Project5 Report

Cheng Ma 105033453, **Jinxi Zou** 605036454,
Shuo Bai 505032786, **Xiaoxi Gong 705034355**

March 20, 2018

1 introduction

For this project, useful practice in social network analysis is to predict future popularity of a subject or event. Twitter, with its public discussion model, is a good platform to perform such analysis. 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. We should use data to train regression model to make prediction for other hash tag set. For the test set, we used trained model to predict the number of tweets after the given time window. And we could infer from the text of the tweet to analyze where the author from. At last, we should define our own work and show how it works. We decided to analyze the sentiment of the tweet to know which team won the game.

2 Popularity Prediction

2.1 Problem 1.1

In this part, we roughly get some basic statistics from the given data, and plot the number of tweets in hour over time for #SuperBowl and #NFL.

Table 1: statistics from the given data

tags	number of tweets per hour	avg number of followers	avg number of retweets
gohawks	325.49	2203.93	0.21
gopatriots	45.70	1401.90	0.03
sb49	1420.88	10267.32	0.18
superbowl	2301.65	8858.97	0.14
nfl	442.02	4653.25	0.05
patriots	835.69	3309.98	0.09

And we can also get the number of tweets in hour over time for #SuperBowl and #NFL easily. The plots are shown below.

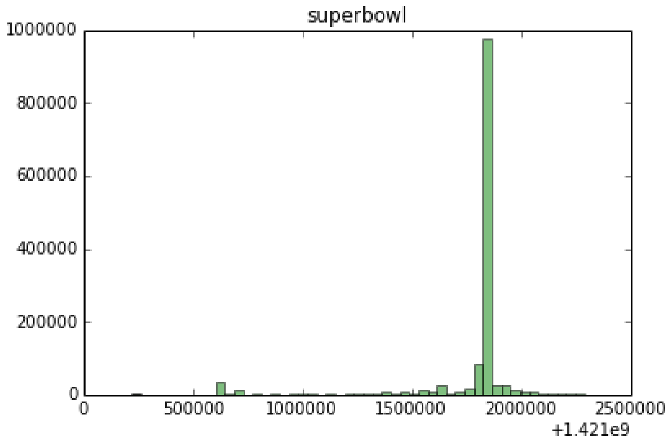


Figure 1: superbowl: the number of tweets in hour

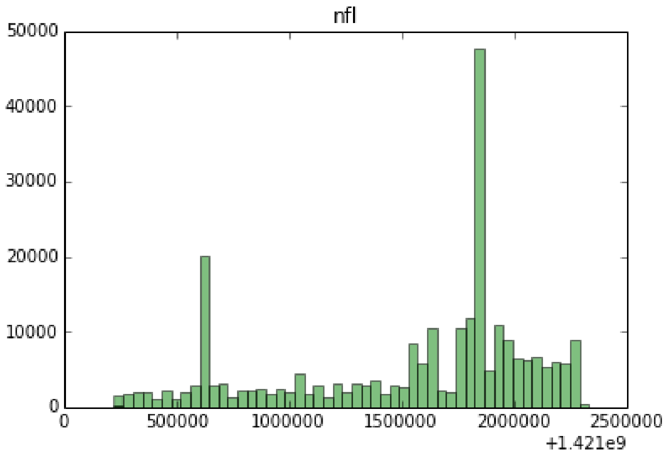


Figure 2: nfl: the number of tweets in hour

2.2 Problem 1.2

In this problem, the main task is to perform a Linear Regression model with five feature to make a prediction of the total tweets in next hour. We use 1 hour as the time window. We make prediction for next hour by the features from the previous one hour window. We should compute the following five values for each kind of hashtag.

Table 2: Feature and Compute

Total Tweets	Index sum in an hour
Total Retweets	Sum of [metrics][citations][total]
Total Followers	Sum of [author][followers]
Max Followers	Max of [author][followers]
Time	datetime(citation_date)

And for these five hashtags, we use the same way to anlyze them. We will show the result respectively.

