EE239AS Project 4

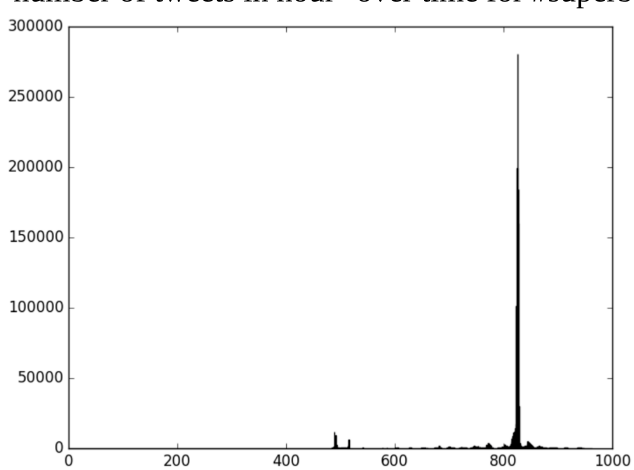
Muchen Xu, Hao Wu, Yuyin Zhou

1) Calculate these statistics for each hashtag: average number of tweets per hour, average number of followers of users posting the tweets, and average number of retweets. Plot "number of tweets in hour" over time for #SuperBowl and #NFL

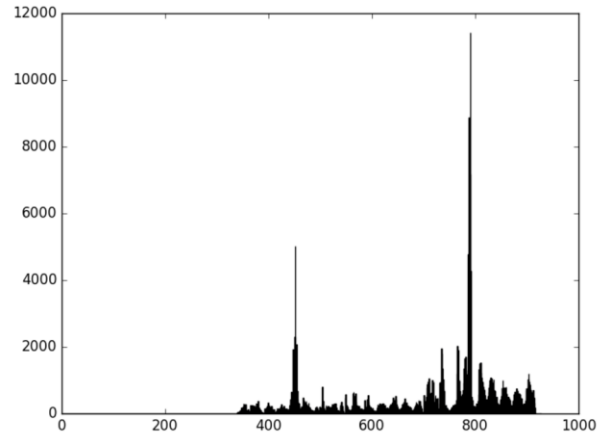
	average number of tweets per hour	number of followers of users posting the tweets*	average number of retweets**
#gohawks	193.36	1593.83	2.01
#gopatriots	38.35	1324.11	1.40
#nfl	279.42	4122.14	1.54
#patriots	498.69	1830.33	1.78
#sb49	1418.44	2379.09	2.51
#superbowl	1400.59	3983.87	2.39

*Number of followers of users: ['author']['followers']
**Number of retweets: ['metrics']["citations"]["total"]

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



"number of tweets in hour" over time for #NFL



2)

Process:

While reading every line of the file, based on the time of the tweet, we put the target variables into different element of target arrays. For example, the first element of *num_tweet* array stores the number of tweets in first hour, every time we ready a tweet belong to the first hour, we increase the first element of *num_tweet* array by one.

After obtaining arrays contains target value in each hour, we use 1 to n-1 row for X and 2 to n row for Y.

Statsmodels.OLS is used for linear regression and analysis. x1, x2, x3, x4 and x5 represent total tweet number, total retweet number, total follower, max follower and hour of the day respectively.

Results:

#gohawks

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.625
Model:                            OLS   Adj. R-squared:                  0.623
Method:                 Least Squares   F-statistic:                     321.4
Date:                Fri, 18 Mar 2016   Prob (F-statistic):         1.26e-202
Time:                        17:37:28   Log-Likelihood:                -7610.8
No. Observations:                 972   AIC:                         1.523e+04
Df Residuals:                     966   BIC:                         1.526e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         -0.3289     40.016     -0.008      0.993       -78.858    78.200
x1             1.1189      0.097     11.520      0.000         0.928     1.310
x2            -0.1802      0.036     -4.940      0.000        -0.252    -0.109
x3          2.428e-06   6.64e-05      0.037      0.971        -0.000     0.000
x4            -0.0002      0.000     -1.642      0.101        -0.000  3.67e-05
x5             0.1321      0.071      1.867      0.062        -0.007     0.271
==============================================================================
Omnibus:                     1023.673   Durbin-Watson:                   2.207
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2074156.023
Skew:                           3.759   Prob(JB):                         0.00
Kurtosis:                     229.180   Cond. No.                     3.35e+06
==============================================================================
```

The R-squared number is low so the training accuracy is low. The total tweet number (x1) has the largest coefficient, t and zero p-value. So it is the most significate feature. Total follower and max follower have p-value > 0.05 so the null hypothesis cannot be rejected. In terms of significance, tweet number > total retweet number > hour of the day > max follower total > total follower

#gopatriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.506
Model:                            OLS   Adj. R-squared:                  0.502
Method:                 Least Squares   F-statistic:                     138.6
Date:                Fri, 18 Mar 2016   Prob (F-statistic):         4.13e-101
Time:                        17:26:13   Log-Likelihood:                -4615.3
No. Observations:                 683   AIC:                             9243.
Df Residuals:                     677   BIC:                             9270.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         -8.8065     16.474     -0.535      0.593       -41.152    23.539
x1            -2.0493      0.255     -8.035      0.000        -2.550    -1.549
x2             2.8243      0.286      9.864      0.000         2.262     3.386
x3            -0.0010      0.000     -3.915      0.000        -0.001    -0.000
x4             0.0005      0.000      2.320      0.021      8.42e-05     0.001
x5             0.0268      0.041      0.657      0.511        -0.053     0.107
==============================================================================
Omnibus:                     1063.589   Durbin-Watson:                   2.379
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1157808.922
Skew:                           8.449   Prob(JB):                         0.00
Kurtosis:                     203.995   Cond. No.                     7.15e+05
==============================================================================
```

The R-squared number is low so the training accuracy is low. The total retweet number (x2) has the largest coefficient, t and zero p-value. So it is the most significate feature. Hour of the day has p-value > 0.05 so the null hypothesis cannot be rejected. In terms of significance, total retweet number > tweet number > total follower > max follower > hour of the day

