EE 219 PROJECT 5
POPULARITY PREDICTION ON TWITTER

TEAM MEMBERS-
Omkar Patil: 904760474
Shikhar Malhotra: 504741656
Twinkle Gupta: 804740325

Social network services have become a viable source of information for users. One such social network is Twitter. Twitter allows its users to 'tweet' their message, restricted to 140 characters. Users may subscribe to other users' tweets—this is known as "following" and subscribers are known as "followers". Individual tweets can be forwarded by other users to their own feed, a process known as a "retweet". Users can also "like" (formerly "favorite") individual tweets. In Twitter, information deemed important by the community propagates through retweets and user mentions. A user can mention another user in the tweet using the '@' symbol. Tweets with the same "hashtags" are grouped together. All these characteristics of a tweet can be used to predict the future popularity of the tweet based on past data.

In this project, we try to predict the popularity of a topic on Twitter. More formally, knowing the previous and current tweet activity for a hashtag, we try to predict its tweet activity in the future and aim to determine whether it gets more or less popular and by how much. For this, we use Regression Models.

DATASET

The available Twitter data is collected by querying popular hashtags related to the 2015 Super Bowl spanning a period starting from 2 weeks before the game to a week after the game. The data is grouped according to 6 hashtags. The tweets are stored in separate files for different hashtags and files are named as tweet_[#hashtag].txt. The tweet file contains one tweet in each line and tweets are sorted with respect to their posting time.
Given the trends of the tweets belonging to different hashtags over a period of time, our task is to predict the popularity of each hashtag in the future.

PART 1 :

In this part, we intend to analyze the dataset and calculate some important statistics such as average number of tweets per hour, average number of followers of users posting the tweets, and average number of retweets.

Each tweet is a JSON string that we can load in Python as a dictionary. We parsed each line in the data and loaded them as JSON object called 'tweet_dict'. To find the above statistics, we found the required values as follows-

user_id = tweet_dict["tweet"]["user"]["id"]
totalFollowers += tweet_dict["author"]["followers"]
retweets += tweet_dict["metrics"]["citations"]["total"]

To find the time elapsed, we calculated the difference between the time the first tweet was posted and the time the last tweet was posted and converted it into hours.

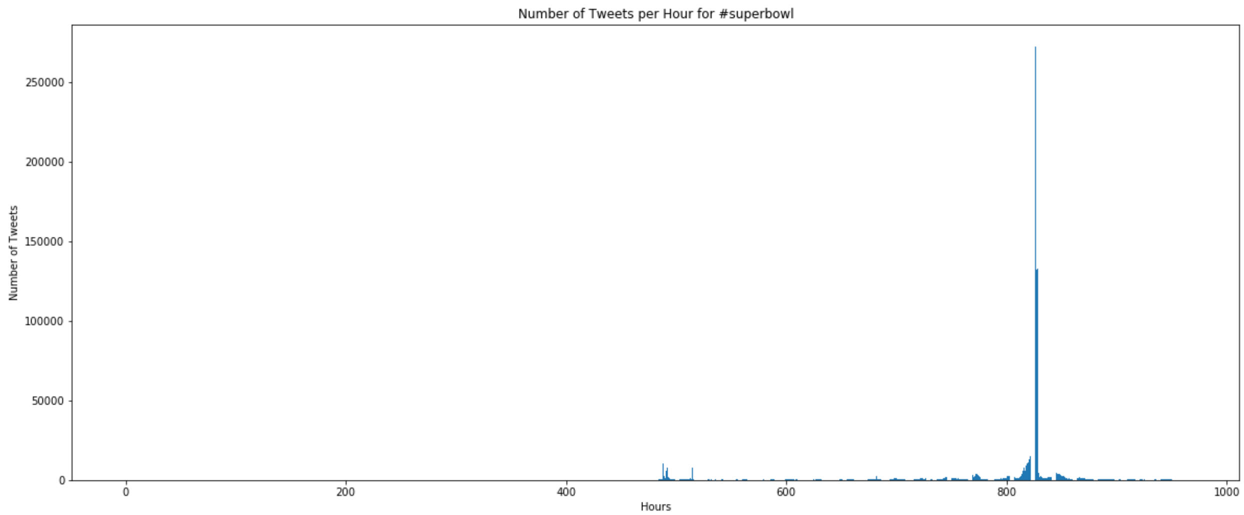
The statistics were then calculated as –
Average number of followers = totalFollowers / total number of unique users
Average number of tweets per hour = total number of tweets / time elapsed in hours
Average number of retweets = number of retweets / total number of tweets

The following results were obtained –

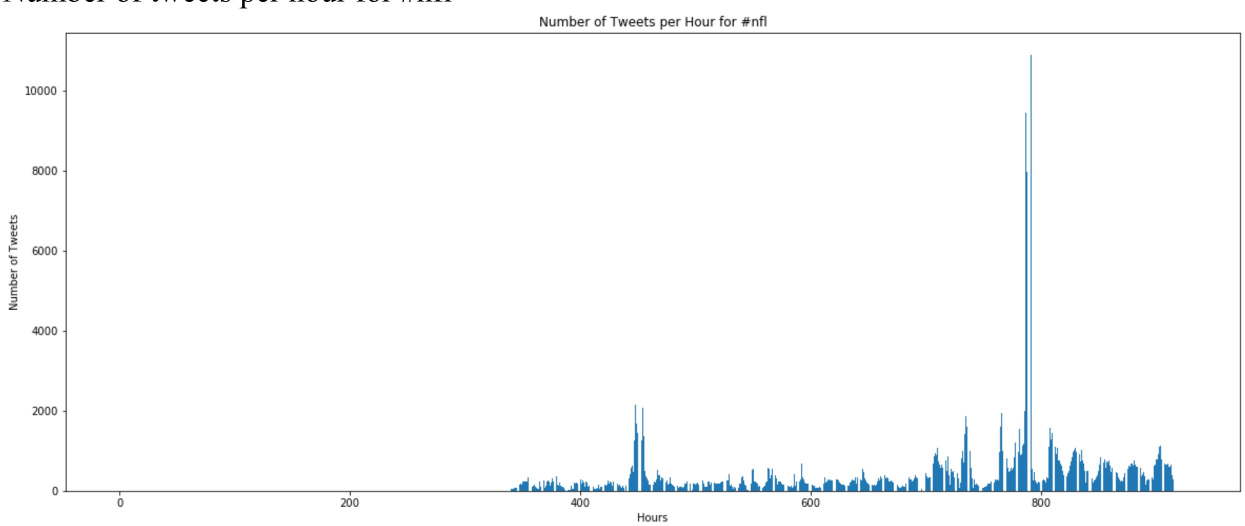
HASHTAG	AVERAGE NUMBER OF TWEETS PER HOUR	AVERAGE NUMBER OF FOLLOWERS	AVERAGE NUMBER OF RETWEETS
#gohawks	193.55556	1544.96979	2.01461
# gopatriots	38.40703	1298.82427	1.40008
#nfl	279.42179	4289.74661	1.53853
#patriots	499.19775	1650.32198	1.78281
#sb49	1420.87800	2235.16367	2.51115
#superbowl	1400.58878	3591.60447	2.38827

The number of tweets per hour for #superbowl and #nfl were plotted with time and the following histograms were obtained –

Number of tweets per hour for #superbowl



Number of tweets per hour for #nfl



We observe sudden burst in number of tweets in both the histograms around the 800th hour. This was during the Superbowl and NFL. For Superbowl, sudden increase in number of tweets is observed during the 830th hour and for NFL around the 800th hour.

PART 2:

In this part, we aim to fit a Linear Regression model using 5 features to predict the number of tweets in the next hour, with features extracted from the tweet data from the previous hour. The features used are – 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 time of the day (which could take 24 values that represent hours of the day with respect to a given time reference). We created time windows of 1 hour and calculated the values of the above mentioned features hourwise. The value to be predicted is the number of tweets in the next hour.

We use the OLS model available in statsmodel library in Python for this purpose.