2.2.1 GoHawks

For GoHawks tweets. we use linear regression and get the distribution plot as following

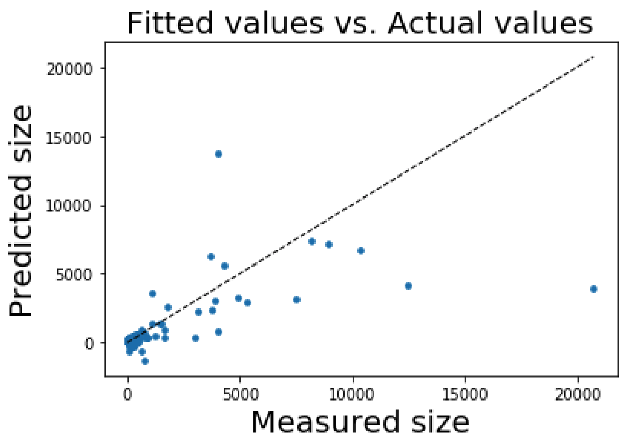


Figure 3: GoHawks distributed plot

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.500
Model:                            OLS   Adj. R-squared:                  0.496
Method:                 Least Squares   F-statistic:                     113.4
Date:                Mon, 12 Mar 2018   Prob (F-statistic):           6.70e-83
Time:                        21:22:31   Log-Likelihood:                -4743.0
No. Observations:                 571   AIC:                             9496.
Df Residuals:                     566   BIC:                             9518.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.2314      0.171      7.206      0.000       0.896       1.567
x2            -0.1286      0.044     -2.899      0.004      -0.216      -0.041
x3            -0.0002   8.57e-05     -2.043      0.041      -0.000   -6.78e-06
x4          2.793e-05      0.000      0.173      0.863      -0.000       0.000
x5             8.8307      3.332      2.650      0.008       2.286      15.375
==============================================================================
Omnibus:                      897.585   Durbin-Watson:                   2.220
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           749641.637
Skew:                           8.522   Prob(JB):                         0.00
Kurtosis:                     179.686   Cond. No.                     2.33e+05
==============================================================================
```

Figure 4: GoHawks distributed plot

R-squared: **0.4724817508090412**
RMSE:**957184.6939752336**

2.2.2 GoPatriots

For GoPatriots tweets. we use linear regression and get the distribution plot as following:

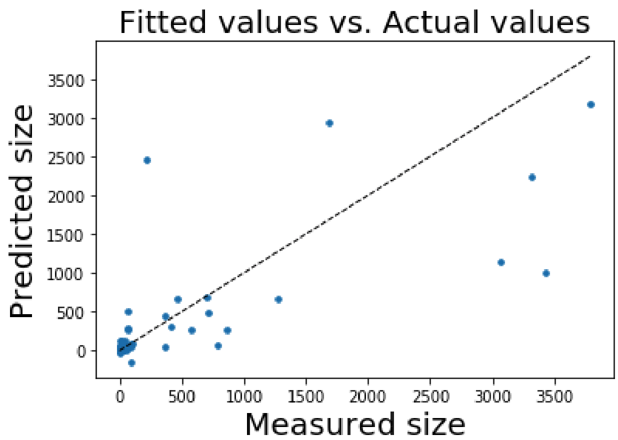


Figure 5: GoPatriots distributed plot

OLS Regression Results

Dep. Variable:	y	R-squared:	0.640
Model:	OLS	Adj. R-squared:	0.635
Method:	Least Squares	F-statistic:	156.5
Date:	Mon, 12 Mar 2018	Prob (F-statistic):	2.44e-95
Time:	21:44:14	Log-Likelihood:	-3017.5
No. Observations:	446	AIC:	6045.
Df Residuals:	441	BIC:	6065.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.0788	0.290	-0.272	0.786	-0.649	0.491
x2	0.5012	0.254	1.971	0.049	0.001	1.001
x3	0.0003	0.000	1.112	0.267	-0.000	0.001
x4	-0.0004	0.000	-1.699	0.090	-0.001	6e-05
x5	0.6618	0.787	0.841	0.401	-0.886	2.209

Omnibus:	367.255	Durbin-Watson:	1.951
Prob(Omnibus):	0.000	Jarque-Bera (JB):	138888.641
Skew:	2.403	Prob(JB):	0.00
Kurtosis:	89.318	Cond. No.	3.75e+04

Figure 6: GoPatriots distributed plot

R-squared **0.6298299526540944**
RMSE **44019.54600896945**

2.2.3 NFL

For NFL tweets. we use linear regression and get the distribution plot as following:

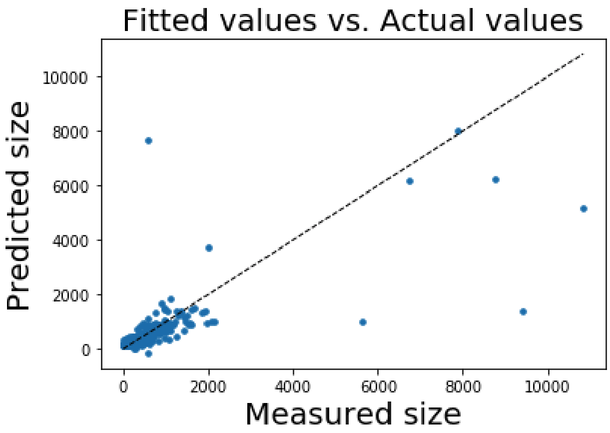


Figure 7: NFL distributed plot

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.644
Method:                 Least Squares   F-statistic:                     211.7
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          5.69e-128
Time:                        22:13:19   Log-Likelihood:                -4536.2
No. Observations:                 582   AIC:                             9082.
Df Residuals:                     577   BIC:                             9104.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.7393      0.134      5.527      0.000       0.477       1.002
x2            -0.1781      0.064     -2.773      0.006      -0.304      -0.052
x3          7.903e-05   2.64e-05      2.994      0.003    2.72e-05       0.000
x4         -7.276e-05   3.61e-05     -2.016      0.044      -0.000   -1.87e-06
x5             7.5271      2.209      3.407      0.001       3.188      11.867
==============================================================================
Omnibus:                      562.394   Durbin-Watson:                   2.326
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           341811.410
Skew:                           3.267   Prob(JB):                         0.00
Kurtosis:                     121.544   Cond. No.                     4.26e+05
==============================================================================
```

