EE219 Project 5

Popularity Prediction on Twitter

Winter 2018

Jianfeng He (005025694)
Shouhan Gao (304944056)
ZhengXu Xia(104250792)
Tairan Zhu(605031908)

03/19/2018

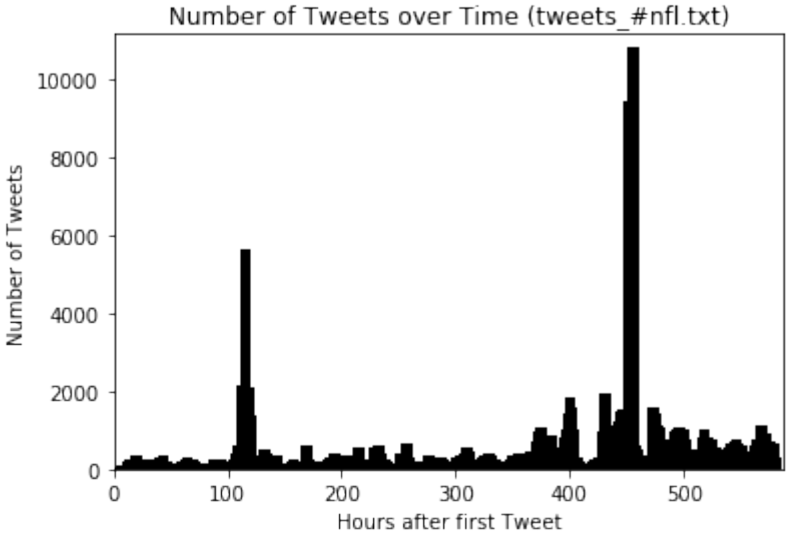
Introduction and Problem Statement:

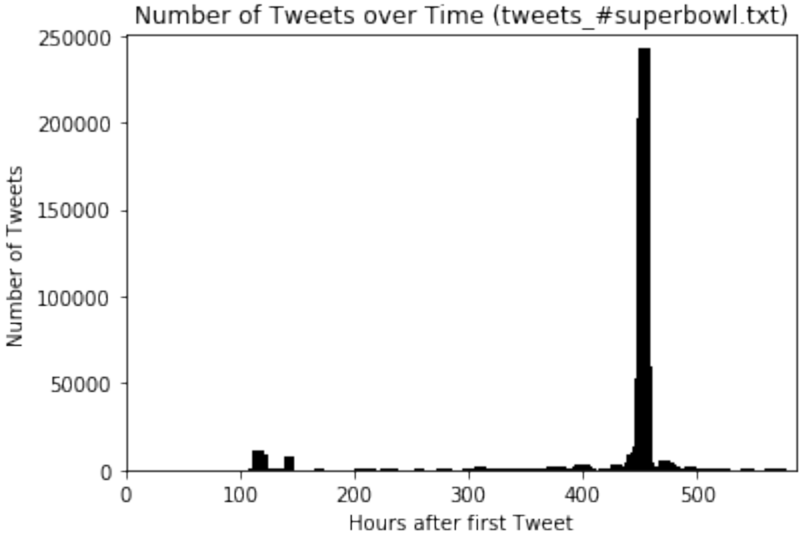
A useful practice in social network analysis is to predict future popularity of a subject or event. Twitter, with its public discussion model, is a good platform to perform such analysis. With Twitter's topic structure in mind, the problem can be stated as: knowing current (and previous) tweet activity for a hashtag, can we predict its tweet activity in the future? More specifically, can we predict if it will become more popular and if so by how much? In this project, we will try to formulate and solve an instance of such problems.

Part 1

Problem 1.1:

```
tweets_#gohawks.txt avg data:
Average number of tweets per hour: 325.371591304
Average number of followers: 1657.4274086
Average number of retweets: 0.20916252073

tweets_#gopatriots.txt avg data:
Average number of tweets per hour: 45.6945105736
Average number of followers: 1325.25823646
Average number of retweets: 0.0268374504422

tweets_#nfl.txt avg data:
Average number of tweets per hour: 441.323431137
Average number of followers: 4126.75479096
Average number of retweets: 0.0509373648774

tweets_#patriots.txt avg data:
Average number of tweets per hour: 834.555509164
Average number of followers: 1820.00749046
Average number of retweets: 0.0914617337093

tweets_#sb49.txt avg data:
Average number of tweets per hour: 1419.88790749
Average number of followers: 2459.03259365
Average number of retweets: 0.178012965702


tweets_#superbowl.txt avg data:
Average number of tweets per hour: 2302.50040188
Average number of followers: 3956.44738677
Average number of retweets: 0.136685580237
```





Problem 1.2:

Features are sorted as:

'Time', 'tweets_total', 'retweets_total', 'followers_total', 'max_followers'

All values of features are counted within one hour window.

Result from tweets_#gohawks.txt:

```
                    OLS Regression Results
==============================================================================
=========
Dep. Variable:                  y   R-squared:                   0.506
Model:                        OLS   Adj. R-squared:              0.502
Method:             Least Squares   F-statistic:                 115.9
Date:            Sat, 10 Mar 2018   Prob (F-statistic):        3.08e-84
Time:                    21:03:33   Log-Likelihood:             -4731.8
No. Observations:             570   AIC:                         9474.
Df Residuals:                 565   BIC:                         9495.
Df Model:                       5
Covariance Type:        nonrobust
==============================================================================
=========
           coef    std err        t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1        10.8334     3.296     3.287     0.001      4.360      17.307
x2         0.4809     0.155     3.102     0.002      0.176       0.785
x3        -0.1030     0.044    -2.322     0.021     -0.190      -0.016
x4         0.0004     0.000     3.301     0.001      0.000       0.001
x5        -0.0007     0.000    -3.799     0.000     -0.001      -0.000
==============================================================================
=========
Omnibus:                  967.837   Durbin-Watson:                2.340
Prob(Omnibus):              0.000   Jarque-Bera (JB):        871763.289
Skew:                      10.037   Prob(JB):                      0.00
Kurtosis:                 193.533   Cond. No.                  1.70e+05
==============================================================================
=========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.7e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[0.00107525 0.00201643 0.02058249 0.00102449 0.00016118]
```

```
Mean squared error = 950839.235
=====================================================================
=========
```

R-squared value: 0.506 **Mean squared error:** 950839.235

Based on P values, the significance of each feature is sorted as below(from high to low)::
max_followers, followers_total, time, tweets_total, retweets_total

Result from tweets_#gopatriots.txt:

```
                            OLS Regression Results
=====================================================================
=========
Dep. Variable:                      y   R-squared:                       0.686
Model:                            OLS   Adj. R-squared:                  0.682
Method:                 Least Squares   F-statistic:                     191.9
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          3.85e-108
Time:                        21:03:41   Log-Likelihood:                -2981.0
No. Observations:                 445   AIC:                             5972.
Df Residuals:                     440   BIC:                             5993.
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================
=========
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.7173      0.734      2.341      0.020       0.276       3.159
x2             0.2588      0.263      0.984      0.326      -0.258       0.776
x3            -0.6968      0.265     -2.626      0.009      -1.218      -0.175
x4             0.0018      0.000      8.093      0.000       0.001       0.002
x5            -0.0018      0.000     -8.426      0.000      -0.002      -0.001
=====================================================================
=========
Omnibus:                      468.998   Durbin-Watson:                   2.127
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           122126.296
Skew:                           3.979   Prob(JB):                         0.00
Kurtosis:                      83.767   Cond. No.                     3.20e+04
=====================================================================
=========
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.2e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[1.96655171e-02 3.25803025e-01 8.93755350e-03 5.75298663e-15
 5.13515590e-16]

Mean squared error = 38557.317
=====================================================================
=========
```

R-squared value: 0.686 **Mean squared error:** 38557.317
Based on P values, the significance of each feature is sorted as below(from high to low)::
max_followers, followers_total, retweets_total, time, tweets_total

Result from tweets_#nfl.txt:

```
                            OLS Regression Results
=====================================================================
=========
Dep. Variable:                      y   R-squared:                       0.651
Model:                            OLS   Adj. R-squared:                  0.648
Method:                 Least Squares   F-statistic:                     214.6
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          5.53e-129
Time:                        21:05:14   Log-Likelihood:                -4526.3
No. Observations:                 581   AIC:                             9063.
Df Residuals:                     576   BIC:                             9084.
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================
=========
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             5.0941      2.127      2.394      0.017       0.916       9.273
x2             1.2095      0.122      9.890      0.000       0.969       1.450
```

```
x3           -0.1486      0.063     -2.363      0.018      -0.272      -0.025
x4        -9.683e-05   2.57e-05     -3.774      0.000      -0.000   -4.64e-05
x5            0.0001   3.28e-05      3.547      0.000     5.2e-05       0.000
=====================================================================
=========
Omnibus:                      481.433   Durbin-Watson:                   2.215
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           314588.545
Skew:                           2.403   Prob(JB):                         0.00
Kurtosis:                     116.894   Cond. No.                     2.96e+05
=====================================================================
=========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.96e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[1.69618149e-02 2.07682569e-21 1.84432660e-02 1.77498327e-04
 4.21540620e-04]

Mean squared error = 342203.373
=====================================================================
=========
```

R-squared value: 0.651 **Mean squared error:** 342203.373
Based on P values, the significance of each feature is sorted as below(from high to low)::
tweets_total,followers_total, max_followers,time, retweets_total

Result from tweets_#patriots.txt:

```
                            OLS Regression Results
=====================================================================
=========
Dep. Variable:                      y   R-squared:                       0.682
Model:                            OLS   Adj. R-squared:                  0.679
Method:                 Least Squares   F-statistic:                     248.4
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          1.47e-141
```

```
Time:                  21:07:51   Log-Likelihood:            -5413.3
No. Observations:               585   AIC:                    1.084e+04
Df Residuals:                 580   BIC:                    1.086e+04
Df Model:                      5
Covariance Type:           nonrobust
=====================================================================
=========
              coef    std err         t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           4.1131      8.737      0.471      0.638     -13.046      21.272
x2           0.9316      0.072     12.947      0.000       0.790       1.073
x3          -0.0496      0.061     -0.816      0.415      -0.169       0.070
x4       -6.031e-05   4.79e-05     -1.260      0.208      -0.000    3.37e-05
x5           0.0003      0.000      2.407      0.016    5.02e-05       0.000
=====================================================================
=========
Omnibus:                     905.513   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           692136.960
Skew:                           8.317   Prob(JB):                         0.00
Kurtosis:                     170.686   Cond. No.                     5.27e+05
=====================================================================
=========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.27e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[6.37969593e-01 7.34635578e-34 4.14706527e-01 2.08241882e-01
 1.63807329e-02]

Mean squared error = 6383819.240
=====================================================================
=========
```