#nfl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.700
Model:                            OLS   Adj. R-squared:                  0.698
Method:                 Least Squares   F-statistic:                     428.5
Date:                Fri, 18 Mar 2016   Prob (F-statistic):          2.36e-237
Time:                        17:42:19   Log-Likelihood:                -6847.0
No. Observations:                 926   AIC:                         1.371e+04
Df Residuals:                     920   BIC:                         1.373e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        -34.5568     27.148     -1.273      0.203       -87.837    18.723
x1             0.8388      0.102      8.215      0.000         0.638     1.039
x2             0.0151      0.057      0.266      0.791        -0.096     0.126
x3         -2.164e-05   2.36e-05     -0.915      0.360       -6.8e-05  2.48e-05
x4          9.736e-06   3.04e-05      0.320      0.749         -5e-05  6.94e-05
x5             0.1880      0.054      3.492      0.001         0.082     0.294
==============================================================================
Omnibus:                      870.752   Durbin-Watson:                   1.932
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           472504.198
Skew:                           3.325   Prob(JB):                         0.00
Kurtosis:                     113.463   Cond. No.                     5.51e+06
==============================================================================
```

The R-squared number is higher than previous two cases so the training accuracy is higher. The total tweet number (x1) has the largest coefficient, t and zero p-value. So it is the most significate feature. Total number of retweets, total follower and max follower have p-value > 0.05 so the null hypothesis cannot be rejected. In terms of significance, total tweet number > hour of the day > total follower > max follower > total retweet number

#patriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.721
Model:                            OLS   Adj. R-squared:                  0.719
Method:                 Least Squares   F-statistic:                     503.1
Date:                Fri, 18 Mar 2016   Prob (F-statistic):          6.15e-267
Time:                        17:39:44   Log-Likelihood:                -8741.9
No. Observations:                 980   AIC:                         1.750e+04
Df Residuals:                     974   BIC:                         1.753e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        -36.0310    118.675     -0.304      0.761      -268.920   196.858
x1             1.7815      0.090     19.792      0.000         1.605     1.958
x2            -0.8256      0.086     -9.559      0.000        -0.995    -0.656
x3             0.0002    4.6e-05      3.890      0.000      8.86e-05     0.000
x4         -7.705e-05   9.38e-05     -0.821      0.412        -0.000     0.000
x5             0.3621      0.212      1.705      0.089        -0.055     0.779
==============================================================================
Omnibus:                     1689.614   Durbin-Watson:                   1.801
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1391405.015
Skew:                          11.167   Prob(JB):                         0.00
Kurtosis:                     186.239   Cond. No.                     1.06e+07
==============================================================================
```

The R-squared number is higher than previous three cases so the training accuracy is higher. The total tweet number (x1) has the largest coefficient, t and zero p-value. So it is the most significate feature. Max follower and hour of the day have p-value > 0.05 so the null hypothesis cannot be rejected. In terms of significance, total tweet number > total retweet number > total follower > hour of the day > max follower

#sb49

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.852
Model:                            OLS   Adj. R-squared:                  0.850
Method:                 Least Squares   F-statistic:                     661.1
Date:                Fri, 18 Mar 2016   Prob (F-statistic):          6.82e-236
Time:                        17:36:11   Log-Likelihood:                -5633.5
No. Observations:                 582   AIC:                         1.128e+04
Df Residuals:                     576   BIC:                         1.131e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         73.1629    327.728      0.223      0.823      -570.524   716.850
x1             1.1678      0.048     24.233      0.000         1.073     1.262
x2            -0.3485      0.039     -8.837      0.000        -0.426    -0.271
x3             0.0002   2.76e-05      7.079      0.000         0.000     0.000
x4            -0.0002   6.97e-05     -2.562      0.011        -0.000 -4.17e-05
x5            -0.7862      1.045     -0.752      0.452        -2.840     1.267
==============================================================================
Omnibus:                      888.041   Durbin-Watson:                   1.487
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           573029.125
Skew:                           8.141   Prob(JB):                         0.00
Kurtosis:                     155.856   Cond. No.                     6.25e+07
==============================================================================
```

The R-squared number is higher than previous four cases so the training accuracy is higher. The total tweet number (x1) has the largest coefficient, t and zero p-value. So it is the most significate feature. Hour of the day has p-value > 0.05 so the null hypothesis cannot be rejected. In terms of significance, total tweet number > total retweet number > total follower > hour of the day > max follower

#superbowl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.713
Model:                            OLS   Adj. R-squared:                  0.711
Method:                 Least Squares   F-statistic:                     474.5
Date:                Fri, 18 Mar 2016   Prob (F-statistic):          5.18e-256
Time:                        17:48:26   Log-Likelihood:                -9950.2
No. Observations:                 962   AIC:                         1.991e+04
Df Residuals:                     956   BIC:                         1.994e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        334.6198    506.203      0.661      0.509      -658.777  1328.017
x1             2.9367      0.308      9.534      0.000         2.332     3.541
x2            -0.9485      0.152     -6.249      0.000        -1.246    -0.651
x3           1.99e-06   3.46e-05      0.058      0.954     -6.59e-05  6.99e-05
x4             0.0007      0.000      3.988      0.000         0.000     0.001
x5            -1.7830      1.037     -1.719      0.086        -3.819     0.253
==============================================================================
Omnibus:                     1063.801   Durbin-Watson:                   1.720
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           912756.093
Skew:                           4.405   Prob(JB):                         0.00
Kurtosis:                     153.645   Cond. No.                     9.08e+07
==============================================================================
```

The R-squared number is high so the training accuracy is high. The total tweet number (x1) has the largest coefficient, t and zero p-value. So it is the most significate feature. Total follower has p-value > 0.05 so the null hypothesis cannot be rejected. In terms of significance, total tweet number > total retweet number > max follower> hour of the day > total follower

3)

Four new features are used, they are:

Author count (A):
Number of authors ([‘author’][‘name’]). This feature can be used to recognize those hashtags automatically posted by some fake accounts.

Mention count (B):
Sum of times been mentioned of each tweet (([‘tweet’][‘entities’][‘user_mentions’])). If a user was mentioned in a tweet with a hashtag, he probably took part in the topic, especially when this mention came from his friends.

Co-occurrence times of other hashtags (C):
Number of tweets that has one or more hashtags ([‘tweet’][‘entities’][‘hashtags’]). More hashtag together may indicate higher popularity.

Url ratio (D):

Number of tweets that has URL/total number of tweets (['tweet']['entities']['urls']). High ratio of tweets with urls may indicate an interesting topic

Together with features in 2), all 9 features are used.

