EE219 Project 5 Report

Xiao Shi (404253039)

Weikun Han (804774358)

INTRODUCTION

In this project, we will make popularity prediction on Twitter. 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 [1]. Twitter is an online news and social networking service. In Twitter, there are some concepts are defined as follows:

- **Tweet:** Tweets are publicly visible by default, but senders can restrict message delivery to just their followers. Users can tweet via the Twitter website, compatible external applications (such as for smartphones), or by Short Message Service (SMS) available in certain countries [2].
- **Follower:** Users may subscribe to other users' tweets—this is known as "following" and subscribers are known as "followers" [3].
- **Retweet:** 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 [4].
- **Hashtag:** Users can group posts together by topic or type by use of hashtags – words or phrases prefixed with a "#" sign. Similarly, the "@" sign followed by a username is used for mentioning or replying to other users [5].

The goal of this project is using sources from Twitter to make predictions for the popularity of certain type hashtag. For provided datasets, it contain the trends for the tweets for difference hashtags over the time that 2015 Super Bowl spanning a period starting from 2 weeks before the game to a week after the game. And we need to predict the amount of popularity for each hashtag will post in the future.

1 DATA PRECESSING

In this problem, we want understand the tweet data, and we calculate some statistics for each hashtag: average number of tweets per hour, average number of followers of users posting the tweets, and average number of retweets.

First, those data are:

- **tweets_#gohawks.txt**
- **tweets_#gopatriots.txt**
- **tweets_#nfl.txt**
- **tweets_#patriots.txt**
- **tweets_#sb49.txt**
- **tweets_#superbowl.txt**

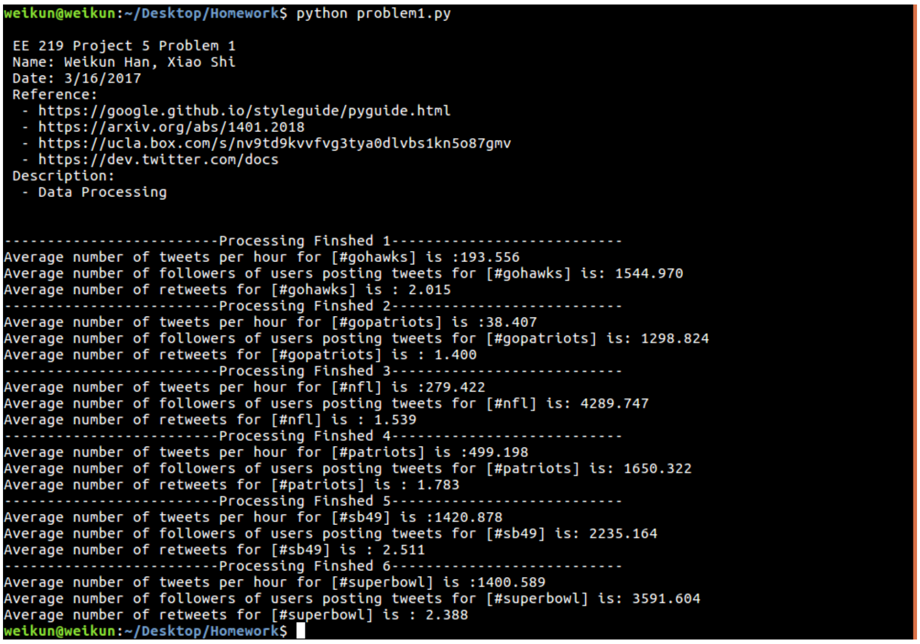
Next, we use Twitter Developer Documentation python API is:

- **retweet = tweet[“metrics”][“citations”][“total”]**
- **user_id = tweet[“tweet”][“user”][“id”]**
- **follower = tweet[“authors”][“followers”]**
- **date = tweet[“firstpost_date”]**

Therefore, we can use formula to calculate question as follows:

- **Average number of tweets per hour =** $\frac{\boldsymbol{Total\ number\ of\ tweets}}{\boldsymbol{Total\ number\ of\ hours}}$
- **Average number of followers of users =** $\frac{\boldsymbol{Total\ number\ of\ followers}}{\boldsymbol{Total\ number\ of\ \ users}}$
- **Average number of retweets =** $\frac{\boldsymbol{Total\ number\ of\ retweets}}{\boldsymbol{Total\ number\ of\ tweets}}$

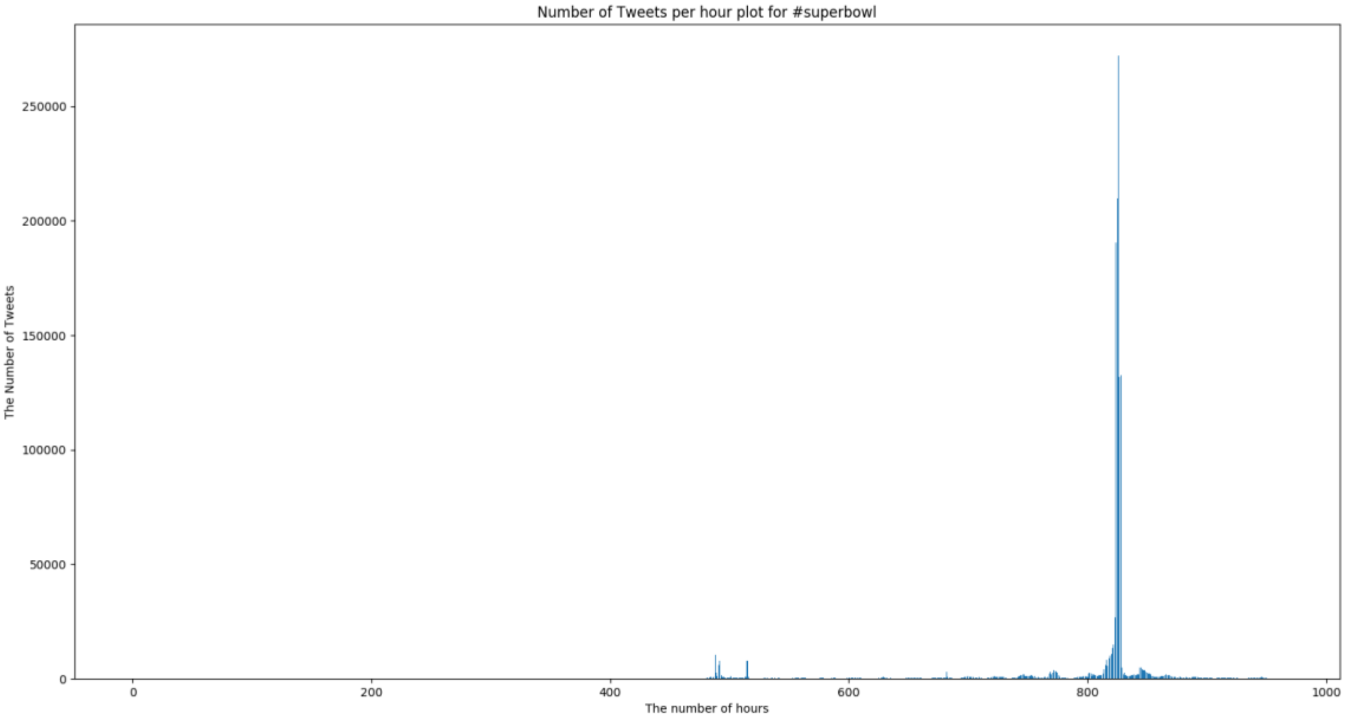
Therefore, we can get the result as follows:



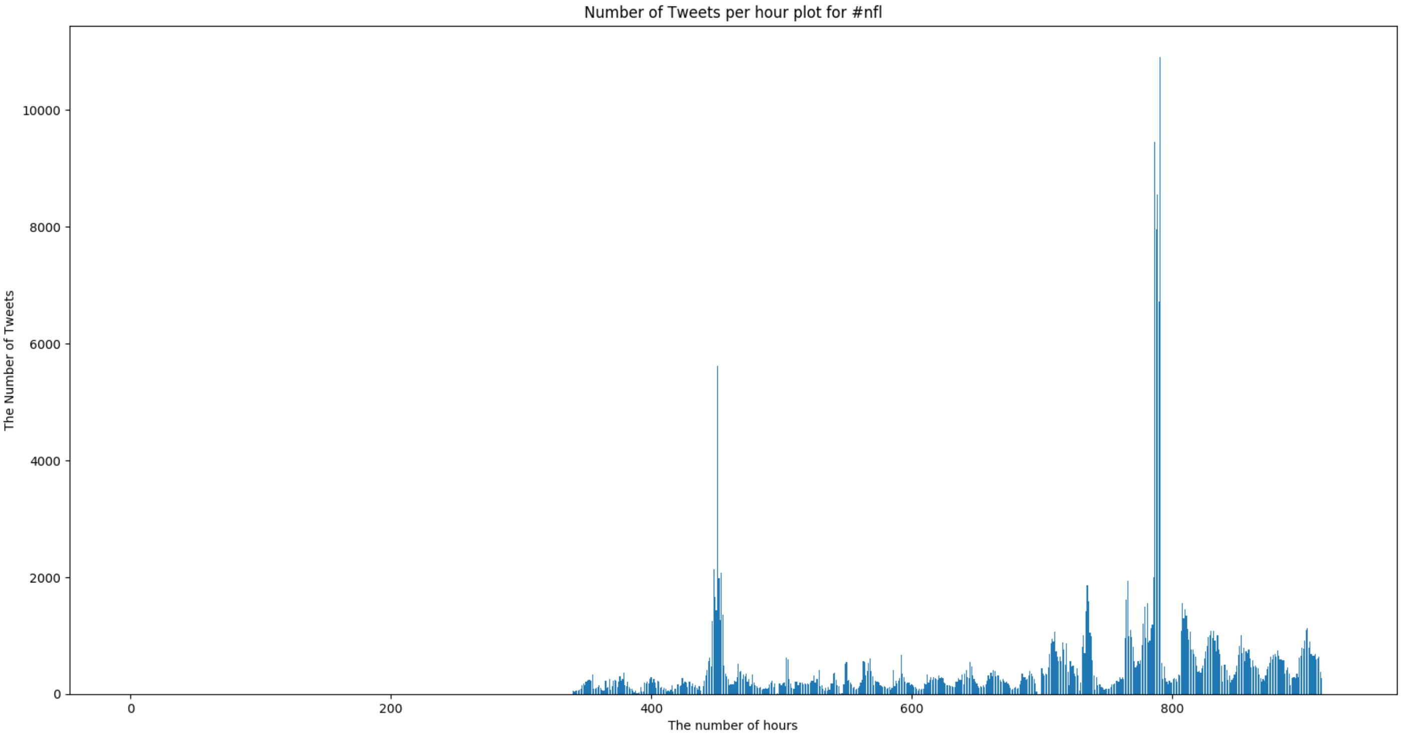
```
weikun@weikun:~/Desktop/Homework$ python problem1.py

 EE 219 Project 5 Problem 1
 Name: Weikun Han, Xiao Shi
 Date: 3/16/2017
 Reference:
  - https://google.github.io/styleguide/pyguide.html
  - https://arxiv.org/abs/1401.2018
  - https://ucla.box.com/s/nv9td9kvvfvg3tya0dlvbs1kn5o87gmv
  - https://dev.twitter.com/docs
 Description:
  - Data Processing


--------------------------Processing Finshed 1---------------------------
Average number of tweets per hour for [#gohawks] is :193.556
Average number of followers of users posting tweets for [#gohawks] is: 1544.970
Average number of retweets for [#gohawks] is : 2.015
--------------------------Processing Finshed 2---------------------------
Average number of tweets per hour for [#gopatriots] is :38.407
Average number of followers of users posting tweets for [#gopatriots] is: 1298.824
Average number of retweets for [#gopatriots] is : 1.400
--------------------------Processing Finshed 3---------------------------
Average number of tweets per hour for [#nfl] is :279.422
Average number of followers of users posting tweets for [#nfl] is: 4289.747
Average number of retweets for [#nfl] is : 1.539
--------------------------Processing Finshed 4---------------------------
Average number of tweets per hour for [#patriots] is :499.198
Average number of followers of users posting tweets for [#patriots] is: 1650.322
Average number of retweets for [#patriots] is : 1.783
--------------------------Processing Finshed 5---------------------------
Average number of tweets per hour for [#sb49] is :1420.878
Average number of followers of users posting tweets for [#sb49] is: 2235.164
Average number of retweets for [#sb49] is : 2.511
--------------------------Processing Finshed 6---------------------------
Average number of tweets per hour for [#superbowl] is :1400.589
Average number of followers of users posting tweets for [#superbowl] is: 3591.604
Average number of retweets for [#superbowl] is : 2.388
weikun@weikun:~/Desktop/Homework$ 
```

Hashtag	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
Average number of tweets per hour	193.556	38.407	279.422	499.197	1420.878	1400.589
Average number of followers of users	1544.969	1298.824	4289.746	1650.321	2235.164	3591.604
Average number of retweets	2.014	1.400	1.538	1.782	2.511	2.388

Therefore, we can get plot "number of tweets in hour" over time for **#SuperBowl** as follows:



Therefore, we can get plot "number of tweets in hour" over time for **#NFL** as follows:



we can get conclustion that Superbowl have the number of tweets had been posted massively at around the 820th hour. On the other hand, NFL have number of tweets had been posted massively at around the 790th hour.

2 LINEAR REGRESSION MODEL WITH 5 FEATRUES

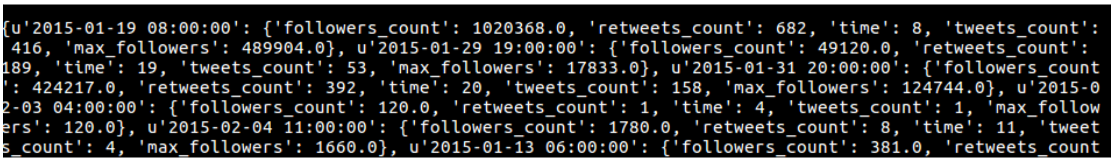
In this part, our task is to make use of 5 features. The features is as follows:

- **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**
- **Time of the day (which could take 24 values that represent hours of the day with respect to a given time reference)**

For the dataset provide, we need to generate the training set. And we first extract each of the above features from the tweet data in only one hour. For every one hour, we first set features in this hour to a default value of 0, and we increment it appropriately each time it appears again in the dataset. Finally, we use training set to get the testing set (in n hour) and predicting set (in n + 1 hour).

We fit our data in to the linear regression model using ordinary least squares (OLS) by StatsModels python API.

By checking training set, we can get our 5 features as follows:

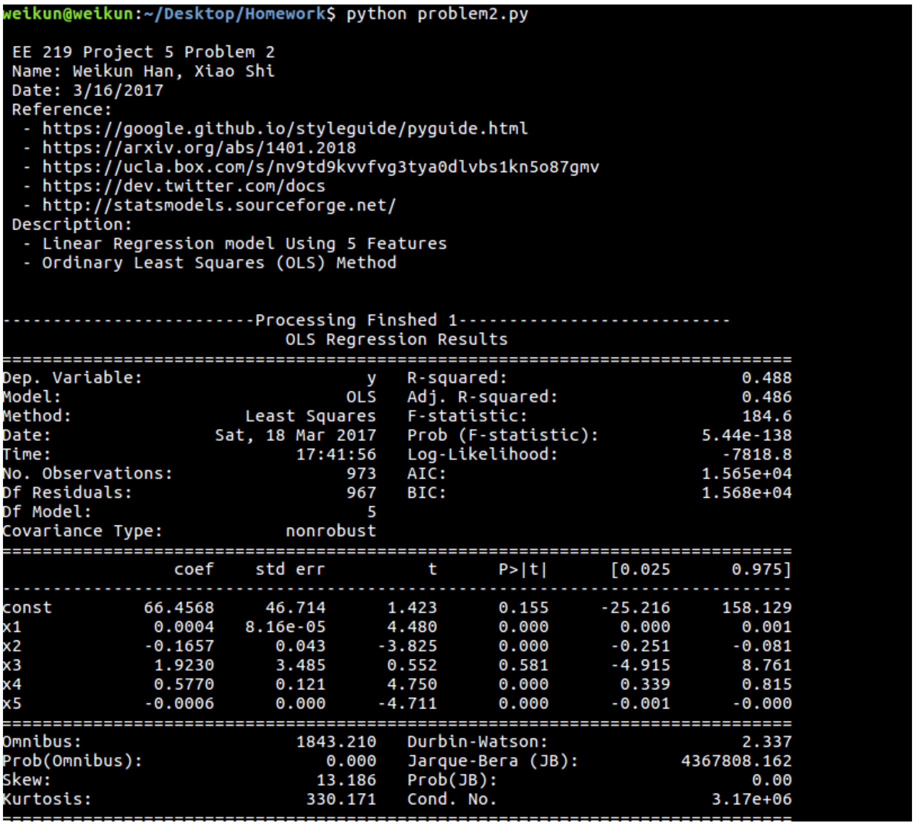


```
{u'2015-01-19 08:00:00': {'followers_count': 1020368.0, 'retweets_count': 682, 'time': 8, 'tweets_count':
 416, 'max_followers': 489904.0}, u'2015-01-29 19:00:00': {'followers_count': 49120.0, 'retweets_count':
189, 'time': 19, 'tweets_count': 53, 'max_followers': 17833.0}, u'2015-01-31 20:00:00': {'followers_count
': 424217.0, 'retweets_count': 392, 'time': 20, 'tweets_count': 158, 'max_followers': 124744.0}, u'2015-0
2-03 04:00:00': {'followers_count': 120.0, 'retweets_count': 1, 'time': 4, 'tweets_count': 1, 'max_follow
ers': 120.0}, u'2015-02-04 11:00:00': {'followers_count': 1780.0, 'retweets_count': 8, 'time': 11, 'tweet
s_count': 4, 'max_followers': 1660.0}, u'2015-01-13 06:00:00': {'followers_count': 381.0, 'retweets_count
```

Note that is result have variables as follows:

- **X1: Sum of the number of followers of the users posting the hashtag**
- **X2: Total number of retweets**
- **X3: Time of the day**
- **X4: Number of tweets**
- **X5: Maximum number of followers of the users posting the hashtag**

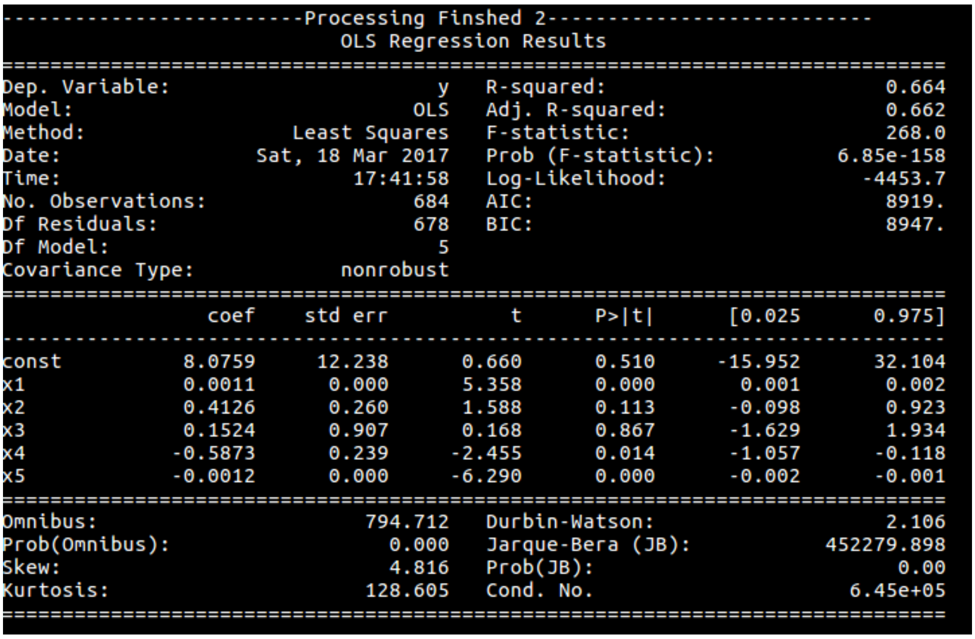
Therefore, we get the result **for #gohawks** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.



```
weikun@weikun:~/Desktop/Homework$ python problem2.py

 EE 219 Project 5 Problem 2
 Name: Weikun Han, Xiao Shi
 Date: 3/16/2017
 Reference:
  - https://google.github.io/styleguide/pyguide.html
  - https://arxiv.org/abs/1401.2018
  - https://ucla.box.com/s/nv9td9kvvfvg3tya0dlvbs1kn5o87gmv
  - https://dev.twitter.com/docs
  - http://statsmodels.sourceforge.net/
 Description:
  - Linear Regression model Using 5 Features
  - Ordinary Least Squares (OLS) Method


-------------------------Processing Finshed 1--------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.488
Model:                            OLS   Adj. R-squared:                  0.486
Method:                 Least Squares   F-statistic:                     184.6
Date:                Sat, 18 Mar 2017   Prob (F-statistic):          5.44e-138
Time:                        17:41:56   Log-Likelihood:                -7818.8
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         66.4568     46.714      1.423      0.155     -25.216     158.129
x1             0.0004   8.16e-05      4.480      0.000       0.000       0.001
x2            -0.1657      0.043     -3.825      0.000      -0.251      -0.081
x3             1.9230      3.485      0.552      0.581      -4.915       8.761
x4             0.5770      0.121      4.750      0.000       0.339       0.815
x5            -0.0006      0.000     -4.711      0.000      -0.001      -0.000
==============================================================================
Omnibus:                     1843.210   Durbin-Watson:                   2.337
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4367808.162
Skew:                          13.186   Prob(JB):                         0.00
Kurtosis:                     330.171   Cond. No.                     3.17e+06
==============================================================================
```

Here, the training accuracy is **R-squared: 0.488**. And the significance of features using the t-test and P-value results is **X1: Sum of the number of followers of the users posting the hashtag, X4: Number of tweets**

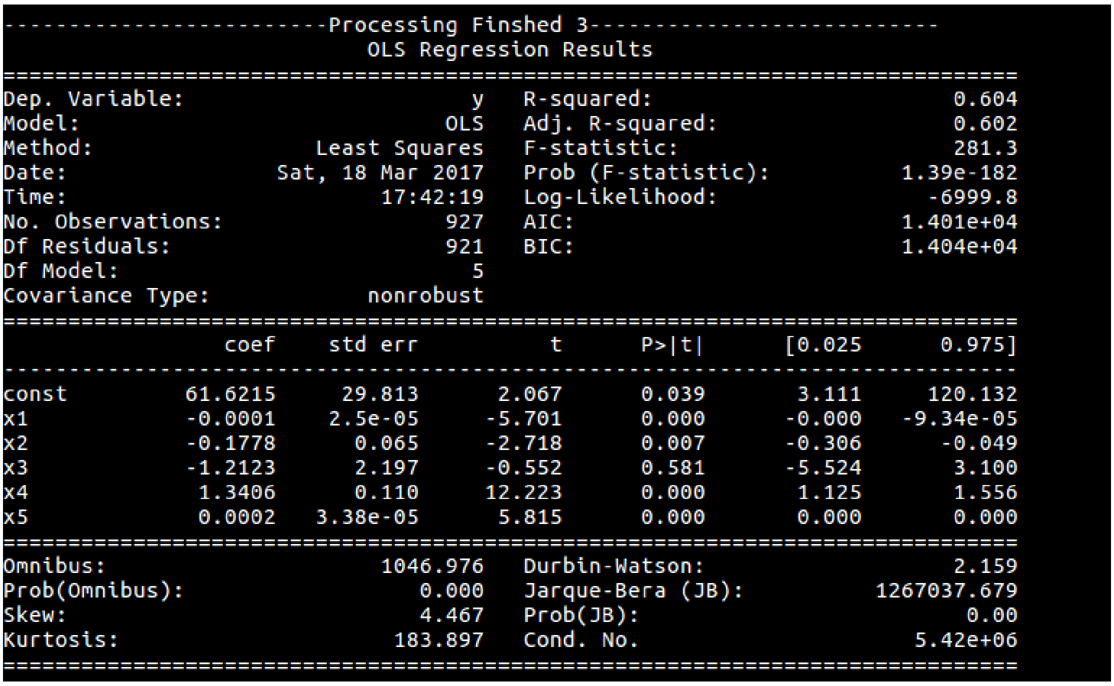
Therefore, we get the result **for #gopatriots** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.



```
-------------------------Processing Finshed 2--------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.664
Model:                            OLS   Adj. R-squared:                  0.662
Method:                 Least Squares   F-statistic:                     268.0
Date:                Sat, 18 Mar 2017   Prob (F-statistic):          6.85e-158
Time:                        17:41:58   Log-Likelihood:                -4453.7
No. Observations:                 684   AIC:                             8919.
Df Residuals:                     678   BIC:                             8947.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.0759     12.238      0.660      0.510     -15.952      32.104
x1             0.0011      0.000      5.358      0.000       0.001       0.002
x2             0.4126      0.260      1.588      0.113      -0.098       0.923
x3             0.1524      0.907      0.168      0.867      -1.629       1.934
x4            -0.5873      0.239     -2.455      0.014      -1.057      -0.118
x5            -0.0012      0.000     -6.290      0.000      -0.002      -0.001
==============================================================================
Omnibus:                      794.712   Durbin-Watson:                   2.106
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           452279.898
Skew:                           4.816   Prob(JB):                         0.00
Kurtosis:                     128.605   Cond. No.                     6.45e+05
==============================================================================
```