R-squared value: 0.682 **Mean squared error:** 6383819.240
Based on P values, the significance of each feature is sorted as below(from high to low)::
tweets_total, max_followers,followers_total,retweets_total,time

Result from tweets_#sb49.txt:

```
                        OLS Regression Results
==============================================================================
=========
Dep. Variable:                    y   R-squared:                      0.829
Model:                          OLS   Adj. R-squared:                 0.828
Method:               Least Squares   F-statistic:                    518.4
Date:              Sat, 10 Mar 2018   Prob (F-statistic):         3.04e-202
Time:                      21:12:14   Log-Likelihood:               -5285.3
No. Observations:               539   AIC:                        1.058e+04
Df Residuals:                   534   BIC:                        1.060e+04
Df Model:                         5
Covariance Type:          nonrobust
==============================================================================
=========
              coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1         -6.8100     15.620     -0.436      0.663     -37.494      23.874
x2          1.1677      0.053     21.985      0.000       1.063       1.272
x3         -0.4065      0.044     -9.227      0.000      -0.493      -0.320
x4          0.0002   3.05e-05      8.152      0.000       0.000       0.000
x5         -0.0003   7.07e-05     -4.603      0.000      -0.000      -0.000
==============================================================================
=========
Omnibus:                     1067.843   Durbin-Watson:                   1.730
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1882729.602
Skew:                          13.681   Prob(JB):                         0.00
Kurtosis:                     291.242   Cond. No.                     2.64e+06
==============================================================================
=========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.64e+06. This might indicate that there are
```

```
strong multicollinearity or other numerical problems.

Pvalues:
[6.63023126e-01 9.07941626e-77 6.46802205e-19 2.56931443e-15
 5.21425755e-06]

Mean squared error = 19263114.200
==============================================================================
```

R-squared value: 0.829 **Mean squared error:** 19263114.200

Based on P values, the significance of each feature is sorted as below(from high to low)::
tweets_total,retweets_total,followers_total,max_followers,time

Result from tweets_#superbowl.txt:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.811
Model:                            OLS   Adj. R-squared:                  0.809
Method:                 Least Squares   F-statistic:                     496.9
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          6.10e-207
Time:                        21:19:04   Log-Likelihood:                -6080.2
No. Observations:                 585   AIC:                         1.217e+04
Df Residuals:                     580   BIC:                         1.219e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           -20.6955     29.055     -0.712      0.477     -77.761      36.370
x2             2.4265      0.084     28.859      0.000       2.261       2.592
x3            -0.4227      0.026    -16.548      0.000      -0.473      -0.373
x4            -0.0002   2.76e-05     -8.197      0.000      -0.000      -0.000
```

```
x5           0.0010      0.000      6.688      0.000       0.001       0.001
=====================================================================
=========
Omnibus:                 1146.321   Durbin-Watson:                 2.094
Prob(Omnibus):              0.000   Jarque-Bera (JB):        2541370.191
Skew:                      13.390   Prob(JB):                       0.00
Kurtosis:                 324.783   Cond. No.                   5.11e+06
=====================================================================
=========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.11e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[4.76572110e-001 3.16841447e-114 1.14555714e-050 1.58267278e-015
 5.33449752e-011]

Mean squared error = 62397153.568
=====================================================================
=========
```

R-squared value: 0.811 **Mean squared error:** 62397153.568
Based on P values, the significance of each feature is sorted as below(from high to low)::
tweets_total,retweets_total,followers_total,max_followers,time

Discussion: Based on R squared values, the performance of model is decent. Every dataset has a R squared value greater than 0.5. Also, based on the number of tweets in each dataset, it can be concluded that larger dataset will yield more accurate result. Although the value of mean squared error is large, it does not imply that the performance is bad. For example, if the actual value is 300 but predicted value is 3000, this one mistake will significantly boost the value of mean squared error.
After analysis the significance of features of each dataset, it can be concluded that tweets_total and followers_total seem to be important features for most of the dataset. These two features could be used in next part.

Problem 1.3:

Feature used: tweets_total, followers_total, length of tweet(avg), favorite_count, number of user_mentioned

Discussion: The team used two features, which were tweets_total and followers_total, from problem 1.2. Why are those two features important? The reason is obvious.They are related to the number and activity of users within a period of time. Based on this observation and analyzing the papers, the team came up with three extra features, which are the average length of tweet, how many “favorite” this tweet has, and how many users this tweet mentions.

Average length of tweet: This feature might be important. For example, if there is a football game, the average length of tweet might be longer before the game since people will post something to support their team. During the game, the average length of tweet will be shorter because some people will post simple tweets like “OMG!”, “What just happened?” to comment some moments. However, the total number of tweets might increase due to the increase of activity of users. After the game, people might start write some longer tweets to express their emotions but there might be less number of twitter during this time period. Based on those assumptions, the team selected this feature.

Favorite_count: This feature might be important. Usually, a user’s tweet will not get much attention from others. However, if at certain period of time, the total number of favorite count becomes very large, there might be some celebrities posting some tweets or people posting some very good tweets. This kind of situation will trigger the discussion, which might increase the number of tweets in next few hour.

Number of user_mentioned: This feature is important because it somehow represents the number of interactions between people. The increasing number of interaction between people might increases the number of tweets.

Result:

Features are sorted as:

tweets_total, followers_total, len of tweet(avg), favorite_count, number of user_mentioned

Result from tweets_#gohawks.txt:

```
                        OLS Regression Results
==============================================================================
Dep. Variable:                  y   R-squared:                       0.511
Model:                        OLS   Adj. R-squared:                  0.506
Method:             Least Squares   F-statistic:                     117.9
Date:            Tue, 13 Mar 2018   Prob (F-statistic):           2.92e-85
Time:                    13:02:55   Log-Likelihood:                -4729.5
```

```
No. Observations:              570   AIC:                        9469.
Df Residuals:                  565   BIC:                        9491.
Df Model:                        5
Covariance Type:         nonrobust
====================================================================
==========
              coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          0.1833      0.131      1.398      0.163      -0.074      0.441
x2       -6.01e-05   6.85e-05     -0.878      0.381      -0.000    7.44e-05
x3         -0.2166      0.471     -0.460      0.646      -1.142      0.709
x4          0.0014      0.022      0.062      0.950      -0.042      0.044
x5          1.6443      0.301      5.472      0.000       1.054      2.235
====================================================================
==========
Omnibus:                   981.502   Durbin-Watson:                  2.214
Prob(Omnibus):               0.000   Jarque-Bera (JB):         1052127.245
Skew:                       10.261   Prob(JB):                        0.00
Kurtosis:                  212.473   Cond. No.                    2.53e+04
====================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.53e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[1.62687274e-01 3.80575090e-01 6.45933568e-01 9.50351546e-01
 6.71242466e-08]

Mean squared error = 942898.113
====================================================================
==========
```

R-squared value: 0.511 **Mean squared error:** 942898.113

Based on P values, top 3 significant features are sorted as below(from high to low)::
number of user_mentioned,tweets_total,followers_total

Result from tweets_#gopatriots.txt:

```
                            OLS Regression Results
======================================================================
==========
Dep. Variable:                      y   R-squared:                   0.811
Model:                            OLS   Adj. R-squared:              0.809
Method:                 Least Squares   F-statistic:                 377.8
Date:                Tue, 13 Mar 2018   Prob (F-statistic):      1.10e-156
Time:                        13:03:04   Log-Likelihood:            -2867.7
No. Observations:                 445   AIC:                         5745.
Df Residuals:                     440   BIC:                         5766.
Df Model:                           5
Covariance Type:            nonrobust
======================================================================
==========
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.2299      0.092     -2.489      0.013      -0.411      -0.048
x2            -0.0001   4.49e-05     -2.965      0.003      -0.000   -4.48e-05
x3            -0.1430      0.081     -1.770      0.077      -0.302       0.016
x4           -17.3317      1.297    -13.359      0.000     -19.882     -14.782
x5             5.9570      0.385     15.470      0.000       5.200       6.714
======================================================================
==========
Omnibus:                      336.594   Durbin-Watson:                   1.944
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            55654.393
Skew:                           2.280   Prob(JB):                         0.00
Kurtosis:                      57.597   Cond. No.                     6.82e+04
======================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.82e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
```

```
[1.31784940e-02 3.19584437e-03 7.74802341e-02 2.08243035e-34
 2.03804017e-43]

Mean squared error = 23171.656
==============================================================================
==========
```

R-squared value: 0.811 **Mean squared error:** 23171.656
Based on P values, top 3 significant features are sorted as below(from high to low)::
number of user_mentioned,favorite_count**,** followers_total

Result from tweets_#nfl.txt:

```
                    OLS Regression Results
==============================================================================
==========
Dep. Variable:                  y   R-squared:                  0.755
Model:                        OLS   Adj. R-squared:              0.753
Method:             Least Squares   F-statistic:                 355.0
Date:            Tue, 13 Mar 2018   Prob (F-statistic):         2.91e-173
Time:                    13:04:41   Log-Likelihood:            -4423.2
No. Observations:             581   AIC:                        8856.
Df Residuals:                 576   BIC:                     8878.
Df Model:                       5
Covariance Type:        nonrobust
==============================================================================
==========
            coef    std err         t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          0.5150      0.127     4.063      0.000       0.266       0.764
x2      -5.911e-06   1.34e-05    -0.442      0.659   -3.22e-05    2.04e-05
x3          0.6528      0.211     3.091      0.002       0.238       1.068
x4         -2.3309      0.154   -15.090      0.000      -2.634      -2.028
x5          2.0508      0.519     3.949      0.000       1.031       3.071
==============================================================================
==========
Omnibus:                  853.369   Durbin-Watson:                2.537
Prob(Omnibus):              0.000   Jarque-Bera (JB):          256755.974
Skew:                       7.886   Prob(JB):                    0.00
```

```
Kurtosis:                 104.771   Cond. No.                 8.46e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.46e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[5.52692399e-05 6.59015156e-01 2.09328059e-03 1.33783693e-43
 8.80751458e-05]