Figure 8: NFL distributed plot

R-squared **0.5637304797580409**
RMSE **340255.5355600498**

2.2.4 Patriots

For Patriots tweets. we use linear regression and get the distribution plot as following:

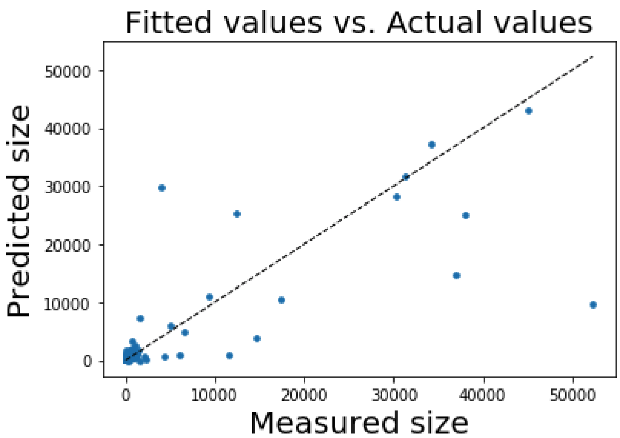


Figure 9: Patriots distributed plot

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.681
Model:                            OLS   Adj. R-squared:                  0.678
Method:                 Least Squares   F-statistic:                     247.8
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          1.88e-141
Time:                        22:19:47   Log-Likelihood:                -5422.9
No. Observations:                 586   AIC:                         1.086e+04
Df Residuals:                     581   BIC:                         1.088e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.9214      0.072     12.878      0.000       0.781       1.062
x2            -0.0870      0.059     -1.476      0.141      -0.203       0.029
x3         -1.184e-06   2.62e-05     -0.045      0.964   -5.27e-05    5.03e-05
x4             0.0002      0.000      1.770      0.077   -1.97e-05       0.000
x5             3.9215      8.737      0.449      0.654     -13.238      21.081
==============================================================================
Omnibus:                      878.853   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           692548.491
Skew:                           7.765   Prob(JB):                         0.00
Kurtosis:                     170.698   Cond. No.                     7.66e+05
==============================================================================
```

Figure 10: Patriots distributed plot

R-squared **0.6696695191747575**
RMSE **6382071.540743866**

2.2.5 SB49

For SB49 tweets. we use linear regression and get the distribution plot as following:

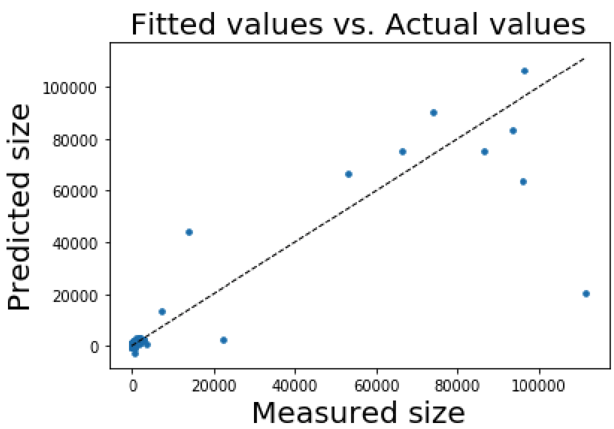


Figure 11: SB49 distributed plot

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.809
Model:                            OLS   Adj. R-squared:                  0.807
Method:                 Least Squares   F-statistic:                     451.8
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          2.26e-189
Time:                        22:27:58   Log-Likelihood:                -5325.5
No. Observations:                 540   AIC:                         1.066e+04
Df Residuals:                     535   BIC:                         1.068e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.1886      0.099     12.031      0.000       0.995       1.383
x2            -0.2151      0.091     -2.362      0.019      -0.394      -0.036
x3          1.869e-05   1.46e-05      1.283      0.200   -9.92e-06    4.73e-05
x4             0.0001   4.95e-05      2.035      0.042    3.47e-06       0.000
x5            -4.0590     16.512     -0.246      0.806     -36.496      28.378
==============================================================================
Omnibus:                     1082.406   Durbin-Watson:                   1.682
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1781152.701
Skew:                          14.109   Prob(JB):                         0.00
Kurtosis:                     282.939   Cond. No.                     7.51e+06
==============================================================================
```