The following results were obtained for the 6 hashtags –
X1 : Maximum number of followers
X2 : Time
X3 : Number of followers
X4 : Number of retweets
X5 : Number of tweets

The **p-value** for each parameter tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that we can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response. On the other hand, the **t-statistic** is useful for making inferences about the regression coefficients. The hypothesis test on coefficient i tests the null hypothesis that it is equal to zero – meaning the corresponding term is not significant – versus the alternate hypothesis that the coefficient is different from zero. There we would want to consider features with high t-test values.

1. #gohawks

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.490
Model:                            OLS   Adj. R-squared:                  0.488
Method:                 Least Squares   F-statistic:                     186.0
Date:                Thu, 16 Mar 2017   Prob (F-statistic):          8.81e-139
Time:                        23:51:50   Log-Likelihood:                -7817.0
No. Observations:                 973   AIC:                         1.565e+04
Df Residuals:                     967   BIC:                         1.568e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         39.0200     35.144      1.110      0.267     -29.947     107.987
x1            -0.0007      0.000     -4.915      0.000      -0.001      -0.000
x2             6.2179      3.122      1.992      0.047       0.091      12.345
x3             0.0004   8.15e-05      4.586      0.000       0.000       0.001
x4            -0.1692      0.043     -3.909      0.000      -0.254      -0.084
x5             0.5719      0.121      4.716      0.000       0.334       0.810
==============================================================================
Omnibus:                     1848.594   Durbin-Watson:                   2.336
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4377767.947
Skew:                          13.277   Prob(JB):                         0.00
Kurtosis:                     330.531   Cond. No.                     2.39e+06
==============================================================================
```

Best features found based on p and t-values are – **Number of followers** and **number of tweets**. However, the R-squared value is only 0.490, which means the model did not fit very well.

2. #gopatriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.664
Model:                            OLS   Adj. R-squared:                  0.662
Method:                 Least Squares   F-statistic:                     268.4
Date:                Thu, 16 Mar 2017   Prob (F-statistic):          4.60e-158
Time:                        23:51:53   Log-Likelihood:                -4453.3
No. Observations:                 684   AIC:                             8919.
Df Residuals:                     678   BIC:                             8946.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.2832      8.891      0.482      0.630     -13.174      21.740
x1            -0.0012      0.000     -6.359      0.000      -0.002      -0.001
x2             0.7502      0.827      0.908      0.364      -0.873       2.373
x3             0.0011      0.000      5.432      0.000       0.001       0.002
x4             0.3815      0.262      1.456      0.146      -0.133       0.896
x5            -0.5687      0.240     -2.369      0.018      -1.040      -0.097
==============================================================================
Omnibus:                      796.993   Durbin-Watson:                   2.103
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           452112.690
Skew:                           4.844   Prob(JB):                         0.00
Kurtosis:                     128.578   Cond. No.                     4.69e+05
==============================================================================
```

Best features found based on p and t-values are – **Number of followers** and **number of retweets**. The R-squared value is only 0.664, which means the model fit fairly well.

3. #nfl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.605
Model:                            OLS   Adj. R-squared:                  0.602
Method:                 Least Squares   F-statistic:                     281.7
Date:                Thu, 16 Mar 2017   Prob (F-statistic):          9.19e-183
Time:                        23:52:27   Log-Likelihood:                -6999.4
No. Observations:                 927   AIC:                         1.401e+04
Df Residuals:                     921   BIC:                         1.404e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.7633     21.497      1.571      0.117      -8.425      75.952
x1             0.0002    3.4e-05      5.622      0.000       0.000       0.000
x2             2.1759      2.036      1.069      0.286      -1.821       6.172
x3            -0.0001    2.5e-05     -5.600      0.000      -0.000   -9.11e-05
x4            -0.1779      0.065     -2.722      0.007      -0.306      -0.050
x5             1.3297      0.110     12.078      0.000       1.114       1.546
==============================================================================
Omnibus:                     1053.806   Durbin-Watson:                   2.146
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1256393.880
Skew:                           4.531   Prob(JB):                         0.00
Kurtosis:                     183.127   Cond. No.                     3.91e+06
==============================================================================
```

Best features found based on p and t-values are – **Maximum Number of followers** and **number of tweets**. However, the R-squared value is only 0.605, which means the model fit decently.

4. #patriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.716
Model:                            OLS   Adj. R-squared:                  0.715
Method:                 Least Squares   F-statistic:                     492.1
Date:                Thu, 16 Mar 2017   Prob (F-statistic):          1.06e-263
Time:                        23:53:30   Log-Likelihood:                -8761.2
No. Observations:                 981   AIC:                         1.753e+04
Df Residuals:                     975   BIC:                         1.756e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         72.1464     83.417      0.865      0.387     -91.550     235.843
x1            -0.0003   9.01e-05     -2.844      0.005      -0.000   -7.94e-05
x2             6.7396      7.839      0.860      0.390      -8.644      22.123
x3             0.0003   4.28e-05      7.792      0.000       0.000       0.000
x4            -0.9539      0.073    -13.071      0.000      -1.097      -0.811
x5             1.7894      0.079     22.523      0.000       1.634       1.945
==============================================================================
Omnibus:                     1877.207   Durbin-Watson:                   1.694
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4075004.536
Skew:                          13.560   Prob(JB):                         0.00
Kurtosis:                     317.577   Cond. No.                     7.10e+06
==============================================================================
```

Best features found based on p and t-values are – **Number of followers** and **number of tweets**. The R-squared value is only 0.716, which means the model fit well.

5. #sb49

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.821
Model:                            OLS   Adj. R-squared:                  0.819
Method:                 Least Squares   F-statistic:                     528.7
Date:                Thu, 16 Mar 2017   Prob (F-statistic):          1.01e-212
Time:                        23:55:19   Log-Likelihood:                -5702.2
No. Observations:                 583   AIC:                         1.142e+04
Df Residuals:                     577   BIC:                         1.144e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        138.7825    323.784      0.429      0.668    -497.156     774.721
x1            -0.0003   6.92e-05     -4.086      0.000      -0.000      -0.000
x2           -15.4646     25.008     -0.618      0.537     -64.582      33.653
x3             0.0002   2.96e-05      7.420      0.000       0.000       0.000
x4            -0.3677      0.043     -8.475      0.000      -0.453      -0.283
x5             1.1410      0.052     21.899      0.000       1.039       1.243
==============================================================================
Omnibus:                     1163.174   Durbin-Watson:                   1.726
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2251333.588
Skew:                          14.042   Prob(JB):                         0.00
Kurtosis:                     306.134   Cond. No.                     5.73e+07
==============================================================================
```

Best features found based on p and t-values are – **Number of followers** and **number of tweets**. However, the R-squared value is only 0.821, which means the model fit very well.

6. #superbowl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.742
Model:                            OLS   Adj. R-squared:                  0.741
Method:                 Least Squares   F-statistic:                     552.3
Date:                Thu, 16 Mar 2017   Prob (F-statistic):          3.39e-279
Time:                        23:58:14   Log-Likelihood:                -9919.2
No. Observations:                 964   AIC:                         1.985e+04
Df Residuals:                     958   BIC:                         1.988e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        136.6962    318.892      0.429      0.668    -489.112     762.504
x1             0.0013      0.000      9.530      0.000       0.001       0.002
x2             0.1737     31.361      0.006      0.996     -61.370      61.717
x3            -0.0004   2.58e-05    -13.814      0.000      -0.000      -0.000
x4             0.0245      0.126      0.195      0.846      -0.222       0.271
x5             1.6751      0.258      6.487      0.000       1.168       2.182
==============================================================================
Omnibus:                     1889.238   Durbin-Watson:                   1.698
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5789800.589
Skew:                          14.125   Prob(JB):                         0.00
Kurtosis:                     381.611   Cond. No.                     6.34e+07
==============================================================================
```

Best features found based on p and t-values are – **Maximum Number of followers** and **number of tweets**. However, the R-squared value is only 0.742, which means the model fit well.

PART 3:

In this part we aim to train the model using features of our own. We selected the following features for this purpose, in addition to the ones mentioned in the previous part :

X1)'totalTweets'
X2)'retweets'
X3)'time'
X4)'followers'
X5)'favorite_count'
X6)'ranking_score'
X7)'urls'
X8)'user_count'
X9)'impressions'

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

1) Favorites count – the total number of times the tweets appearing within a given hourly window have been "liked" by the users.
2) User count – the total number of users tweeting the hashtag
3) Ranking Score – the total amount of influence that the tweets within a given hourly window have on the audience.