Mean squared error = 239996.412
==============================================================================
R-squared value: 0.755  Mean squared error: 239996.412
Based on P values, top 3 significant features are sorted as below(from high to low)::
favorite_count,tweets_total,number of user_mentioned
```

Result from tweets_#patriots.txt:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.733
Model:                            OLS   Adj. R-squared:                  0.731
Method:                 Least Squares   F-statistic:                     318.3
Date:                Tue, 13 Mar 2018   Prob (F-statistic):          1.24e-163
Time:                        13:07:28   Log-Likelihood:                -5362.0
No. Observations:                 585   AIC:                         1.073e+04
Df Residuals:                     580   BIC:                         1.076e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
x1         -0.3998      0.119     -3.369      0.001      -0.633      -0.167
x2          0.0004    4.7e-05      8.487      0.000       0.000       0.000
x3         -1.0786      0.944     -1.143      0.253      -2.932       0.775
x4         -0.2507      0.153     -1.640      0.102      -0.551       0.050
x5          0.5840      0.053     10.921      0.000       0.479       0.689
====================================================================
==========
Omnibus:                     1037.288   Durbin-Watson:                   1.869
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           831065.218
Skew:                          11.164   Prob(JB):                         0.00
Kurtosis:                     186.293   Cond. No.                     6.16e+04
====================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.16e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[8.05157286e-04 1.76981317e-16 2.53427102e-01 1.01584926e-01
 2.22667549e-25]

Mean squared error = 5355562.302
====================================================================
==========
```

R-squared value: 0.733 **Mean squared error:** 5355562.302
Based on P values, top 3 significant features are sorted as below(from high to low)::
number of user_mentioned,followers_total,tweets_total

Result from tweets_#sb49.txt:

```
                            OLS Regression Results
====================================================================
==========
Dep. Variable:                      y   R-squared:                       0.833
```

```
Model:                        OLS   Adj. R-squared:                 0.832
Method:                 Least Squares   F-statistic:                  534.5
Date:             Tue, 13 Mar 2018   Prob (F-statistic):          3.45e-205
Time:                       13:12:24   Log-Likelihood:              -5278.5
No. Observations:                 539   AIC:                      1.057e+04
Df Residuals:                     534   BIC:                      1.059e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
==========
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0227      0.097     -0.235      0.814      -0.213       0.167
x2             0.0002   1.85e-05      8.293      0.000       0.000       0.000
x3            -3.6471      1.864     -1.956      0.051      -7.309       0.015
x4            -0.3015      0.090     -3.337      0.001      -0.479      -0.124
x5             0.3932      0.042      9.370      0.000       0.311       0.476
==============================================================================
==========
Omnibus:                     1047.232   Durbin-Watson:                   1.703
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1605892.700
Skew:                          13.152   Prob(JB):                         0.00
Kurtosis:                     269.108   Cond. No.                     3.18e+05
==============================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.18e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[8.14209861e-01 9.00842330e-16 5.09501480e-02 9.06795800e-04
 2.02494251e-19]

Mean squared error = 18779477.023
==============================================================================
==========
```

R-squared value: 0.833 **Mean squared error:** 18779477.023

Based on P values, top 3 significant features are sorted as below(from high to low)::
number of user_mentioned,followers_total,favorite_count

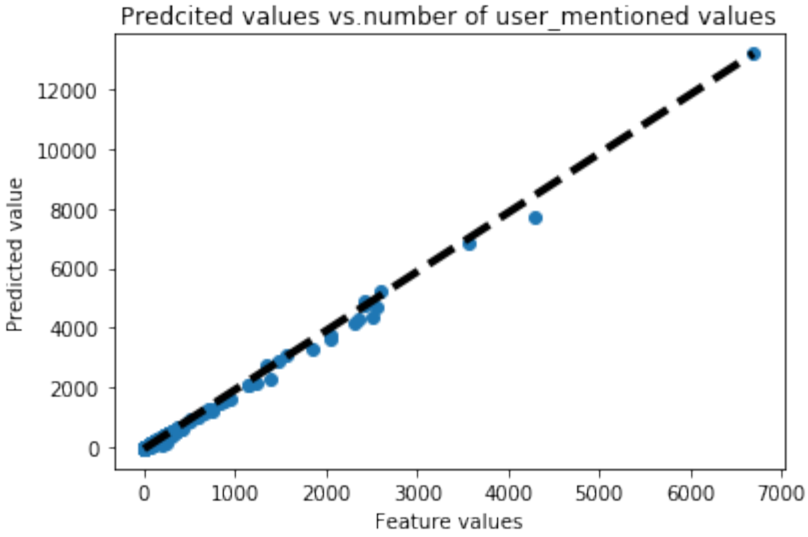
Result from tweets_#superbowl.txt:

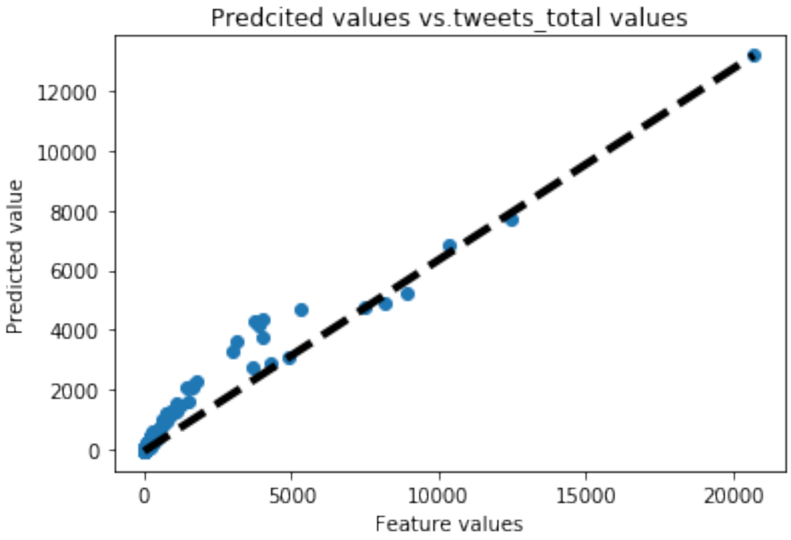
```
                    OLS Regression Results
==============================================================================
==========
Dep. Variable:                  y   R-squared:                  0.834
Model:                        OLS   Adj. R-squared:               0.832
Method:                 Least Squares   F-statistic:                581.4
Date:              Tue, 13 Mar 2018   Prob (F-statistic):         3.40e-223
Time:                      13:20:24   Log-Likelihood:             -6042.4
No. Observations:               585   AIC:                     1.209e+04
Df Residuals:                   580   BIC:                  1.212e+04
Df Model:                         5
Covariance Type:          nonrobust
==============================================================================
==========
              coef    std err        t     P>|t|      [0.025     0.975]
------------------------------------------------------------------------------
x1          0.7533      0.165      4.564      0.000       0.429      1.077
x2      -4.709e-05   1.96e-05     -2.397      0.017   -8.57e-05  -8.51e-06
x3          4.7054      3.081      1.527      0.127      -1.346     10.757
x4         -3.0451      0.144    -21.220      0.000      -3.327     -2.763
x5          2.2736      0.585      3.884      0.000       1.124      3.423
==============================================================================
==========
Omnibus:                   1253.847   Durbin-Watson:                1.969
Prob(Omnibus):                0.000   Jarque-Bera (JB):         2773579.640
Skew:                        16.586   Prob(JB):                      0.00
Kurtosis:                   338.690   Cond. No.                  5.77e+05
==============================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.77e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Pvalues:
[6.14281704e-06 1.68363938e-02 1.27267699e-01 2.16889779e-74
 1.14553139e-04]

Mean squared error = 54834368.407
================================================================================
==========
```

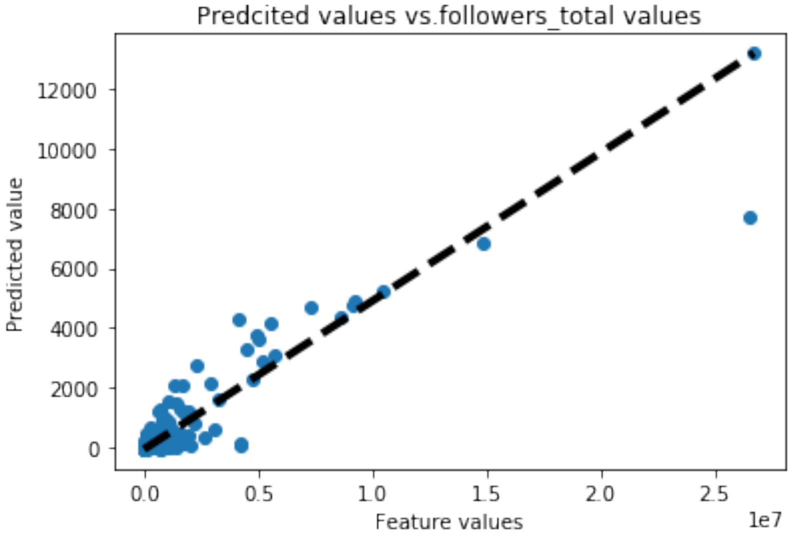
R-squared value: 0.834 **Mean squared error:** 54834368.407
Based on P values, top 3 significant features are sorted as below(from high to low)::
favorite_count,tweets_total,number of user_mentioned

Discussion: Based on R squared values and mean squared error, the performance of model is improved with these five features. Based on the P values, top 3 significant features are extracted from five features for each of dataset. The plots of predicted values versus feature values are shown below:

Result from tweets_#gohawks.txt:

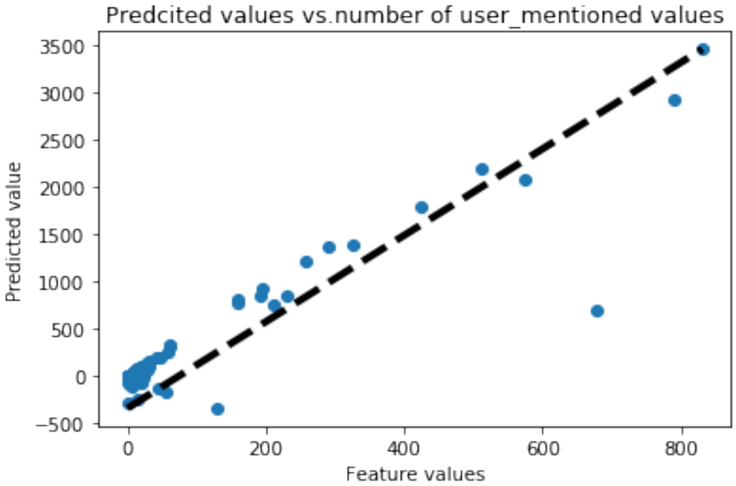


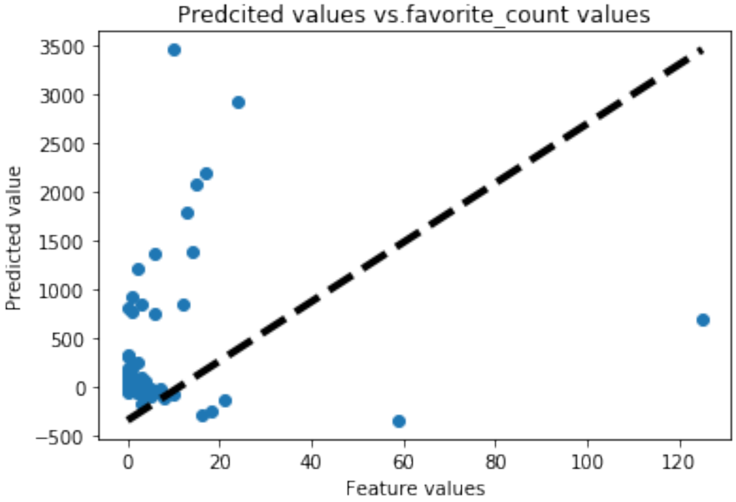

Predcited values vs.tweets_total values

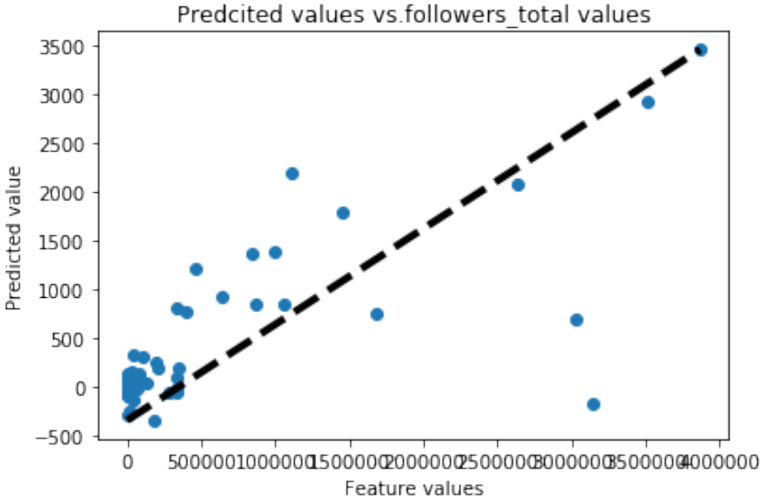

Predcited values vs.followers_total values

===
=======

Result from tweets_#gopatriots.txt:

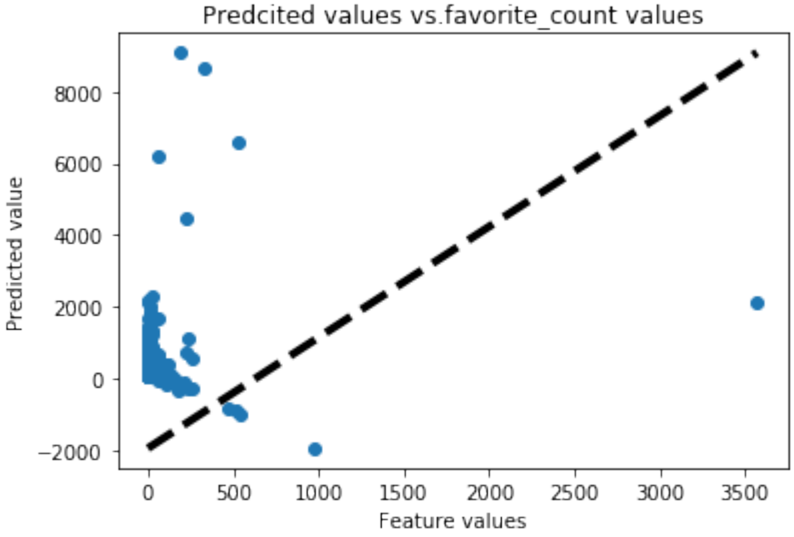

Predcited values vs.number of user_mentioned values
Predicted value
3500
3000
2500
2000
1500
1000
500
0
−500
0
200
400
600
800
Feature values

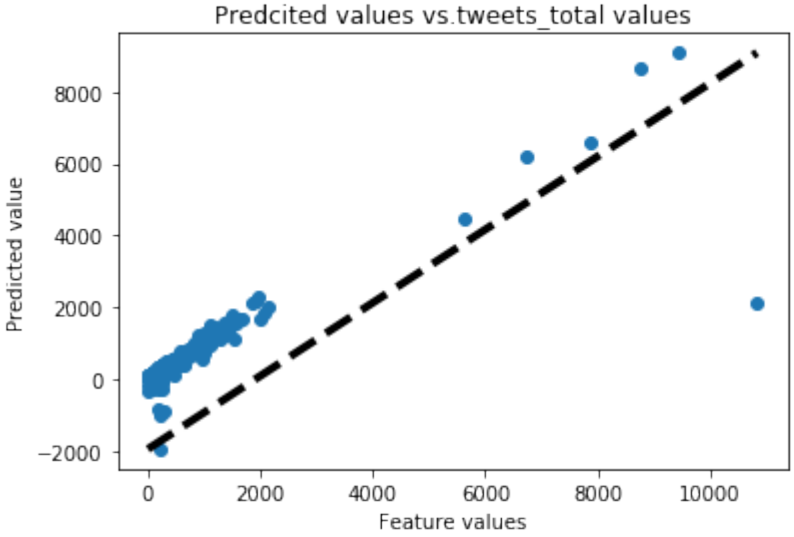

Predcited values vs.favorite_count values
Predicted value
3500
3000
2500
2000
1500
1000
500
0
−500
0
20
40
60
80
100
120
Feature values

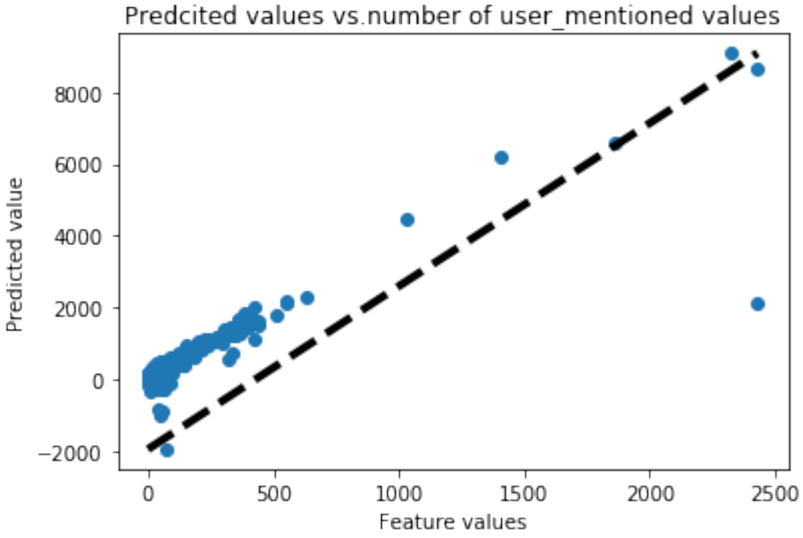

Predcited values vs.followers_total values
Predicted value
3500
3000
2500
2000
1500
1000
500
0
−500
0
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
Feature values

```
======================================================================
=======
```

Result from tweets_#nfl.txt:

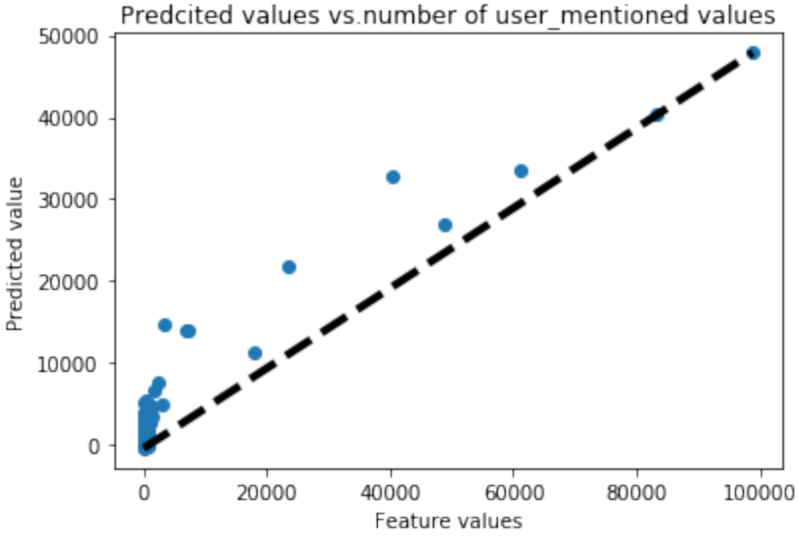


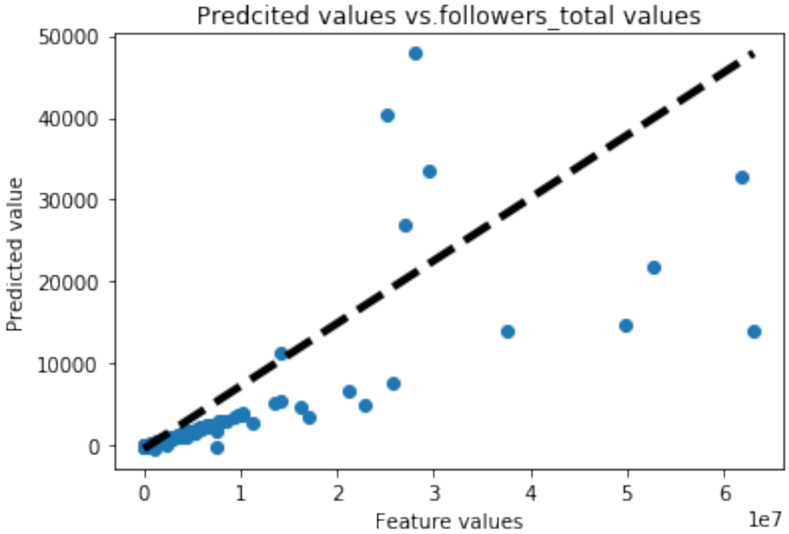


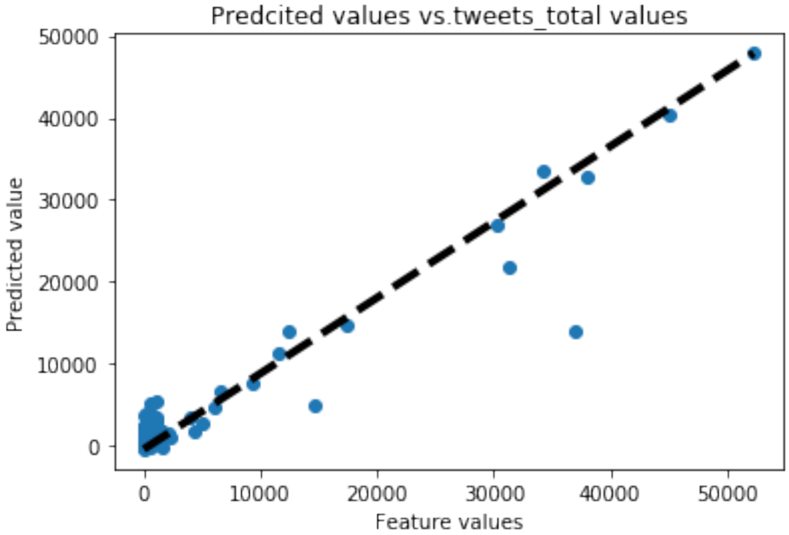


===

Result from tweets_#patriots.txt:

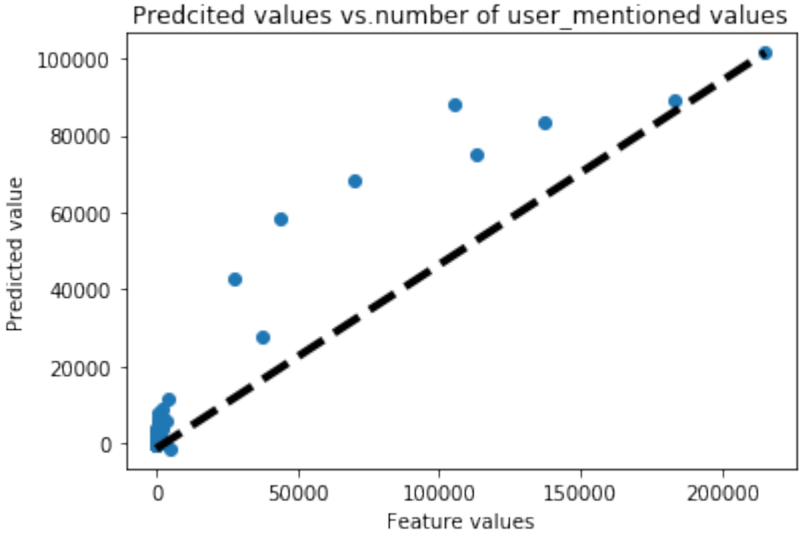


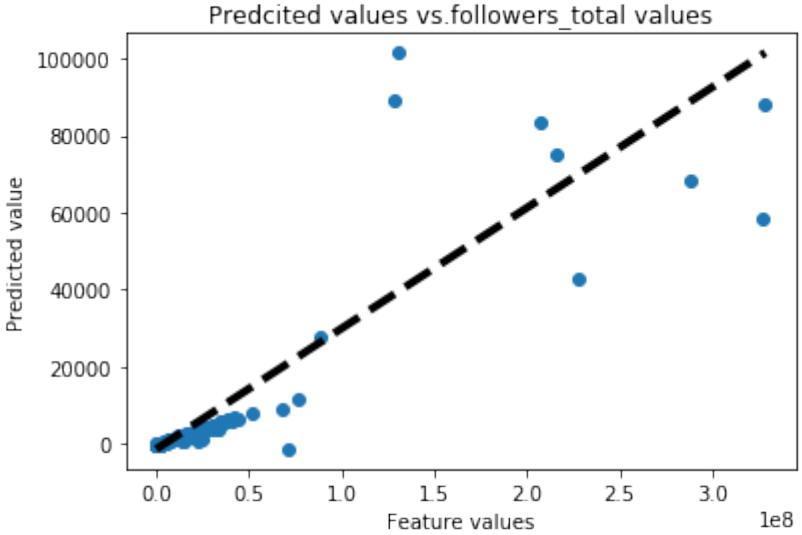

Predcited values vs.followers_total values

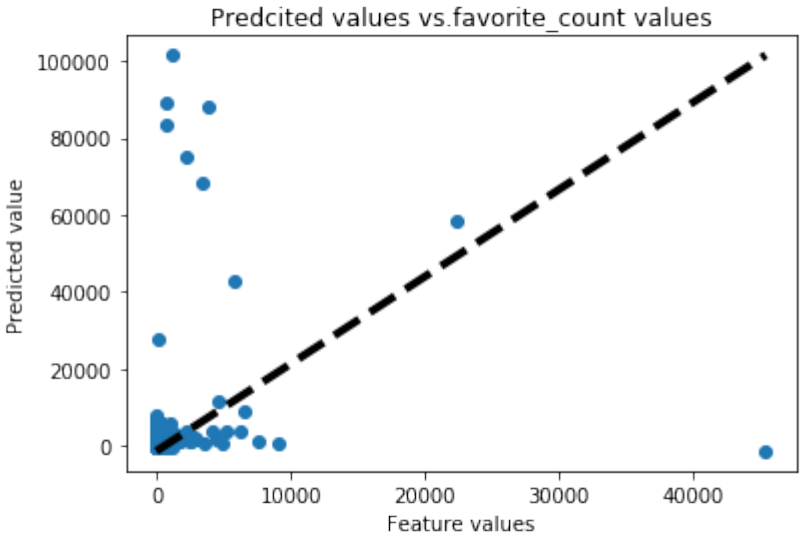

Predcited values vs.tweets_total values

```
===============================================================================
=======
```

Result from tweets_#sb49.txt:

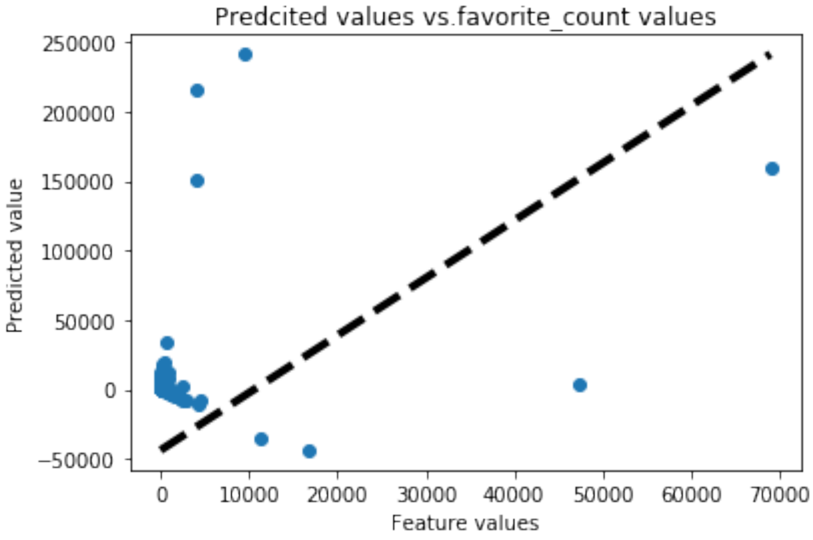


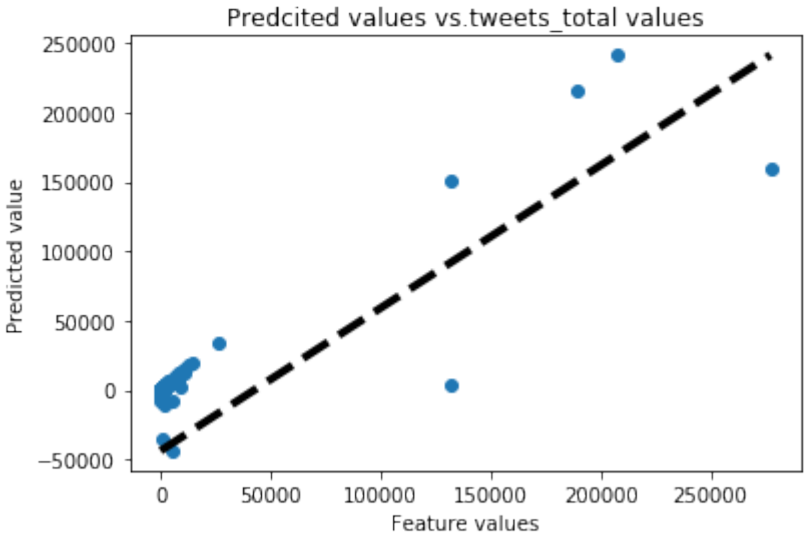


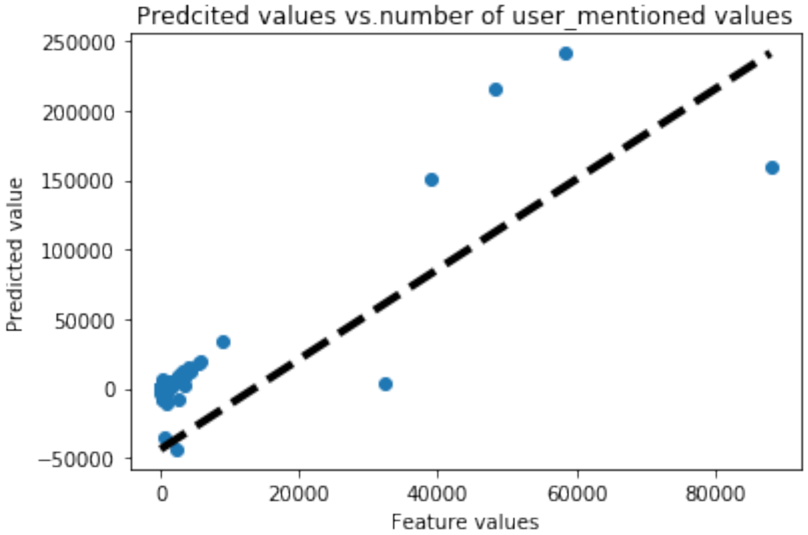


==
=======

Result from tweets_#superbowl.txt:







===
=======

Discussion: It can be observed that most of the top 3 features of each dataset has a relatively linear relationship with predicted value. Also, the performance of the model is improved. The team has designed good features.

Problem 1.4:

In this part, our team implemented the model created in previous problem, in which we chose the features of ***tweets_total, followers_total, length of tweet(avg), favorite_count, number of user_mentioned.*** Firstly, we tried to implement this model to each hashtag ("tweets_#gohawks.txt", "tweets_#gopatriots.txt", "tweets_#nfl.txt", "tweets_#patriots.txt", "tweets_#sb49.txt", "tweets_#superbowl.txt") according to 3 time period.