Figure 12: SB49 distributed plot

R-squared **0.8045921985948535**

RMSE **21537350.76006268**

2.2.6 SuperBowl

For SuperBowl tweets. we use linear regression and get the distribution plot as following:

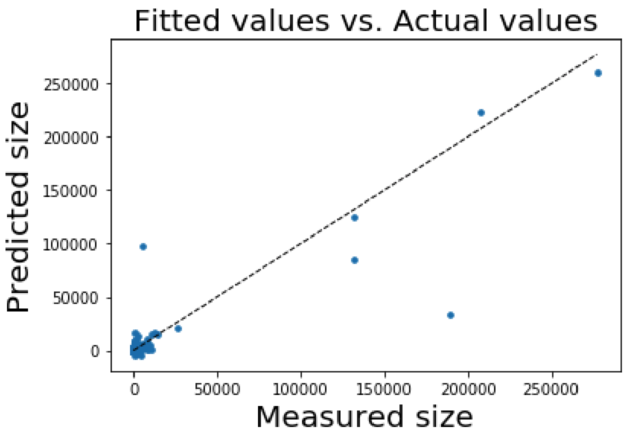


Figure 13: SuperBowl distributed plot

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.804
Method:                 Least Squares   F-statistic:                     480.7
Date:                Mon, 12 Mar 2018   Prob (F-statistic):         9.59e-204
Time:                        22:42:09   Log-Likelihood:                -6098.3
No. Observations:                 586   AIC:                         1.221e+04
Df Residuals:                     581   BIC:                         1.223e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             2.3014      0.079     28.962      0.000       2.145       2.457
x2            -0.2899      0.036     -8.059      0.000      -0.361      -0.219
x3            -0.0001   1.87e-05     -7.020      0.000      -0.000   -9.46e-05
x4             0.0008      0.000      5.457      0.000       0.000       0.001
x5           -38.8599     29.764     -1.306      0.192     -97.319      19.599
==============================================================================
Omnibus:                     1012.645   Durbin-Watson:                   2.317
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1838502.343
Skew:                          10.123   Prob(JB):                         0.00
Kurtosis:                     276.655   Cond. No.                     1.09e+07
==============================================================================
```

Figure 14: SuperBowl distributed plot

R-squared **0.8021867271634405**
RMSE **64056953.49228067**

2.3 Problem 1.3

In this part, we choose five features as the train feature, and they are followers number, favorite_count, citation data, length of the title and the number of the twitters. Different from the citation data in the previous part, we use the data from 1 to 600 instead of mapping the data into 24 hours in a day.

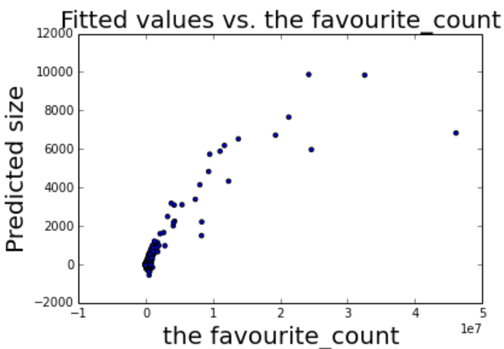
2.3.1 GoHawks

For GoHawks tweets, we use linear regression and get the following statistics.

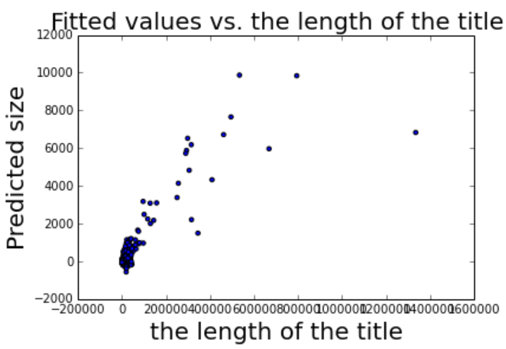
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.641
Model:                            OLS   Adj. R-squared:                  0.638
Method:                 Least Squares   F-statistic:                     204.9
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          5.01e-125
Time:                        19:28:08   Log-Likelihood:                -4702.0
No. Observations:                 578   AIC:                             9414.
Df Residuals:                     573   BIC:                             9436.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0002   4.92e-05     -4.466      0.000      -0.000      -0.000
x2             0.0017      0.000     14.865      0.000       0.001       0.002
x3             0.0184      0.113      0.163      0.870      -0.204       0.241
x4             0.0212      0.004      5.690      0.000       0.014       0.029
x5            -4.3520      0.393    -11.062      0.000      -5.125      -3.579
==============================================================================
Omnibus:                      974.336   Durbin-Watson:                   2.047
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           697873.381
Skew:                          10.022   Prob(JB):                         0.00
Kurtosis:                     172.043   Cond. No.                     4.99e+04
==============================================================================
```