Here, the training accuracy is **R-squared: 0.664**. And the significance of features using the t-test and P-value results is **X1: Sum of the number of followers of the users posting the hashtag**
X2: Total number of retweets

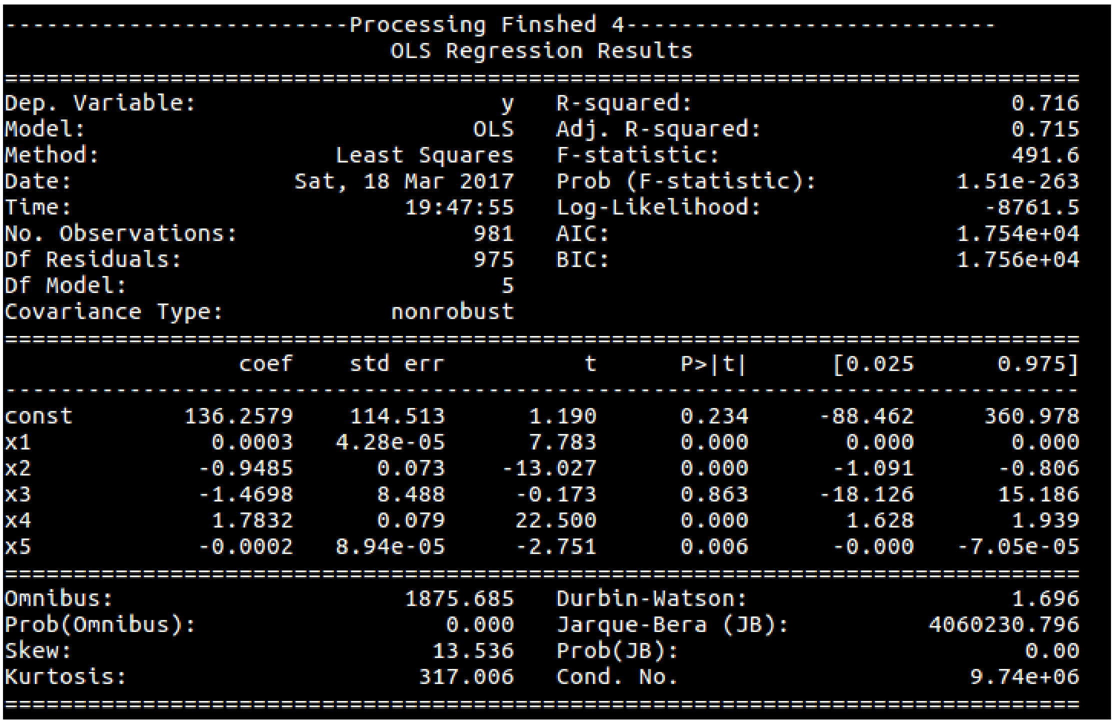
Therefore, we get the result **for # #nfl** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.



```
-------------------------Processing Finshed 3---------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.604
Model:                            OLS   Adj. R-squared:                  0.602
Method:                 Least Squares   F-statistic:                     281.3
Date:                Sat, 18 Mar 2017   Prob (F-statistic):          1.39e-182
Time:                        17:42:19   Log-Likelihood:                -6999.8
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         61.6215     29.813      2.067      0.039       3.111     120.132
x1            -0.0001    2.5e-05     -5.701      0.000      -0.000   -9.34e-05
x2            -0.1778      0.065     -2.718      0.007      -0.306      -0.049
x3            -1.2123      2.197     -0.552      0.581      -5.524       3.100
x4             1.3406      0.110     12.223      0.000       1.125       1.556
x5             0.0002   3.38e-05      5.815      0.000       0.000       0.000
==============================================================================
Omnibus:                     1046.976   Durbin-Watson:                   2.159
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1267037.679
Skew:                           4.467   Prob(JB):                         0.00
Kurtosis:                     183.897   Cond. No.                     5.42e+06
==============================================================================
```

Here, the training accuracy is **R-squared: 0.604**. And the significance of features using the t-test and P-value results is **X4: Number of tweets, X5: Maximum number of followers of the users posting the hashtag**

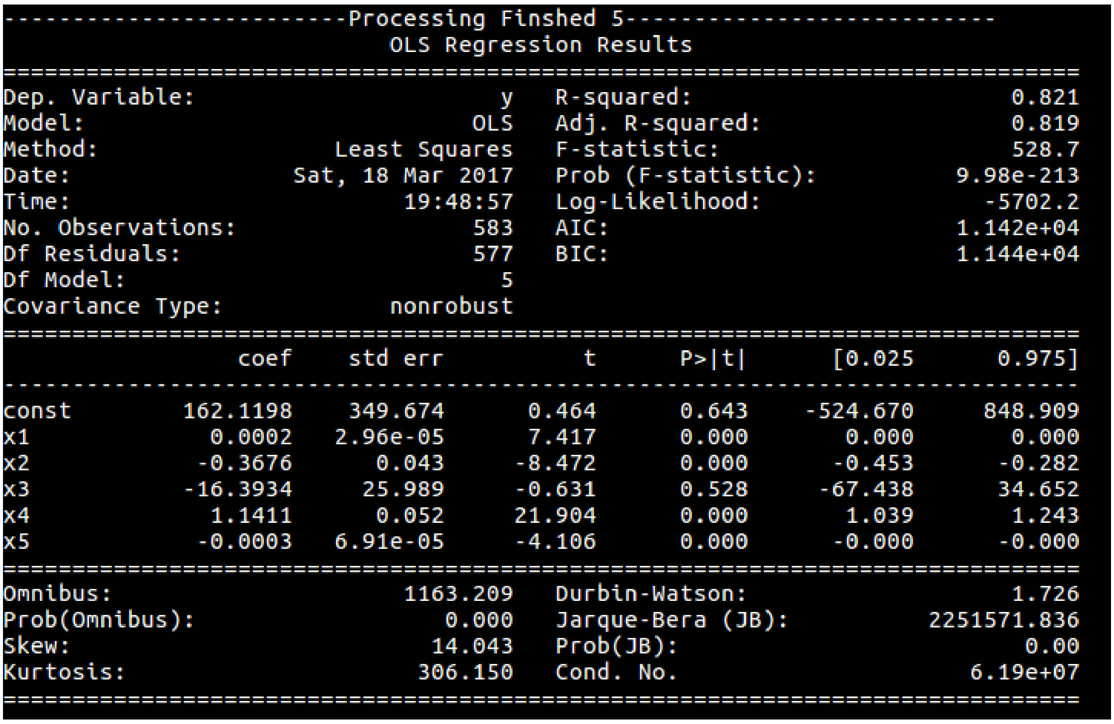
Therefore, we get the result **for # #patriots** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.



```
------------------------Processing Finshed 4--------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.716
Model:                            OLS   Adj. R-squared:                  0.715
Method:                 Least Squares   F-statistic:                     491.6
Date:                Sat, 18 Mar 2017   Prob (F-statistic):          1.51e-263
Time:                        19:47:55   Log-Likelihood:                -8761.5
No. Observations:                 981   AIC:                         1.754e+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        136.2579    114.513      1.190      0.234     -88.462     360.978
x1             0.0003   4.28e-05      7.783      0.000       0.000       0.000
x2            -0.9485      0.073    -13.027      0.000      -1.091      -0.806
x3            -1.4698      8.488     -0.173      0.863     -18.126      15.186
x4             1.7832      0.079     22.500      0.000       1.628       1.939
x5            -0.0002   8.94e-05     -2.751      0.006      -0.000   -7.05e-05
==============================================================================
Omnibus:                     1875.685   Durbin-Watson:                   1.696
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4060230.796
Skew:                          13.536   Prob(JB):                         0.00
Kurtosis:                     317.006   Cond. No.                     9.74e+06
==============================================================================
```

Here, the training accuracy is **R-squared: 0.716**. And the significance of features using the t-test and P-value results is **X1: Sum of the number of followers of the users posting the hashtag, X4: Number of tweets**

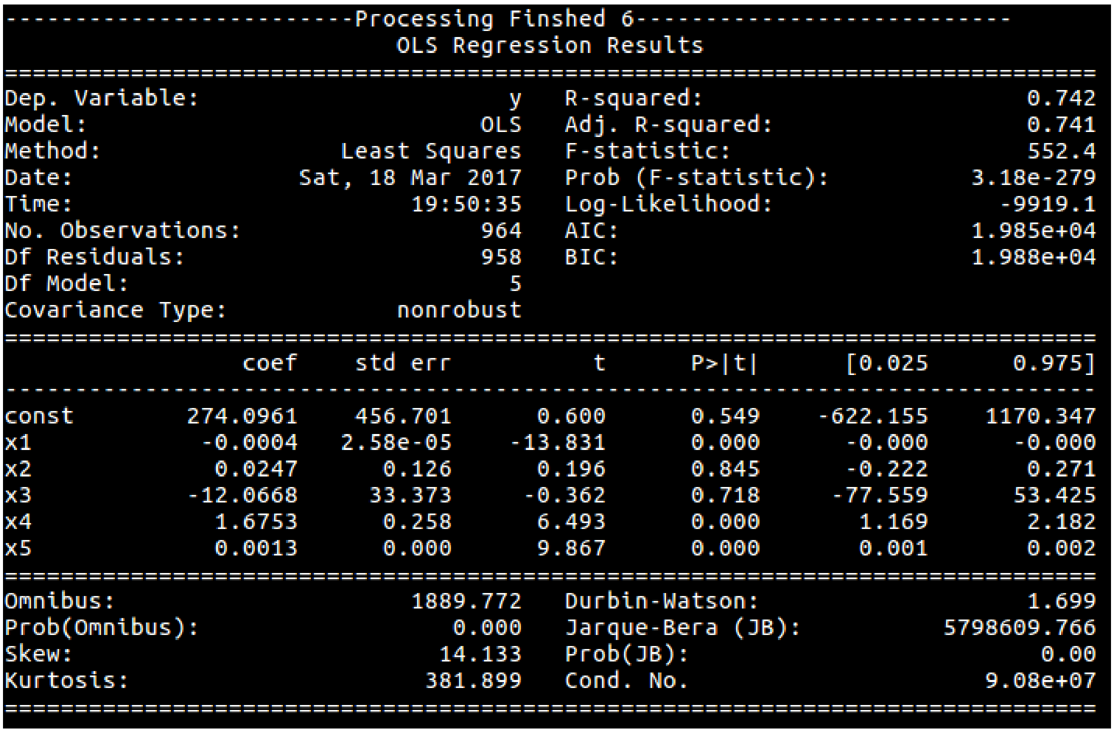
Therefore, we get the result **for #sb49** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.



```
------------------------Processing Finshed 5--------------------------
                            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:                Sat, 18 Mar 2017   Prob (F-statistic):          9.98e-213
Time:                        19:48:57   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        162.1198    349.674      0.464      0.643    -524.670     848.909
x1             0.0002   2.96e-05      7.417      0.000       0.000       0.000
x2            -0.3676      0.043     -8.472      0.000      -0.453      -0.282
x3           -16.3934     25.989     -0.631      0.528     -67.438      34.652
x4             1.1411      0.052     21.904      0.000       1.039       1.243
x5            -0.0003   6.91e-05     -4.106      0.000      -0.000      -0.000
==============================================================================
Omnibus:                     1163.209   Durbin-Watson:                   1.726
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2251571.836
Skew:                          14.043   Prob(JB):                         0.00
Kurtosis:                     306.150   Cond. No.                     6.19e+07
==============================================================================
```

Here, the training accuracy is **R-squared: 0.821**. And the significance of features using the t-test and P-value results is **X1: Sum of the number of followers of the users posting the hashtag, X4: Number of tweets**

Therefore, we get the result **for #superbowl** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.



```
------------------------Processing Finshed 6--------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.742
Model:                            OLS   Adj. R-squared:                  0.741
Method:                 Least Squares   F-statistic:                     552.4
Date:                Sat, 18 Mar 2017   Prob (F-statistic):          3.18e-279
Time:                        19:50:35   Log-Likelihood:                -9919.1
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        274.0961    456.701      0.600      0.549    -622.155    1170.347
x1            -0.0004   2.58e-05    -13.831      0.000      -0.000      -0.000
x2             0.0247      0.126      0.196      0.845      -0.222       0.271
x3           -12.0668     33.373     -0.362      0.718     -77.559      53.425
x4             1.6753      0.258      6.493      0.000       1.169       2.182
x5             0.0013      0.000      9.867      0.000       0.001       0.002
==============================================================================
Omnibus:                     1889.772   Durbin-Watson:                   1.699
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5798609.766
Skew:                          14.133   Prob(JB):                         0.00
Kurtosis:                     381.899   Cond. No.                     9.08e+07
==============================================================================
```

Here, the training accuracy is **R-squared: 0.742**. And the significance of features using the t-test and P-value results is **X4: Number of tweets, X5: Maximum number of followers of the users posting the hashtag**

In the end, all result we get is reasonable, and we know how to use t-test and P-value to determine which features are more important for our model. Also from different model, the dominant features are different.

3 LINEAR REGRESSION MODEL WITH 10 FEATRUES

In this part, we have already used 5 features. The features is as follows:

- **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**
- **Time of the day (which could take 24 values that represent hours of the day with respect to a given time reference)**

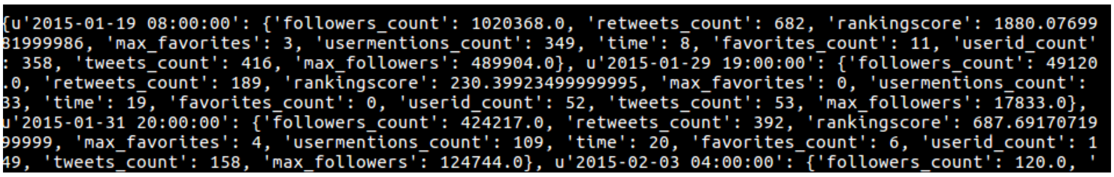
Next, we need define more features to make our model predict more accurcy. The features is as follows:

- **Sum of the number of favorites of the users posting the hashtag**
- **Maximum number of favorites of the users posting the hashtag**
- **Number of user ID**
- **Total number of mentioning others**
- **Sum of the number of ranking scores of the users posting the hashtag**

Here, we want to explain why we define those 5 featues. First, favorites indicate user like is post tweet which mean people think is post important. Second, maximun favorites can show how much it important in people eyes. Next, number of user ID can show how many people is online to dicuss is hashtag which mean is feature is important too. Next, number of mentioning other can show how many people join to this post tweet which mean is also important. Finally, ranking scores is obviously can show how important it is.

We use same way to get the training set, and we also fit our data in to the linear regression model using ordinary least squares (OLS) by StatsModels python API.

By checking training set, we can get our 10 features as follows:



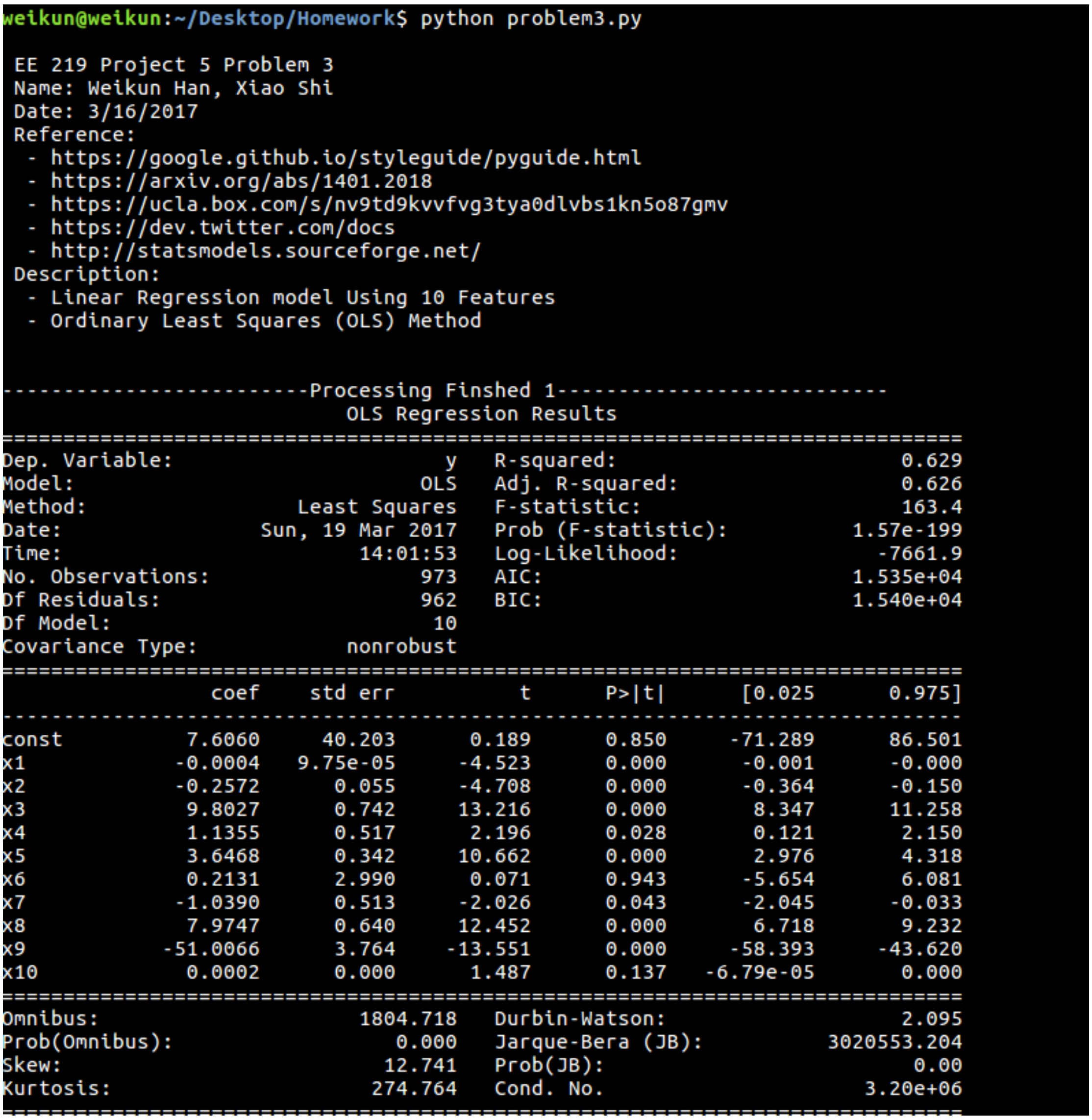
```
{u'2015-01-19 08:00:00': {'followers_count': 1020368.0, 'retweets_count': 682, 'rankingscore': 1880.07699
31999986, 'max_favorites': 3, 'usermentions_count': 349, 'time': 8, 'favorites_count': 11, 'userid_count'
: 358, 'tweets_count': 416, 'max_followers': 489904.0}, u'2015-01-29 19:00:00': {'followers_count': 49120
.0, 'retweets_count': 189, 'rankingscore': 230.39923499999995, 'max_favorites': 0, 'usermentions_count':
33, 'time': 19, 'favorites_count': 0, 'userid_count': 52, 'tweets_count': 53, 'max_followers': 17833.0},
u'2015-01-31 20:00:00': {'followers_count': 424217.0, 'retweets_count': 392, 'rankingscore': 687.69170719
99999, 'max_favorites': 4, 'usermentions_count': 109, 'time': 20, 'favorites_count': 6, 'userid_count': 1
49, 'tweets_count': 158, 'max_followers': 124744.0}, u'2015-02-03 04:00:00': {'followers_count': 120.0, '
```

Note that is result have variables as follows:

- **X1: Sum of the number of followers of the users posting the hashtag**
- **X2: Total number of retweets**
- **X3: Sum of the number of ranking scores of the users posting the hashtag**
- **X4: Maximum number of favorites of the users posting the hashtag**
- **X5: Total number of mentioning others**
- **X6: Time of the day**
- **X7: Sum of the number of favorites of the users posting the hashtag**

- **X8: Number of user ID**
- **X9: Number of tweets**
- **X10: Maximum number of followers of the users posting the hashtag**

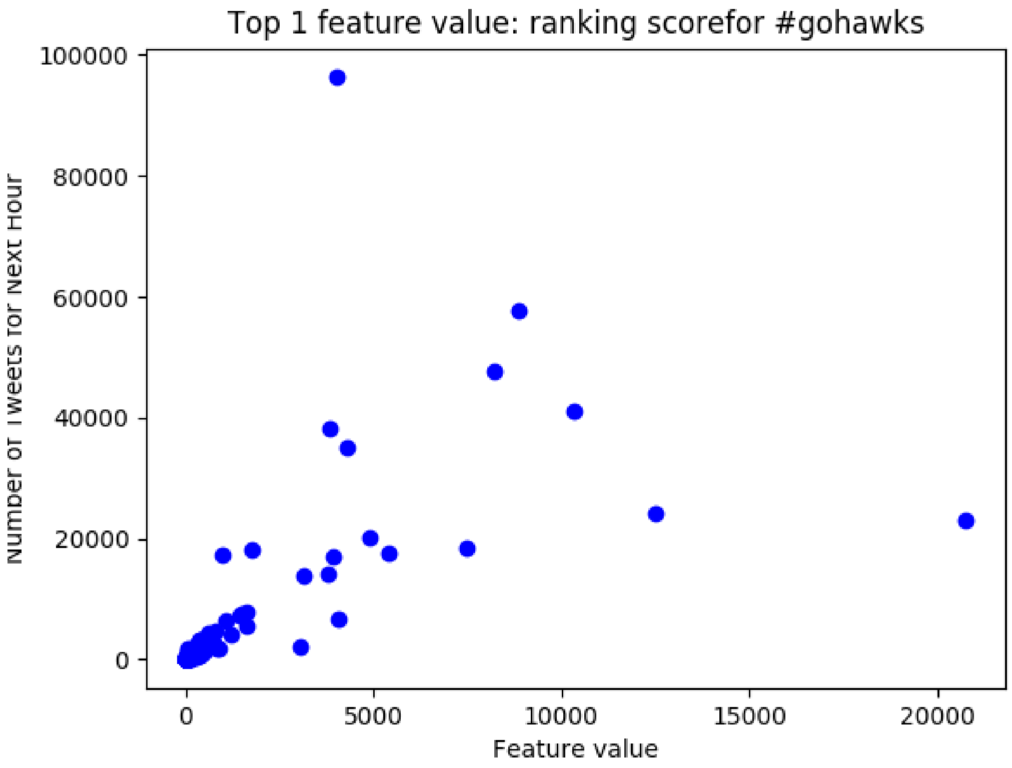
Therefore, we get the result **for #gohawks** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.