4)URLS – the total number of tweets containing a link of a picture, a song, a video, or just some general news.
5)Impressions - The total number of users in whose feed the tweet appeared

Now, we use the above 9 to fit the OLS regression model/
We followed this approach in finding the most significant features in all the files. We trained the model with all the 9 features
As we were already using a lot of features, to avoid overfitting we select the three most important features for every tweet as asked, and plot the scatter plot for each of them

Observations :-

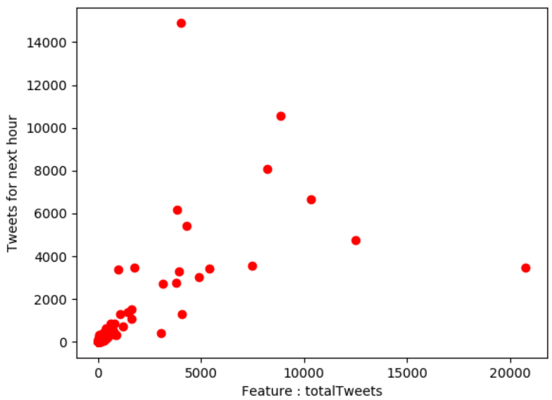
1) #gohawks

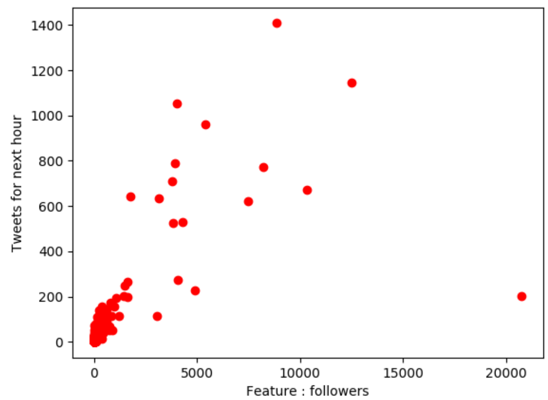
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.639
Model:                            OLS   Adj. R-squared:                  0.636
Method:                 Least Squares   F-statistic:                     189.7
Date:                Sun, 19 Mar 2017   Prob (F-statistic):         2.34e-206
Time:                        00:20:30   Log-Likelihood:                -7648.6
No. Observations:                 973   AIC:                         1.532e+04
Df Residuals:                     963   BIC:                         1.537e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.2665     28.654      0.044      0.965     -54.965      57.497
x1             4.5862      0.737      6.223      0.000       3.140       6.032
x2            -0.0003   4.98e-05     -6.967      0.000      -0.000      -0.000
x3            -0.2011      0.055     -3.676      0.000      -0.308      -0.094
x4             9.1580      0.775     11.824      0.000       7.638      10.678
x5             2.9099      2.563      1.135      0.257      -2.120       7.940
x6         -4.603e-10   1.95e-10     -2.364      0.018   -8.42e-10   -7.82e-11
x7             7.7322      0.587     13.172      0.000       6.580       8.884
x8             0.0881      0.021      4.222      0.000       0.047       0.129
x9           -38.3959      2.963    -12.959      0.000     -44.210     -32.582
==============================================================================
Omnibus:                     1962.042   Durbin-Watson:                   2.216
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5530326.385
Skew:                          15.179   Prob(JB):                         0.00
Kurtosis:                     371.089   Cond. No.                     4.05e+11
==============================================================================
```

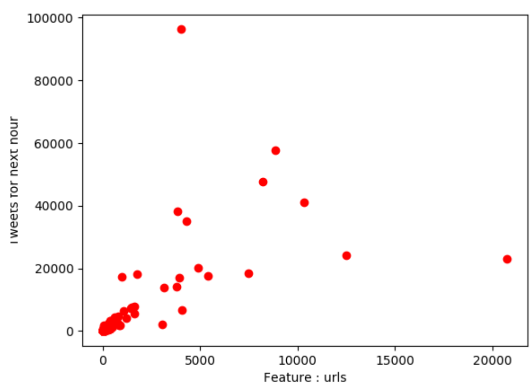
Out of these, 9 features, we select the following highlighted features for plotting the graphs.
We selected these features based on the t values available above.

X1 totalTweets
X4 followers
X7 urls

Following are there scatter plots







2) #gopatriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.792
Model:                            OLS   Adj. R-squared:                  0.789
Method:                 Least Squares   F-statistic:                     284.8
Date:                Sun, 19 Mar 2017   Prob (F-statistic):         5.83e-223
Time:                        00:20:34   Log-Likelihood:                -4290.0
No. Observations:                 684   AIC:                             8600.
Df Residuals:                     674   BIC:                             8645.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
```

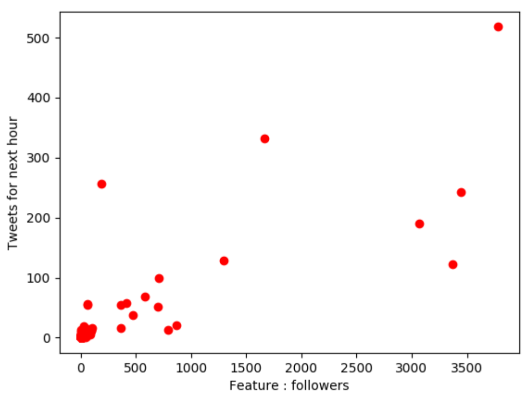
const	1.7719	6.733	0.263	0.792	-11.447	14.991
x1	-2.9745	0.665	-4.472	0.000	-4.280	-1.669
x2	-1.419e-05	4.42e-05	-0.321	0.749	-0.000	7.27e-05
x3	-0.6442	0.233	-2.761	0.006	-1.102	-0.186
x4	10.1974	0.795	12.835	0.000	8.637	11.757
x5	0.8724	0.625	1.395	0.163	-0.355	2.100
x6	-2.664e-09	4.15e-09	-0.642	0.521	-1.08e-08	5.49e-09
x7	0.9229	0.359	2.572	0.010	0.218	1.628
x8	-7.1164	1.765	-4.031	0.000	-10.583	-3.650
x9	-1.0377	1.981	-0.524	0.601	-4.927	2.852

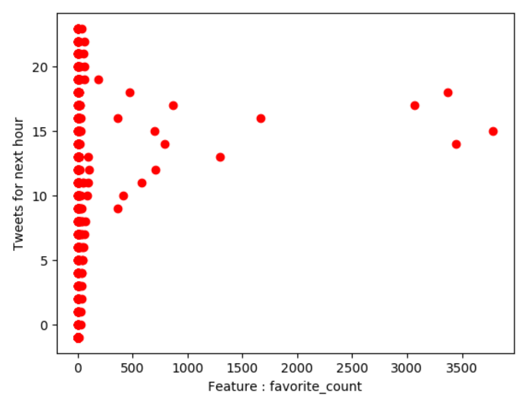
Omnibus:	769.908	Durbin-Watson:	1.944
Prob(Omnibus):	0.000	Jarque-Bera (JB):	316337.484
Skew:	4.641	Prob(JB):	0.00
Kurtosis:	107.945	Cond. No.	1.06e+10

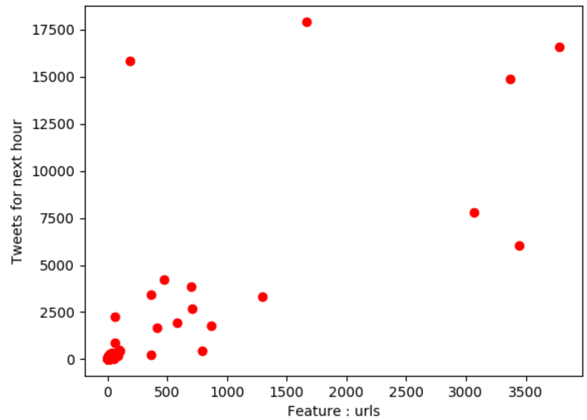
Out of these, 9 features, we select the following highlighted features for plotting the graphs. We selected these features based on the t values available above.