Figure 15: Regression Results

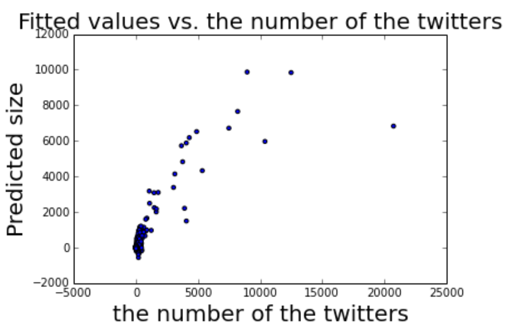
The R-squared is 0.62049. And the top three features are favorite count, length of the tile and the number of twitters. And the scatter plot of predict versus value of the features are given below.



(a)



(b)



(c)

Figure 16: scatter plot of predict versus value of the features

2.3.2 GoPatriots

For GoPatriots tweets, we use linear regression and get the following statistics.

OLS Regression Results

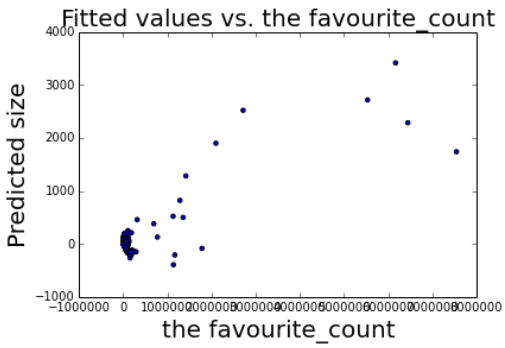
Dep. Variable:	y	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.769
Method:	Least Squares	F-statistic:	383.5
Date:	Mon, 12 Mar 2018	Prob (F-statistic):	1.51e-179
Time:	19:28:10	Log-Likelihood:	-3681.0
No. Observations:	574	AIC:	7372.
Df Residuals:	569	BIC:	7394.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	3.408e-05	4.25e-05	0.801	0.423	-4.94e-05	0.000
x2	-0.0024	0.000	-17.308	0.000	-0.003	-0.002
x3	0.0568	0.025	2.236	0.026	0.007	0.107
x4	0.0136	0.004	3.296	0.001	0.006	0.022
x5	4.2737	0.400	10.697	0.000	3.489	5.058

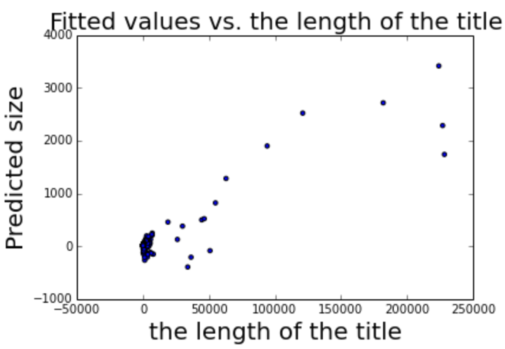
Omnibus:	414.948	Durbin-Watson:	1.848
Prob(Omnibus):	0.000	Jarque-Bera (JB):	218798.303
Skew:	-1.877	Prob(JB):	0.00
Kurtosis:	98.573	Cond. No.	4.50e+04