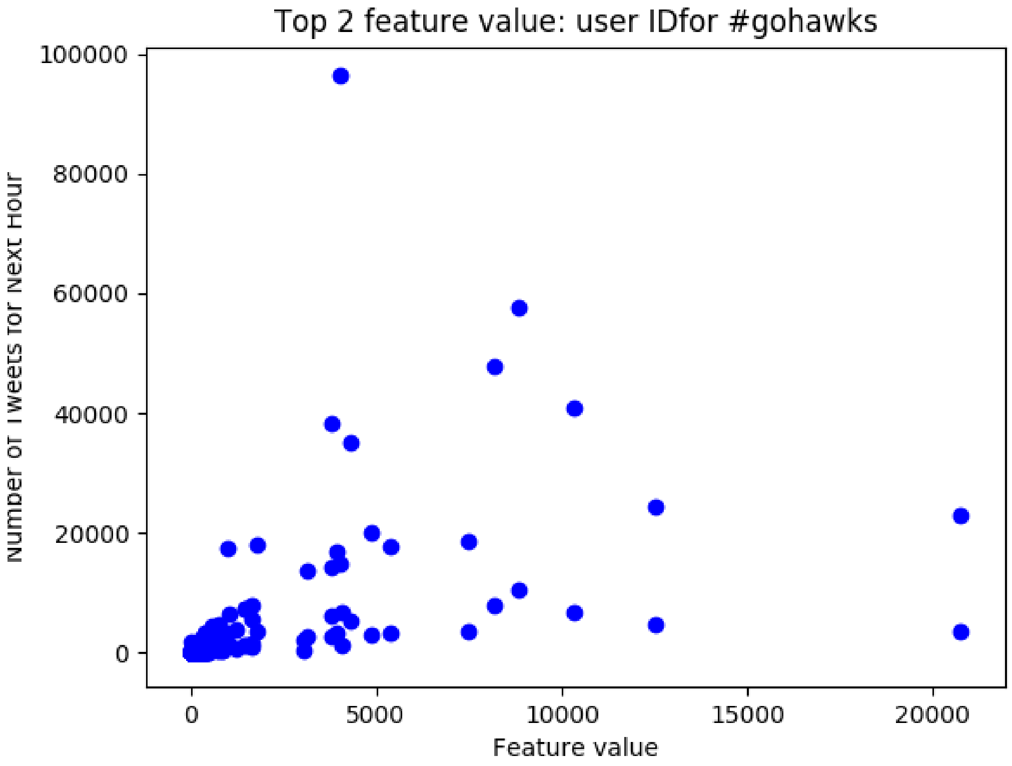
```
weikun@weikun:~/Desktop/Homework$ python problem3.py

 EE 219 Project 5 Problem 3
 Name: Weikun Han, Xiao Shi
 Date: 3/16/2017
 Reference:
  - https://google.github.io/styleguide/pyguide.html
  - https://arxiv.org/abs/1401.2018
  - https://ucla.box.com/s/nv9td9kvvfvg3tya0dlvbs1kn5o87gmv
  - https://dev.twitter.com/docs
  - http://statsmodels.sourceforge.net/
 Description:
  - Linear Regression model Using 10 Features
  - Ordinary Least Squares (OLS) Method


-------------------------Processing Finshed 1--------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.629
Model:                            OLS   Adj. R-squared:                  0.626
Method:                 Least Squares   F-statistic:                     163.4
Date:                Sun, 19 Mar 2017   Prob (F-statistic):          1.57e-199
Time:                        14:01:53   Log-Likelihood:                -7661.9
No. Observations:                 973   AIC:                         1.535e+04
Df Residuals:                     962   BIC:                         1.540e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.6060     40.203      0.189      0.850     -71.289      86.501
x1            -0.0004   9.75e-05     -4.523      0.000      -0.001      -0.000
x2            -0.2572      0.055     -4.708      0.000      -0.364      -0.150
x3             9.8027      0.742     13.216      0.000       8.347      11.258
x4             1.1355      0.517      2.196      0.028       0.121       2.150
x5             3.6468      0.342     10.662      0.000       2.976       4.318
x6             0.2131      2.990      0.071      0.943      -5.654       6.081
x7            -1.0390      0.513     -2.026      0.043      -2.045      -0.033
x8             7.9747      0.640     12.452      0.000       6.718       9.232
x9           -51.0066      3.764    -13.551      0.000     -58.393     -43.620
x10            0.0002      0.000      1.487      0.137   -6.79e-05       0.000
==============================================================================
Omnibus:                     1804.718   Durbin-Watson:                   2.095
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          3020553.204
Skew:                          12.741   Prob(JB):                         0.00
Kurtosis:                     274.764   Cond. No.                     3.20e+06
==============================================================================
```

Here, the training accuracy is **R-squared: 0.629**. And the 3 significance of features using the t-test and P-value results is • **X3: Sum of the number of ranking scores of the users posting the hashtag** • **X8: Number of user ID** • **X5: Total number of mentioning others**

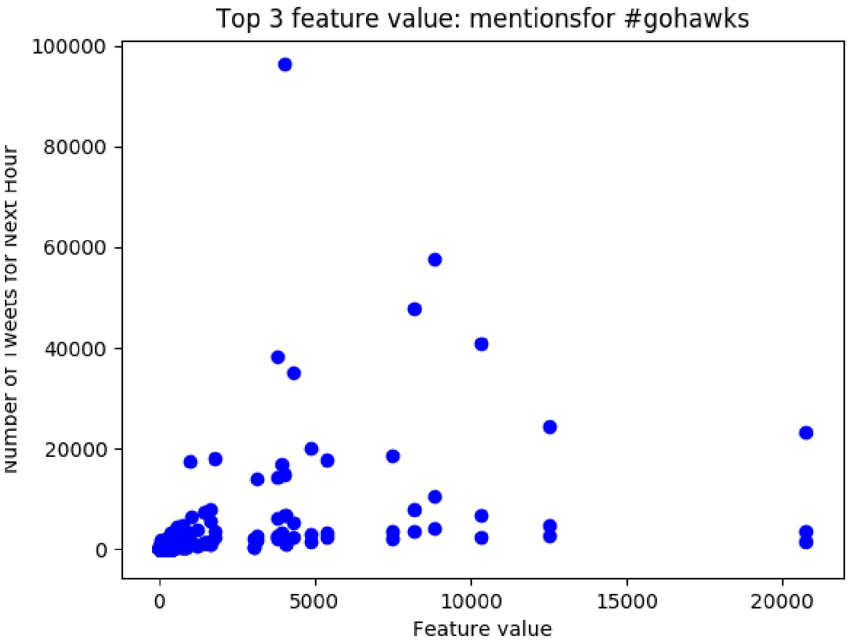
Therefore, we can get plot "prediction (number of tweets for next hour) versus feature value" for **#SuperBowl** as follows:



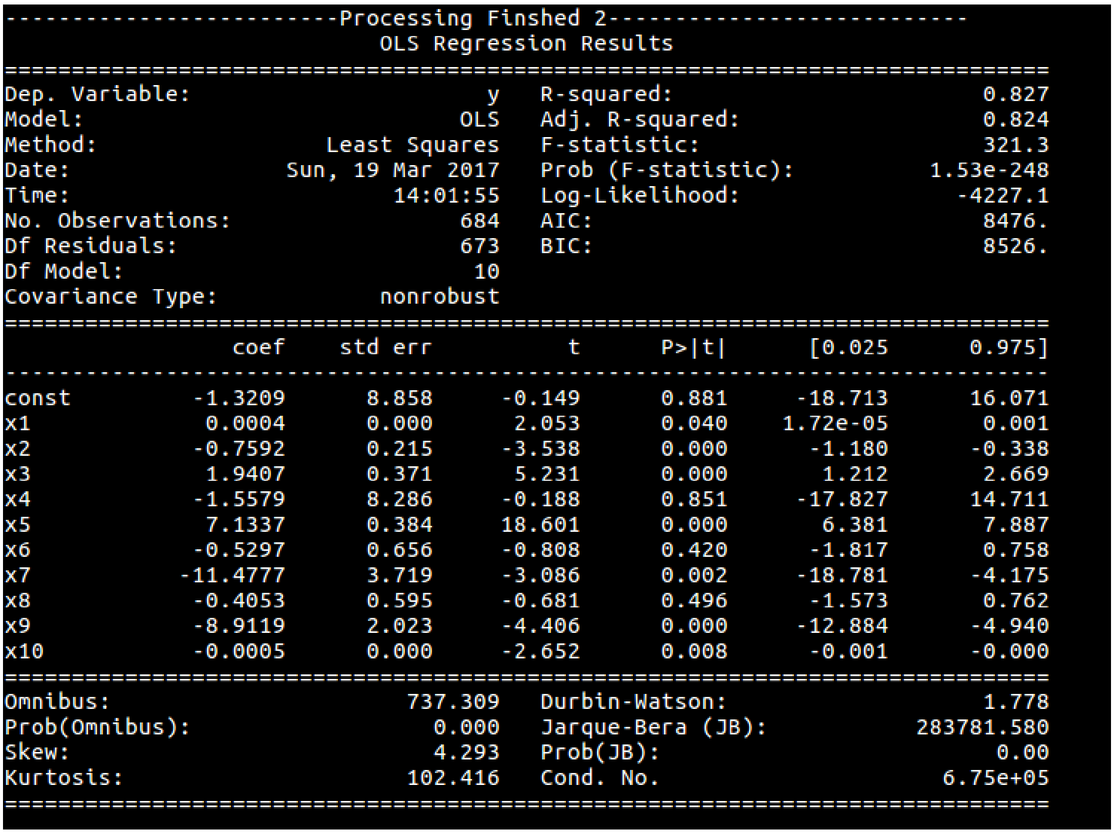
Top 1 feature value: ranking scorefor #gohawks



Top 2 feature value: user IDfor #gohawks



Therefore, we get the result **for** **#gopatriots** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.

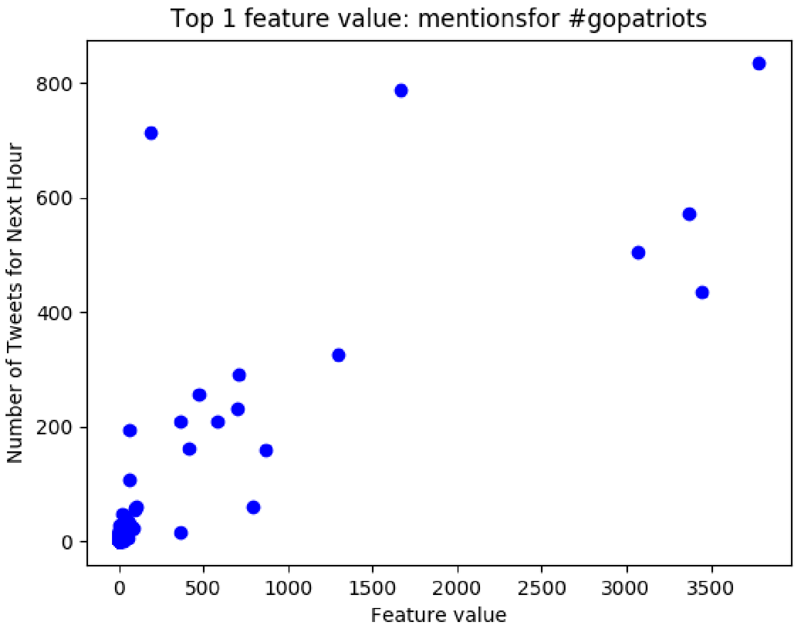


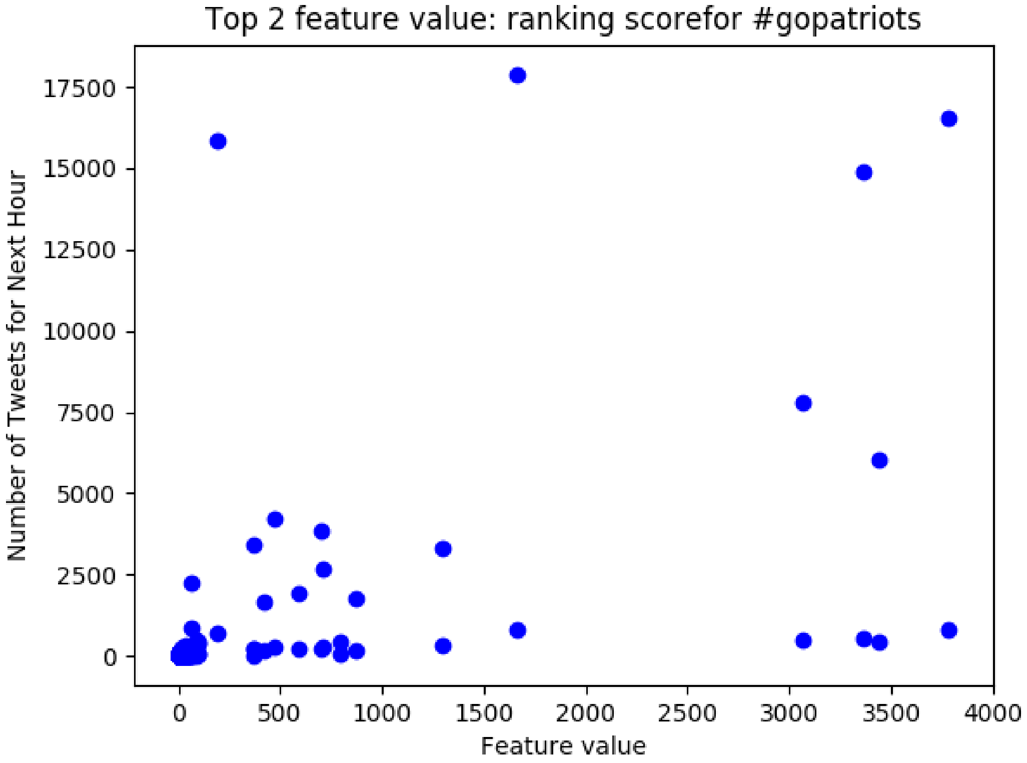
```
-------------------------Processing Finshed 2--------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.827
Model:                            OLS   Adj. R-squared:                  0.824
Method:                 Least Squares   F-statistic:                     321.3
Date:                Sun, 19 Mar 2017   Prob (F-statistic):          1.53e-248
Time:                        14:01:55   Log-Likelihood:                -4227.1
No. Observations:                 684   AIC:                             8476.
Df Residuals:                     673   BIC:                             8526.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.3209      8.858     -0.149      0.881     -18.713      16.071
x1             0.0004      0.000      2.053      0.040    1.72e-05       0.001
x2            -0.7592      0.215     -3.538      0.000      -1.180      -0.338
x3             1.9407      0.371      5.231      0.000       1.212       2.669
x4            -1.5579      8.286     -0.188      0.851     -17.827      14.711
x5             7.1337      0.384     18.601      0.000       6.381       7.887
x6            -0.5297      0.656     -0.808      0.420      -1.817       0.758
x7           -11.4777      3.719     -3.086      0.002     -18.781      -4.175
x8            -0.4053      0.595     -0.681      0.496      -1.573       0.762
x9            -8.9119      2.023     -4.406      0.000     -12.884      -4.940
x10           -0.0005      0.000     -2.652      0.008      -0.001      -0.000
==============================================================================
Omnibus:                      737.309   Durbin-Watson:                   1.778
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           283781.580
Skew:                           4.293   Prob(JB):                         0.00
Kurtosis:                     102.416   Cond. No.                     6.75e+05
==============================================================================
```

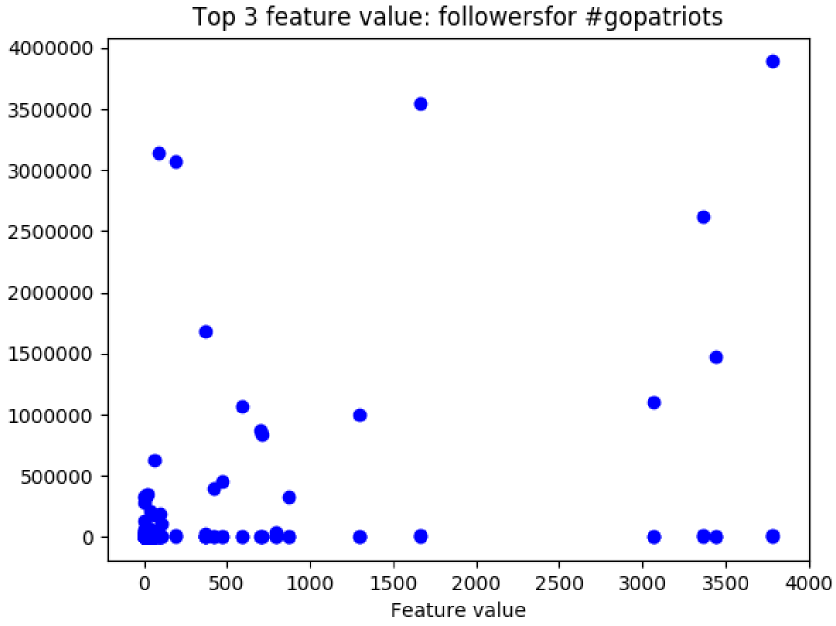
Here, the training accuracy is **R-squared: 0.827**. And the significance of features using the t-test and P-value results is **X5: Total number of mentioning others, X3:**

Sum of the number of ranking scores of the users posting the hashtag, • **X1: Sum of the number of followers of the users posting the hashtag**

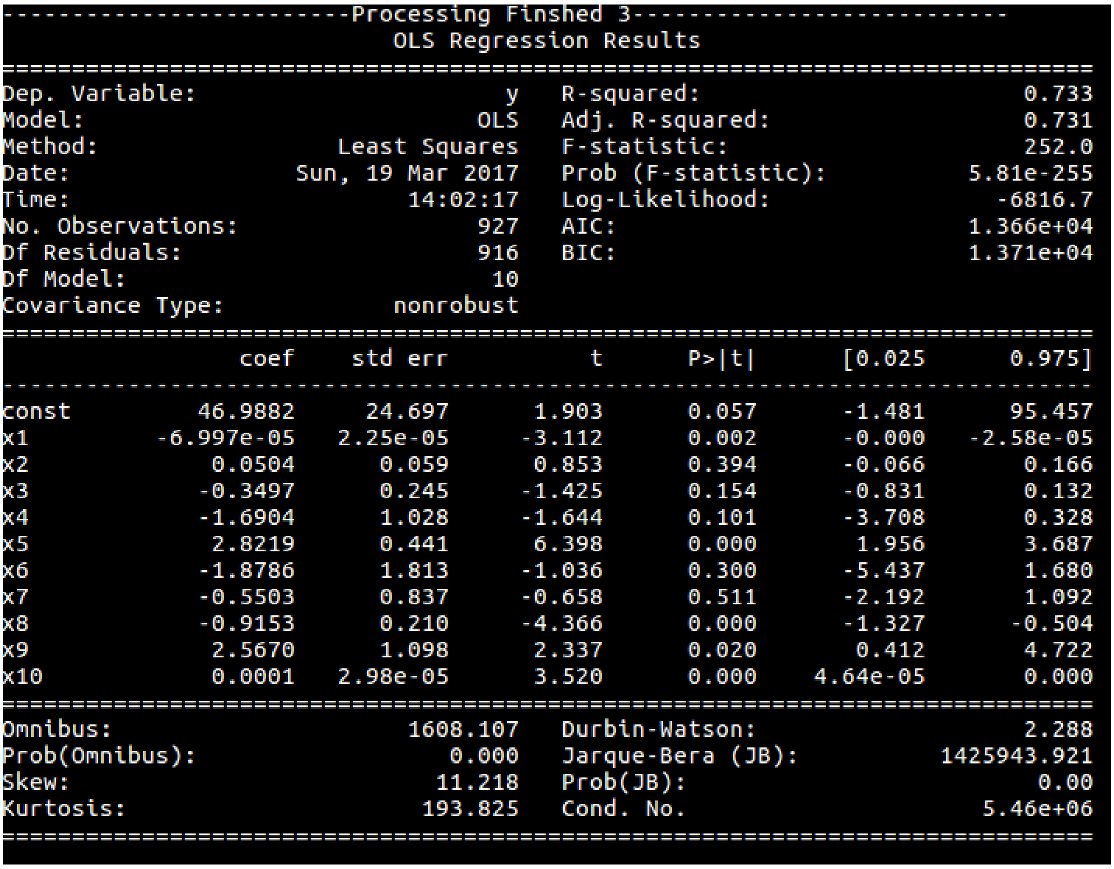
Therefore, we can get plot "prediction (number of tweets for next hour) versus feature value" for **for #gopatriots** as follows:







Therefore, we get the result **for # #nfl** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.

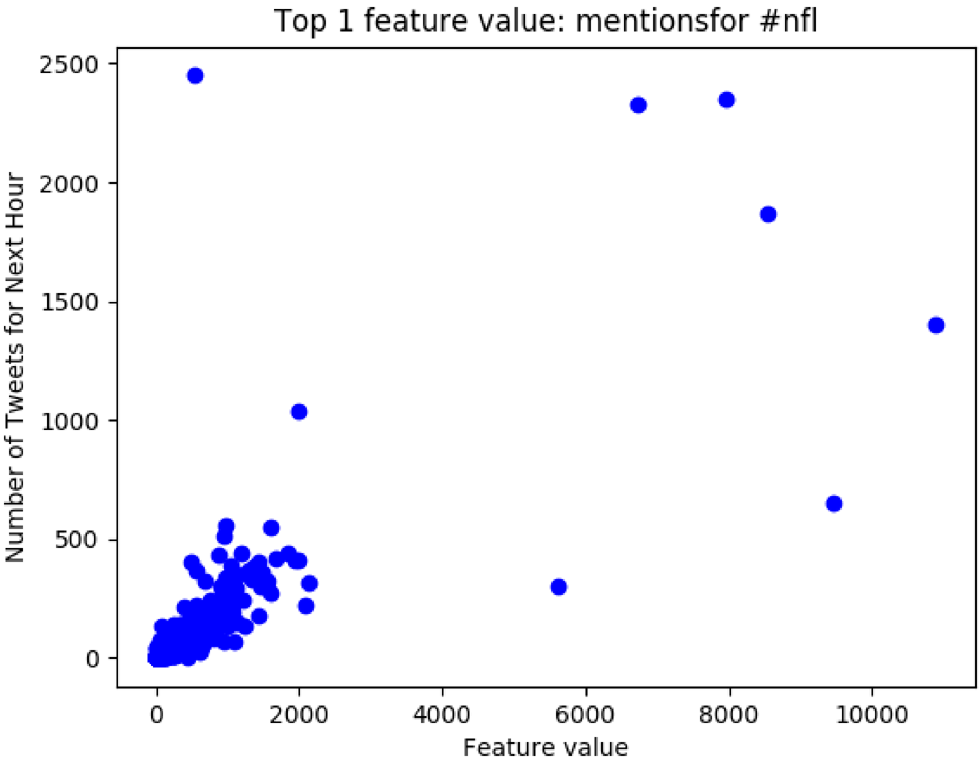


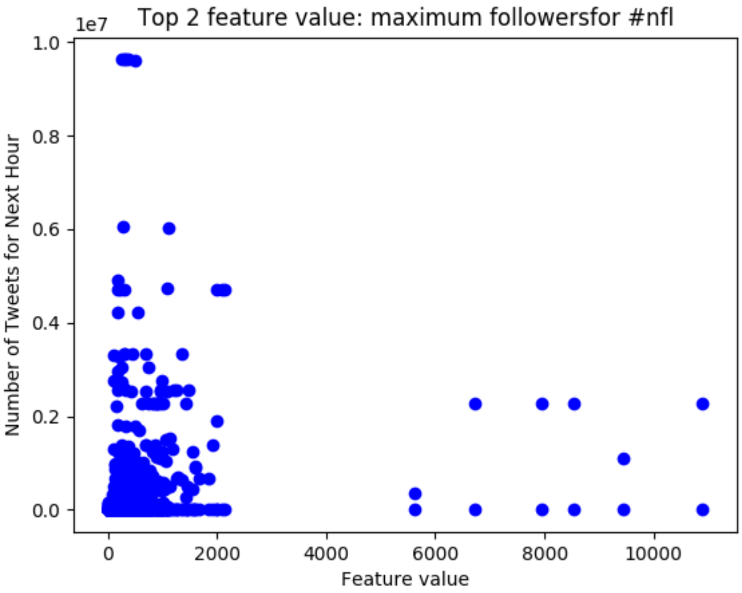
```
------------------------Processing Finshed 3--------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.733
Model:                            OLS   Adj. R-squared:                  0.731
Method:                 Least Squares   F-statistic:                     252.0
Date:                Sun, 19 Mar 2017   Prob (F-statistic):          5.81e-255
Time:                        14:02:17   Log-Likelihood:                -6816.7
No. Observations:                 927   AIC:                         1.366e+04
Df Residuals:                     916   BIC:                         1.371e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         46.9882     24.697      1.903      0.057      -1.481      95.457
x1         -6.997e-05   2.25e-05     -3.112      0.002      -0.000   -2.58e-05
x2             0.0504      0.059      0.853      0.394      -0.066       0.166
x3            -0.3497      0.245     -1.425      0.154      -0.831       0.132
x4            -1.6904      1.028     -1.644      0.101      -3.708       0.328
x5             2.8219      0.441      6.398      0.000       1.956       3.687
x6            -1.8786      1.813     -1.036      0.300      -5.437       1.680
x7            -0.5503      0.837     -0.658      0.511      -2.192       1.092
x8            -0.9153      0.210     -4.366      0.000      -1.327      -0.504
x9             2.5670      1.098      2.337      0.020       0.412       4.722
x10            0.0001   2.98e-05      3.520      0.000    4.64e-05       0.000
==============================================================================
Omnibus:                     1608.107   Durbin-Watson:                   2.288
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1425943.921
Skew:                          11.218   Prob(JB):                         0.00
Kurtosis:                     193.825   Cond. No.                     5.46e+06
==============================================================================
```