Random Forrest Tree model is used as regression model for this problem for better accuracy compared with linear regression model.

x1, x2, x3, x4, x5, x6, x7, x8, x9 represent author number, mentioned number, hashtag co-occurrence, URL ratio, total tweet number, total retweet number, total follower, max follower, hour of the day

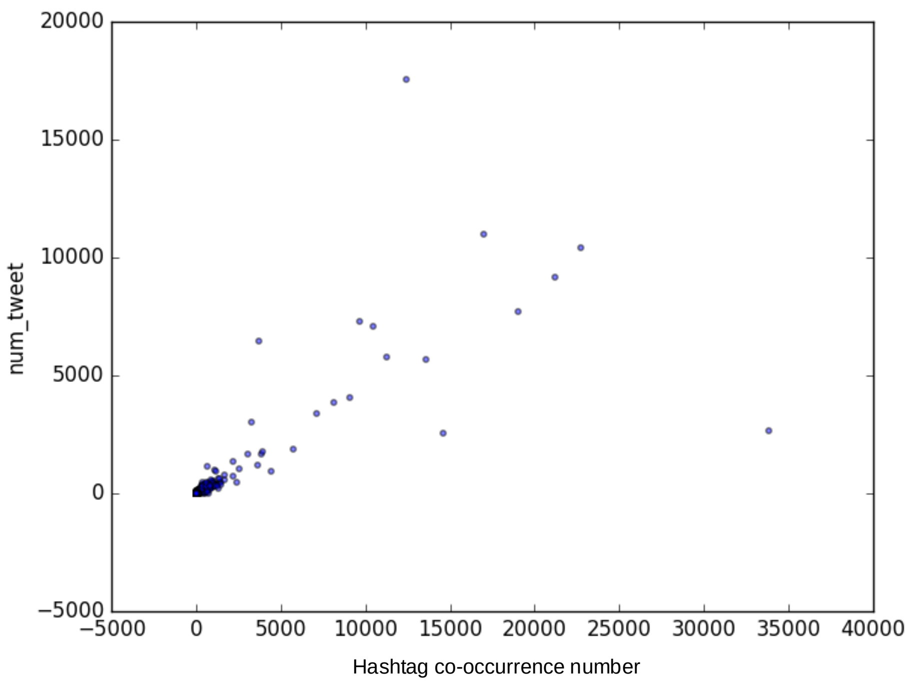
sklearn.ensemble.RandomForestClassifier was used and its feature_importances_ and score functions are used to evaluate feature importance and regression accuracy.

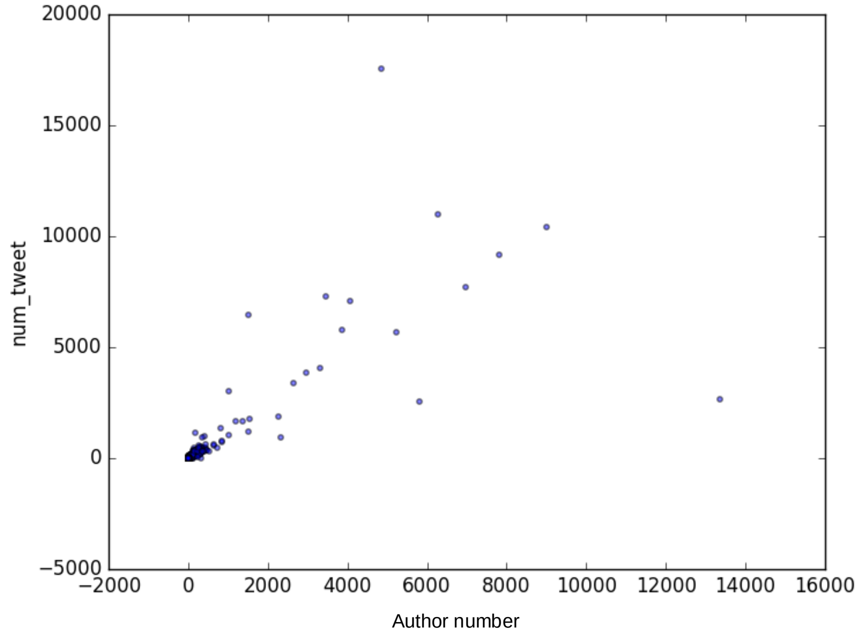
Results:
#gohawks
Accuracy:
Mean accuracy score is 0.94953120160.

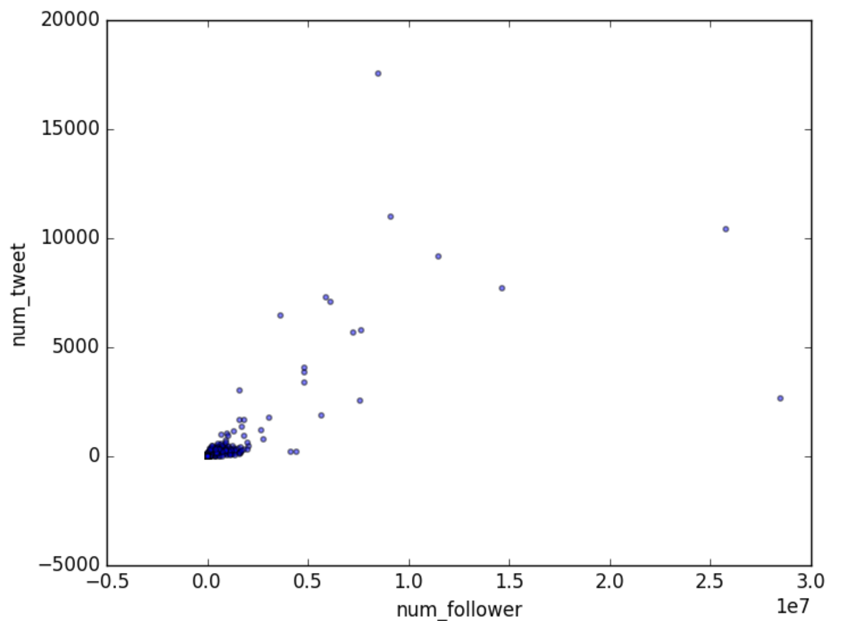
Feature importance:

x1	x2	x3	x4	x5	x6	x7	x8	x9
1.59e-1	1.39e-1	1.88e-1	3.16e-9	1.04e-1	5.80e-2	2.78e-1	3.70e-2	3.65e-2

Follower number, author number, mentioned number, hashtag co-occurrence are the top three features. URL ratio is the least importance feature.







#gopatriots

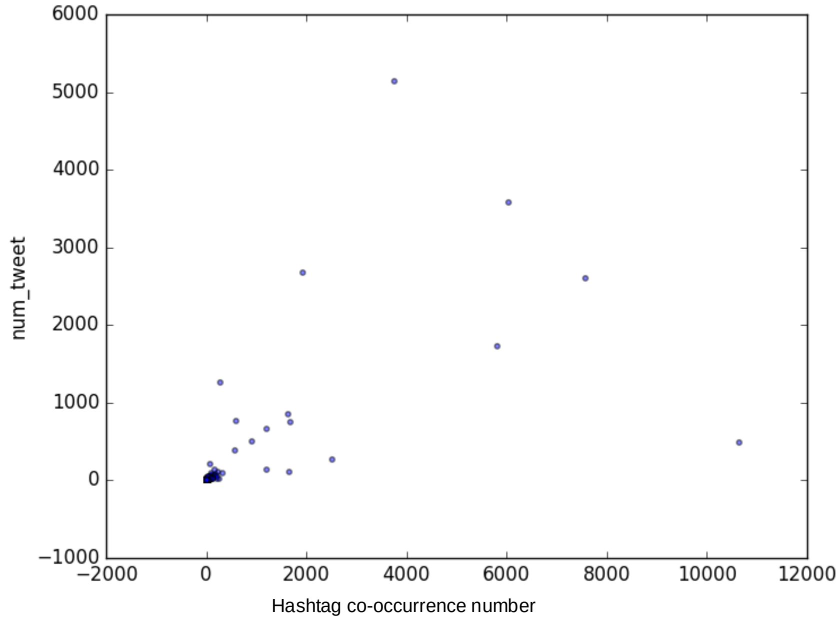
Accuracy:

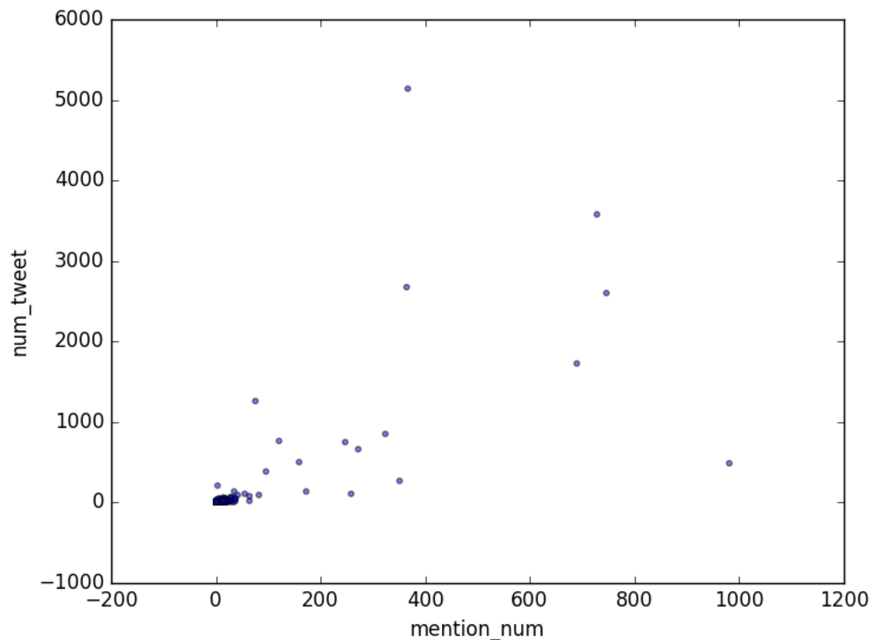
Mean accuracy score is 0.933574203546

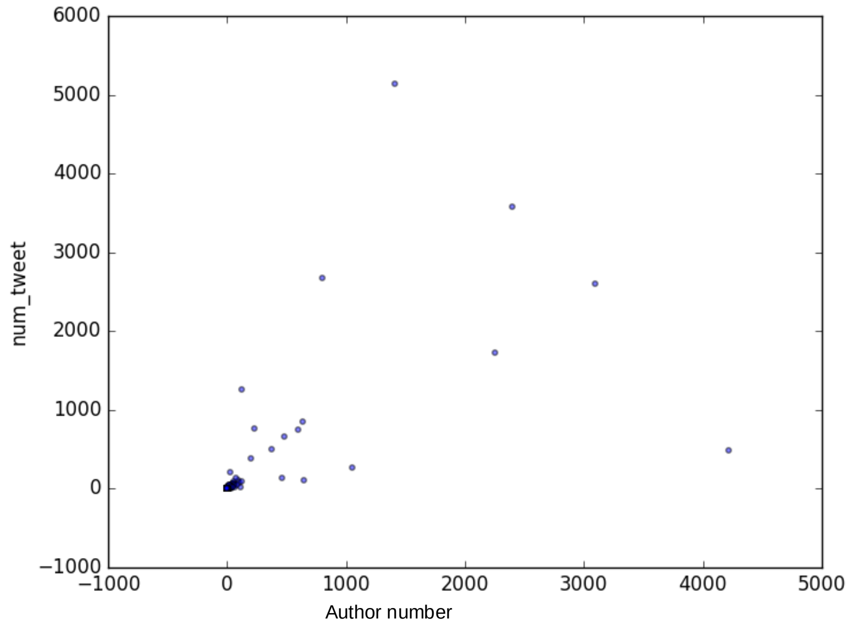
Feature importance:

x1	x2	x3	x4	x5	x6	x7	x8	x9
1.19e-1	4.30e-1	1.82e-1	2.61e-7	8.44e-2	6.95e-2	1.47e-2	5.25e-2	4.74e-2

Author number, mentioned number, hashtag co-occurrence are the top three features. URL ratio is the least importance feature.







#nfl

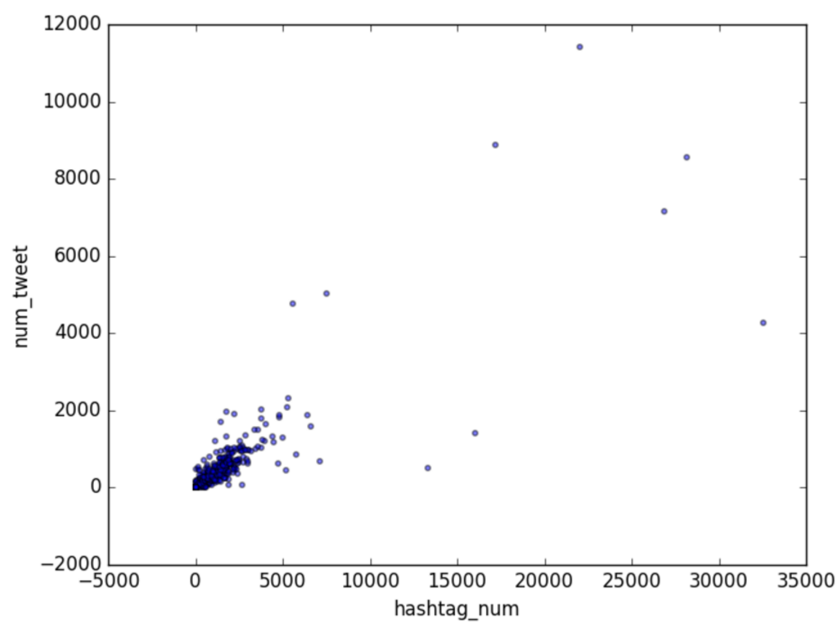
Accuracy:

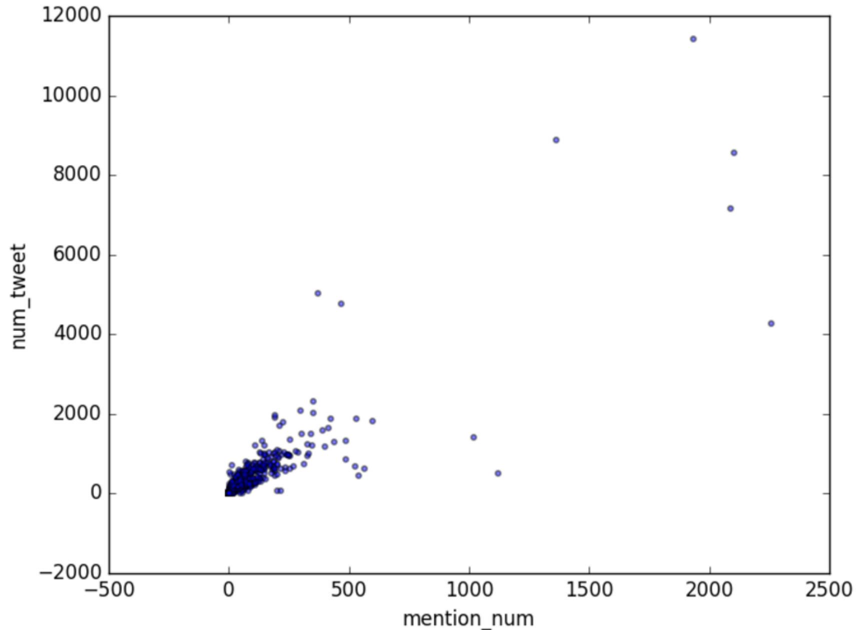
Mean accuracy score is 0.949531201604

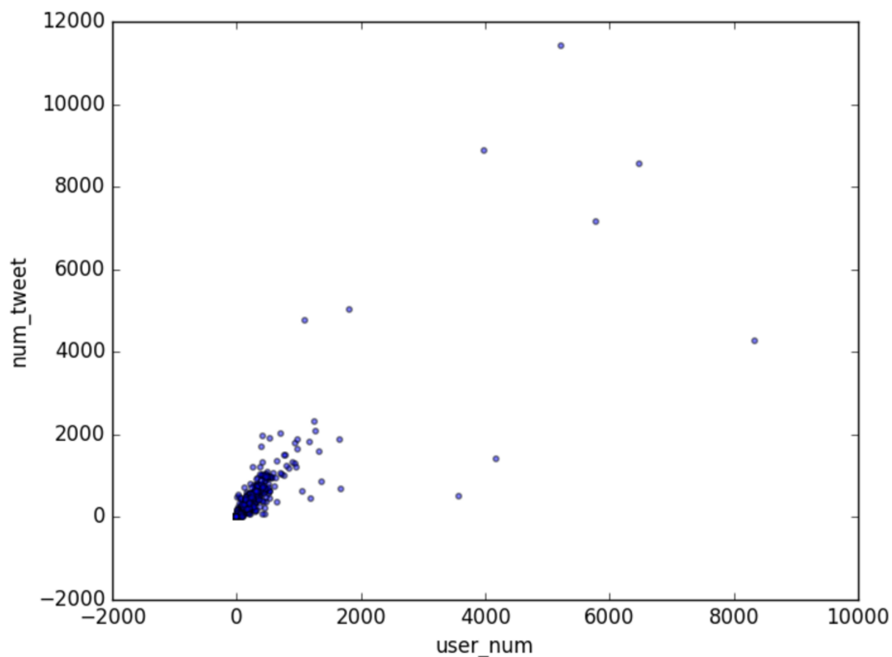
Feature importance:

x1	x2	x3	x4	x5	x6	x7	x8	x9
2.1e-1	1.97e-1	2.6e-1	2.69e-8	1.08e-1	1.27e-1	4.91e-2	2.22e-2	2.55e-2

Hashtag co-occurrence, author number, mentioned number, hashtag co-occurrence are the top three features. URL ratio is the least importance feature.