X4 followers
X5 favourite_count
X7 urls

Following are there scatter plots







3) #nfl

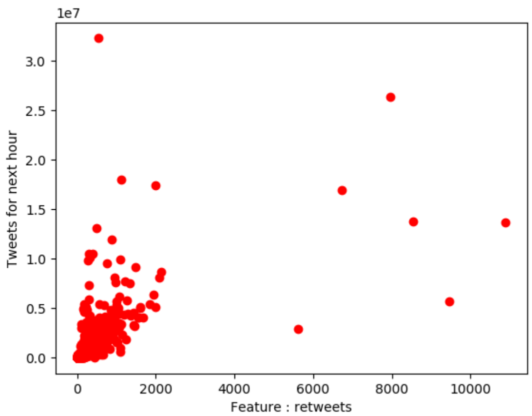
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.720
Model:                            OLS   Adj. R-squared:                  0.717
Method:                 Least Squares   F-statistic:                     261.5
Date:                Sun, 19 Mar 2017   Prob (F-statistic):         3.55e-246
Time:                        00:21:12   Log-Likelihood:                -6840.1
No. Observations:                 927   AIC:                         1.370e+04
Df Residuals:                     917   BIC:                         1.375e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         30.6815     18.207      1.685      0.092      -5.051      66.414
x1            -0.4702      0.325     -1.448      0.148      -1.107       0.167
x2          1.572e-05    1.1e-05      1.431      0.153   -5.84e-06    3.73e-05
x3             0.1184      0.059      1.990      0.047       0.002       0.235
x4            -0.0539      0.140     -0.384      0.701      -0.330       0.222
x5             1.0294      1.834      0.561      0.575      -2.570       4.629
x6          1.877e-10   7.82e-11      2.402      0.016    3.44e-11    3.41e-10
x7            -0.2160      0.266     -0.812      0.417      -0.738       0.306
x8            -2.3785      0.162    -14.654      0.000      -2.697      -2.060
x9             1.9385      1.283      1.511      0.131      -0.579       4.456
==============================================================================
Omnibus:                     1603.653   Durbin-Watson:                   2.219
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1404625.182
Skew:                          11.155   Prob(JB):                         0.00
Kurtosis:                     192.388   Cond. No.                     8.79e+11
==============================================================================
```

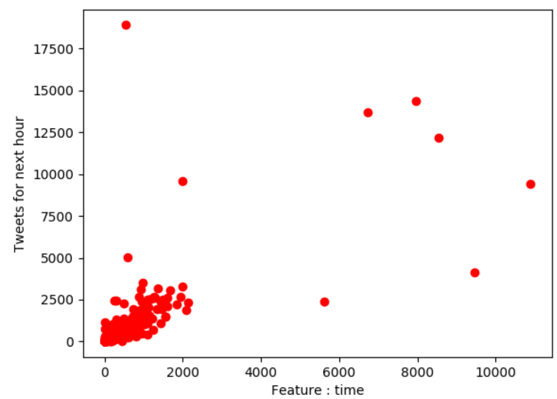
Out of these, 9 features, we select the following highlighted features for plotting the graphs. We selected these features based on the t values available above.

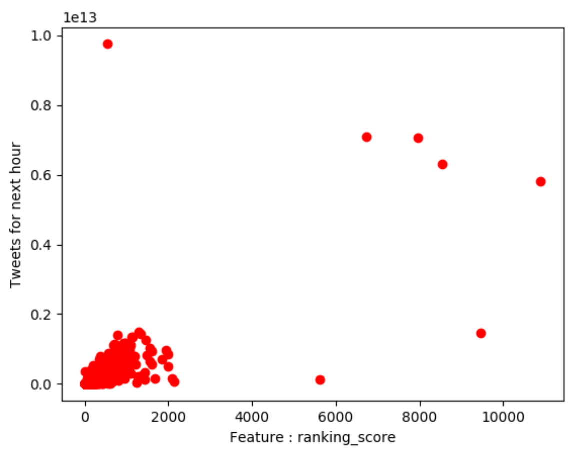
X2 retweets

X3 **time**
X6 **ranking_score**

Following are there scatter plots







4) #patriots

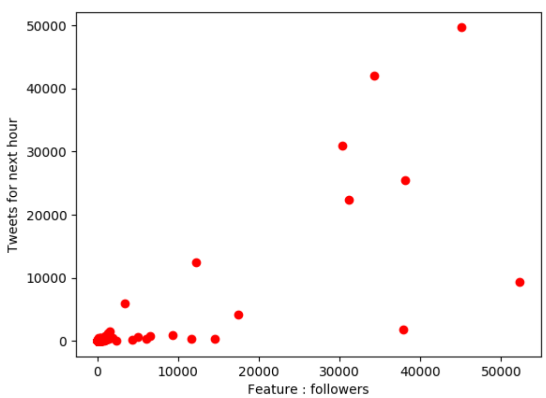
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.778
Model:                            OLS   Adj. R-squared:                  0.776
Method:                 Least Squares   F-statistic:                     377.2
Date:                Sun, 19 Mar 2017   Prob (F-statistic):          1.03e-309
Time:                        00:22:21   Log-Likelihood:                -8641.6
No. Observations:                 981   AIC:                         1.730e+04
Df Residuals:                     971   BIC:                         1.735e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -65.2799     71.623     -0.911      0.362    -205.834      75.274
x1             2.4032      0.797      3.017      0.003       0.840       3.966
x2          7.794e-05     3.6e-05      2.167      0.031    7.34e-06       0.000
```

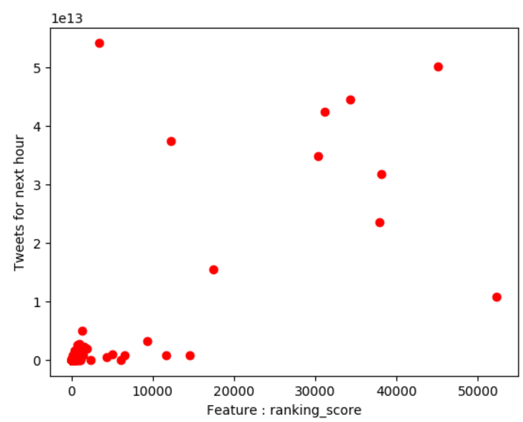
```
x3            -0.4334      0.118     -3.672      0.000      -0.665      -0.202
x4             5.7093      0.376     15.203      0.000       4.972       6.446
x5           -10.3892      6.750     -1.539      0.124     -23.636       2.858
x6          5.705e-10   9.82e-11      5.807      0.000    3.78e-10    7.63e-10
x7            10.4597      0.825     12.672      0.000       8.840      12.079
x8            -0.1429      0.180     -0.795      0.427      -0.496       0.210
x9           -49.8016      4.203    -11.848      0.000     -58.050     -41.553
==============================================================================
Omnibus:                     1989.575   Durbin-Watson:                   1.637
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5662244.732
Skew:                          15.397   Prob(JB):                         0.00
Kurtosis:                     373.915   Cond. No.                     5.19e+12
==============================================================================
```

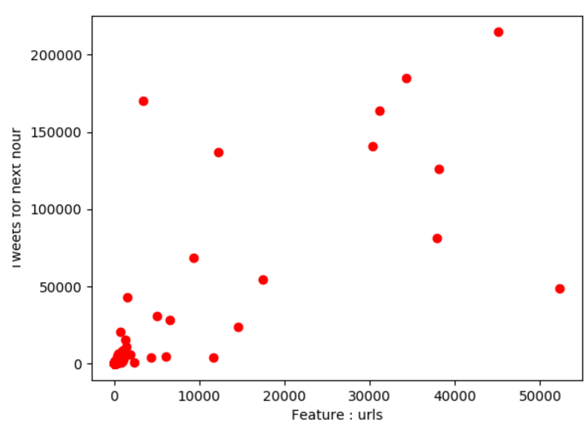
Out of these, 9 features, we select the following highlighted features for plotting the graphs. We selected these features based on the t values available above.