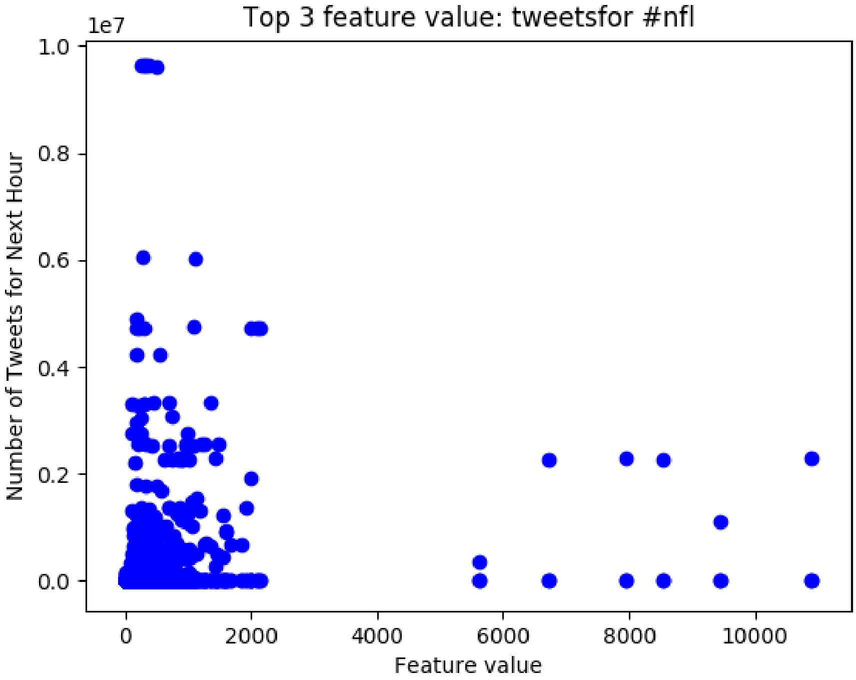
Here, the training accuracy is **R-squared: 0.7333**. And the significance of features using the t-test and P-value results is

- **X5: Total number of mentioning others**
- **X9: Number of tweets**
- **X10: Maximum number of followers of the users posting the hashtag**

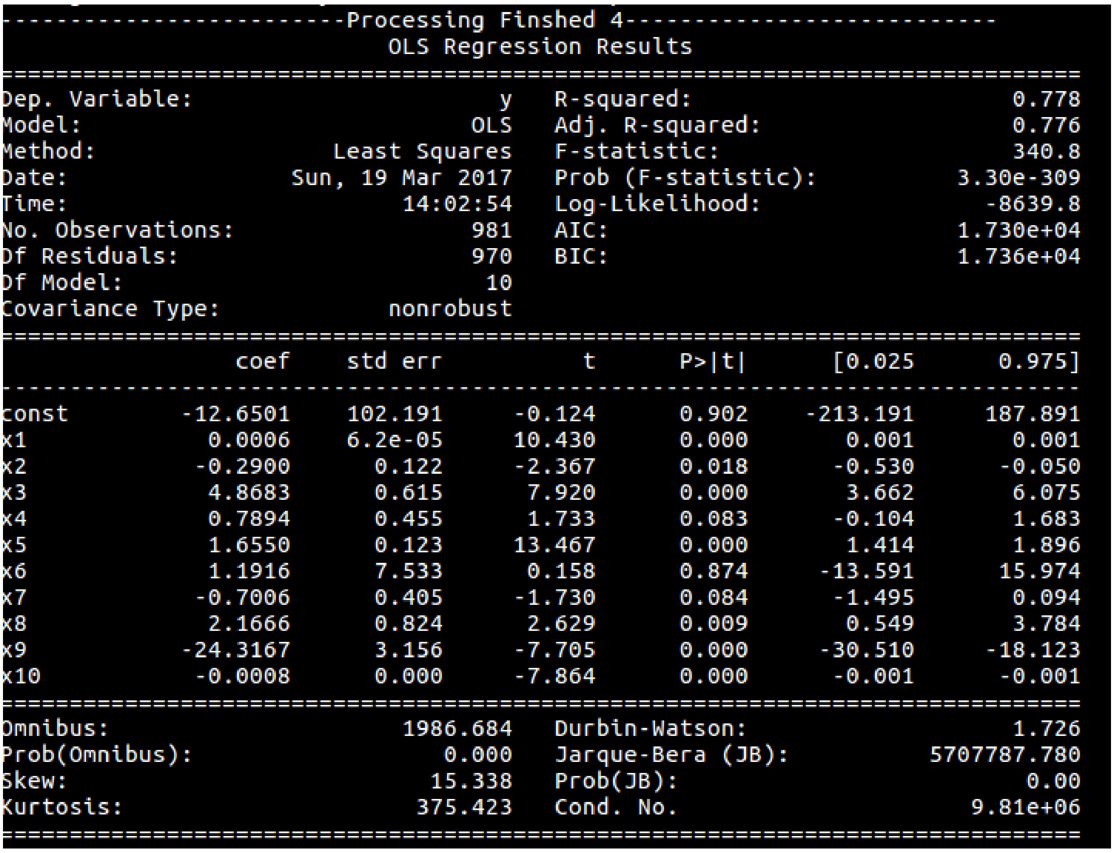
Therefore, we can get plot "prediction (number of tweets for next hour) versus feature value" for **for # #nfl** to as follows:







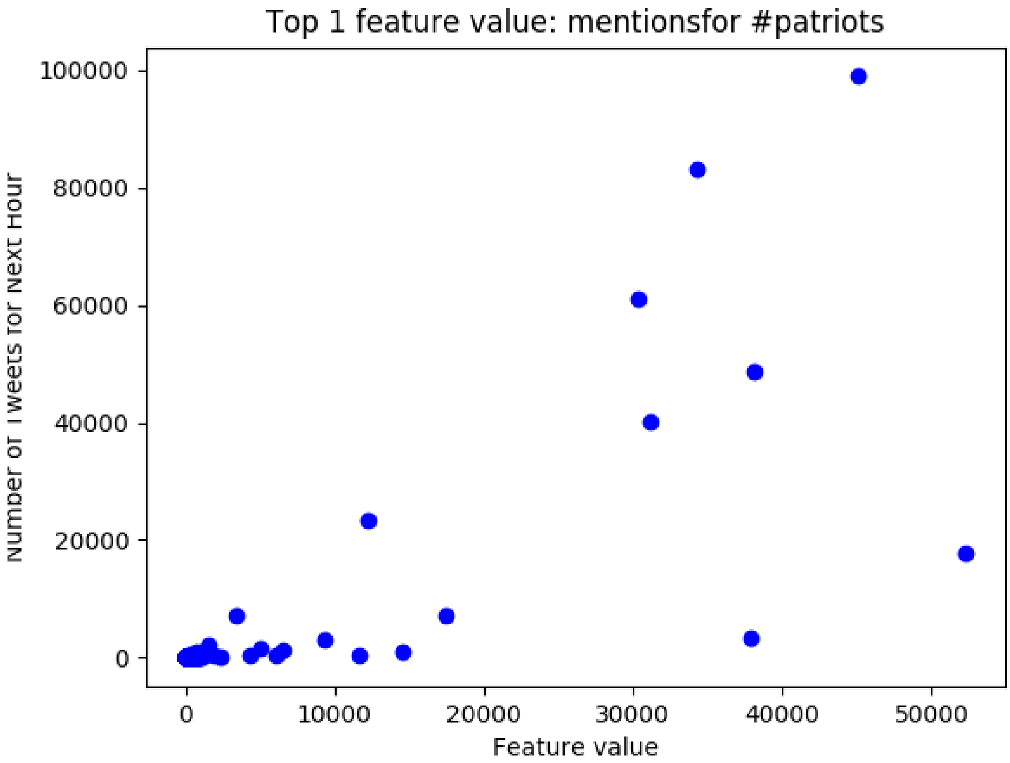
Therefore, we get the result **for #** **#patriots** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.

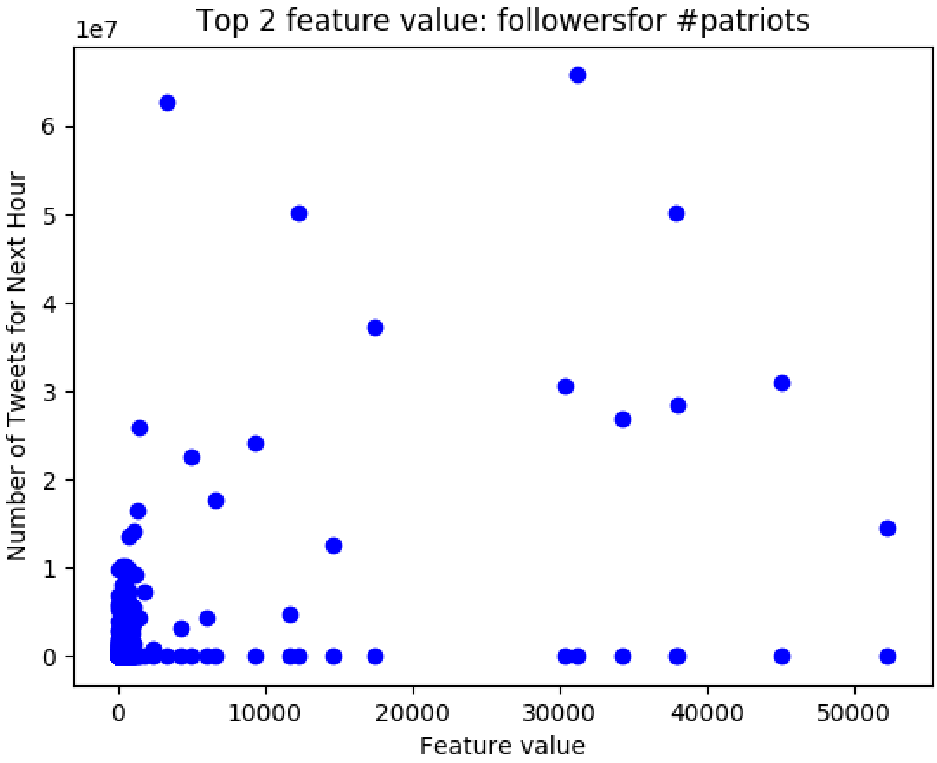

```
------------------------Processing Finshed 4--------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.778
Model:                            OLS   Adj. R-squared:                  0.776
Method:                 Least Squares   F-statistic:                     340.8
Date:                Sun, 19 Mar 2017   Prob (F-statistic):          3.30e-309
Time:                        14:02:54   Log-Likelihood:                -8639.8
No. Observations:                 981   AIC:                         1.730e+04
Df Residuals:                     970   BIC:                         1.736e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -12.6501    102.191     -0.124      0.902    -213.191     187.891
x1             0.0006    6.2e-05     10.430      0.000       0.001       0.001
x2            -0.2900      0.122     -2.367      0.018      -0.530      -0.050
x3             4.8683      0.615      7.920      0.000       3.662       6.075
x4             0.7894      0.455      1.733      0.083      -0.104       1.683
x5             1.6550      0.123     13.467      0.000       1.414       1.896
x6             1.1916      7.533      0.158      0.874     -13.591      15.974
x7            -0.7006      0.405     -1.730      0.084      -1.495       0.094
x8             2.1666      0.824      2.629      0.009       0.549       3.784
x9           -24.3167      3.156     -7.705      0.000     -30.510     -18.123
x10           -0.0008      0.000     -7.864      0.000      -0.001      -0.001
==============================================================================
Omnibus:                     1986.684   Durbin-Watson:                   1.726
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5707787.780
Skew:                          15.338   Prob(JB):                         0.00
Kurtosis:                     375.423   Cond. No.                     9.81e+06
==============================================================================
```

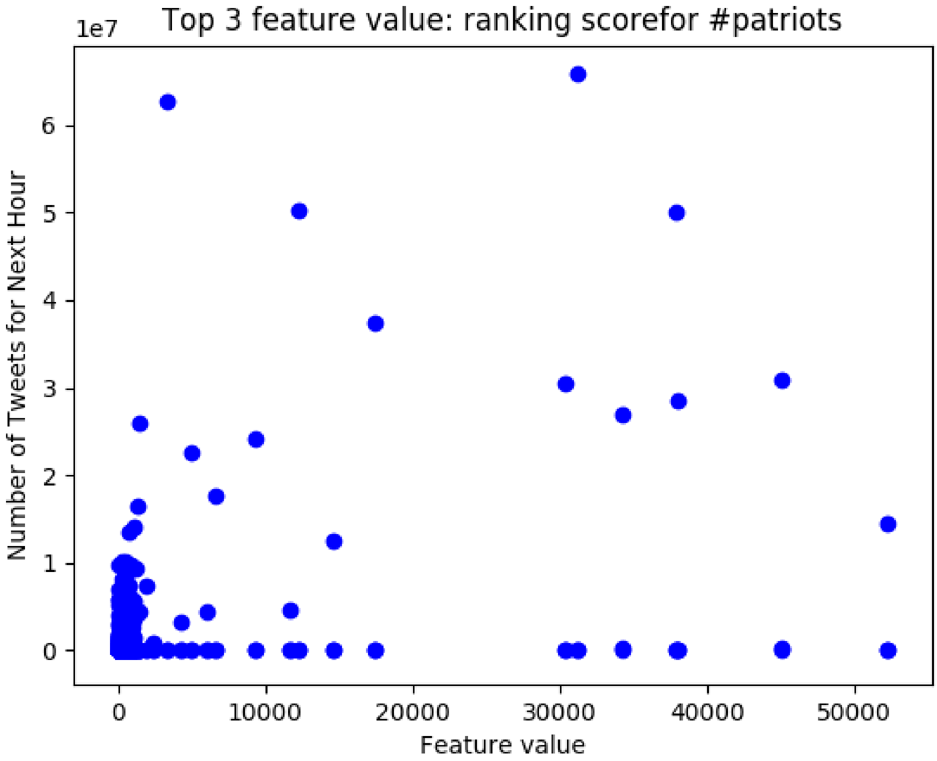
Here, the training accuracy is **R-squared: 0.778**. And the significance of features using the t-test and P-value results is

- **X1: Sum of the number of followers of the users posting the hashtag**
- **X3: Sum of the number of ranking scores of the users posting the hashtag**
- **X5: Total number of mentioning others**

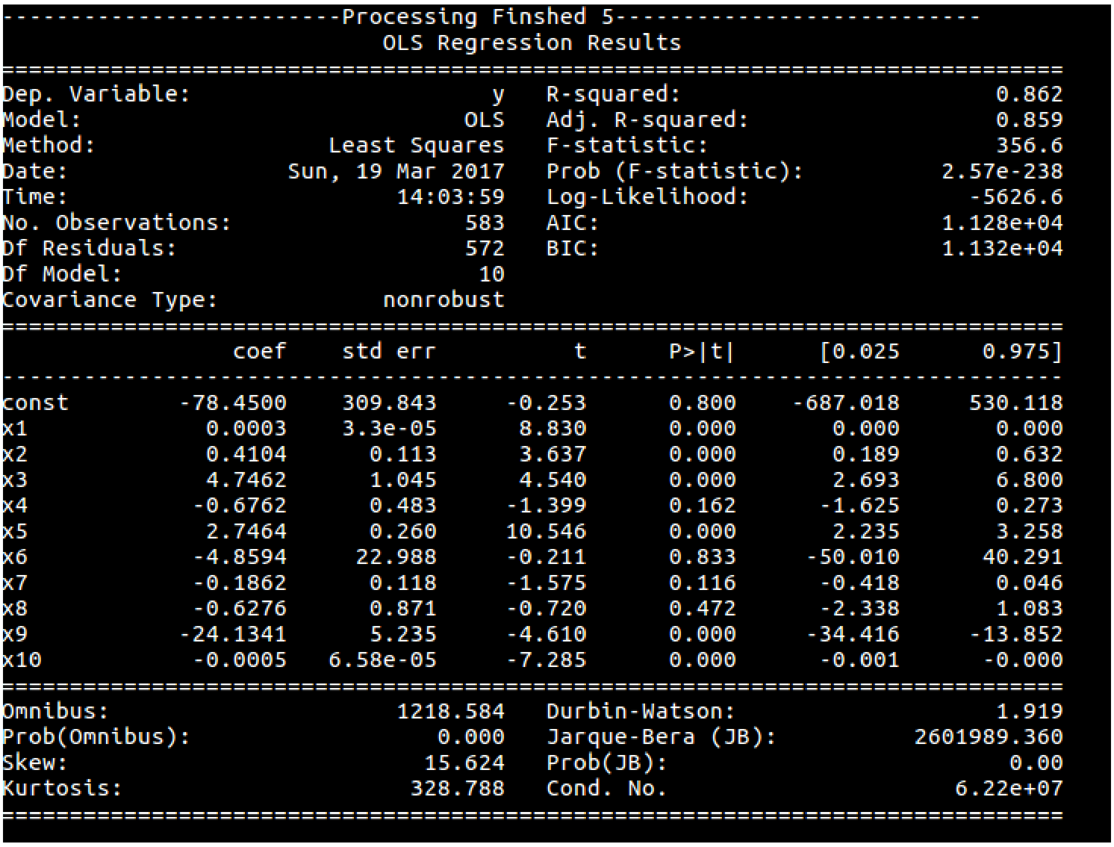
Therefore, we can get plot "prediction (number of tweets for next hour) versus feature value" for **for** **#patriots** to as follows:







Therefore, we get the result **for #sb49** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.

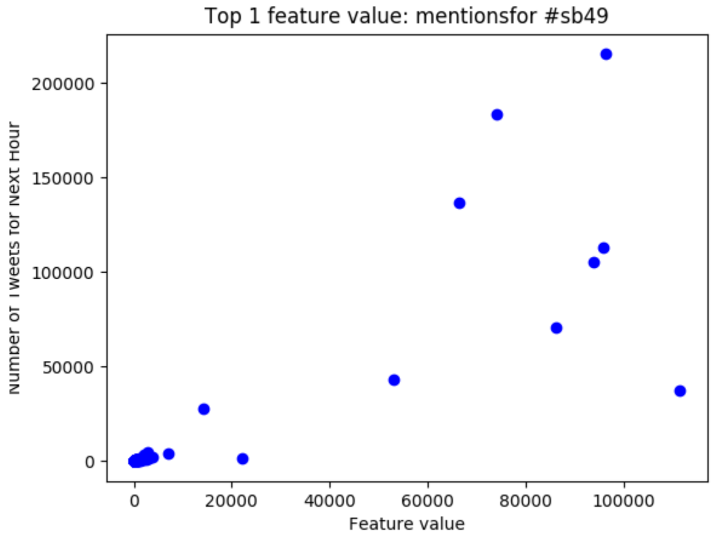


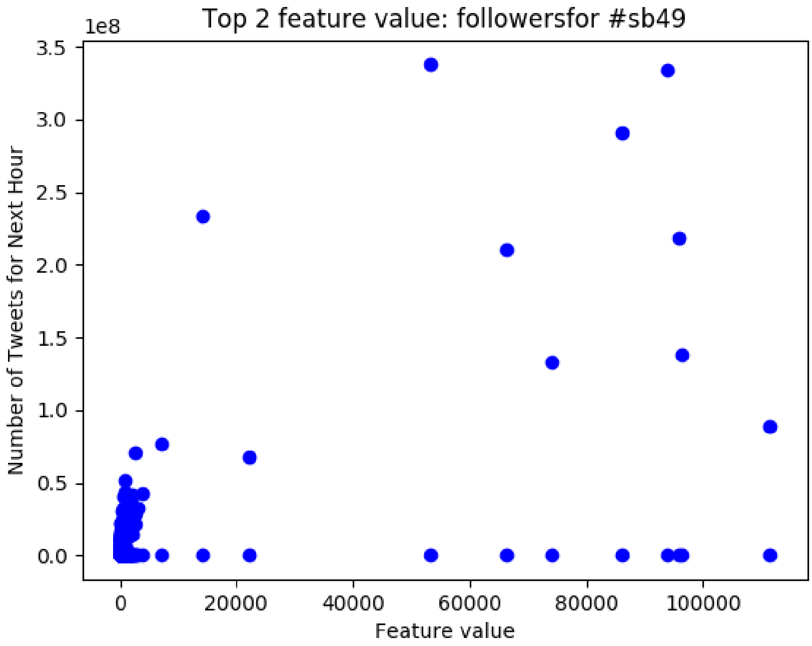
```
------------------------Processing Finshed 5-------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.862
Model:                            OLS   Adj. R-squared:                  0.859
Method:                 Least Squares   F-statistic:                     356.6
Date:                Sun, 19 Mar 2017   Prob (F-statistic):          2.57e-238
Time:                        14:03:59   Log-Likelihood:                -5626.6
No. Observations:                 583   AIC:                         1.128e+04
Df Residuals:                     572   BIC:                         1.132e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -78.4500    309.843     -0.253      0.800    -687.018     530.118
x1             0.0003    3.3e-05      8.830      0.000       0.000       0.000
x2             0.4104      0.113      3.637      0.000       0.189       0.632
x3             4.7462      1.045      4.540      0.000       2.693       6.800
x4            -0.6762      0.483     -1.399      0.162      -1.625       0.273
x5             2.7464      0.260     10.546      0.000       2.235       3.258
x6            -4.8594     22.988     -0.211      0.833     -50.010      40.291
x7            -0.1862      0.118     -1.575      0.116      -0.418       0.046
x8            -0.6276      0.871     -0.720      0.472      -2.338       1.083
x9           -24.1341      5.235     -4.610      0.000     -34.416     -13.852
x10           -0.0005   6.58e-05     -7.285      0.000      -0.001      -0.000
==============================================================================
Omnibus:                     1218.584   Durbin-Watson:                   1.919
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2601989.360
Skew:                          15.624   Prob(JB):                         0.00
Kurtosis:                     328.788   Cond. No.                     6.22e+07
==============================================================================
```

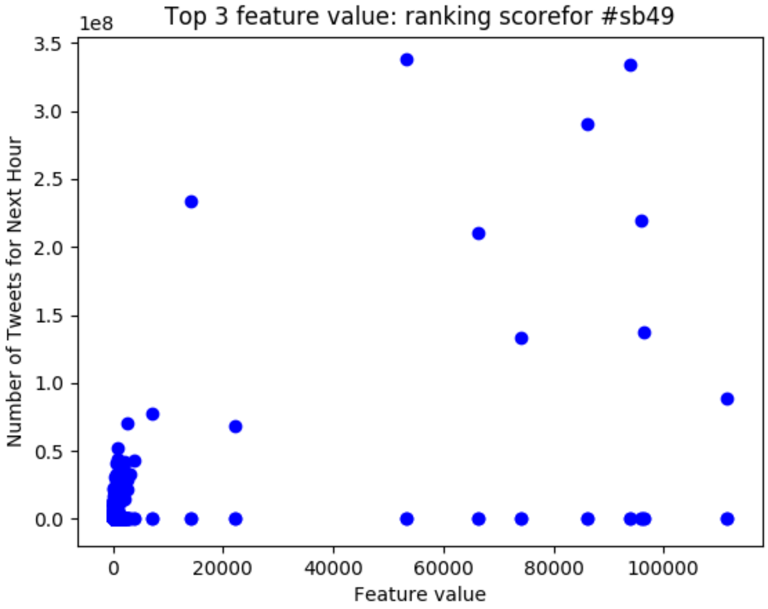
Here, the training accuracy is **R-squared: 0.862**. And the significance of features using the t-test and P-value results is

- **X1: Sum of the number of followers of the users posting the hashtag**
- **X3: Sum of the number of ranking scores of the users posting the hashtag**
- **X5: Total number of mentioning others**

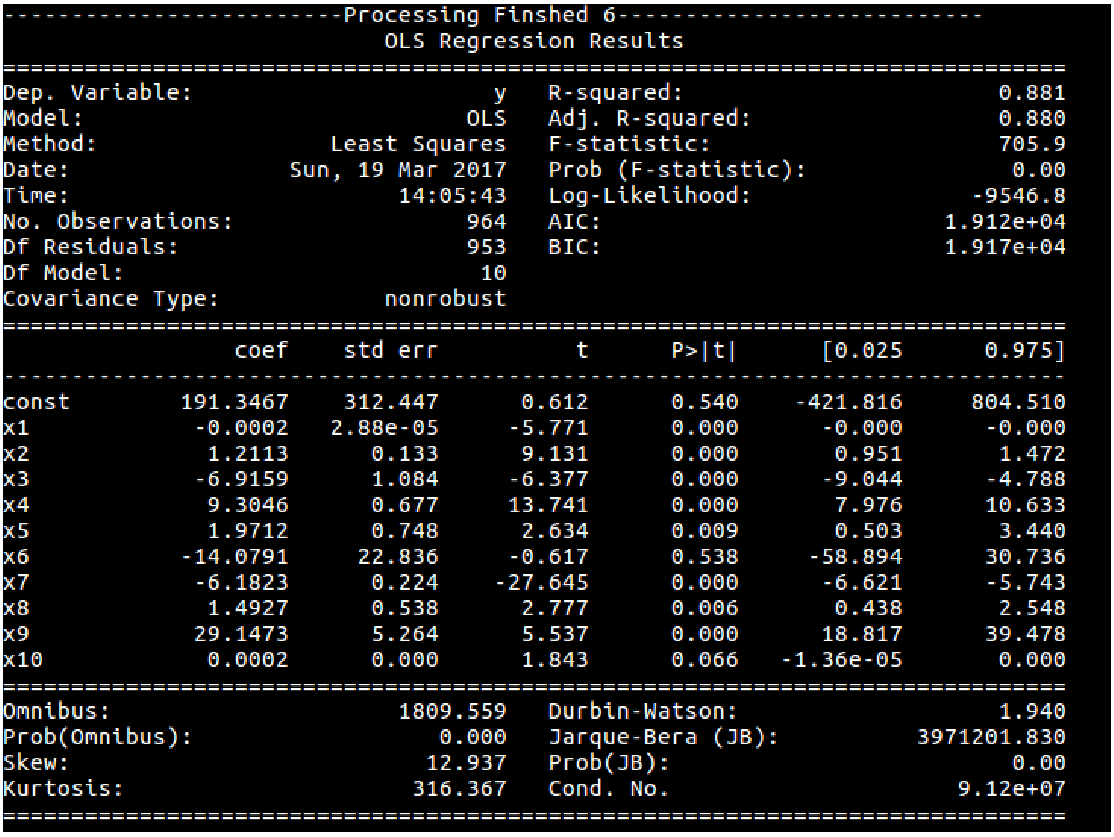
Therefore, we can get plot "prediction (number of tweets for next hour) versus feature value" for **#sb49** to as follows:


Top 1 feature value: mentionsfor #sb49





Therefore, we get the result **for #superbowl** to explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.