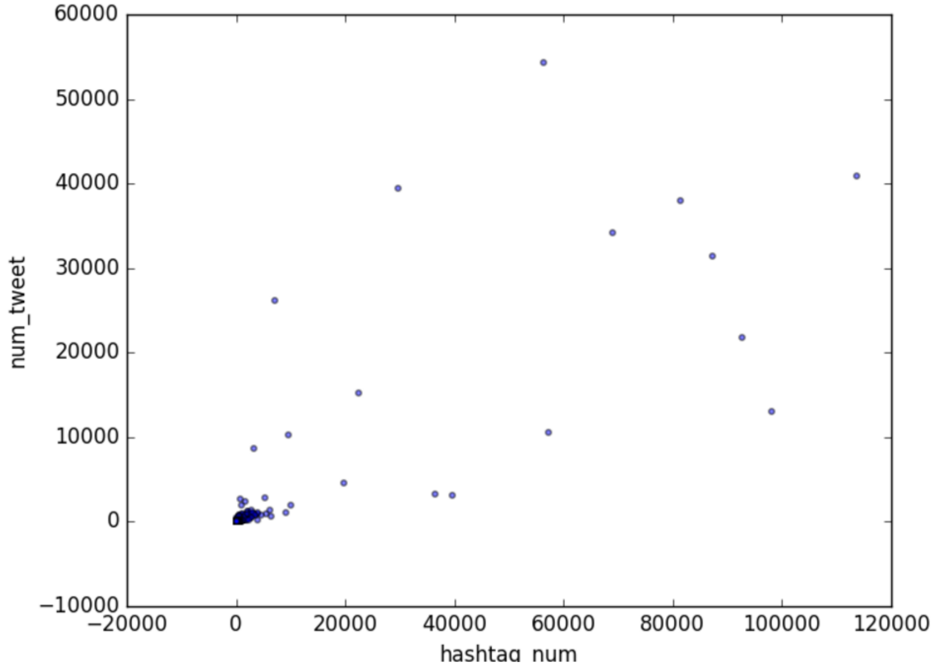


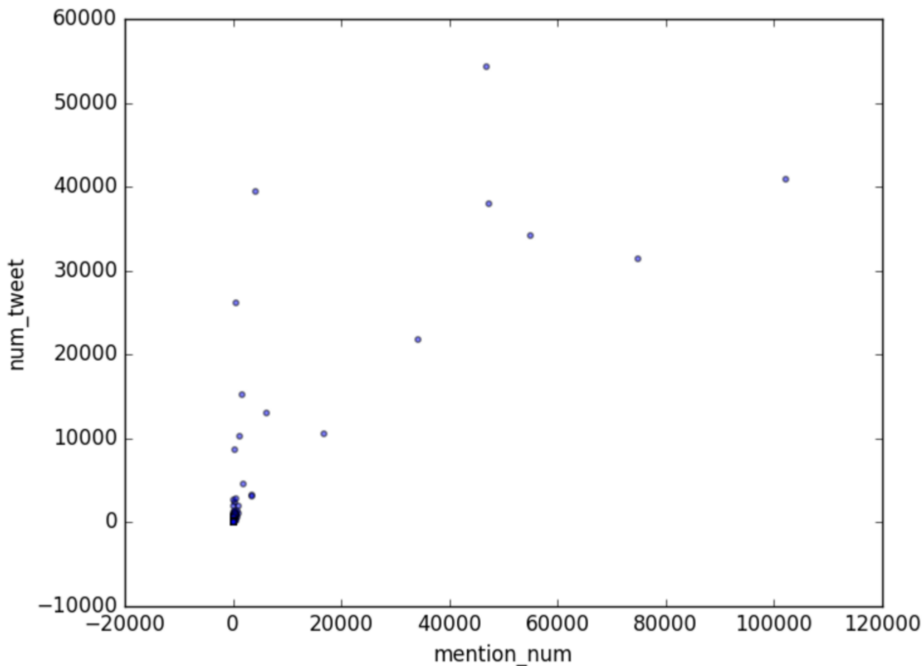
#patriots
Accuracy:
Mean accuracy score is 0.950926015404

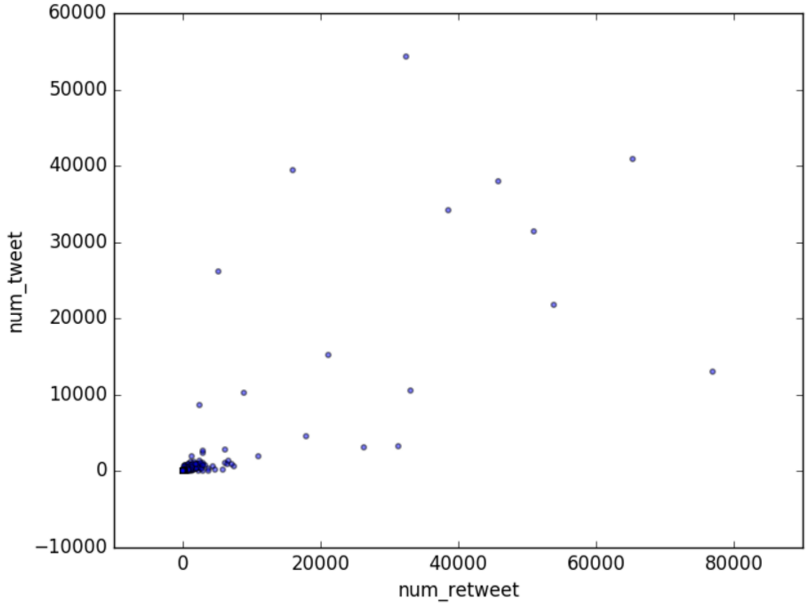
Feature importance:

x1	x2	x3	x4	x5	x6	x7	x8	x9
3.83e-2	6.58e-1	9.59e-2	5.8e-12	4.94e-2	7.89e-2	4.7e-2	2.09e-2	1.15e-2

Hashtag co-occurrence, number of retweets, mentioned number, hashtag co-occurrence are the top three features. URL ratio is the least importance feature.


60000
50000
40000
30000
20000
10000
0
−10000
num_tweet
−20000
0
20000
40000
60000
80000
100000
120000
hashtag_num


60000
50000
40000
30000
20000
10000
0
−10000
num_tweet
−20000
0
20000
40000
60000
80000
100000
120000
mention_num

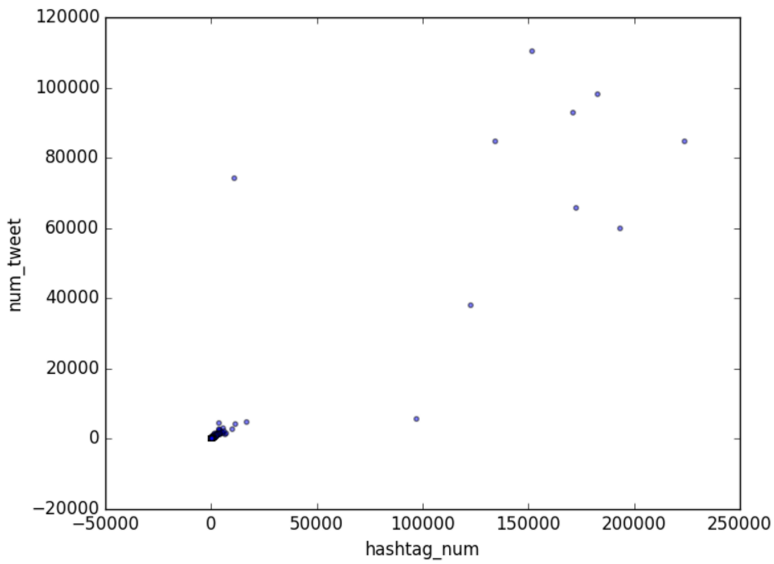


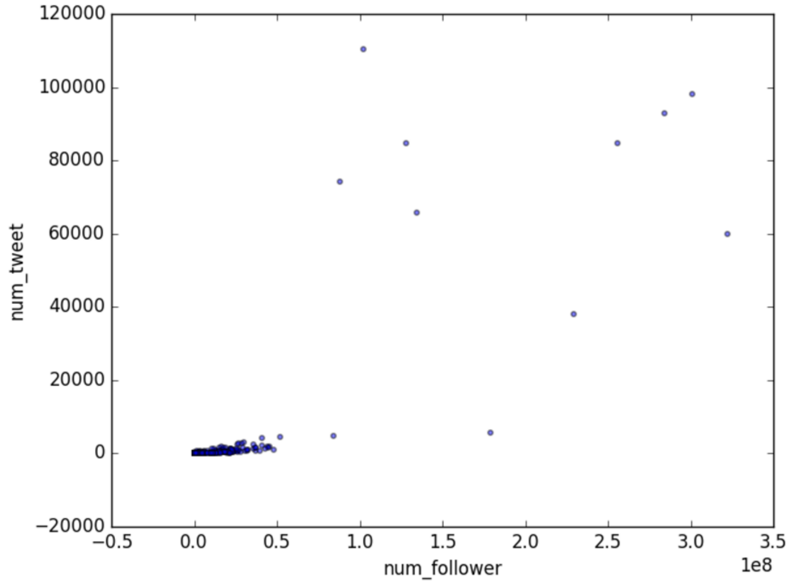
#sb49
Accuracy:
Mean accuracy score is 0.964792397543