X4 followers
X6 ranking_score
X7 urls

Following are there scatter plots







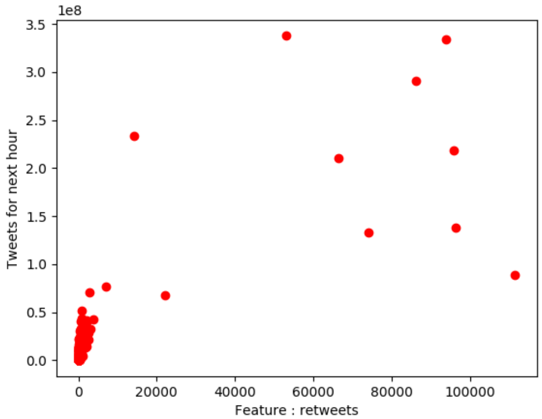
5) #sb49

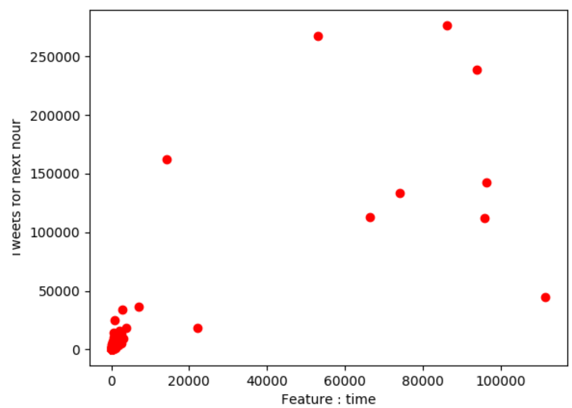
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.863
Model:                            OLS   Adj. R-squared:                  0.861
Method:                 Least Squares   F-statistic:                     402.7
Date:                Sun, 19 Mar 2017   Prob (F-statistic):          3.69e-241
Time:                        00:24:13   Log-Likelihood:                -5623.0
No. Observations:                 583   AIC:                         1.127e+04
Df Residuals:                     573   BIC:                         1.131e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -95.5994    281.680     -0.339      0.734    -648.852     457.653
x1             0.3526      0.828      0.426      0.670      -1.273       1.978
x2             0.0001   2.33e-05      5.891      0.000    9.13e-05       0.000
x3             0.3104      0.108      2.872      0.004       0.098       0.523
x4            -2.4355      1.030     -2.365      0.018      -4.458      -0.413
x5           -22.4972     21.573     -1.043      0.297     -64.870      19.876
x6         -4.415e-10    4.2e-11    -10.503      0.000   -5.24e-10   -3.59e-10
x7            -3.2295      1.811     -1.783      0.075      -6.787       0.328
x8            -0.2443      0.089     -2.755      0.006      -0.419      -0.070
x9            16.5686      8.743      1.895      0.059      -0.603      33.740
==============================================================================
Omnibus:                     1208.618   Durbin-Watson:                   1.906
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2467329.362
Skew:                          15.347   Prob(JB):                         0.00
Kurtosis:                     320.221   Cond. No.                     7.13e+13
==============================================================================
```

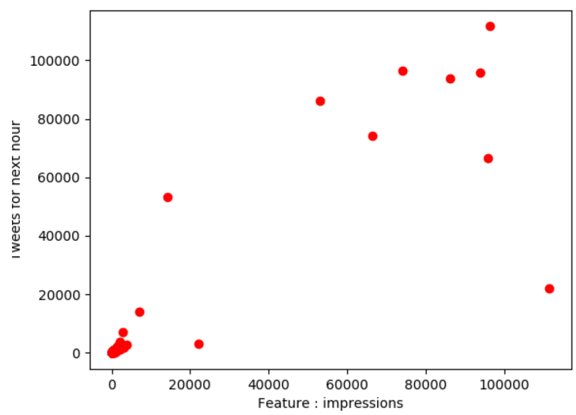
Out of these, 9 features, we select the following highlighted features for plotting the graphs. We selected these features based on the t values available above.

X2 retweets
X3 time
X9 impressions

Following are there scatter plots







6) #superbowl

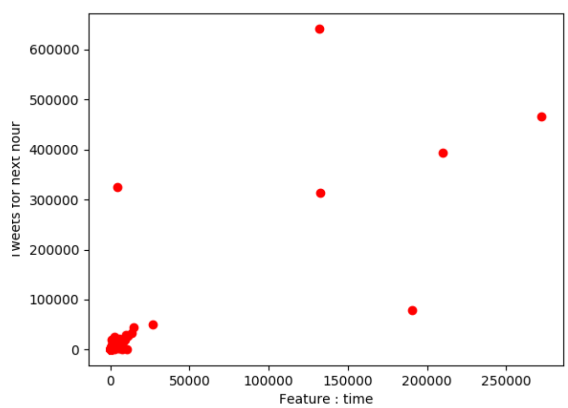
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.891
Model:                            OLS   Adj. R-squared:                  0.890
Method:                 Least Squares   F-statistic:                     863.1
Date:                Sun, 19 Mar 2017   Prob (F-statistic):               0.00
Time:                        00:27:08   Log-Likelihood:                -9506.3
No. Observations:                 964   AIC:                         1.903e+04
Df Residuals:                     954   BIC:                         1.908e+04
Df Model:                           9                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -140.0020    199.909     -0.700      0.484    -532.315     252.311
x1            -1.3614      0.543     -2.505      0.012      -2.428      -0.295
x2            -0.0001   3.46e-05     -4.178      0.000      -0.000   -7.66e-05
x3             1.6410      0.100     16.356      0.000       1.444       1.838
x4             5.9035      0.955      6.184      0.000       4.030       7.777
x5           -12.3165     19.397     -0.635      0.526     -50.383      25.750
x6         -1.558e-10   1.35e-11    -11.501      0.000   -1.82e-10   -1.29e-10
x7            -3.9606      1.102     -3.594      0.000      -6.123      -1.798
```

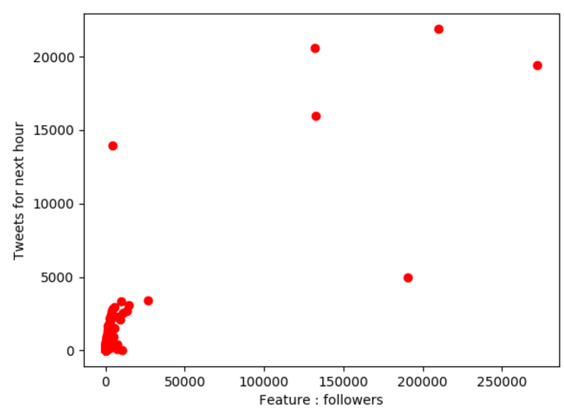
```
x8            -2.1247       0.187     -11.384      0.000      -2.491      -1.758
x9            17.3249       5.427       3.192      0.001       6.674      27.975
==============================================================================
Omnibus:                     1823.521   Durbin-Watson:                   2.067
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4710966.094
Skew:                          13.084   Prob(JB):                         0.00
Kurtosis:                     344.469   Cond. No.                     1.28e+14
==============================================================================
```

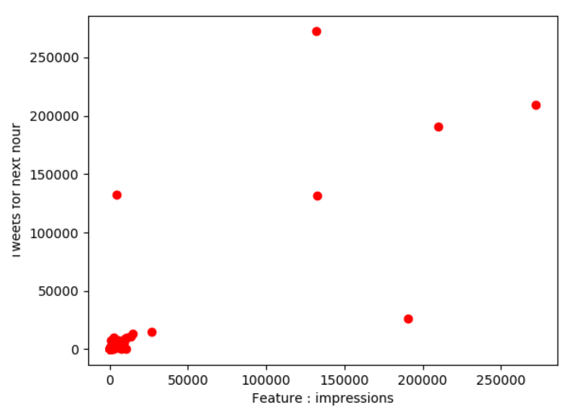
Out of these, 9 features, we select the following highlighted features for plotting the graphs. We selected these features based on the t values available above.

X3 time
X4 followers
X9 impressions

Following are there scatter plots







PART 4:

a.

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 $|Npredicted - Nreal|$ for each fold, and then take the average over the 10 folds.

Observations:

Hashtag	Average Prediction error
#gohawks	7.23854571533
#gopatriots	0.662027684759
#nfl	2.79884372525
#patriots	13.3817974166
#sb49	641.125037341
#superbowl	62.8921707714

b.