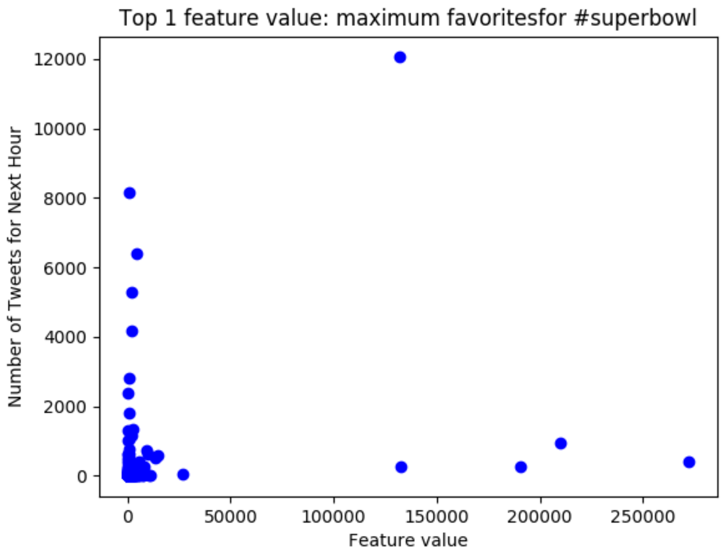


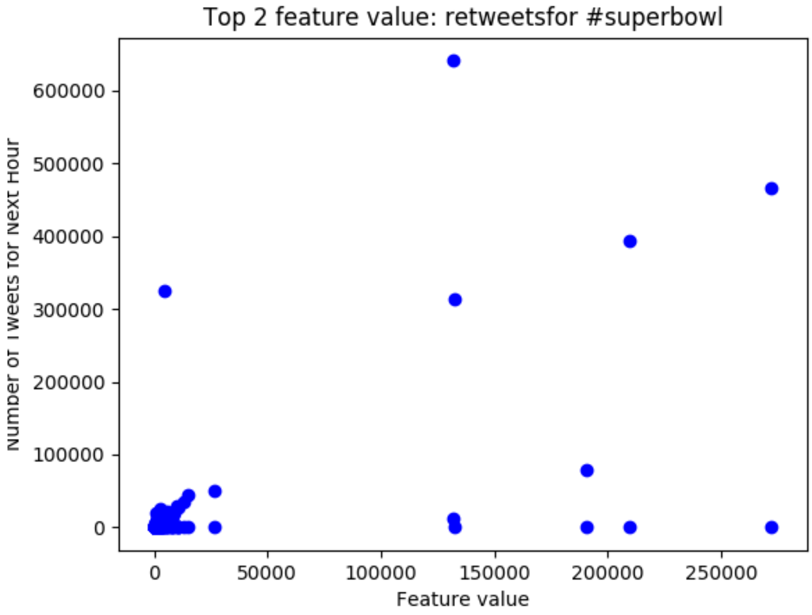
```
-------------------------Processing Finshed 6--------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.881
Model:                            OLS   Adj. R-squared:                  0.880
Method:                 Least Squares   F-statistic:                     705.9
Date:                Sun, 19 Mar 2017   Prob (F-statistic):               0.00
Time:                        14:05:43   Log-Likelihood:                -9546.8
No. Observations:                 964   AIC:                         1.912e+04
Df Residuals:                     953   BIC:                         1.917e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        191.3467    312.447      0.612      0.540    -421.816     804.510
x1            -0.0002   2.88e-05     -5.771      0.000      -0.000      -0.000
x2             1.2113      0.133      9.131      0.000       0.951       1.472
x3            -6.9159      1.084     -6.377      0.000      -9.044      -4.788
x4             9.3046      0.677     13.741      0.000       7.976      10.633
x5             1.9712      0.748      2.634      0.009       0.503       3.440
x6           -14.0791     22.836     -0.617      0.538     -58.894      30.736
x7            -6.1823      0.224    -27.645      0.000      -6.621      -5.743
x8             1.4927      0.538      2.777      0.006       0.438       2.548
x9            29.1473      5.264      5.537      0.000      18.817      39.478
x10            0.0002      0.000      1.843      0.066   -1.36e-05       0.000
==============================================================================
Omnibus:                     1809.559   Durbin-Watson:                   1.940
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          3971201.830
Skew:                          12.937   Prob(JB):                         0.00
Kurtosis:                     316.367   Cond. No.                     9.12e+07
==============================================================================
```

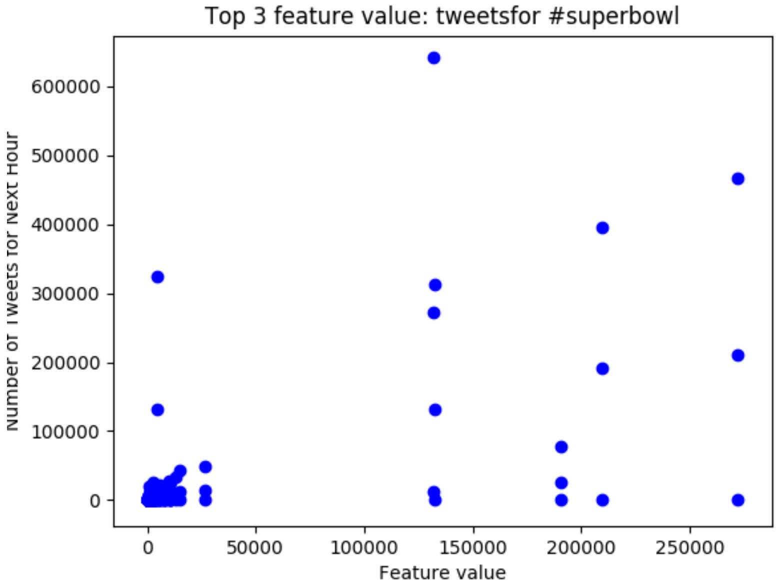
Here, the training accuracy is **R-squared: 0.881**. And the significance of features using the t-test and P-value results is

- **X2: Total number of retweets**
- **X4: Maximum number of favorites of the users posting the hashtag**
- **X9: Number of tweets**

Therefore, we can get plot "prediction (number of tweets for next hour) versus feature value" for **#superbowl** to as follows:







In the end, all result we get is better than question 3, and we know how to use t-test and P-value to determine which features are more important for our model. Here, we can also get that the features we created can make our model to make prediction more accuracy.

4 LINEAR REGRESSION MODEL WITH 10 FEATRUES USING 10-FOLD CROSS-VALIDATION

In this part, we have already used 10 features. And we create 10-fold cross-validation follow the requirement by:

"Split the feature data (your set of (features,predictant) pairs for windows) into 10 parts to perform cross-validation. Run 10 tests, each time fitting your model on 9 parts and predicting the number of tweets for the 1 remaining part. Calculate the average prediction error over samples in the remaining part, and then average these values over the 10 tests."

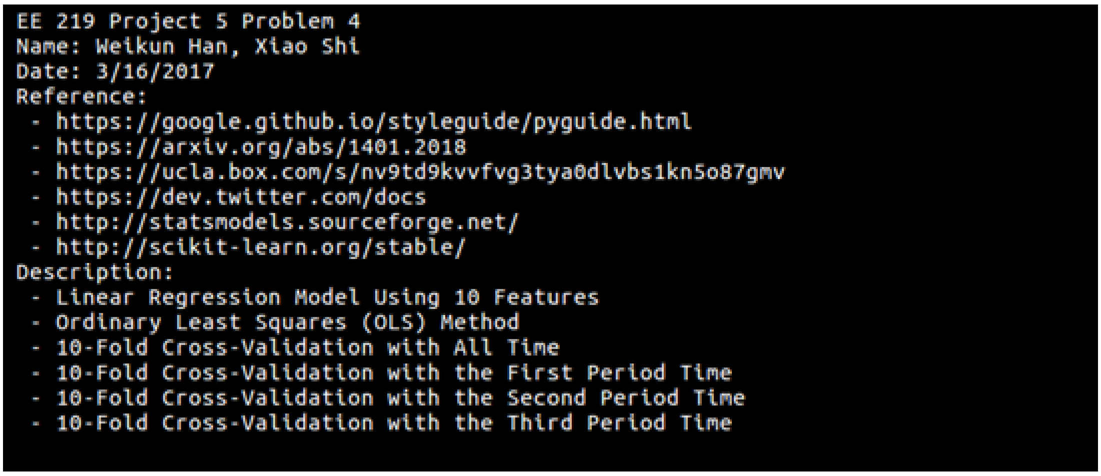
Moreover, base on the requirement:

"Since we know the Super Bowl's date and time, we can create different regression models for different periods of time. Train 3 regression models for these time periods (The times are all in PST):
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."

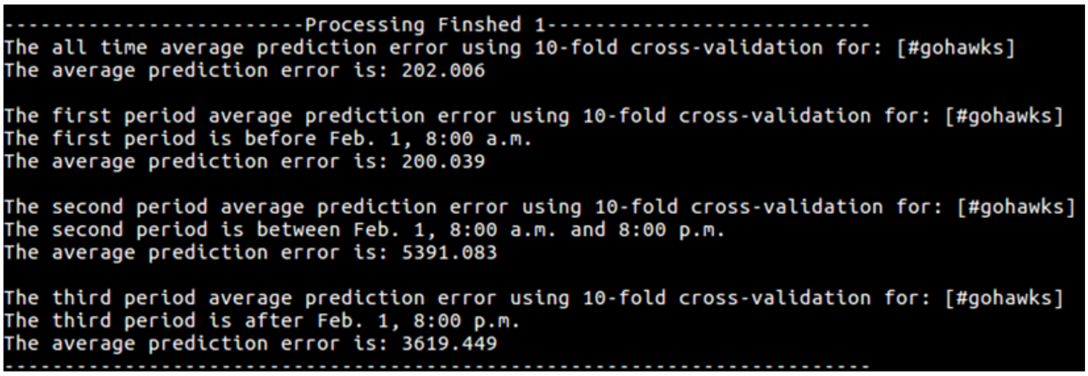
Therefore, we have another there model can use same way (10-fold cross-validation) to do it. For 10-fold cross-validation, we use scikit-learn python API. And we can get our models as follows:

- **All time model (from 2 weeks before the game to a week after the game)**
- **The first period time model (before Feb. 1, 8:00 a.m)**
- **The second period time model (between Feb. 1, 8:00 a.m. and 8:00 p.m.)**
- **The third period time model (after Feb. 1, 8:00 p.m.)**

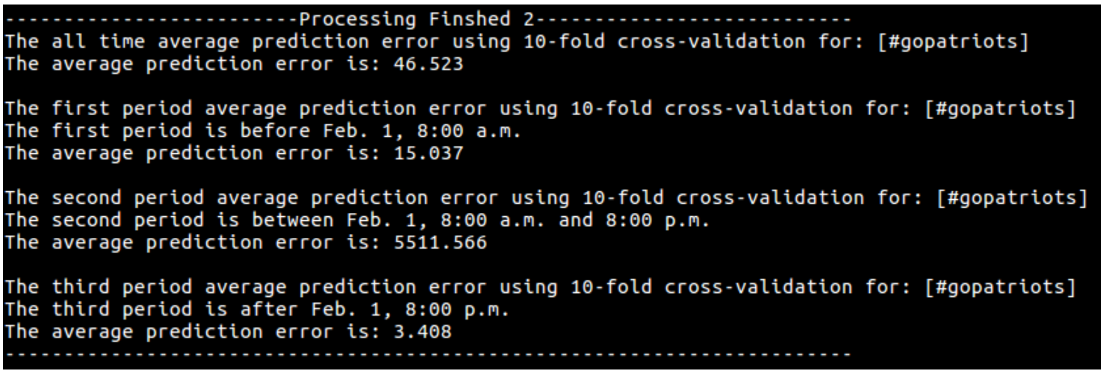
Therefore, we can get the result as follows:



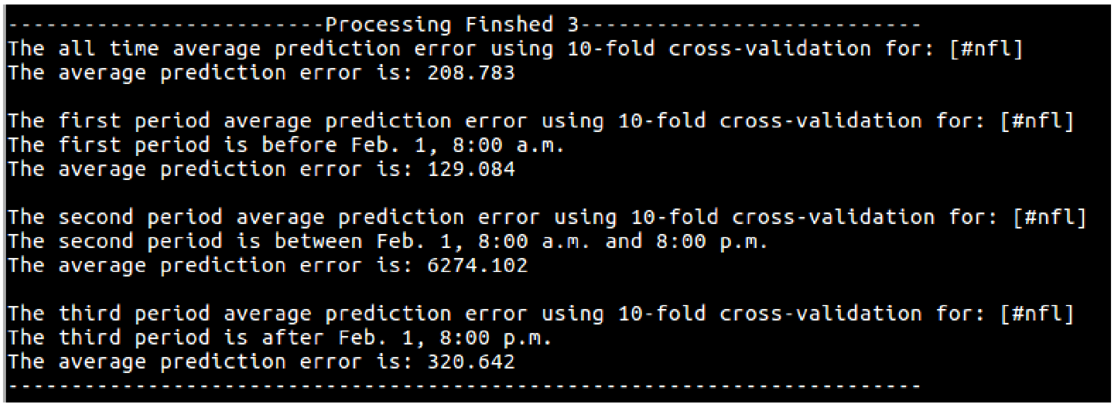
```
EE 219 Project 5 Problem 4
Name: Weikun Han, Xiao Shi
Date: 3/16/2017
Reference:
 - https://google.github.io/styleguide/pyguide.html
 - https://arxiv.org/abs/1401.2018
 - https://ucla.box.com/s/nv9td9kvvfvg3tya0dlvbs1kn5o87gmv
 - https://dev.twitter.com/docs
 - http://statsmodels.sourceforge.net/
 - http://scikit-learn.org/stable/
Description:
 - Linear Regression Model Using 10 Features
 - Ordinary Least Squares (OLS) Method
 - 10-Fold Cross-Validation with All Time
 - 10-Fold Cross-Validation with the First Period Time
 - 10-Fold Cross-Validation with the Second Period Time
 - 10-Fold Cross-Validation with the Third Period Time
```



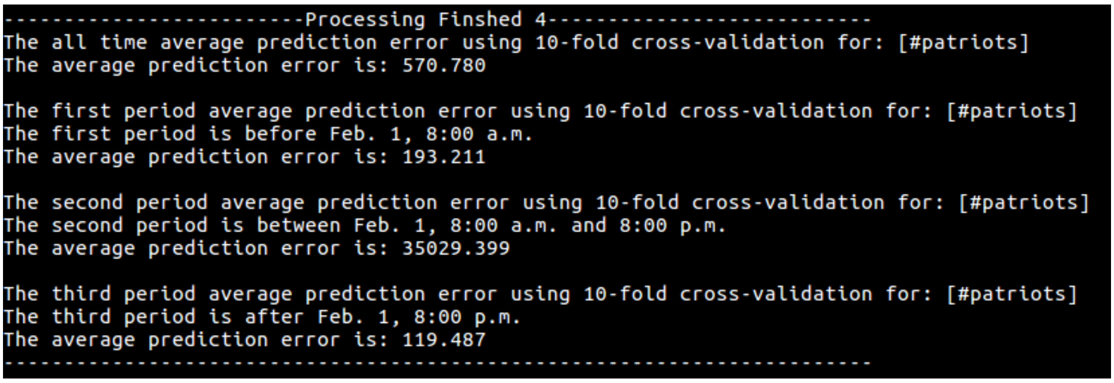
```
-------------------------Processing Finshed 1--------------------------
The all time average prediction error using 10-fold cross-validation for: [#gohawks]
The average prediction error is: 202.006

The first period average prediction error using 10-fold cross-validation for: [#gohawks]
The first period is before Feb. 1, 8:00 a.m.
The average prediction error is: 200.039

The second period average prediction error using 10-fold cross-validation for: [#gohawks]
The second period is between Feb. 1, 8:00 a.m. and 8:00 p.m.
The average prediction error is: 5391.083

The third period average prediction error using 10-fold cross-validation for: [#gohawks]
The third period is after Feb. 1, 8:00 p.m.
The average prediction error is: 3619.449
-----------------------------------------------------------------------
```



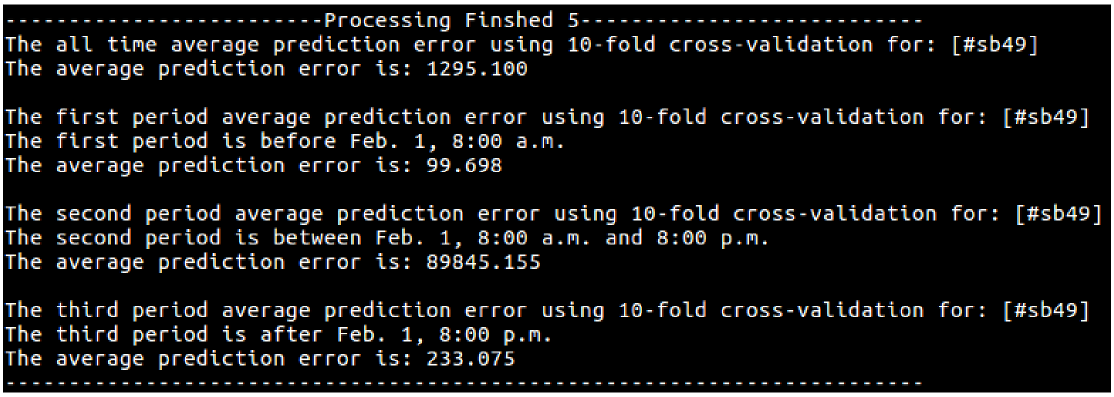
```
-------------------------Processing Finshed 2--------------------------
The all time average prediction error using 10-fold cross-validation for: [#gopatriots]
The average prediction error is: 46.523

The first period average prediction error using 10-fold cross-validation for: [#gopatriots]
The first period is before Feb. 1, 8:00 a.m.
The average prediction error is: 15.037

The second period average prediction error using 10-fold cross-validation for: [#gopatriots]
The second period is between Feb. 1, 8:00 a.m. and 8:00 p.m.
The average prediction error is: 5511.566

The third period average prediction error using 10-fold cross-validation for: [#gopatriots]
The third period is after Feb. 1, 8:00 p.m.
The average prediction error is: 3.408
-----------------------------------------------------------------------
```



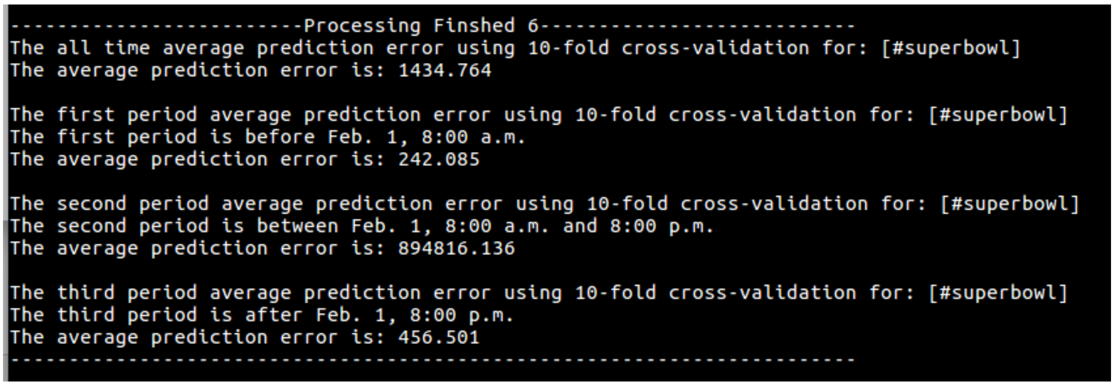
```
-------------------------Processing Finshed 3--------------------------
The all time average prediction error using 10-fold cross-validation for: [#nfl]
The average prediction error is: 208.783

The first period average prediction error using 10-fold cross-validation for: [#nfl]
The first period is before Feb. 1, 8:00 a.m.
The average prediction error is: 129.084

The second period average prediction error using 10-fold cross-validation for: [#nfl]
The second period is between Feb. 1, 8:00 a.m. and 8:00 p.m.
The average prediction error is: 6274.102

The third period average prediction error using 10-fold cross-validation for: [#nfl]
The third period is after Feb. 1, 8:00 p.m.
The average prediction error is: 320.642
-----------------------------------------------------------------------
```



```
-------------------------Processing Finshed 4--------------------------
The all time average prediction error using 10-fold cross-validation for: [#patriots]
The average prediction error is: 570.780

The first period average prediction error using 10-fold cross-validation for: [#patriots]
The first period is before Feb. 1, 8:00 a.m.
The average prediction error is: 193.211

The second period average prediction error using 10-fold cross-validation for: [#patriots]
The second period is between Feb. 1, 8:00 a.m. and 8:00 p.m.
The average prediction error is: 35029.399

The third period average prediction error using 10-fold cross-validation for: [#patriots]
The third period is after Feb. 1, 8:00 p.m.
The average prediction error is: 119.487
-----------------------------------------------------------------------
```



```
-------------------------Processing Finshed 5--------------------------
The all time average prediction error using 10-fold cross-validation for: [#sb49]
The average prediction error is: 1295.100

The first period average prediction error using 10-fold cross-validation for: [#sb49]
The first period is before Feb. 1, 8:00 a.m.
The average prediction error is: 99.698

The second period average prediction error using 10-fold cross-validation for: [#sb49]
The second period is between Feb. 1, 8:00 a.m. and 8:00 p.m.
The average prediction error is: 89845.155

The third period average prediction error using 10-fold cross-validation for: [#sb49]
The third period is after Feb. 1, 8:00 p.m.
The average prediction error is: 233.075
-----------------------------------------------------------------------
```



```
------------------------Processing Finshed 6--------------------------
The all time average prediction error using 10-fold cross-validation for: [#superbowl]
The average prediction error is: 1434.764

The first period average prediction error using 10-fold cross-validation for: [#superbowl]
The first period is before Feb. 1, 8:00 a.m.
The average prediction error is: 242.085

The second period average prediction error using 10-fold cross-validation for: [#superbowl]
The second period is between Feb. 1, 8:00 a.m. and 8:00 p.m.
The average prediction error is: 894816.136

The third period average prediction error using 10-fold cross-validation for: [#superbowl]
The third period is after Feb. 1, 8:00 p.m.
The average prediction error is: 456.501
----------------------------------------------------------------------
```

Hashtag	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
Average prediction error for all time model	202.006	46.532	208.783	570.780	1295.100	1434.764
Average prediction error for the first period time model	200.039	15.037	129.084	193.211	99.698	242.085
Average prediction error for the second period time model	5391.083	5511.566	6274.102	35029.399	89846.155	894816.136
Average prediction error for the third period time model	3619.449	3.408	320.642	119.487	233.075	456.501

In the end, we use 10-fold cross-validation to verify keep verify our model. From the result, we can see basically the error we got is reason. But the second period time model is seems not good, because we had 12 training data and it hard to get good predication.

