had trained in question 4 and try predicting the values for the next hour. There were 10 files in all, each of them corresponding to one of the three time periods. However, unlike before, the files had a mixture of all hashtags. But the models we had trained were specific to a specific hashtag. So we found the most dominant hashtag in each of the ten files. The dominant hashtags were:

Test File	Dominant Hashtag	Model Used
Sample1_period1	#superbowl	Superbowl model for period 1
Sample2_period2	#superbowl	Superbowl model for period 2
Sample3_period3	#superbowl	Superbowl model for period 3
Sample4_period1	#nfl	Nfl model for period 1
Sample5_period1	#nfl	Nfl model for period 1
Sample6_period2	#superbowl	Superbowl model for period 2
Sample7_period3	#nfl	Nfl model for period 3
Sample8_period1	#nfl	Nfl model for period 1
Sample9_period2	#superbowl	Superbowl model for period 2
Sample10_period3	#nfl	Nfl model for period 2

For each tag we had data give for 6 hours. We had to predict the value for next hour. So given the data from hour 1 to hour 6, we had to predict from hour 2 to hour 7. Here are our results

Test File	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	Hour 7	Error
Sample1_period1	115.12	49.78	176.68	265.27	463.99	650.02	213.71
Sample2_period2	614679.8	68409.27	503125	412958	3331211	1805309	1124174
Sample3_period3	509.03	623.35	705.78	628.06	646.21	653.34	197.78
Sample4_period1	1375.94	562.02	221.95	342.30	134.77	86.02	332.01
Sample5_period1	491.76	542.83	397.72	308.70	448.62	263.73	253.39
Sample6_period2	11855.12	10885539	66174686	5643991.7	4233358.1	347051.3	3512142.
Sample7_period3	86.61	69.31	60.58	51.63	54.21	68.96	31.34
Sample8_period1	NA	57647.17	47250.27	58692.12	72259.96	101448.2	67423.56
Sample9_period2	907629	936522	790894	750649	1019	895972	715378
Sample10_period	43.57	41.00	38.55	36.31	35.28	32.25	25.27

The values in the Hour 2 to Hour 7 are the predicted values using the data from the previous hour. The error column is the difference between the actual and predicted values. For hour 7, the data was not available. Hence the error term excludes hour 7. It's only calculated from hours 2 to hour 7.

Problem 6

Problem definition:

- Tweet data is a rich source of information and provides us insights about factors like the content of the tweet, the users' information who posted the tweet and a lot of other Meta data.
- One such interesting information that can be extracted from the tweets is the geolocation that is the place where the user who posted the tweet.
- The most popular application of tweets is sentiment analysis. The task is to predict the sentiment of a particular tweet
- We can apply the concept of sentiment analysis on the tweet data in such a way that could help us find the sentiment of different regions towards a certain event given the tweet data originating from several different regions

How the approach is implemented:

- Since the time at our disposal was less, we couldn't develop a sentiment analyzer on our own. We used a free API available. The name of the API was 'text-processing'.
- Here is a gist of our process:
    1. We choose the #gohawks file, since it was the smallest file available in the dataset. Since the API only allowed us to scan a limited number of tweets, we couldn't use the other files.
    2. We filtered the tweets from each file based on the location. We implemented for 'Boston'.
    3. If the tweet was posted by a user who was from Boston, we stored his tweet for further processing.
    4. Now all the Boston tweets were subject to text cleaning and pre-processing before giving them to the sentiment analyzer.
    5. We used a tool called as 'TextBlob' for enabling this. We had to clean these items:
        a. Emoticons
        b. URL's
        c. User mentions
        d. Hashtags
        e. Stop Words
    6. After cleaning, we did a POST request to the API.
    7. The API returned us three value: positive, negative and neutral. The three values reflect the sentiment of the tweet. We took the positive and negative sentiment into consideration.
    8. We then plotted the sentiment values of the tweet for Boston city.

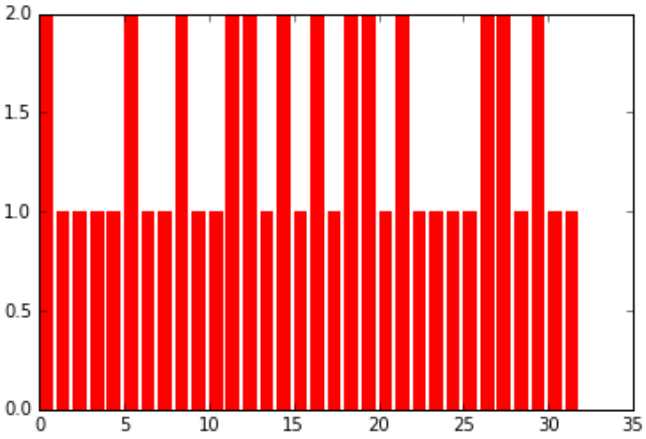
Results:

We just analyzed the #gohawks file for "Boston" city.

Boston and #gohawks	
Total Positive Sentiment	15.7636726721
Total Negative Sentiment	16.2363273279
Total Neutral Sentiment	11.906505983

The negative sentiments seem to dominate the Boston city.

The graph below confirms this result.



Analysis of the graph:

- The peaks are the positive sentiment and the lower values are the negative sentiments.
- #gohawks is team from Seattle.
- We analyzed #gohawks from Boston, where we expected the sentiment would be negative. And the results reflect that.

Limitations:

- The API had restricted access, so we had a limit on the number of tweets we could analyze.

Application of the approach

- In the context of this project, this approach can be used to visualize the changing trends of the sentiments of the audience during the course of the games played at Superbowl. This could possibly help in clustering and differentiating amongst the different regions based on their attitude towards the event.
- However, this approach is pretty open-ended and can be applied to several other applications.
- For example: during the presidential election period, it could be useful for a presidential candidate to know the attitude of the twitter users based in different regions towards him, so that he can modify he can strategize and campaign accordingly.
- The entertainment industry could use this approach to analyze how different regions react to particular content that is broadcast on television to filter out their content appropriately in regions where it is negatively received.

References

[1] https://dev.twitter.com/overview/api/tweets

[2] Shoubin Kong, Qiaozhu Mei, Ling Feng, Zhe Zhao, Fei Ye, “On the Real-time Prediction Problems of Bursting Hashtags in Twitter”, arXiv:1401.2018v2 [cs.SI] 3 Jun 2014.

[3] S. Petrovic, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. Prof. of AAAI on Weblogs and Social Media, 2011.

[4] http://text-processing.com/api/sentiment/