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.

Hashtag	Period1	Period2	Period3
#gohawks	6.66358764162	26930.592272	3921.00642844
#gopatriots	0.373799273238	671.134110061	17.2043735787
#nfl	3.32135896744	7496.3000768	166.987414963
#patriots	2.74245505624	64506.8881789	492.26330596
#sb49	18.6666616237	55083.1990926	722.955491941
#superbowl	12.3941055678	145900.920872	912.096081266

PART 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
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 given 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	181.313	143.802	490.336	173.892	348.952	402.641	194.962
Sample2_period2	55806.04	42395.119	30165.78	35233.64	142149.31	225682.43	122685
Sample3_period3	440.37	577.90	617.08	820.78	748.24	711.65	233.71
Sample4_period1	1079.35	425.2	227.04	265.46	287.56	175.76	231.57
Sample5_period1	206.80	142.25	260.13	60.73	204.87	132.56	216.36
Sample6_period2	72247	86683	249129	258167	184732	136458	198665
Sample7_period3	68.120	55.28	97.47	84.46	75.35	32.28	35.30
Sample8_period1	35460	29661	23510	17080	10563	101448	23291
Sample9_period2	79758.93	74457.20	65721.41	9165.99	42732.098	63921.35	715378
Sample10_period3	47.12	39.85	41.72	29.83	27.895	25.9680	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.

PART 6:

In this part, the objective is to predict the location of the author of the tweet by analyzing the textual content.

First, all the tweets in the file with #superbowl were taken into consideration. This was followed by preprocessing of the tweets. Only the tweets considering the following words in the "location" tag are stored –

a. MA
b. Massachusetts
c. WA
d. Washington

In all, there were 7643 users from Massachusetts and 11563 from Washington. Once the location of these users was known, the next task was to store the content of the **"text"** tag to predict the labels of the locations of these users. For this, all the tweets for each user was loaded into the TF-IDF matrix. This was achieved by using the inbuilt functions of Python.

The training and testing set was divided in the ratio of 3:1, wherein 75% of data belonging to the Washington and Massachusetts class was put into the training set, and the remaining was used to test the model.

After forming the tweets vs terms matrix, the number of terms came out to be – 36242. This caused the matrix to be huge and sparse. To prune down the dimensions, truncated SVD was applied with number of components as 50. Next, a variety of machine learning models were applied to predict the location of the users based on the textual content of the tweets they sent out. 5 classification algorithms were applied – Support Vector Machine, Logistic Regression, L2 regularized Logistic Regression, Neural Network and Naïve Bayes. For each of the algorithm, the precision, recall, confusion matrix and ROC was calculated and plotted –

a. Support Vector Machine

The results were -

'1' is Washington and '0' is Massachusetts
The accuracy for the model is 0.764546
The precision and recall values are:

	precision	recall	f1-score	support
0	0.84	0.51	0.63	1905
1	0.74	0.94	0.83	2890
avg / total	0.78	0.76	0.75	4795

The confusion matrix is:
[[963 942]
 [187 2703]]

b. Logistic Regression

The results were -

'1' is Washington and '0' is Massachusetts
The accuracy for the model is 0.775600

The precision and recall values are:

	precision	recall	f1-score	support
0	0.83	0.55	0.66	1905
1	0.76	0.92	0.83	2890
avg / total	0.78	0.78	0.76	4795

The confusion matrix is:
[[1049 856]
 [220 2670]]

c. L2 regularized Logistic Regression

The results were -

'1' is Washington and '0' is Massachusetts
The accuracy for the model is 0.774557

The precision and recall values are:

	precision	recall	f1-score	support
0	0.82	0.55	0.66	1905
1	0.76	0.92	0.83	2890
avg / total	0.78	0.77	0.76	4795

The confusion matrix is:
[[1050 855]
 [226 2664]]

d. Neural Network

For Neural Network, the ideal number of neurons in the hidden layer was a major issue. For this a list of fixed neuron values were taken and the model was run iteratively on each element of the list. The number of neurons for which the model performed best was 100 and this was the number taken for building the final model.

The results were -

'1' is Washington and '0' is Massachusetts
The accuracy for the model is 0.757039

The precision and recall values are:

	precision	recall	f1-score	support
0	0.73	0.62	0.67	1905
1	0.77	0.85	0.81	2890
avg / total	0.75	0.76	0.75	4795

The confusion matrix is:
[[1172 733]
 [432 2458]]

e. Naïve Bayes

The results were -

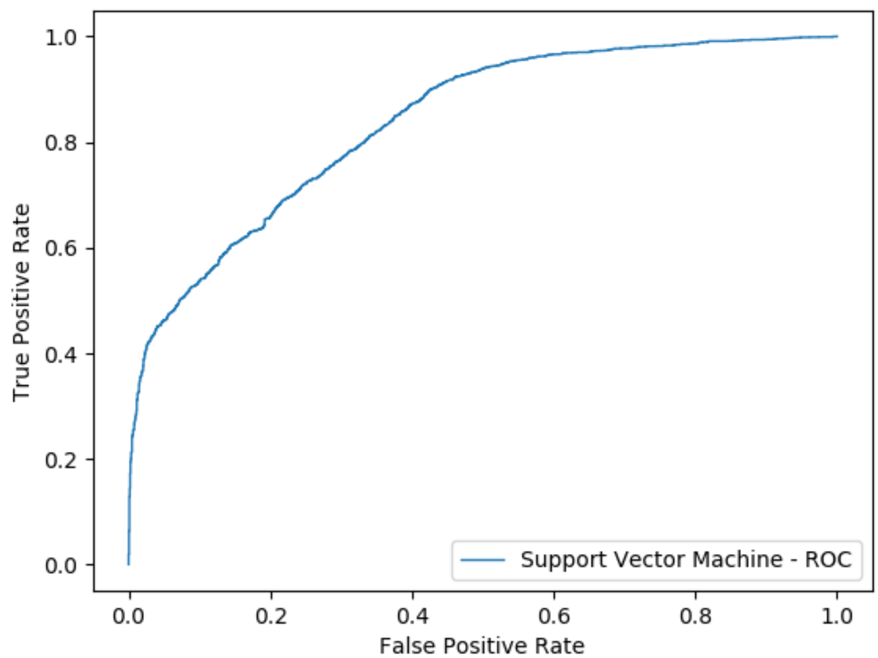
'1' is Washington and '0' is Massachusetts
The accuracy for the model is 0.694056

The precision and recall values are:

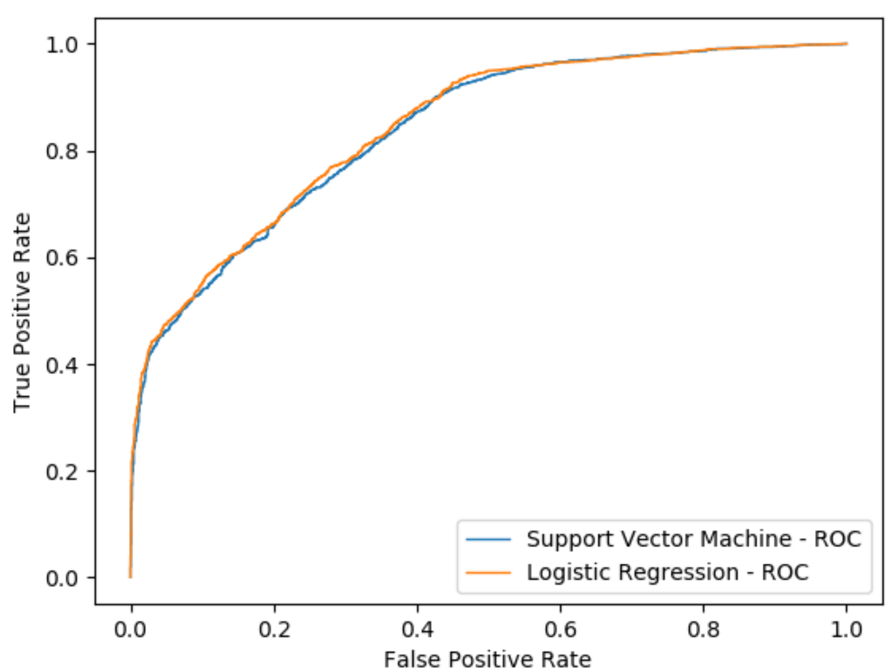
	precision	recall	f1-score	support
0	0.60	0.68	0.64	1905
1	0.77	0.70	0.74	2890
avg / total	0.70	0.69	0.70	4795