Furthermore, in addition to ***Linear Regression Model***, our group decided to implement ***Linear Support Vector Regression Model*** and ***Random Forest Regressor Model.*** The reason we chose linear SVM model is that it has a relatively good performance in high dimensional spaces and effective in cases where number of dimensions is greater than the number of samples. To predict future popularity of a certain event (in our case Superbowl), we utilized the features from Tweets (in our case 5 features), which is likely to be treated as a high dimension. Therefore, a prediction model with good performance at high dimension will likely to be working better in prediction according to Tweets' features. Additionally, we implemented Random Forest Regressor Model. It fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. It works in a manner that the error between the value of fitted regression model at each point and the actual value is minimized. Also, we observed that Tweets' data is likely to be nonlinear. In this case, Random Forest Regression Model would have a better behavior in a nonlinear fashion.

Hashtag gohawks:

first period:

```
('RMSE for linear regression model is', 1671.571821081425)
('RMSE for Random Forest Regressor model is', 1096.1852083005842)
('RMSE for LinearSVR model is', 1100.2600887786016)
```

second period:

```
('RMSE for linear regression model is', 20266.900027210446)
('RMSE for Random Forest Regressor model is', 2769.1677585780044)
('RMSE for LinearSVR model is', 6514.048278880562)
```

third period:

```
('RMSE for linear regression model is', 214.7525014340775)
('RMSE for Random Forest Regressor model is', 66.92900398846412)
('RMSE for LinearSVR model is', 383.0070664688258)
```

Hashtag gopatriots:

first period

```
('RMSE for linear regression model is', 60.64625904551412)
```

```
('RMSE for Random Forest Regressor model is', 61.618926891704106)
('RMSE for LinearSVR model is', 82.30775393437325)
second period
('RMSE for linear regression model is', 2217.6928222474935)
('RMSE for Random Forest Regressor model is', 1128.6984757287598)
('RMSE for LinearSVR model is', 1455.8720110671686)
third period
('RMSE for linear regression model is', 22.895030583627182)
('RMSE for Random Forest Regressor model is', 8.919631297604576)
('RMSE for LinearSVR model is', 98.13851963200935)
```

Hashtag nfl:

```
first period
('RMSE for linear regression model is', 318.1609315476443)
('RMSE for Random Forest Regressor model is', 284.64714221884753)
('RMSE for LinearSVR model is', 1630.519554138228)
second period
('RMSE for linear regression model is', 4326.9571395384555)
('RMSE for Random Forest Regressor model is', 3158.7069904439204)
('RMSE for LinearSVR model is', 7457.270071841277)
third period
('RMSE for linear regression model is', 156.25726821547462)
('RMSE for Random Forest Regressor model is', 166.93016726413404)
('RMSE for LinearSVR model is', 1733.4765280817935)
```

Hashtag patriots:

```
first period
('RMSE for linear regression model is', 756.3088025731204)
('RMSE for Random Forest Regressor model is', 749.8751142851752)
('RMSE for LinearSVR model is', 1437.1633280867065)
second period
('RMSE for linear regression model is', 26082.288753840166)
('RMSE for Random Forest Regressor model is', 18435.08080132765)
('RMSE for LinearSVR model is', 21335.436806684756)
third period
('RMSE for linear regression model is', 319.02870954874817)
('RMSE for Random Forest Regressor model is', 153.14366527634579)
('RMSE for LinearSVR model is', 3043.9832472180296)
```

Hashtag sb49:

```
first period
('RMSE for linear regression model is', 99.92309893343163)
('RMSE for Random Forest Regressor model is', 118.65700915657258)
('RMSE for LinearSVR model is', 743.363651130486)
second period
('RMSE for linear regression model is', 319765.74488015106)
('RMSE for Random Forest Regressor model is', 33098.93104544145)
('RMSE for LinearSVR model is', 33622.279871172745)
third period
('RMSE for linear regression model is', 1166.0111578620972)
('RMSE for Random Forest Regressor model is', 230.5987330029301)
('RMSE for LinearSVR model is', 9429.138141534404)
```

Hashtag superbowl:

```
first period
('RMSE for linear regression model is', 800.3889383965305)
('RMSE for Random Forest Regressor model is', 785.0638115117295)
('RMSE for LinearSVR model is', 34070.959436975594)
second period
('RMSE for linear regression model is', 627468.4500797167)
('RMSE for Random Forest Regressor model is', 68842.09401419864)
('RMSE for LinearSVR model is', 142623.71478833762)
third period
('RMSE for linear regression model is', 630.0717465460622)
('RMSE for Random Forest Regressor model is', 447.0603729858627)
('RMSE for LinearSVR model is', 5889.797743610945)
```

Then I aggregate data and fits them using three models and get the following results.

All hashtags:

```
first period
('RMSE for linear regression model is', 2551.3706912759853)
('RMSE for Random Forest Regressor model is', 2685.010457186324)
('RMSE for LinearSVR model is', 18282.941331870403)
second period
```

```
('RMSE for linear regression model is', 540809.654549252)
('RMSE for Random Forest Regressor model is', 97771.44100768653)
('RMSE for LinearSVR model is', 160041.66435283766)
third period
('RMSE for linear regression model is', 1589.7368697188265)
('RMSE for Random Forest Regressor model is', 691.5662886364768)
('RMSE for LinearSVR model is', 12830.72981868729)
```

Observation:

We found that the second period always has the largest RMSE compared to other 2 time periods no matter what models we used. We know that the second time period corresponds to hashtag active period. The reason for that is probably during the game time, because of some random events such as turnovers, short time score leading, and half time show, etc., people’s reaction and tweets are really showing an unstable and fluctuating pattern. And that’s why the error between the prediction and actual data is relatively large. Moreover, we figured that Random Forest Regressor has a better performance across all hashtags and time period. This leads to our conclusion that the advantage of RF Regressor in predicting the future events by using nonlinear data is the most essential factor that we need to take consideration when choosing the most accurate model.

Additionally, we found that the aggregate data has a relatively worse predicting performance compared to the prediction for each individual hashtag. The reason is that more data that combined with different hashtags is easy to cause the cross evaluation error which degrade the prediction accuracy. Furthermore, because of the features we chose, for example, favorite counts, the aggregated data would not have the same accuracy compared to the individual hashtag. Favorite counts would show the fact that during a certain time period the tweets become very hot and popular for individual hashtag, but this doesn’t apply to the combined data.

Problem 1.5:

In this problem, we used ***Random Forest Regressor*** to predict the number of tweets. We separated aggregated training data into 3 time periods according to ‘firstpost_date’ and trained the model in different time periods respectively. Then we applied testing data in different periods to predict the total number of tweets in last hour. The result is shown below. We noticed that for some samples of testing data, for instance sample8_period1, it only has a 5-hour span, which contradicts to the question description that says to predict number of tweets using test data that has a span of 6 hours.

test_data/sample1_period1.txt

```
[215.03846154 181.32564103 170.63589744 298.58974359]
```

test_data/sample4_period1.txt

```
[526.40769231 384.93205128 244.7481685  237.97530825]
```

test_data/sample5_period1.txt

```
[ 518.54807692 1393.42655678  519.68296703  250.78579882]
```

test_data/sample8_period1.txt

```
[200.05128205 154.71794872 148.91575092]
```

test_data/sample2_period2.txt

```
[133388.53846154 133388.53846154 133388.53846154 135944.        ]
```

test_data/sample6_period2.txt

```
[133388.53846154 135944.         184159.38461538 171287.30769231]
```

test_data/sample9_period2.txt

```
[133388.53846154 133388.53846154 133388.53846154 133388.53846154]
```

test_data/sample3_period3.txt

```
[ 761.46153846  867.80769231 1004.57692308  991.11965812]
```

test_data/sample7_period3.txt

```
[65.69230769 64.53846154 63.         57.23076923]
```

test_data/sample10_period3.txt

```
[57.23076923 57.23076923 57.23076923 59.07692308]
```

Part 2:

In this part, we will predict the location of the author of a tweet (either Washington or Massachusetts) based on the content of a tweet. In order to the best algorithm, we will implement three classification algorithms, such as support vector machine (SVM), Naïve Bayes, and Logistic.

Below is the table of accuracy, precision, and recall for three algorithms. The results for SVM and logistic are very similar, because they are both good at predicting binary result.

	Accuracy	Precision	Recall
SVM	0.7065	0.8284	0.3905
Naïve Bayes	0.6639	0.6204	0.5399
Logistic	0.7070	0.8331	0.3886

The following section will talk about how these three algorithms are approached. The roc and confusion matrix are presented as well.

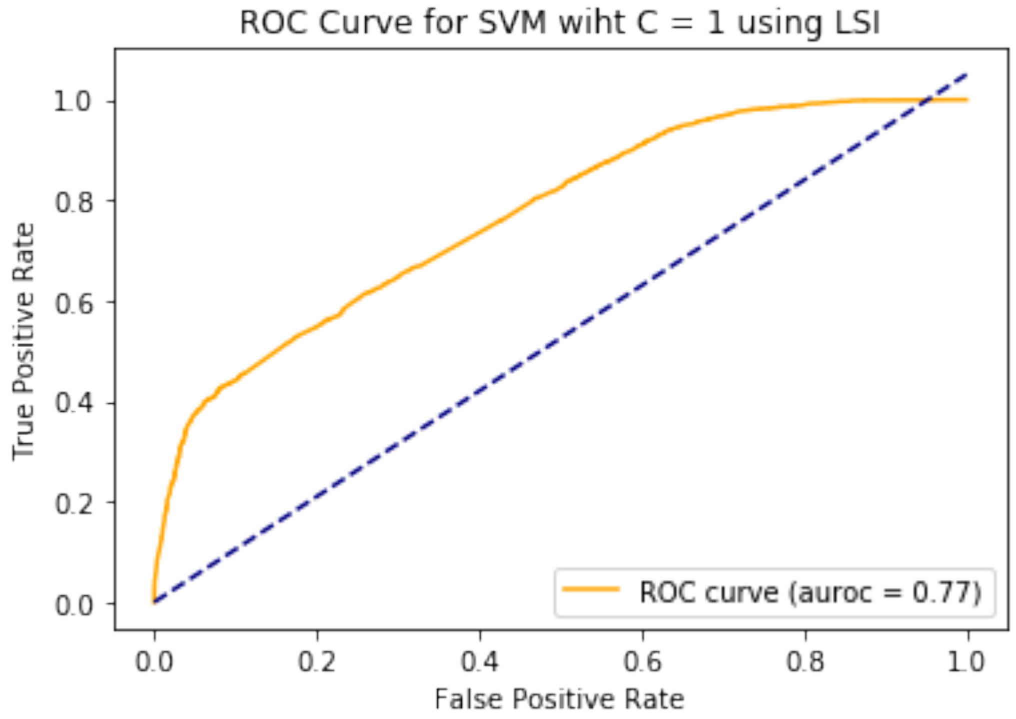
There are soft and hard SVM, based on gamma's value. In order to investigate the best gamma, cross validation is implemented and is used to find the score for each gamma.