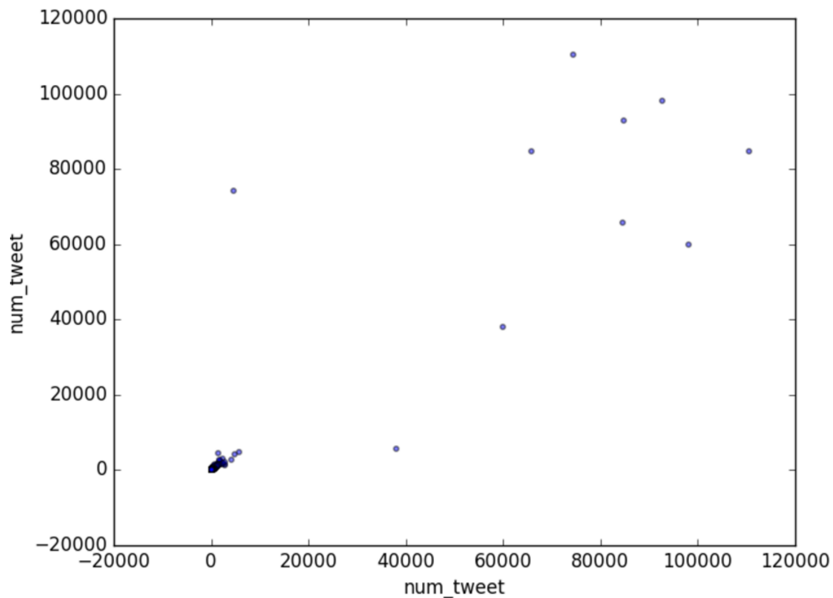
Feature importance:

x1	x2	x3	x4	x5	x6	x7	x8	x9
5.02e-2	7.81e-2	1.45e-1	3.6e-11	2.31e-1	1.2e-1	4.41e-1	1.54e-2	1.94e-2

Hashtag co-occurrence, number of follower, number of tweet, hashtag co-occurrence are the top three features. URL ratio is the least importance feature.





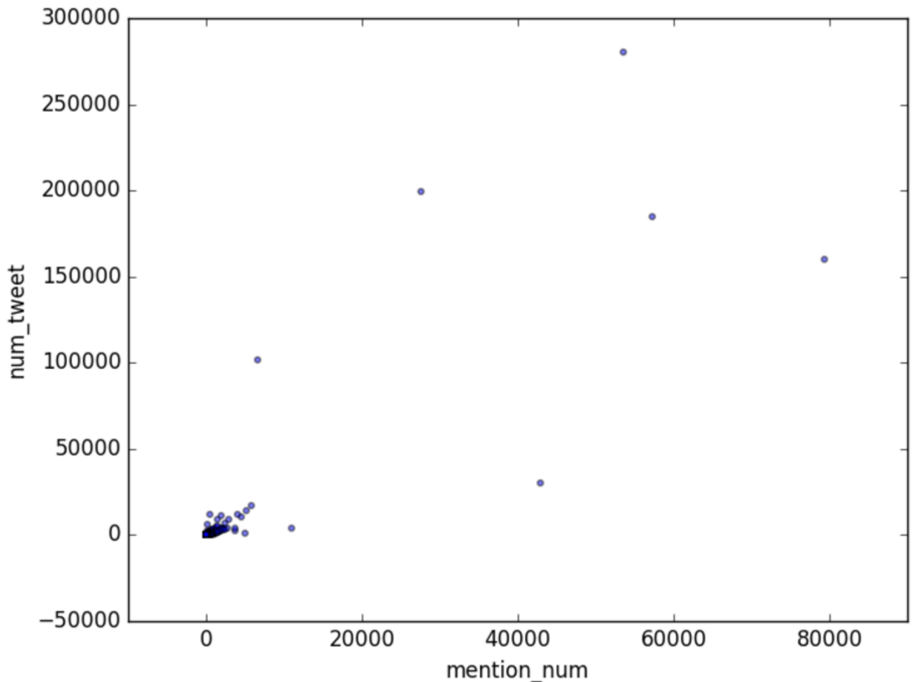


#superbowl
Accuracy:
Mean accuracy score is 0.965857619053

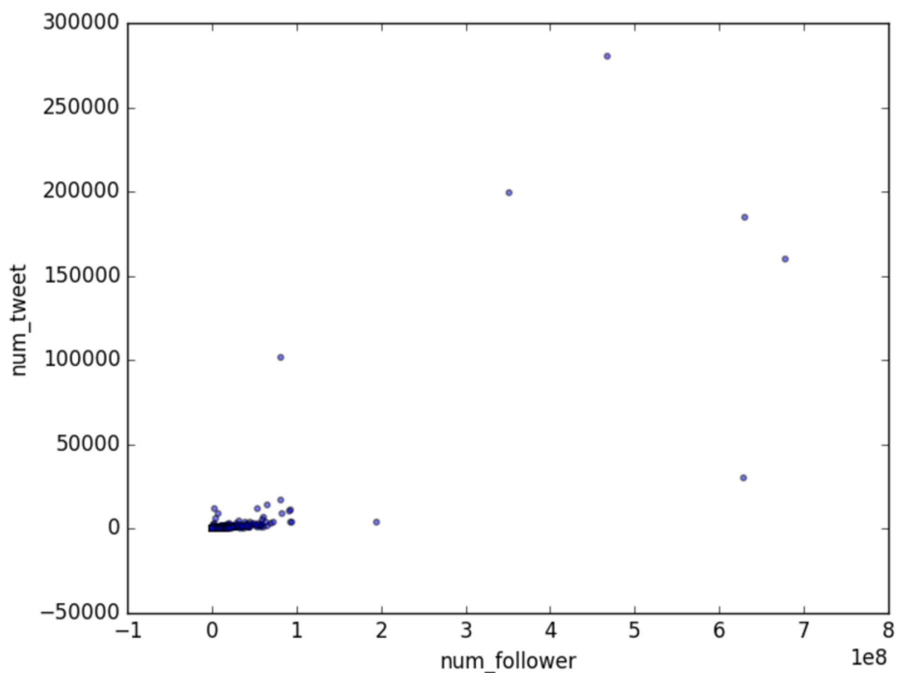
Feature importance:

x1	x2	x3	x4	x5	x6	x7	x8	x9
0.103	0.162	0.113	0	0.214	0.103	0.143	0.088	0.073

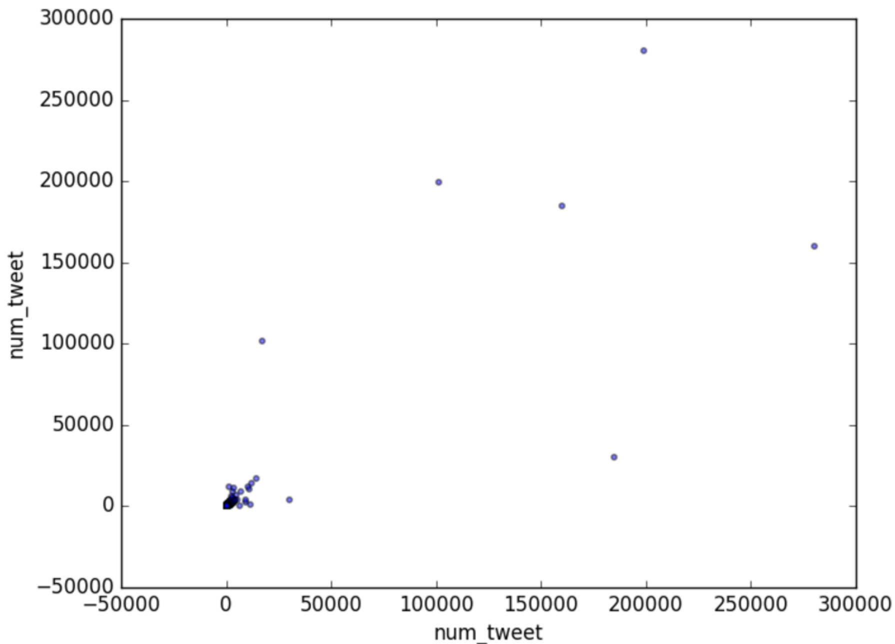
Mention number, number of follower, number of tweet, hashtag co-occurrence are the top three features. URL ratio is the least importance feature.



300000
250000
200000
150000
100000
50000
0
−50000
num_tweet
0
20000
40000
60000
80000
mention_num



300000
250000
200000
150000
100000
50000
0
−50000
num_tweet
−1
0
1
2
3
4
5
6
7
8
1e8
num_follower



4)

10 part cross validation:
Cross_validation in sklean is used, for example:
X_train, X_test, y_train, y_test = cross_validation.train_test_split(X, Y, test_size=0.1)

To avoid overfitting, URL ratio feature is removed since it is not related from previous problem.

Cross-validation error:

#gohawks	93.4550819439
#gopatriots	19.1288128799
#nfl	92.4687694997
#patriots	244.781287531
#sb49	490.957822444
#superbowl	836.040754656

The error is about half of the average value.

Cross-validation error in different time range:

Before Feb. 1, 8:00 a.m.

#gohawks	58.8963910282
#gopatriots	11.0605621615
#nfl	60.2011197581
#patriots	86.2837751873
#sb49	42.7552721268
#superbowl	173.75410161

Between Feb. 1, 8:00 a.m. and 8:00 p.m.

#gohawks	193.36

#gopatriots	1430.1275
#nfl	1155.68875
#patriots	11561.355
#sb49	18296.5725
#superbowl	35740.46375

After Feb. 1, 8:00 p.m.

#gohawks	15.8492142378
#gopatriots	1.53445339069
#nfl	134.013681186
#patriots	46.0447061913
#sb49	107.301734095
#superbowl	198.991091879

5)

For each time period, we have already obtained 6 models in previous problem. In this problem, first we need to decide which model to use.
We tested each sample in corresponding 6 models of its period and calculate the mean accuracy* score. Then we pick the model that produce the largest score as the model to be used to predict the number in 7th hour.

*mean accuracy = 1-u/v, where u = regression sum of squares: ((y_true - y_pred) ** 2).sum() and v = residual sum of squares: ((y_true - y_true.mean()) ** 2).sum(). Best possible value is 1.

	Model to use	Mean accuracy	hour 7 predict
sample1_period1	#sb49	-1.41997031919	157.7960901
sample2_period2	#superbowl	0.385764909431	142108.8
sample3_period3	#superbowl	-1.38192630731	717.45
sample4_period1	#sb49	-3.20574389085	217.06458333
sample5_period1	#gopatriots	-0.827388461522	429.15294118
sample6_period2	# patriots	0.10577786725	29402.625
sample7_period3	#gohawk	0.0378589322475	205.475
sample8_period1	#gopatriots	0.0656369334582	12.97919207
sample9_period2	# patriots	-0.927428431502	19482.724
sample10_period3	#gohawk	-9.03937851124	72.95

Due to the limitation of our random forest model and feature selection, the mean accuracy is relatively low.

6)

Problem: sentiment analysis of fans in both teams
Description: before, during and after the game, we can see the emotion of the fans based on the contents of their tweets. The ideal process is:

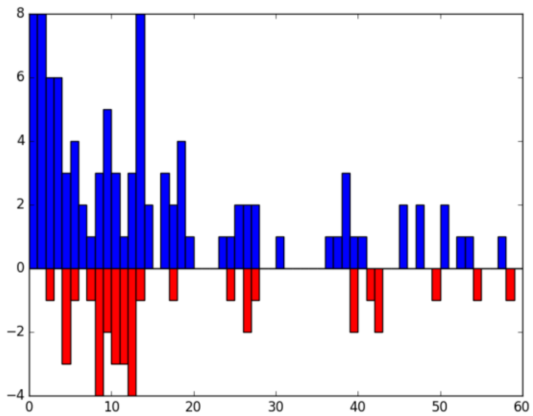
1. Scan all the tweets, find all the tweets with positive and negative emoticon

2. Summarize all the useful words in positive emoticon as positive words collection, and all the useful words in negative emoticon as negative words collection
3. Scan all the tweets, for each tweets, based on how many positive words and negative words, rate the emotion score

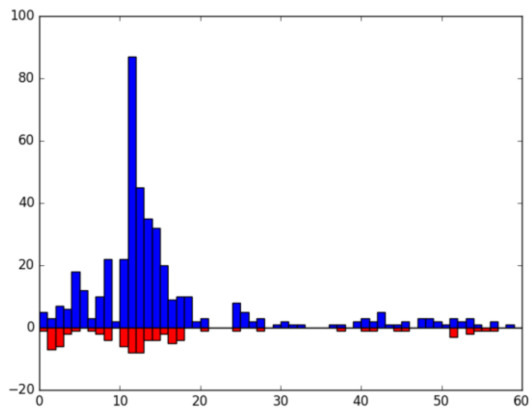
In this project, we simplified this process and only focused on the game night. For the game period, we summarize all the number of positive emoticon and negative emoticon in each ten minutes.

Emotion over the game period (red is negative, blue is positive), starting time is Feb 01, 6pm and ending time is Feb 02, 0 am

Seahawks fan



Patriots fan



As can be see, the seahawks fan were clearly more exited at the beginning of the game, however, since they were behind for most of the game, so there are more negative emotions. When the game ended at around 20, the seahawks fan were clearly more disappointed that patriots fans,