5 LINEAR REGRESSION MODEL WITH 10 FEATRUES USING ANOTHER 10 DIFFERENT DATA VALIDATION

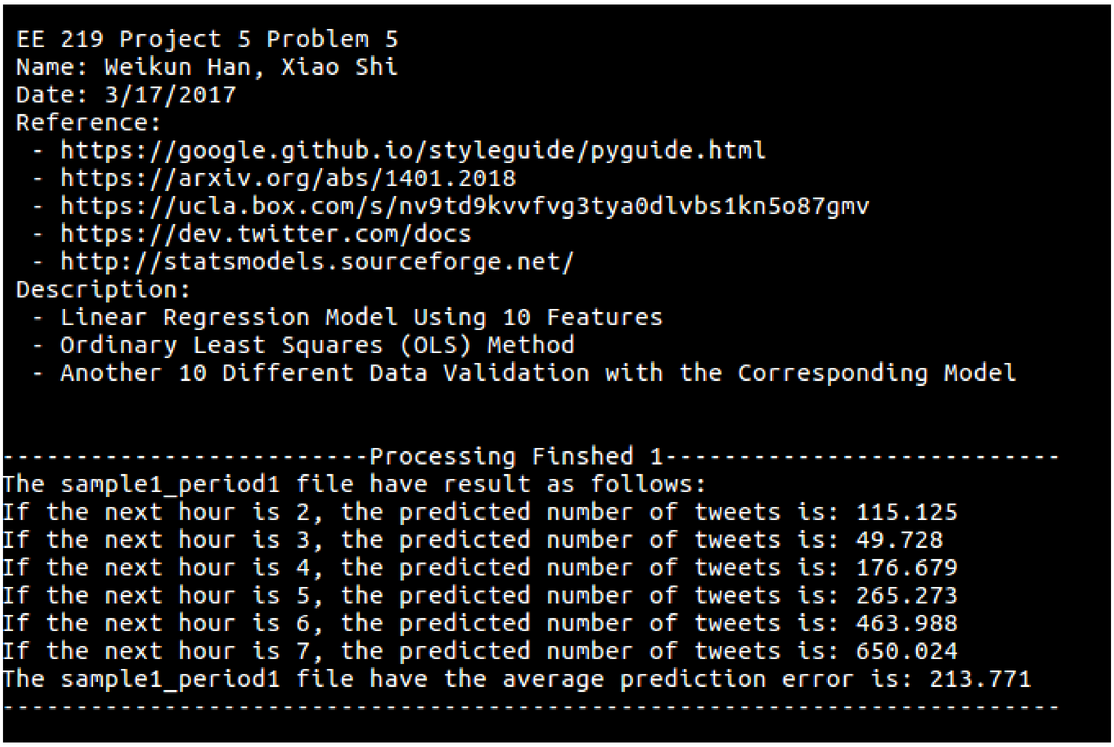
In this part, we have already used 10 features. And we create 10 different data is provided is follows [6]:

- **Sample1_period1**
- **Sample2_period2**
- **Sample3_period3**
- **Sample4_period1**
- **Sample5_period1**
- **Sample6_period2**
- **Sample7_period3**
- **Sample8_period1**
- **Sample9_period2**
- **Sample10_period3**

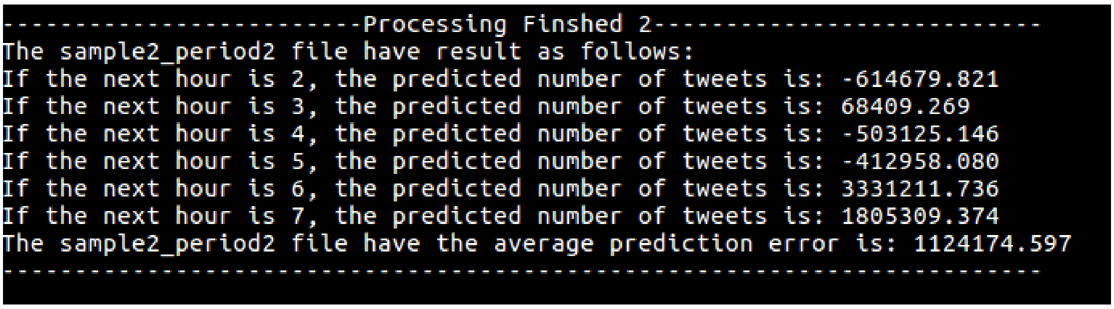
Because of we have no idea those data belong to which hashtag; it is problem for use to use question 4 model. However, we can open each file and use search to almost check the dominant hashtag for each file, and then we can fit the corresponding model to make prediction. Therefore, we can get the result as follow:

- **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**

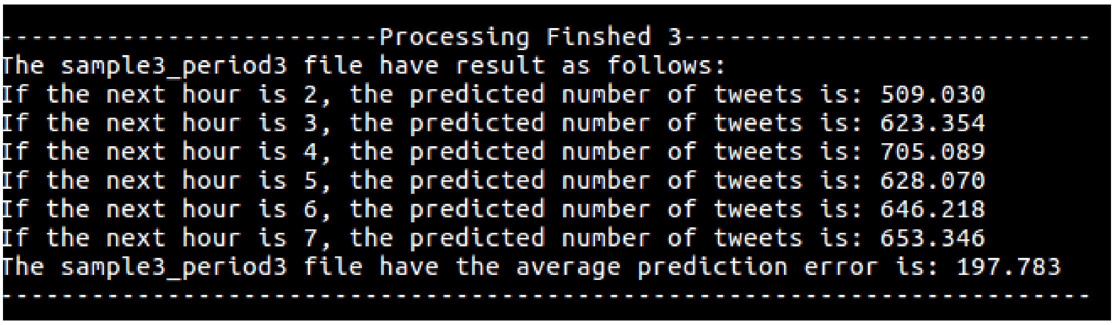
Therefore, we can get the result as follows:



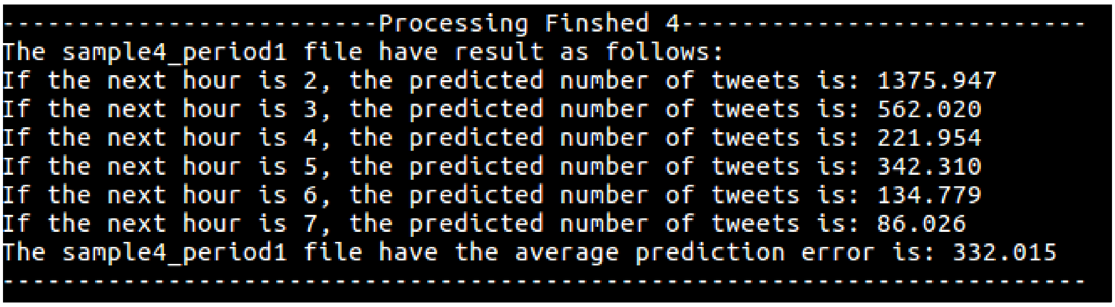
```
 EE 219 Project 5 Problem 5
 Name: Weikun Han, Xiao Shi
 Date: 3/17/2017
 Reference:
  - https://google.github.io/styleguide/pyguide.html
  - https://arxiv.org/abs/1401.2018
  - https://ucla.box.com/s/nv9td9kvvfvg3tya0dlvbs1kn5o87gmv
  - https://dev.twitter.com/docs
  - http://statsmodels.sourceforge.net/
 Description:
  - Linear Regression Model Using 10 Features
  - Ordinary Least Squares (OLS) Method
  - Another 10 Different Data Validation with the Corresponding Model


------------------------Processing Finshed 1--------------------------
The sample1_period1 file have result as follows:
If the next hour is 2, the predicted number of tweets is: 115.125
If the next hour is 3, the predicted number of tweets is: 49.728
If the next hour is 4, the predicted number of tweets is: 176.679
If the next hour is 5, the predicted number of tweets is: 265.273
If the next hour is 6, the predicted number of tweets is: 463.988
If the next hour is 7, the predicted number of tweets is: 650.024
The sample1_period1 file have the average prediction error is: 213.771
---------------------------------------------------------------------
```



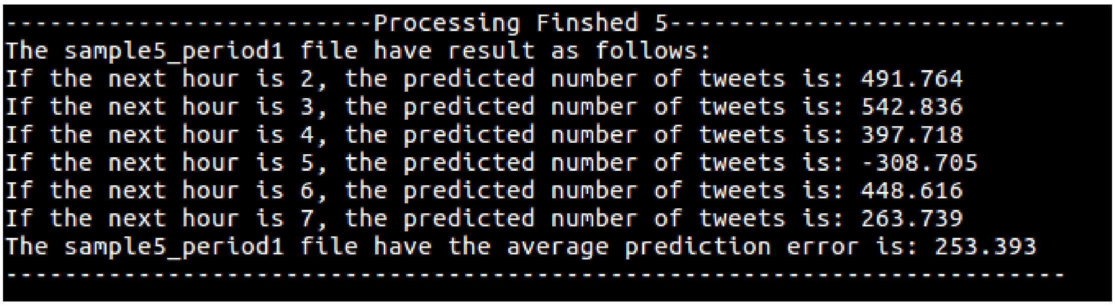
```
------------------------Processing Finshed 2--------------------------
The sample2_period2 file have result as follows:
If the next hour is 2, the predicted number of tweets is: -614679.821
If the next hour is 3, the predicted number of tweets is: 68409.269
If the next hour is 4, the predicted number of tweets is: -503125.146
If the next hour is 5, the predicted number of tweets is: -412958.080
If the next hour is 6, the predicted number of tweets is: 3331211.736
If the next hour is 7, the predicted number of tweets is: 1805309.374
The sample2_period2 file have the average prediction error is: 1124174.597
---------------------------------------------------------------------
```



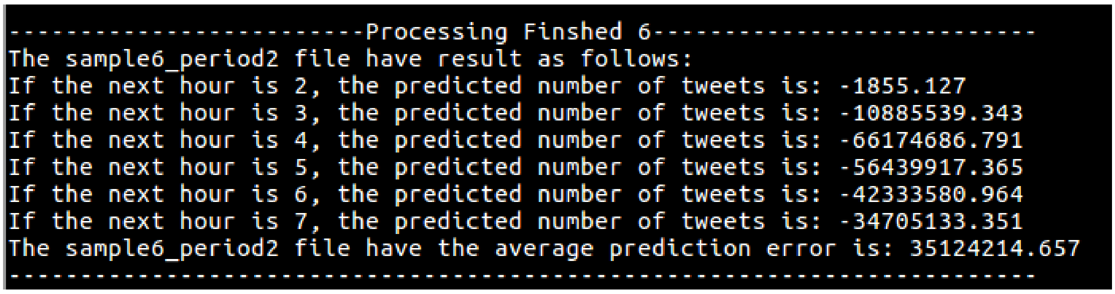
```
------------------------Processing Finshed 3--------------------------
The sample3_period3 file have result as follows:
If the next hour is 2, the predicted number of tweets is: 509.030
If the next hour is 3, the predicted number of tweets is: 623.354
If the next hour is 4, the predicted number of tweets is: 705.089
If the next hour is 5, the predicted number of tweets is: 628.070
If the next hour is 6, the predicted number of tweets is: 646.218
If the next hour is 7, the predicted number of tweets is: 653.346
The sample3_period3 file have the average prediction error is: 197.783
---------------------------------------------------------------------
```



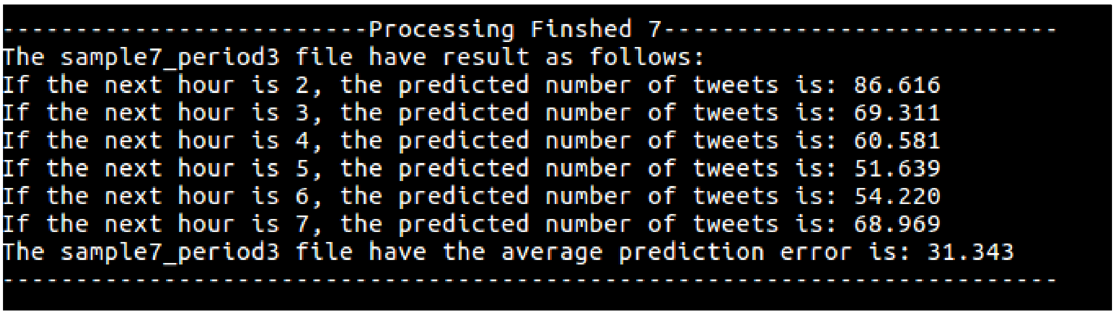
```
------------------------Processing Finshed 4--------------------------
The sample4_period1 file have result as follows:
If the next hour is 2, the predicted number of tweets is: 1375.947
If the next hour is 3, the predicted number of tweets is: 562.020
If the next hour is 4, the predicted number of tweets is: 221.954
If the next hour is 5, the predicted number of tweets is: 342.310
If the next hour is 6, the predicted number of tweets is: 134.779
If the next hour is 7, the predicted number of tweets is: 86.026
The sample4_period1 file have the average prediction error is: 332.015
---------------------------------------------------------------------
```



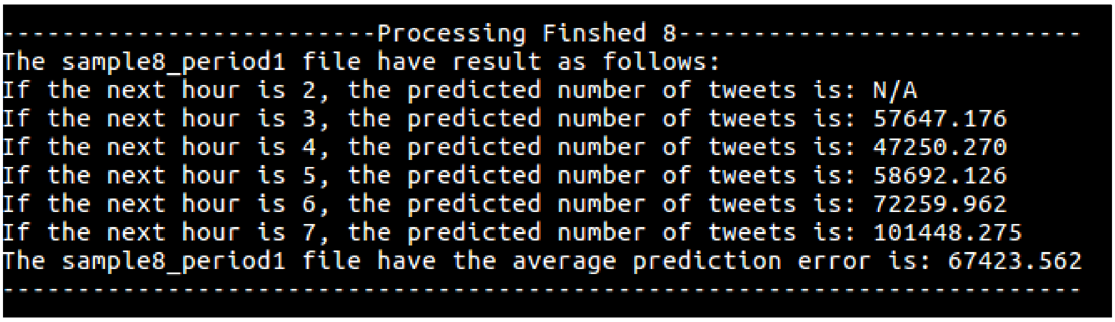
```
-------------------------Processing Finshed 5--------------------------
The sample5_period1 file have result as follows:
If the next hour is 2, the predicted number of tweets is: 491.764
If the next hour is 3, the predicted number of tweets is: 542.836
If the next hour is 4, the predicted number of tweets is: 397.718
If the next hour is 5, the predicted number of tweets is: -308.705
If the next hour is 6, the predicted number of tweets is: 448.616
If the next hour is 7, the predicted number of tweets is: 263.739
The sample5_period1 file have the average prediction error is: 253.393
-----------------------------------------------------------------------
```



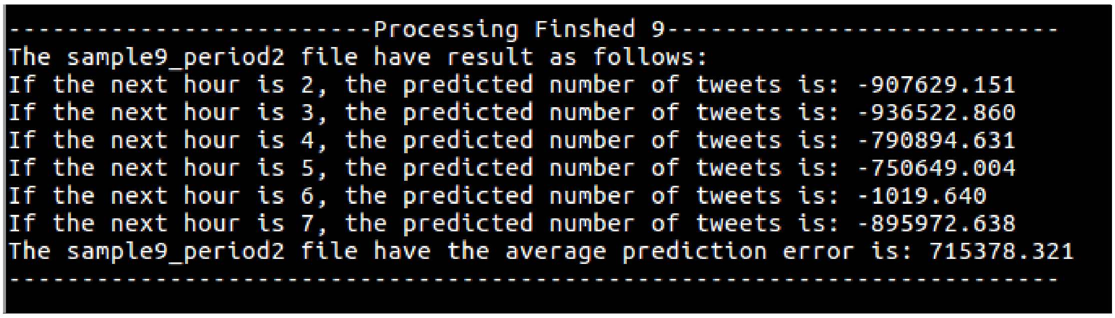
```
-------------------------Processing Finshed 6--------------------------
The sample6_period2 file have result as follows:
If the next hour is 2, the predicted number of tweets is: -1855.127
If the next hour is 3, the predicted number of tweets is: -10885539.343
If the next hour is 4, the predicted number of tweets is: -66174686.791
If the next hour is 5, the predicted number of tweets is: -56439917.365
If the next hour is 6, the predicted number of tweets is: -42333580.964
If the next hour is 7, the predicted number of tweets is: -34705133.351
The sample6_period2 file have the average prediction error is: 35124214.657
-----------------------------------------------------------------------
```



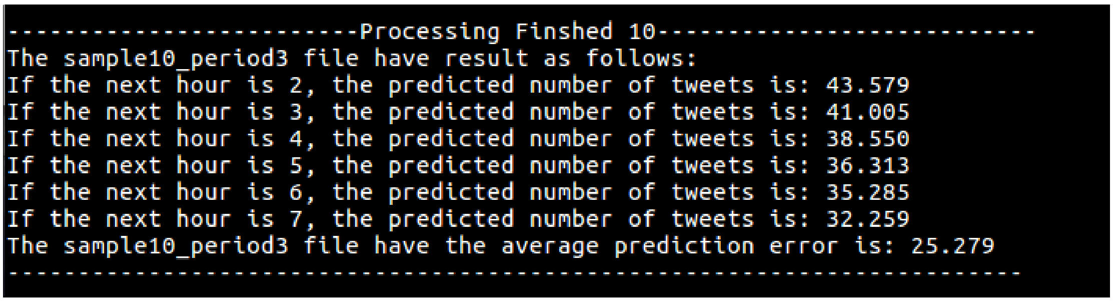
```
-------------------------Processing Finshed 7--------------------------
The sample7_period3 file have result as follows:
If the next hour is 2, the predicted number of tweets is: 86.616
If the next hour is 3, the predicted number of tweets is: 69.311
If the next hour is 4, the predicted number of tweets is: 60.581
If the next hour is 5, the predicted number of tweets is: 51.639
If the next hour is 6, the predicted number of tweets is: 54.220
If the next hour is 7, the predicted number of tweets is: 68.969
The sample7_period3 file have the average prediction error is: 31.343
-----------------------------------------------------------------------
```



```
-------------------------Processing Finshed 8--------------------------
The sample8_period1 file have result as follows:
If the next hour is 2, the predicted number of tweets is: N/A
If the next hour is 3, the predicted number of tweets is: 57647.176
If the next hour is 4, the predicted number of tweets is: 47250.270
If the next hour is 5, the predicted number of tweets is: 58692.126
If the next hour is 6, the predicted number of tweets is: 72259.962
If the next hour is 7, the predicted number of tweets is: 101448.275
The sample8_period1 file have the average prediction error is: 67423.562
-----------------------------------------------------------------------
```



```
-------------------------Processing Finshed 9--------------------------
The sample9_period2 file have result as follows:
If the next hour is 2, the predicted number of tweets is: -907629.151
If the next hour is 3, the predicted number of tweets is: -936522.860
If the next hour is 4, the predicted number of tweets is: -790894.631
If the next hour is 5, the predicted number of tweets is: -750649.004
If the next hour is 6, the predicted number of tweets is: -1019.640
If the next hour is 7, the predicted number of tweets is: -895972.638
The sample9_period2 file have the average prediction error is: 715378.321
-----------------------------------------------------------------------
```



```
-------------------------Processing Finshed 10--------------------------
The sample10_period3 file have result as follows:
If the next hour is 2, the predicted number of tweets is: 43.579
If the next hour is 3, the predicted number of tweets is: 41.005
If the next hour is 4, the predicted number of tweets is: 38.550
If the next hour is 5, the predicted number of tweets is: 36.313
If the next hour is 6, the predicted number of tweets is: 35.285
If the next hour is 7, the predicted number of tweets is: 32.259
The sample10_period3 file have the average prediction error is: 25.279
----------------------------------------------------------------------
```

File Name	Sample1_period1	Sample2_period2	Sample3_period3	Sample4_period1	Sample5_period1	Sample6_period2	Sample7_period3	Sample8_period1	Sample9_period2	Sample10_period
Next Hour is 2	115.12	614679.8	509.03	1375.94	491.76	11855.12	86.61	N/A	907629	43.57
Next Hour is 3	49.78	68409.27	623.35	562.02	542.83	10885539	69.31	57647.17	936522	41.00
Next Hour is 4	176.68	503125	705.78	221.95	397.72	66174686	60.58	47250.27	790894	38.55
Next Hour is 5	265.27	412958	412958	342.30	308.70	5643991.7	51.63	58692.12	750649	36.31
Next Hour is 6	463.99	3331211	646.21	134.77	448.62	4233358.1	54.21	72259.96	1019	35.28
Next Hour is 7	650.02	1805309	653.34	86.02	263.73	347051.3	68.96	101448.2	895972	32.25
Average Prediction Error	213.71	1124174	197.78	332.01	253.39	3512142.	131.34	67423.56	715378	25.27

In the end, the 10 different data validations to verify keep verify our model. From the result, we can see basically the error we got is reason. Also, the method we use to determine error is use hour from 2 to 6, because we have no test date in hour 7. (Each file in the test data contains a hashtag's tweets for a 6-hour window.)

6 FAN BASE PREDICTIONS

In this part of project, we need deal with data first because it:

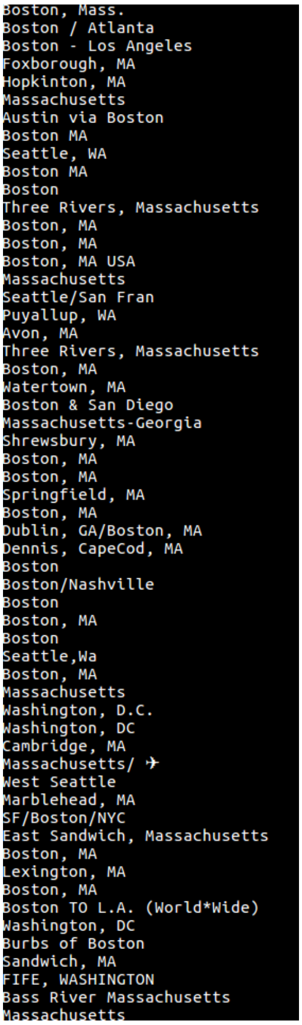
“Recognizing that supporting a sport team has a lot to do with the user location, we try to use the textual content of the tweet posted by a user to predict her location. In order to make the problem more specific, let us consider all the tweets including #superbowl, posted by the users whose specified location is either in the state of Washington or Massachusetts.”