The confusion matrix is:
[[1292 613]
 [854 2036]]

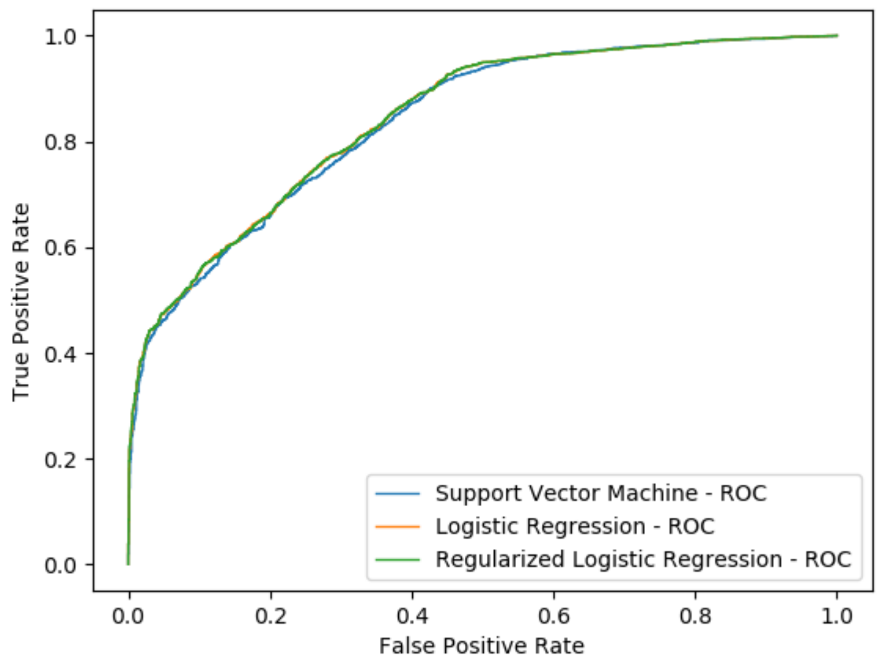
The following 5 graphs show the ROC for each of the algorithms in a cumulative fashion –

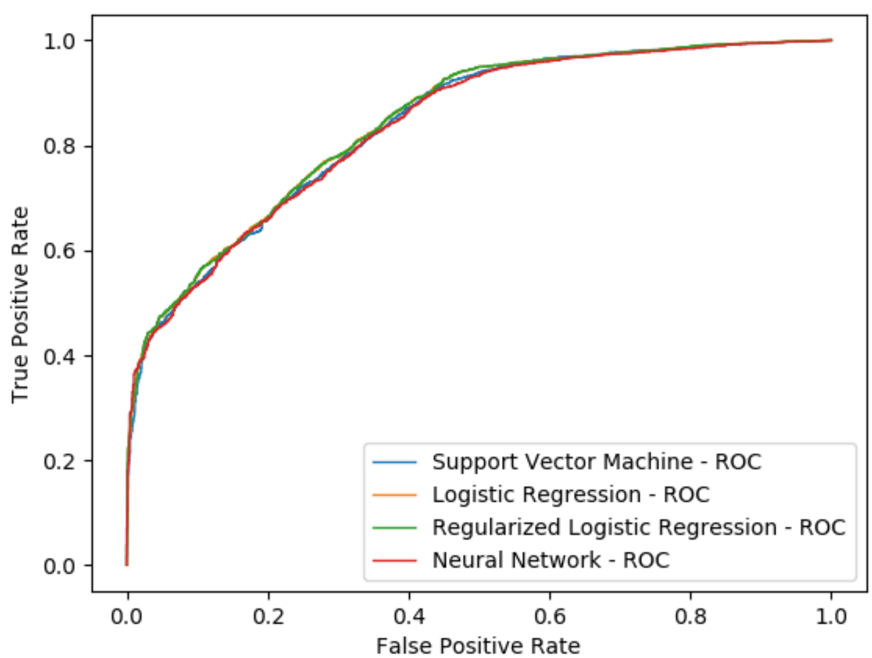


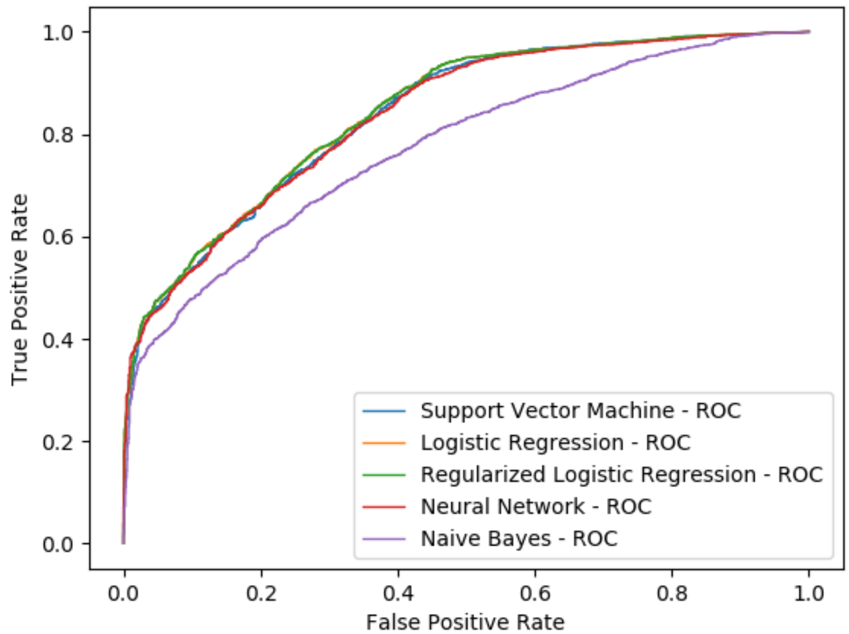
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
Support Vector Machine - ROC
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate



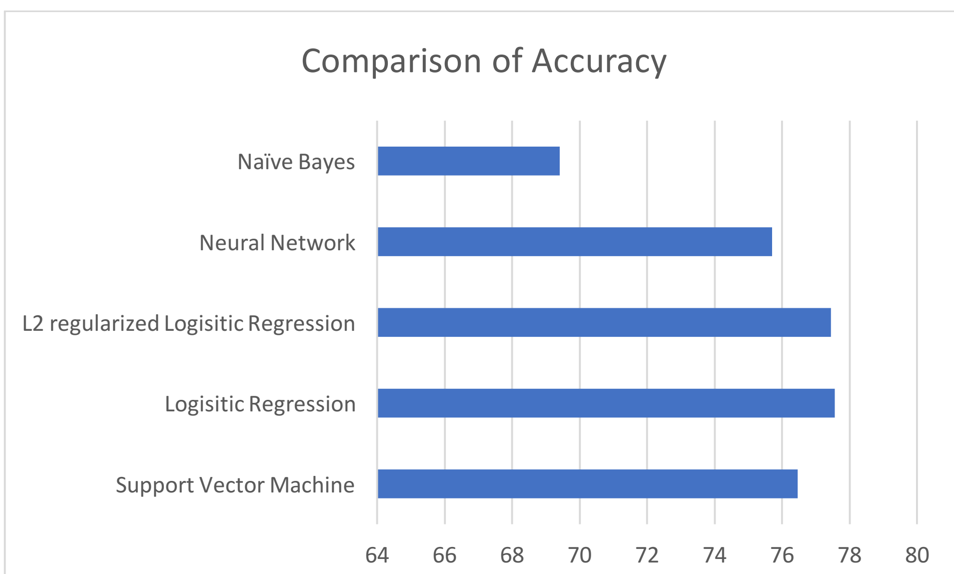
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
Support Vector Machine - ROC
Logistic Regression - ROC
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate


True Positive Rate
False Positive Rate
Support Vector Machine - ROC
Logistic Regression - ROC
Regularized Logistic Regression - ROC


True Positive Rate
False Positive Rate
Support Vector Machine - ROC
Logistic Regression - ROC
Regularized Logistic Regression - ROC
Neural Network - ROC



The following figure shows the comparison between the 5 algorithms for the metric of accuracy–



This clearly indicates that the best performing model for given problem set is logistic regression with an accuracy of 77.56%, followed closely by L2 regularized Logistic Regression. The worst performing model was Naïve Bayes with an accuracy of 69.40.