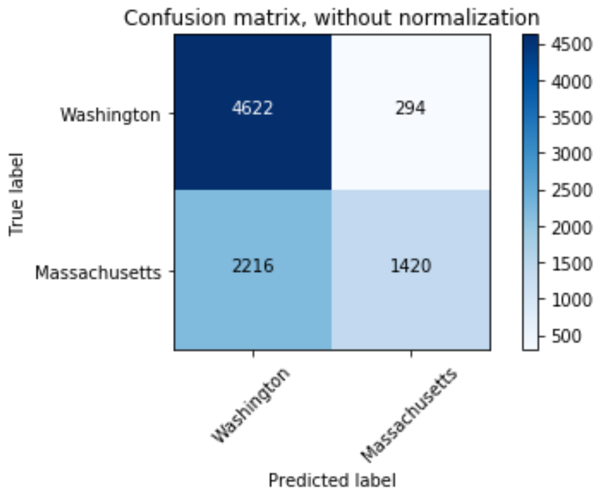
Gamma	10e-3	10e-2	10e-1	0	1	10	100	1000
Score	0.6996	0.7078	0.709	0.7085	0.7091	0.7090	0.7089	0.6235

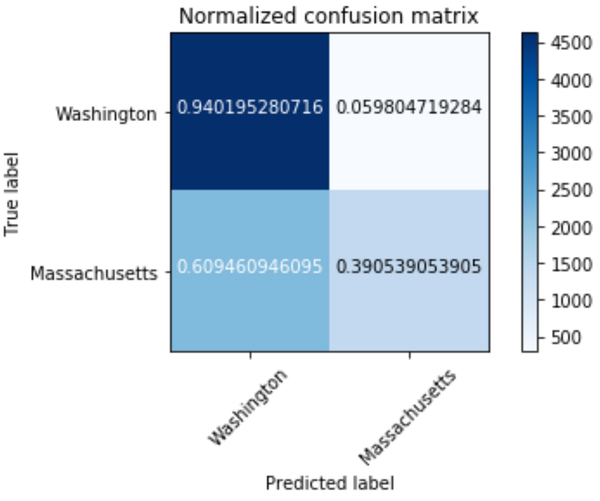
Table: Score for each Gamma

As table shows, the best gamma is 1, because it has highest score.

Latent semantic analysis is utilized to reduce data dimensionality. SVM with gamma equal to 1 is used to train our classifier. Below is the roc curve for SVM with gamma equal to 1. The auc is 0.77

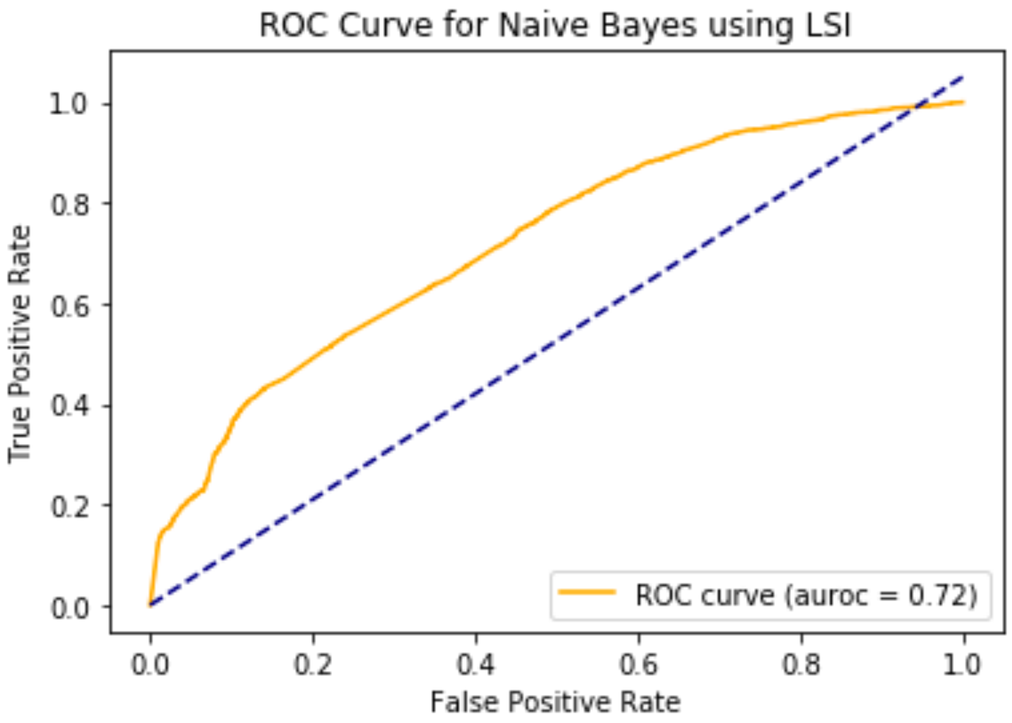


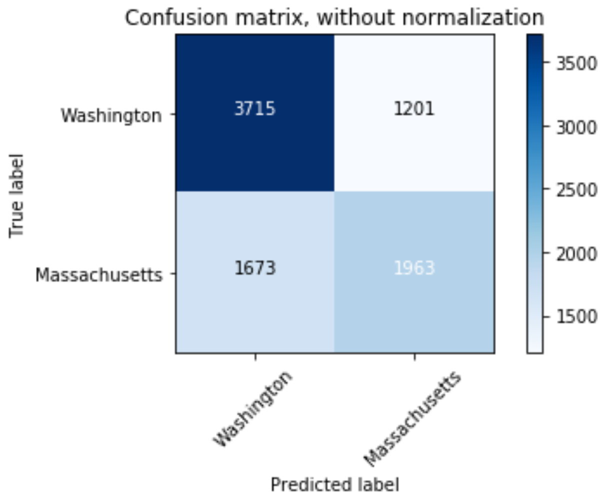


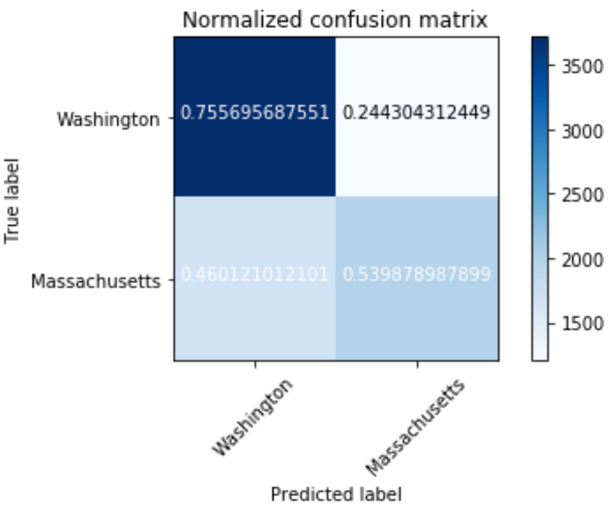


Accuracy = 0.706501, Precision = 0.828471, Recall= 0.390539

The second algorithm is Naïve Bayes. The ROC curve for Naïve Bayes is flatter than that for SVM.







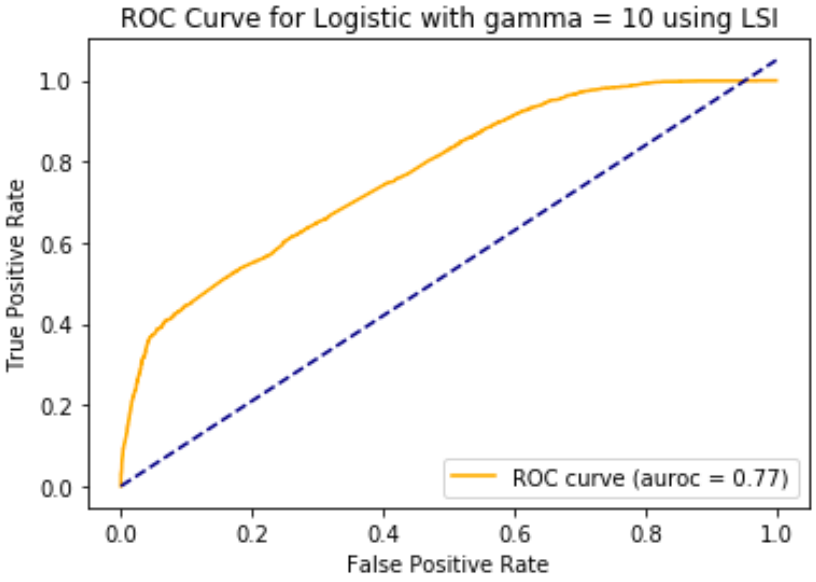
Accuracy = 0.663938, Precision = 0.620417, Recall= 0.539879

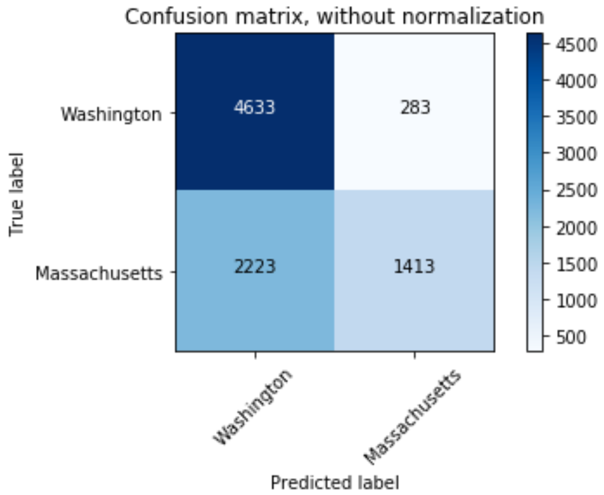
The third algorithm is logistic regression. In logistic regression, penalty equal to l1 is used to regularize the classifier. Different values of gamma for l1 will generate various results. Therefore, cross validation is implemented to find out best gamma.