Therefore, we split the part is two subpart, the two as follows:

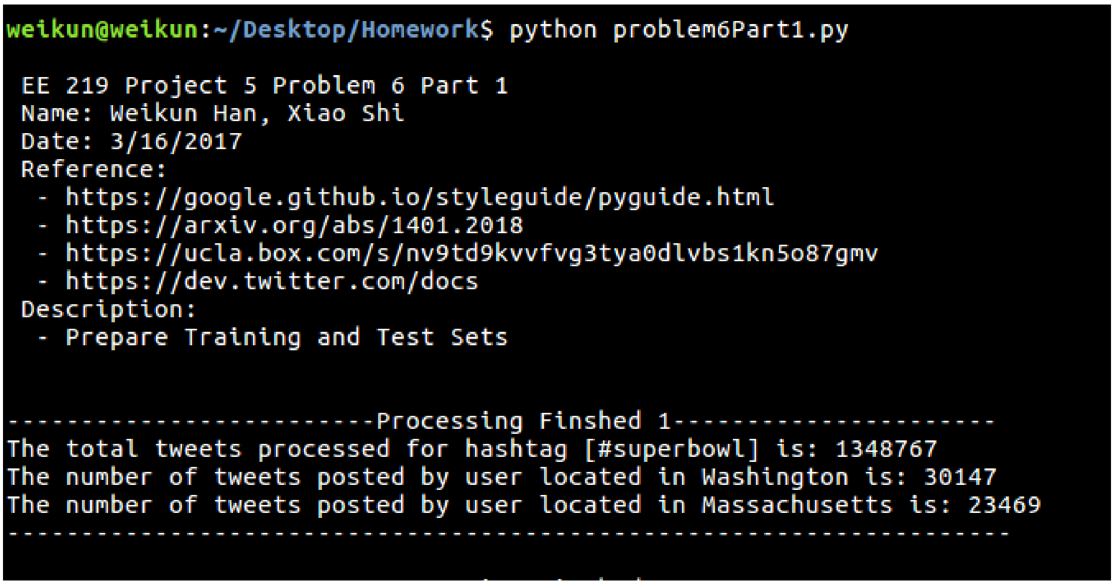
- **First, we need filter some tweet in hashtag #superbowl to make all data is belong to Washington or Massachusetts**
- **Second, we need create the model to make prediction**

We start at first, we want to know how many tweets are in hashtag #superbowl, and then we want to know how many tweets are belong to state of Washington or Massachusetts”, which we can use twitter Developer Documentation python API is:

- **Location = tweet[“tweet”][“user”][“location”]**



```
Boston, Mass.
Boston / Atlanta
Boston - Los Angeles
Foxborough, MA
Hopkinton, MA
Massachusetts
Austin via Boston
Boston MA
Seattle, WA
Boston MA
Boston
Three Rivers, Massachusetts
Boston, MA
Boston, MA
Boston, MA USA
Massachusetts
Seattle/San Fran
Puyallup, WA
Avon, MA
Three Rivers, Massachusetts
Boston, MA
Watertown, MA
Boston & San Diego
Massachusetts-Georgia
Shrewsbury, MA
Boston, MA
Boston, MA
Springfield, MA
Boston, MA
Dublin, GA/Boston, MA
Dennis, CapeCod, MA
Boston
Boston/Nashville
Boston
Boston, MA
Boston
Seattle,Wa
Boston, MA
Massachusetts
Washington, D.C.
Washington, DC
Cambridge, MA
Massachusetts/ ✈
West Seattle
Marblehead, MA
SF/Boston/NYC
East Sandwich, Massachusetts
Boston, MA
Lexington, MA
Boston, MA
Boston TO L.A. (World*Wide)
Washington, DC
Burbs of Boston
Sandwich, MA
FIFE, WASHINGTON
Bass River Massachusetts
Massachusetts
```



```
weikun@weikun:~/Desktop/Homework$ python problem6Part1.py

 EE 219 Project 5 Problem 6 Part 1
 Name: Weikun Han, Xiao Shi
 Date: 3/16/2017
 Reference:
  - https://google.github.io/styleguide/pyguide.html
  - https://arxiv.org/abs/1401.2018
  - https://ucla.box.com/s/nv9td9kvvfvg3tya0dlvbs1kn5o87gmv
  - https://dev.twitter.com/docs
 Description:
  - Prepare Training and Test Sets


-------------------------Processing Finshed 1----------------------
The total tweets processed for hashtag [#superbowl] is: 1348767
The number of tweets posted by user located in Washington is: 30147
The number of tweets posted by user located in Massachusetts is: 23469
-------------------------------------------------------------------
```

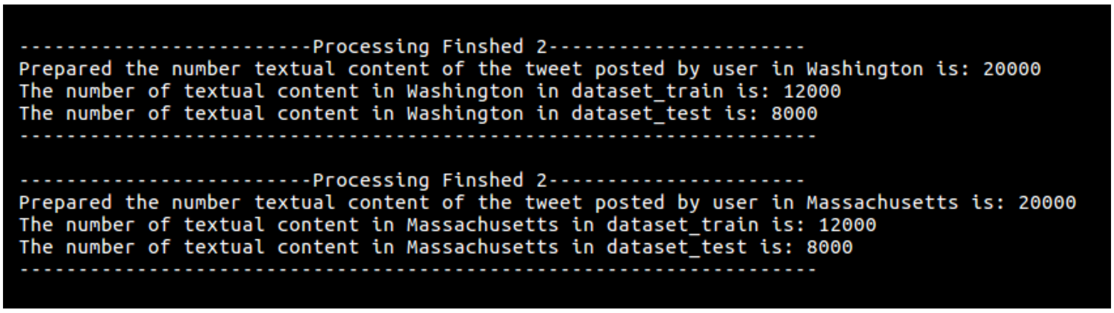
Once we successfully know how many tweets in hashtag #superbowl and we know how many belong to Washington and Massachusetts. Next, we use Twitter Developer Documentation python API is:

- **textualcontent = tweet[“tweet”][“text”]**

to get the textual content of the tweet posted by a user in location Washington or Massachusetts.

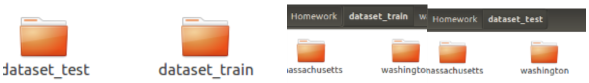
To make the creating model much simple, we still need work a lot in those dataset. Therefore, we want to create is same as the 20 Newsgroups data set. In the 20 Newsgroups data set, we already familiar with 20news-bydate.tar.gz - 20 Newsgroups sorted by date; duplicates and some headers removed (18846 documents). This dataset is is sorted by date into training(60%) and test(40%) sets, does not include cross-posts (duplicates) and does not include newsgroup-identifying headers (Xref, Newsgroups, Path, Followup-To, Date).

Therefore, if we can create the same dataset as above, it will help us much easy to deal with the problem is question. Since we already practice a lot in project 2, we do really want make the dataset like that. Therefore, after lot of trying, we can the dataset as blow:

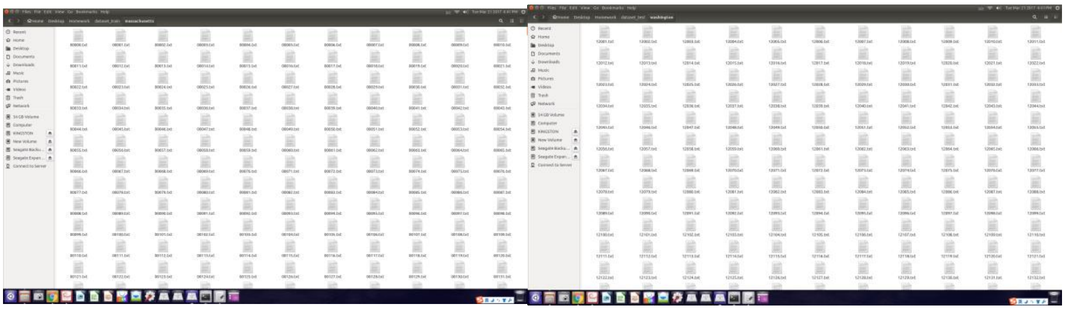


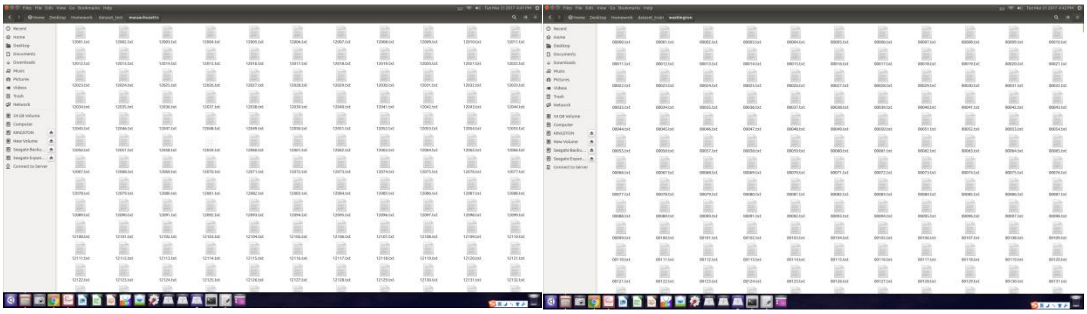
```
------------------------Processing Finshed 2---------------------
Prepared the number textual content of the tweet posted by user in Washington is: 20000
The number of textual content in Washington in dataset_train is: 12000
The number of textual content in Washington in dataset_test is: 8000
-------------------------------------------------------------------

------------------------Processing Finshed 2---------------------
Prepared the number textual content of the tweet posted by user in Massachusetts is: 20000
The number of textual content in Massachusetts in dataset_train is: 12000
The number of textual content in Massachusetts in dataset_test is: 8000
-------------------------------------------------------------------
```

Here, we can keep check the dataset we get, which is same as the 20 Newsgroups data set folder path.



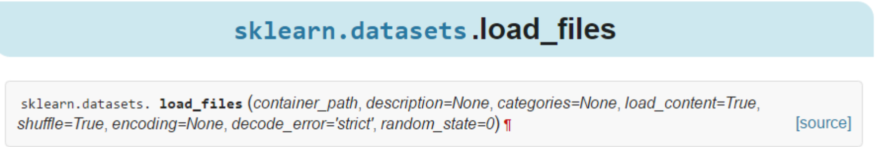
Next, we check data.txt under each folder.





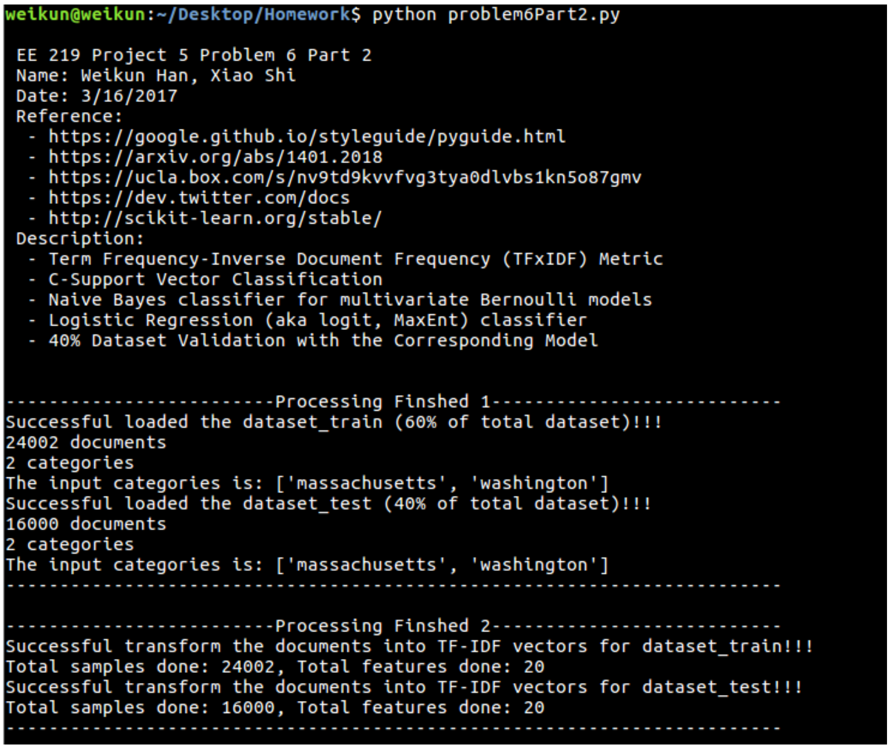
Therefore, with is perfect select dataset, it can much easy help us to do classification.

In part 2, we use we use scikit-learn python API. To load data, is API is show as blow:



`sklearn.datasets`.load_files

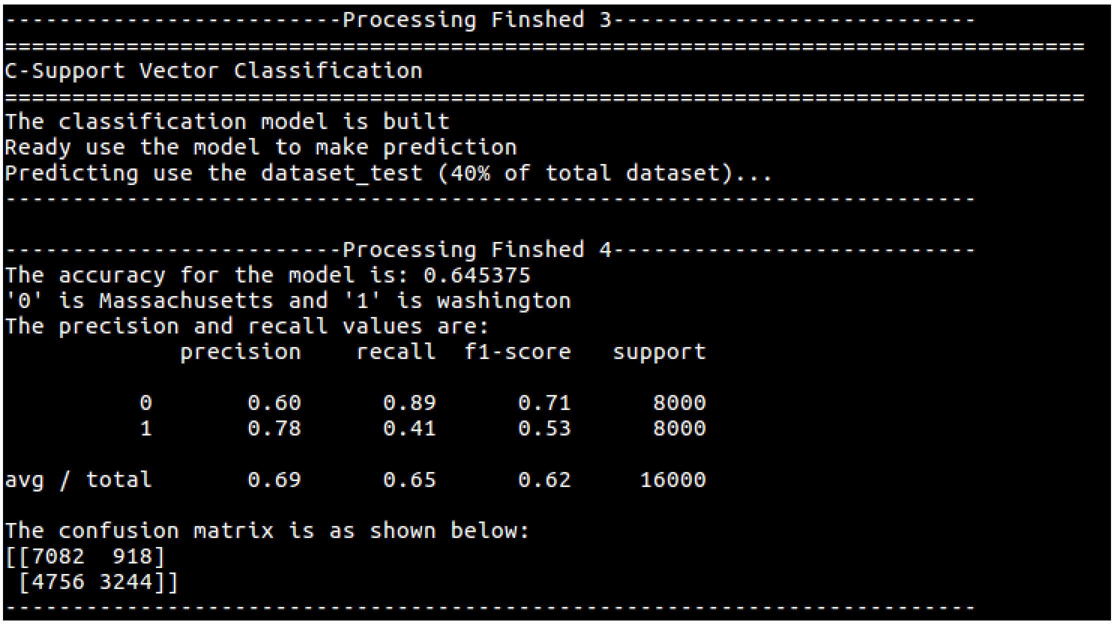
sklearn.datasets. **load_files** (*container_path*, *description=None*, *categories=None*, *load_content=True*, *shuffle=True*, *encoding=None*, *decode_error='strict'*, *random_state=0*) ¶ [source]



```
weikun@weikun:~/Desktop/Homework$ python problem6Part2.py

 EE 219 Project 5 Problem 6 Part 2
 Name: Weikun Han, Xiao Shi
 Date: 3/16/2017
 Reference:
  - https://google.github.io/styleguide/pyguide.html
  - https://arxiv.org/abs/1401.2018
  - https://ucla.box.com/s/nv9td9kvvfvg3tya0dlvbs1kn5o87gmv
  - https://dev.twitter.com/docs
  - http://scikit-learn.org/stable/
 Description:
  - Term Frequency-Inverse Document Frequency (TFxIDF) Metric
  - C-Support Vector Classification
  - Naive Bayes classifier for multivariate Bernoulli models
  - Logistic Regression (aka logit, MaxEnt) classifier
  - 40% Dataset Validation with the Corresponding Model


-------------------------Processing Finshed 1-------------------------
Successful loaded the dataset_train (60% of total dataset)!!!
24002 documents
2 categories
The input categories is: ['massachusetts', 'washington']
Successful loaded the dataset_test (40% of total dataset)!!!
16000 documents
2 categories
The input categories is: ['massachusetts', 'washington']
----------------------------------------------------------------------

-------------------------Processing Finshed 2-------------------------
Successful transform the documents into TF-IDF vectors for dataset_train!!!
Total samples done: 24002, Total features done: 20
Successful transform the documents into TF-IDF vectors for dataset_test!!!
Total samples done: 16000, Total features done: 20
----------------------------------------------------------------------
```

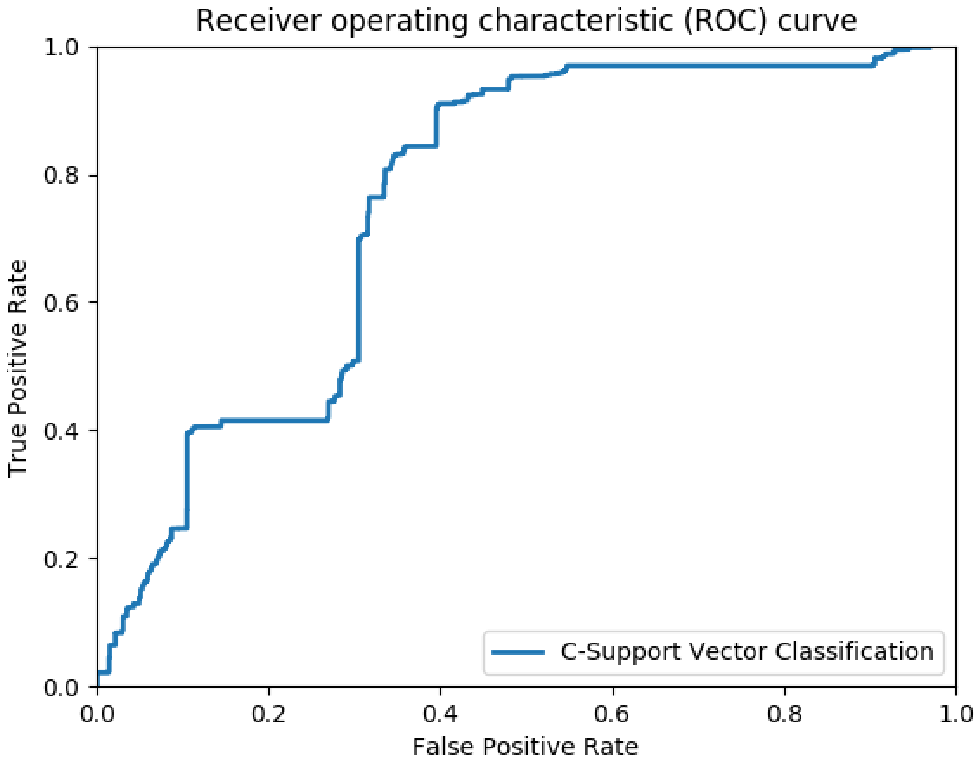
And then we use same scikit-learn python API in project 2 to do traing and predication. Therefore, we get the reuslt as follows:

For the C-Support Vector Classification

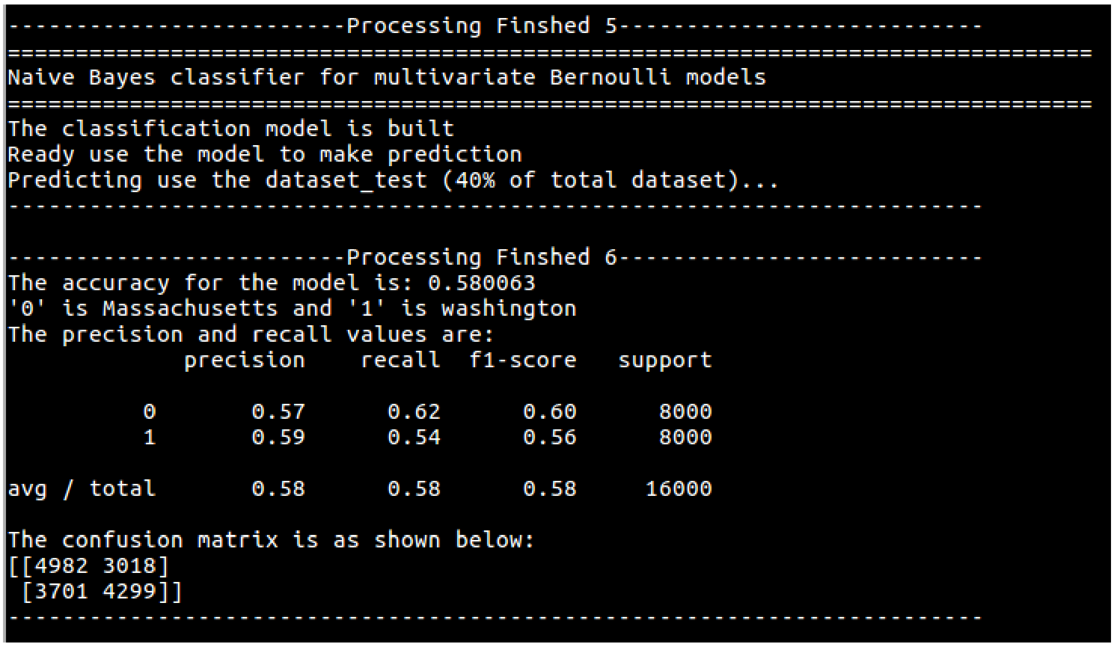


```
-------------------------Processing Finshed 3--------------------------
================================================================================
C-Support Vector Classification
================================================================================
The classification model is built
Ready use the model to make prediction
Predicting use the dataset_test (40% of total dataset)...
------------------------------------------------------------------------

-------------------------Processing Finshed 4--------------------------
The accuracy for the model is: 0.645375
'0' is Massachusetts and '1' is washington
The precision and recall values are:
             precision    recall  f1-score   support

          0       0.60      0.89      0.71      8000
          1       0.78      0.41      0.53      8000

avg / total       0.69      0.65      0.62     16000

The confusion matrix is as shown below:
[[7082  918]
 [4756 3244]]
------------------------------------------------------------------------
```

The accuracy is 0.6464
Precision and recall see result above
Confusion matrixes see result above
ROC is blow:



For Naive Bayes classifier for multivariate Bernoulli models



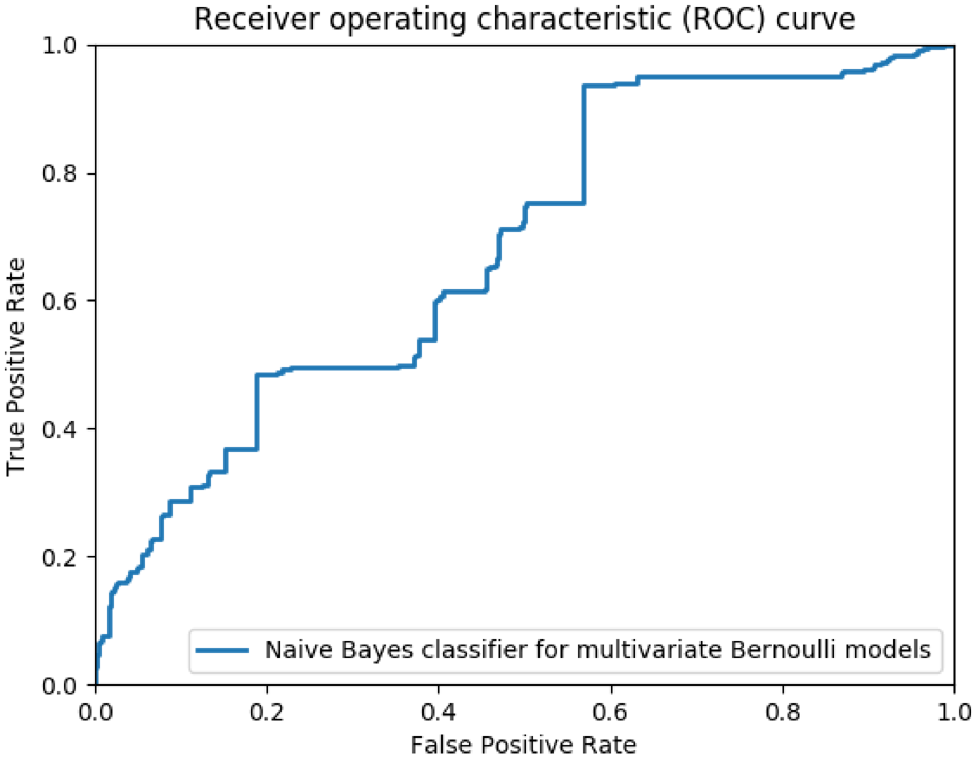
```
------------------------Processing Finshed 5-------------------------
===============================================================================
Naive Bayes classifier for multivariate Bernoulli models
===============================================================================
The classification model is built
Ready use the model to make prediction
Predicting use the dataset_test (40% of total dataset)...
----------------------------------------------------------------------

------------------------Processing Finshed 6-------------------------
The accuracy for the model is: 0.580063
'0' is Massachusetts and '1' is washington
The precision and recall values are:
             precision    recall  f1-score   support

          0       0.57      0.62      0.60      8000
          1       0.59      0.54      0.56      8000

avg / total       0.58      0.58      0.58     16000

The confusion matrix is as shown below:
[[4982 3018]
 [3701 4299]]
----------------------------------------------------------------------
```

The accuracy is 0.58

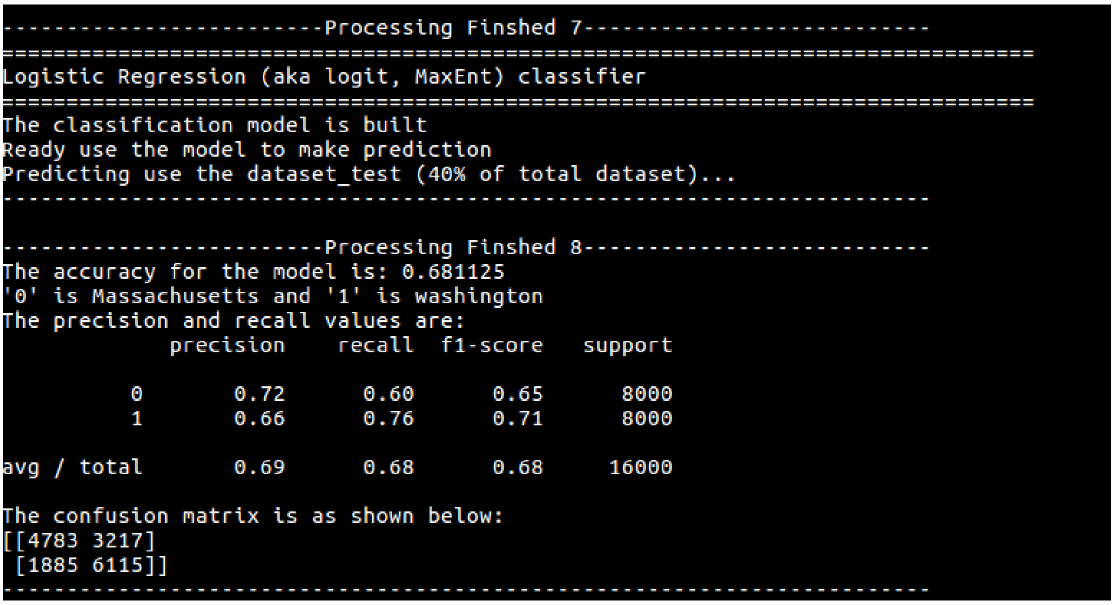
Precision and recall see result above

Confusion matrixes see result above

ROC is blow:



For Logistic Regression (aka logit, MaxEnt) classifier



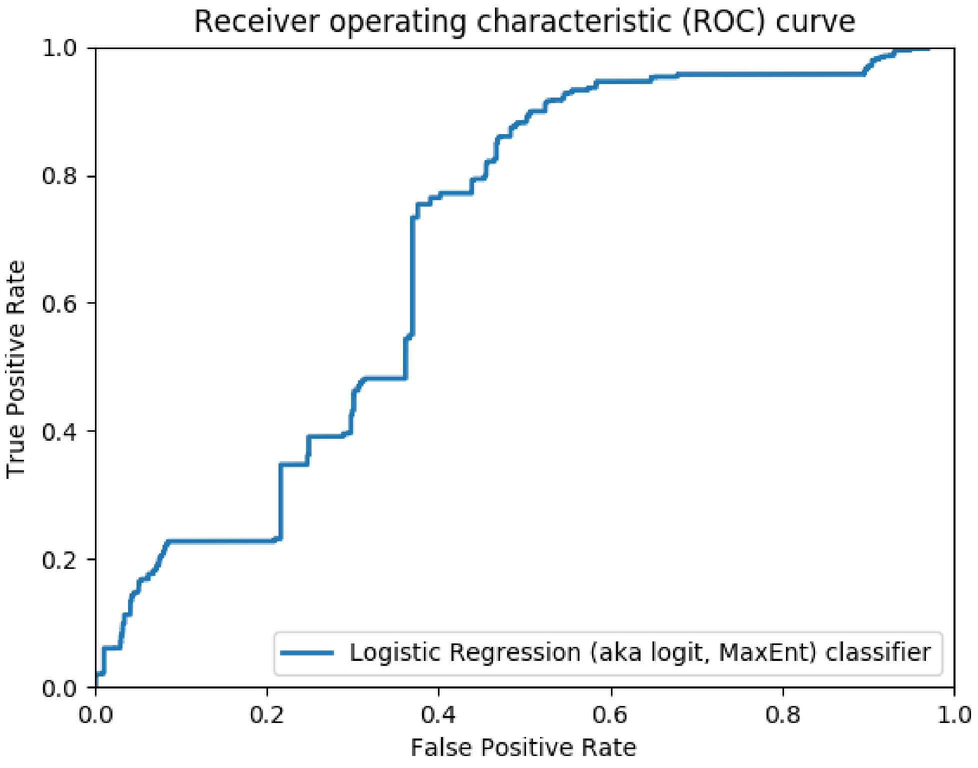
```
-------------------------Processing Finshed 7--------------------------
================================================================================
Logistic Regression (aka logit, MaxEnt) classifier
================================================================================
The classification model is built
Ready use the model to make prediction
Predicting use the dataset_test (40% of total dataset)...
------------------------------------------------------------------------

-------------------------Processing Finshed 8--------------------------
The accuracy for the model is: 0.681125
'0' is Massachusetts and '1' is washington
The precision and recall values are:
             precision    recall  f1-score   support

          0       0.72      0.60      0.65      8000
          1       0.66      0.76      0.71      8000

avg / total       0.69      0.68      0.68     16000

The confusion matrix is as shown below:
[[4783 3217]
 [1885 6115]]
------------------------------------------------------------------------
```

The accuracy is 0.68

Precision and recall see result above

Confusion matrixes see result above

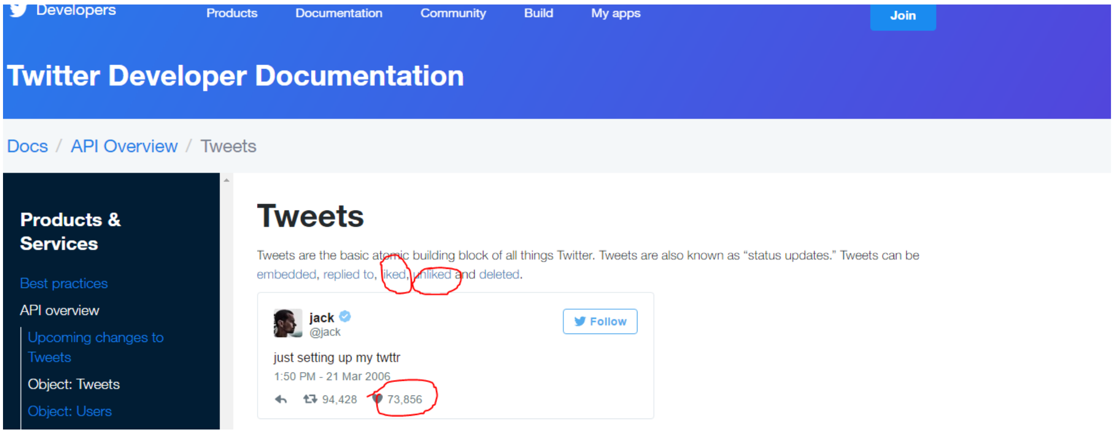
ROC is blow:



In the end, we can get conclusion that the different classification model can get different accuracy. Second, we can the features size using TF-IDT will also impact the accuracy. Third, Using regulation (L1 or L2) also will cause accuracy change.

7 SENTIMENT PREDICTIONS

In this part, we need create our own project. Since we have practice in problem 6, we can predicate the location. That can be part as we want to do in our project. Also, we can see that



Twitter Developer Documentation

Tweets

Tweets are the basic atomic building block of all things Twitter. Tweets are also known as "status updates." Tweets can be embedded, replied to, liked, unliked and deleted.

jack @jack

just setting up my twttr

1:50 PM - 21 Mar 2006

94,428 73,856

Why not we make some sentiment combine with location detection?

Therefore, we want to do the location sentiment data analyze. It can help people to see with city have positive attitude or negative attitude for special thing. That is really meaningful!. If we can do it, we can predict like U.S. presidential election. And we can predict which state will win or not.

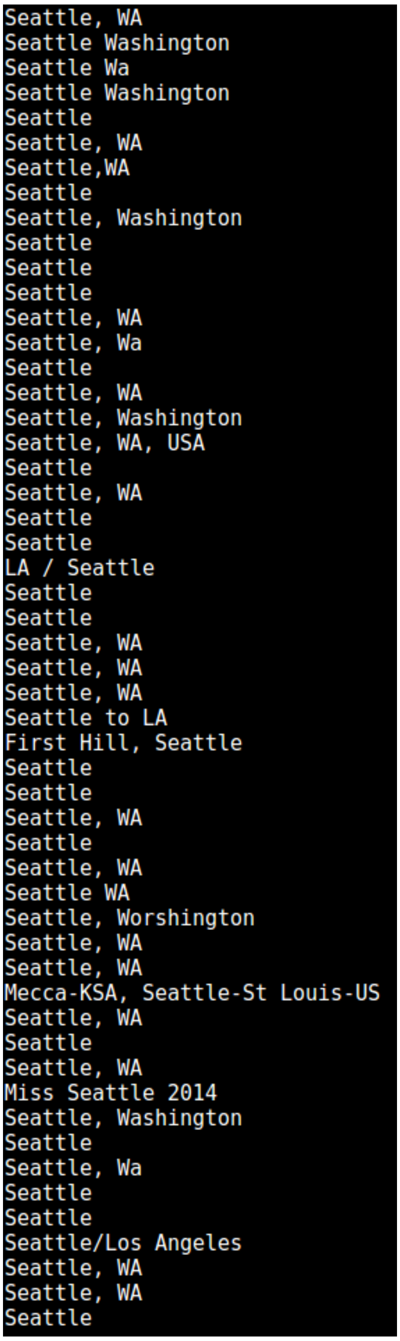
How we do it? Because there are lot of API to select sentiment word from document, we can use TextBlob python API to do it.

“TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.”

Next, we need choose what data we want use. Because we want to use people attitude to predict which team will win in the final. Therefore, if we want use the dataset tweets_#gohawks, then we can use Settle to analyze it because the Seattle Seahawks are a professional American football franchise based in Seattle, Washington.

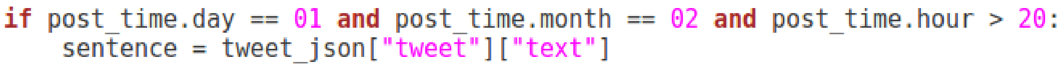
Therefore, if people in Settle have much more positive sentiment, the Seattle Seahawks are more possible to win. If people in Settle have much more negative sentiment, the Seattle Seahawks are more possible to loose.

As same, if we want predict the location sentiment, we can use same method in last problem to select user location who posted tweet. And let us use it to select for the city Settle.



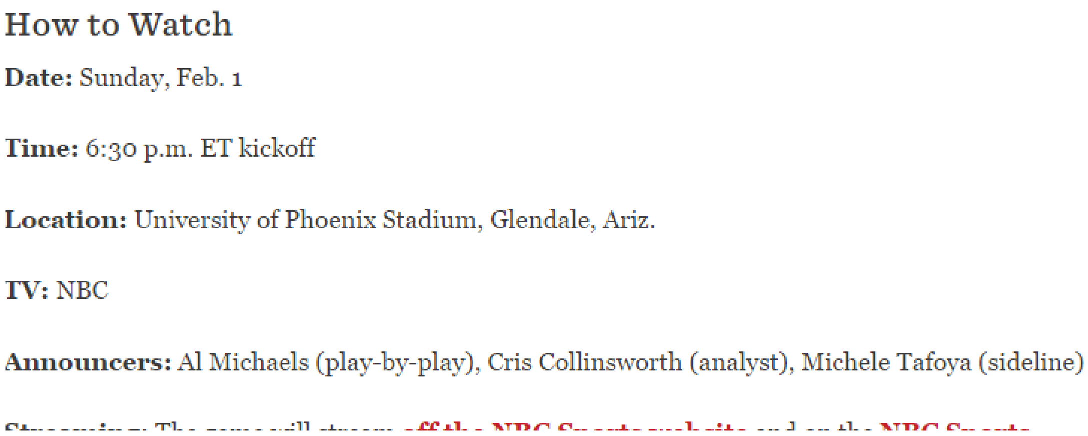
```
Seattle, WA
Seattle Washington
Seattle Wa
Seattle Washington
Seattle
Seattle, WA
Seattle,WA
Seattle
Seattle, Washington
Seattle
Seattle
Seattle
Seattle, WA
Seattle, Wa
Seattle
Seattle, WA
Seattle, Washington
Seattle, WA, USA
Seattle
Seattle, WA
Seattle
Seattle
LA / Seattle
Seattle
Seattle
Seattle, WA
Seattle, WA
Seattle, WA
Seattle to LA
First Hill, Seattle
Seattle
Seattle
Seattle, WA
Seattle
Seattle, WA
Seattle WA
Seattle, Worshington
Seattle, WA
Seattle, WA
Mecca-KSA, Seattle-St Louis-US
Seattle, WA
Seattle
Seattle, WA
Miss Seattle 2014
Seattle, Washington
Seattle
Seattle, Wa
Seattle
Seattle
Seattle/Los Angeles
Seattle, WA
Seattle, WA
Seattle
```

After, we prepare data, we can use same way to get get the textual content of the tweet posted by a user in location Seattle. Next, we install TextBlob python API to do analyze. Because of local user limitation, we cannot train much data. Therefore, we select data as blow:



```
if post_time.day == 01 and post_time.month == 02 and post_time.hour > 20:
    sentence = tweet_json["tweet"]["text"]
```

This code can only pick tweet from 02/01/2015 after 8 p.m. data. The reason we choose is data is because



How to Watch

Date: Sunday, Feb. 1

Time: 6:30 p.m. ET kickoff

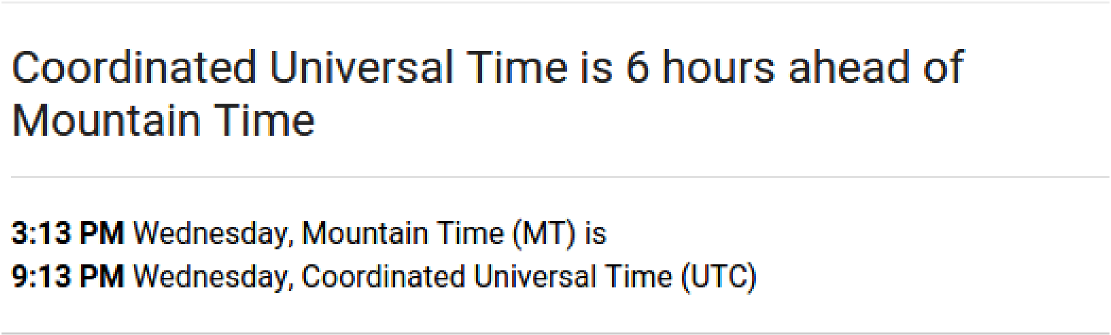
Location: University of Phoenix Stadium, Glendale, Ariz.

TV: NBC

Announcers: Al Michaels (play-by-play), Cris Collinsworth (analyst), Michele Tafoya (sideline)

Streaming: The game will stream off the NBC Sports website and on the NBC Sports

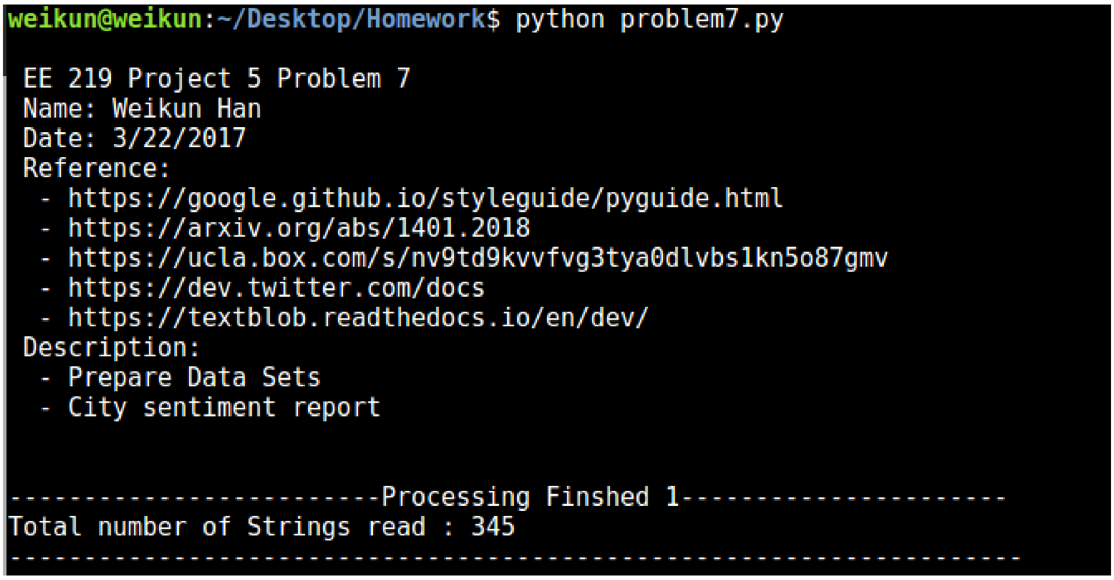
And the time zone difference, the actually start super bowl is 02/02/2015 in UTC.



Coordinated Universal Time is 6 hours ahead of Mountain Time

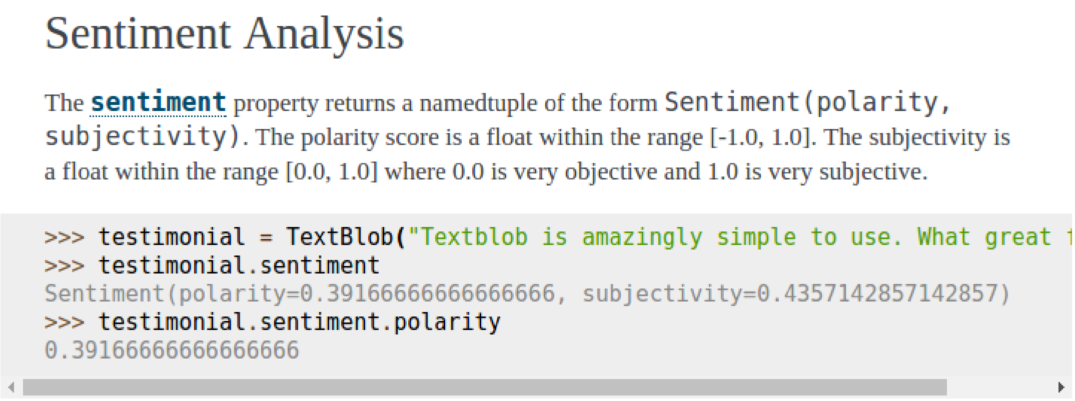
3:13 PM Wednesday, Mountain Time (MT) is
9:13 PM Wednesday, Coordinated Universal Time (UTC)

Now, we select data is before the game and let look up how those data can produce what kind of result.



```
weikun@weikun:~/Desktop/Homework$ python problem7.py

 EE 219 Project 5 Problem 7
 Name: Weikun Han
 Date: 3/22/2017
 Reference:
  - https://google.github.io/styleguide/pyguide.html
  - https://arxiv.org/abs/1401.2018
  - https://ucla.box.com/s/nv9td9kvvfvg3tya0dlvbs1kn5o87gmv
  - https://dev.twitter.com/docs
  - https://textblob.readthedocs.io/en/dev/
 Description:
  - Prepare Data Sets
  - City sentiment report


-------------------------Processing Finshed 1----------------------
Total number of Strings read : 345
-------------------------------------------------------------------
```

Here, we can get total string to be classification positive and negative sentiment. And we use

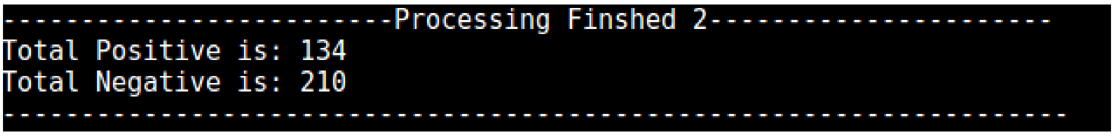


Sentiment Analysis

The `sentiment` property returns a namedtuple of the form `Sentiment(polarity, subjectivity)`. The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

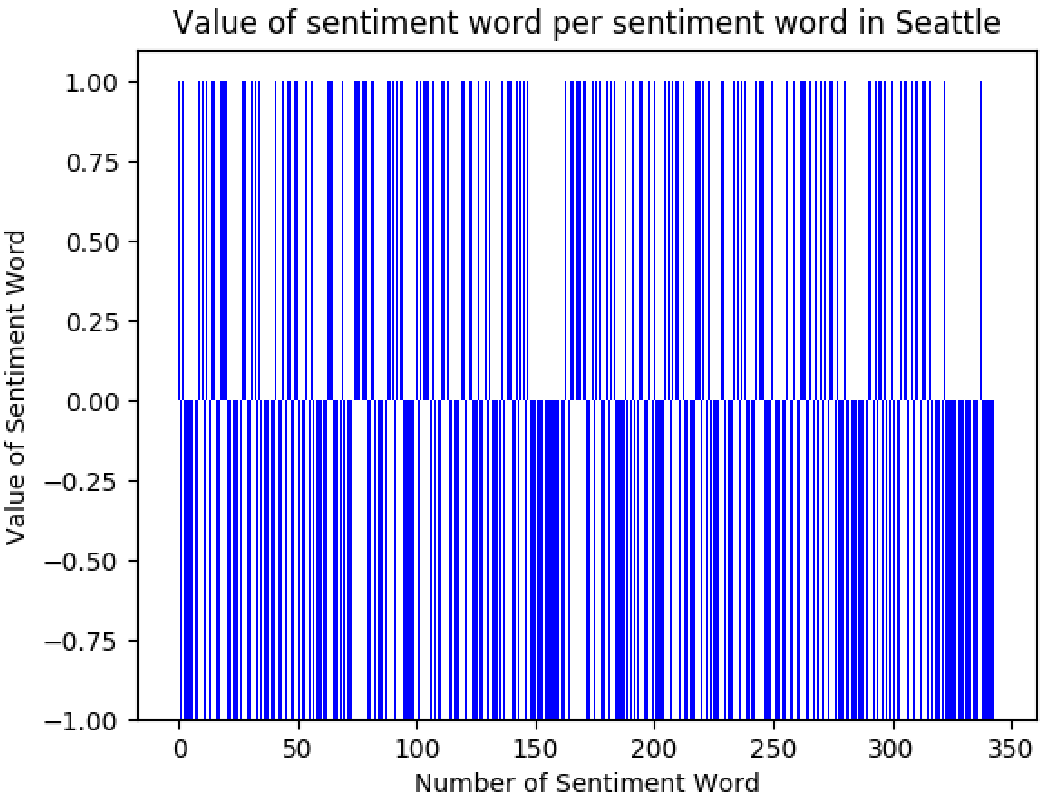
```
>>> testimonial = TextBlob("Textblob is amazingly simple to use. What great t
>>> testimonial.sentiment
Sentiment(polarity=0.39166666666666666, subjectivity=0.4357142857142857)
>>> testimonial.sentiment.polarity
0.39166666666666666
```

To get result as follows:



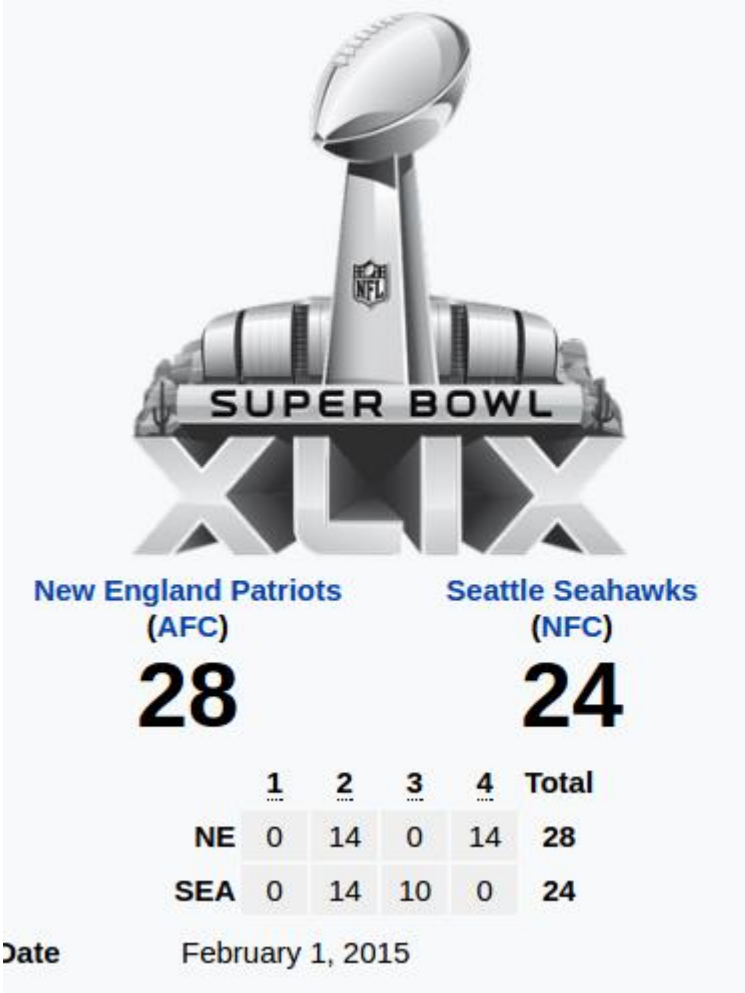
```
------------------------Processing Finshed 2----------------------
Total Positive is: 134
Total Negative is: 210
-------------------------------------------------------------------
```

We can also plot is result, which is more clearly to read. Is value is 1 means positive and -1 means negative.



Therefore, people in Settle have much more negative sentiment, the Seattle Seahawks are more possible to loose.

And let take look at the result



New England Patriots (AFC) 28

Seattle Seahawks (NFC) 24

	1	2	3	4	Total
NE	0	14	0	14	28
SEA	0	14	10	0	24

Date February 1, 2015

Now, we can confirm our result is correctly.

Since we select data period and hashtag is like we pick special features. The predict result may be exit fortuity. However, if we train much data not only in single period, the result can be much more reliable.

In the end, from the result, we can get people a real time feedback to make prople predict who going to win. Like I said in begin, we can predict like U.S. presidential election. And we can predict which state will win or not. Secondly, we can use it to predict like stock price, if people use some social media to show their attitude.

8 REFERENCE

[1] https://ucla.box.com/s/nv9td9kvvfvg3tya0dlvbs1kn5o87gmv

[2] "Using Twitter with Your Phone". Twitter Support. Retrieved June 1, 2010. We currently support 2-way (sending and receiving) Twitter SMS via short codes and one-way (sending only) via long codes.

[3] Stone, Biz (October 30, 2009). "There's a List for That". blog.twitter.com. Retrieved February 1, 2010.

[4] "Twitter officially kills off favorites and replaces them with likes". The Verge. Vox Media. Retrieved November 4, 2015.

[5] Strachan, Donald (February 19, 2009). "Twitter: How To Set Up Your Account". The Daily Telegraph. London. Retrieved February 13, 2011.

[6] https://ucla.box.com/s/ojvvthudugp9d2gze5nuep9ogwjydnur