Part 7:

The data given to us is indeed rich and provides a lot of insights regarding different factors such as favorite tweets, most popular hashtags, user's information and other metadata. An important aspect which can be mined from the tweets is the sentiment. This can provide a keen insight into the mood of the users and can affect the marketing strategies of various companies.

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker.

There are many benefits of Sentiment Analysis –

a. In business and marketing fields, companies need to develop strategies according to customer's feelings. For instance, in our case, during the superbowl, if the sentiment is more towards Seattle Seahawks, the companies can sell their products by orienting them towards this trend. This also helps in gauging how people respond to their campaigns and product launches.
b. Analyzing the tweets also helps in determining if some widespread social phenomena is yet to occur. For instance, if during the superbowl the negative sentiments of supporters of the losing side continue to rise, this could lead to potential dangerous situations. There might be protests, marches and social unrest, which if predicted timely can be controlled.

Problem and Implementation

We plan to analyze the tweets of users with hashtags - #gopatriots and #gohawks. Using all these tweets, we iterate over half hour windows from 3:00 P.M. PST – 8:00 P.M. PST on February 1, 2015 (The timings of Super bowl, 2015). In each of these 10 timestamps, we aim to do sentiment analysis of people posting tweets with the respective hashtags. The objective is to analyze the mood and sentiments of people during the grueling Super bowl game. Receiving the trends of the sentiments of supporters of both the teams in real time can help the companies to target their marketing campaigns in a way that could rake more profits.

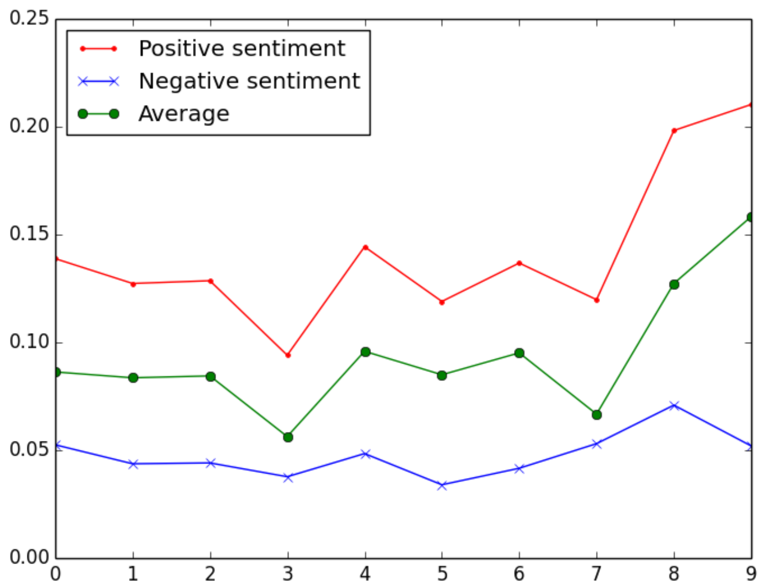
Our approach is as follows –

1. We take the files of #gohawks and #gopatriots separately.
2. The tweets are then loaded and preprocessed to remove any stop words and special characters.
3. Based on the half hour intervals, from 3:00 P.M. PST – 8:00 P.M. PST on February 1, 2015, the tweets are divided into 10 parts.
4. All the tweets are then passed to *TextBlob.* Before this step, all emoticons, URLs, user mentions starting with @ and stop words are removed. This helps us in getting significant words which attribute to the sentiments.
5. The sentiments of the tweets are calculated as either positive or negative using the inbuilt function of *TextBlob.analysis.sentiment.polarity.* A positive value returned by this

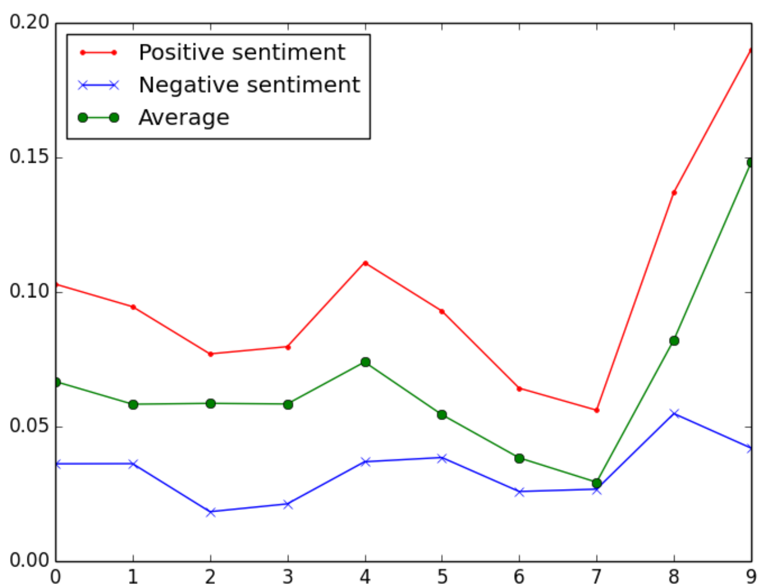
function indicates the degree of positive sentiment and a negative value indicated degree of negative sentiment.

6. We then plot the normalized sentiment values over the 10 30 minute windows for both #gohawks and #gopatriots. The following results are obtained.

#gohawks –

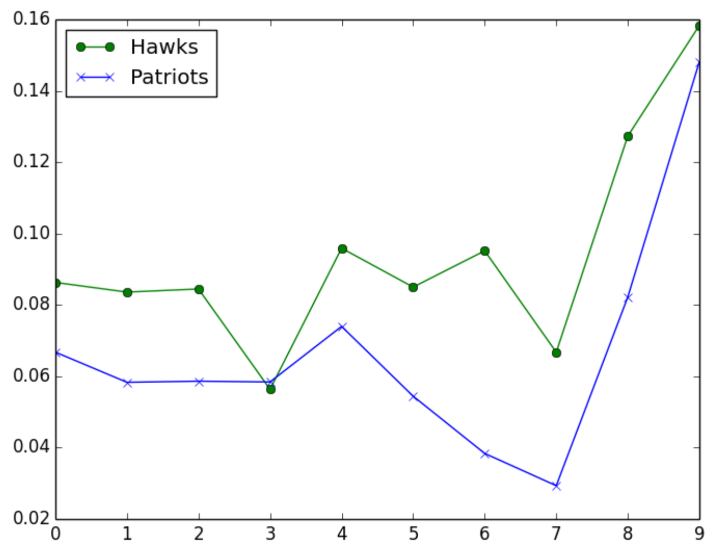


#gopatriots –



A sharp increase in positive sentiment and decrease in negative sentiment is seen in #gopatriots, as the Patriots won the game. As the game progresses , we see a decrease in positive sentiment and increase in negative sentiment for #gohawks. This is due to the fact the Seahawks were losing the game.

We also plot the average trend in sentiments observed for the two teams over time and following graph is obtained. The average value depicts the overall degree of positive sentiment observed.



We see a sharp drop in positive sentiment for Hawks in the middle and end of the game, and a gradual increase in positive sentiment for the Patriots. However, we also notice that Hawks remains the more 'positively' favored team throughout. We can infer that Hawks had more support than Patriots and thus we do not observe a very sharp drop in the above graph for Hawks. To reaffirm our observation, we pulled up statistics of people support and observed that indeed there were more supporters for Hawks than Patriots. The following map reaffirms this observation-

Conclusion:

As proposed, we implement sentiment analysis of tweets over time and analyze the results obtained as graphs, which are shown above. The scope of this problem can further be spread to advertisements by displaying advertisements according to user's sentiment at that time.