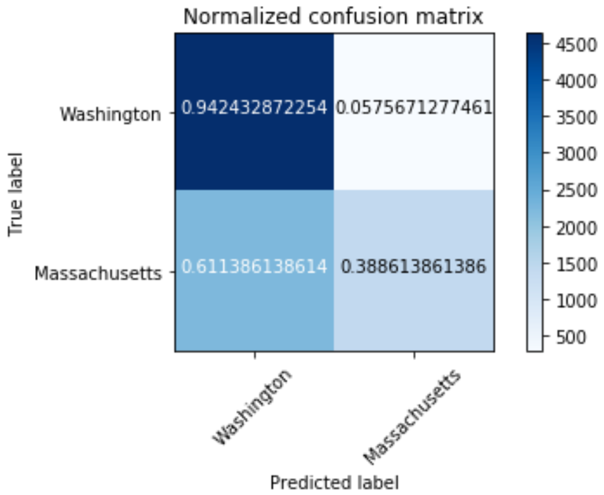
The best gamma according to table below is 10.

Gamma	10e-6	10e-5	10e-4	10e-3	10e-2	0.05	0.1	0.5
Score	0.5888	0. 5888	0. 5888	0.5901	0.6817	0.7065	0.7089	0.7102

Gamma	1	10	100	1000
Score	0.7102	0. 7104	0. 7103	0.7103







Accuracy = 0.706969, Precision = 0.833137, Recall= 0.388614

Part 3

Problem definition

Twitter has become one of the most popular social medias all around the world, people tweet their ideas, opinions, and judges about everything they see, they listen, they feel, and they think. Since humans' words have polarities and can reflect feelings, tweets are not out of the box. Super bowl is always a lot topic that people would like to tweet. As one the most dramatic super bowl game in history, super bowl 2015 has been the most tweeted football game in NFL history. Because intense competitions always have direct connection to fan's feelings, fans of the losing side may have a lot of negative judges against their own team while fans of the winning side may have a lot of positive reviews about their own team. Based on this assumption, we are trying to use unsupervised learning methods to find the clusters of the fans for both team and draw the connections of between the development of the game and people's tweets' polarities.

Data Loading

We are also specially interested in the English-spoken fans from Massachusetts and Washington. Therefore, we loaded tweets with hashtags "#gopatriots", "#patriots", "#gohawks", and "#superbowl" and then filtered out the users who did not speak English. We also dropped tweets which the users did not provided a valid address or did not reside in Massachusetts and Washington.

When loading the tweet data, we were specifically interested in the tweets which were posted in the duration of Super Bowl 2015. In order to do so, we encountered the difficulties of lacking real timestamps of major events of the game because most of sport websites only provide the game timestamps in play-by-play summary. We overcame the difficulties by exhaustively searching online and finding a game video which has all commercials and timeouts. So it can provide the real duration of the game. By watching it, we could manually calculated the unix timestamp of all major events in the game by comparing the video time to Super Bowl 2015 kickoff time(6:30pm PST on February 1st 2015).

After filtering, we obtained 24036 tweets which satisfied our conditions.

Features Extraction

Our feature vector contains game time, user location, retweet count, tweet polarity, if the user is a football fan, real-time game score difference(Patriots - Seahawks).

Game time feature is calculated by taking the difference of tweet citation time and super bowl kickoff time.

Tweet polarity is provided by NLTK VADER sentiment analysis tool. The sentiment analysis tool determines the 4 polarity scores by consulting the positive word corpus, negative word corpus, and English syntax. We took the sign of compound score returned from sentiment analysis tool to determine whether the tweet is positive, negative, or neutral. In our case, we only considered positive and negative tweets and all neutral tweets were dropped in order to improve the clustering results.

Real-time game score differences were obtained from watching the game replay and comparing to the actual kickoff time. We determined the feature based on tweet's citation time.

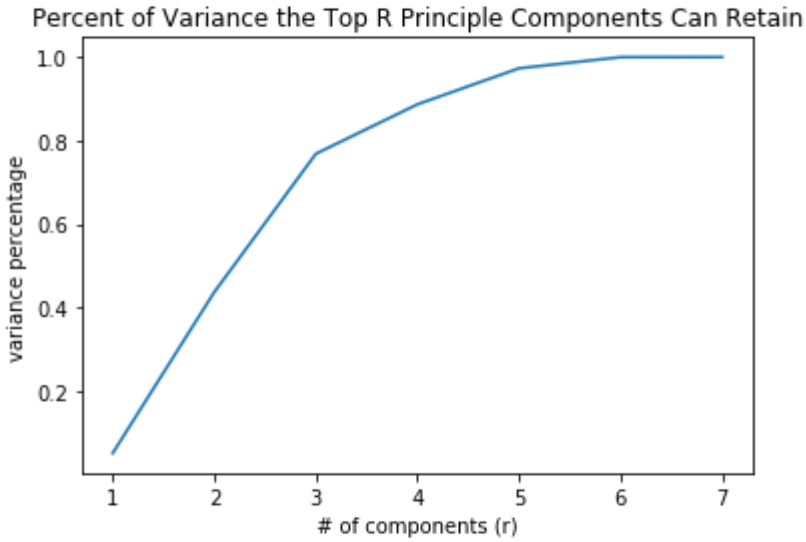
To determine whether the user is a football fan, we kept a set of football game keywords and tried search those keywords in users description.

Other features were provided by twitter dataset.

We performed one hot encoding on user location and tweet polarity and min-max scaler over all feature vector columns so that all feature columns were normalized in the feature matrix.

Dimension Reduction

We performed singular value decomposition on the obtained feature matrix so that we can reduce its dimension and drop useless values.
The figure below describes the each component' contribution to the total variance of the dataset.

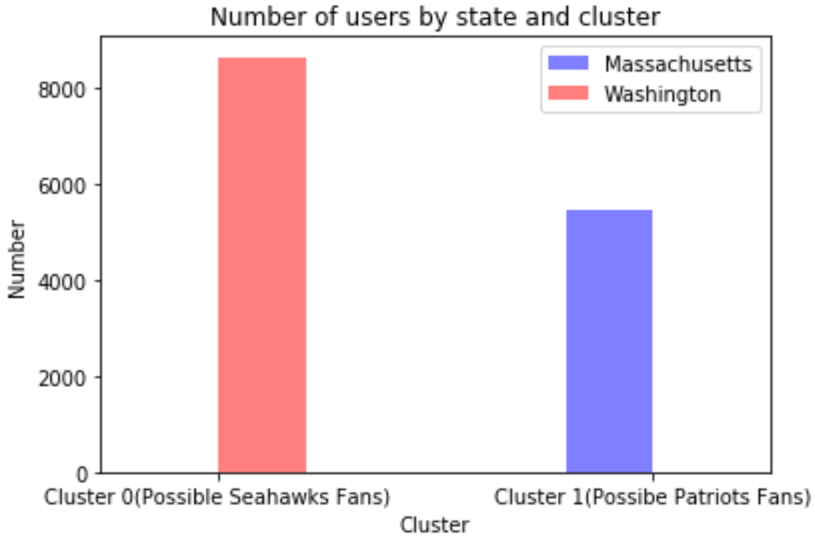


From the figure above, we can see that when the number of component reaches to 5 and the decomposed matrix can basically replace the old feature matrix without major information loss. Therefore, we decided to keep the top 5 components.

Unsupervised Learning with Kmean

We trained Kmean classifier from sklearn package and tried to separate the feature samples to two clusters. We assumed that Seahawks fans and Patriots fans should behave differently during the super bowl game and their twitter features should naturally separate them into two clusters.

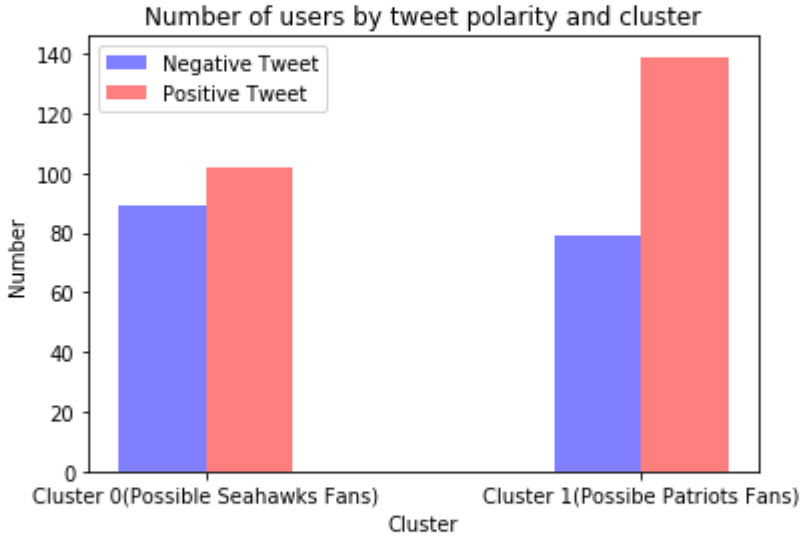
After obtaining the predicted labels, we plotted user locations against the cluster labels to see whether the clusters were interleaved by the user locations.



The figure above shows that cluster 0 mainly contains the users from Washington and cluster 1 mainly contains users from Massachusetts. The figure above makes sense because Kmean classifier tries to group people behave similarly in the same cluster. Due to scale issue, there are also a small number of Massachusetts users in cluster 0 but they are not shown in the figure. The table below gives more details.

	Washington	Massachusetts
Cluster 0	8648	3
Cluster 1	0	5471

We also tried to analyze how the key play of the game impacted twitter users' tweet polarity. As we all know, Seahawks Quarterback Wilson did not choose to let Marshawn Lynch to rush the final play but passed the football which were intercepted just at the end zone line. The bad decision led to their loss of championship and really disappointed the fans and other audiences. We extracted all the tweets which were posted within 4 minutes from the key interception by Patriots player Malcolm Butler and tried to analyze the number of positive and negative tweets in both clusters. What we found are shown in the figure below.



Right after this game-change play, we found that there are more negative reviews in cluster 0, which possibly contains mostly Seahawks fans, than that in cluster 1, which possibly contain mostly Patriots fans. This makes sense because Seahawks fans must be much more disappointed with the stupid decision Wilson made.

We also found that there are much more positive tweets in cluster 1, which possibly contains mostly Patriots fans, than that in cluster 0, which possibly contain mostly Patriots fans.

Conclusion

In conclusion, we did found connections between game key plays and twitter polarities as shown in the results above. The key point of this problem is that we need background knowledge (game scores in this case) other than the tweet datasets. The context information sometimes are more interesting than the dataset itself.

However, this approach contains a lot of problems. Firstly, determining tweets' polarities using nltk sentiment analysis tool is very inaccurate. This determination highly depends on literal meaning of words in tweets and lacks background knowledge of users and the context of the real world. For example, this method fails to understand sarcasm. Secondly, this model may not work well on totally neutral game watchers. Thirdly, this approach suffers from the disadvantages of kmean algorithm. It requires number of clusters to be initialized. It may also suffer from initialization problem of centroids.