Figure 17: Regression Results

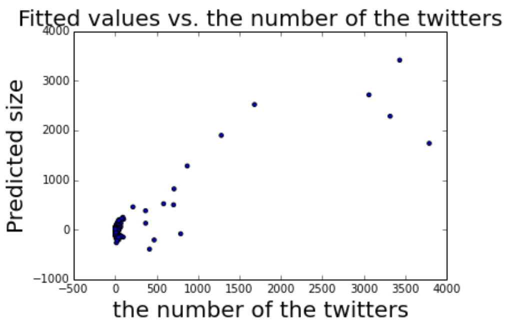
The R-squared is 0.766665. And the top three features are favorite count, length of the tile and the number of twitters. And the scatter plot of predict versus value of the features are given below.



(a)



(b)



(c)

Figure 18: scatter plot of predict versus value of the features

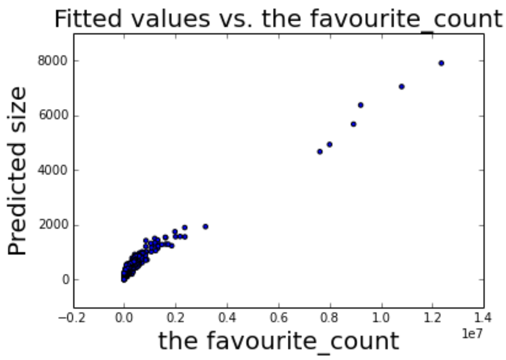
2.3.3 NFL

For NFL tweets, we use linear regression and get the following statistics.

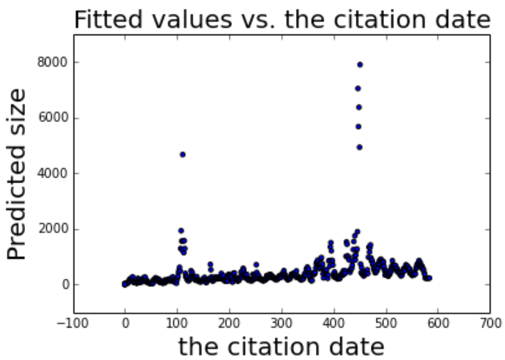
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.651
Model:                            OLS   Adj. R-squared:                  0.648
Method:                 Least Squares   F-statistic:                     216.8
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          2.90e-130
Time:                        19:28:45   Log-Likelihood:                -4562.4
No. Observations:                 586   AIC:                             9135.
Df Residuals:                     581   BIC:                             9157.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          2.136e-05   1.44e-05      1.482      0.139   -6.96e-06    4.97e-05
x2             0.0004      0.000      2.615      0.009       0.000       0.001
x3             0.4042      0.128      3.153      0.002       0.152       0.656
x4             0.0103      0.005      2.106      0.036       0.001       0.020
x5            -0.7964      0.573     -1.390      0.165      -1.922       0.329
==============================================================================
Omnibus:                      531.243   Durbin-Watson:                   2.319
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           310334.235
Skew:                           2.901   Prob(JB):                         0.00
Kurtosis:                     115.589   Cond. No.                     1.11e+05
==============================================================================
```

Figure 19: Regression Results

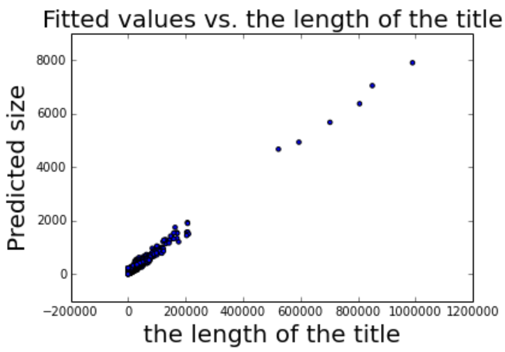
The R-squared is 0.56326. And the top three features are favorite count, length of the tile and the number of twitters. And the scatter plot of predict versus value of the features are given below.



(a)



(b)



(c)

Figure 20: scatter plot of predict versus value of the features

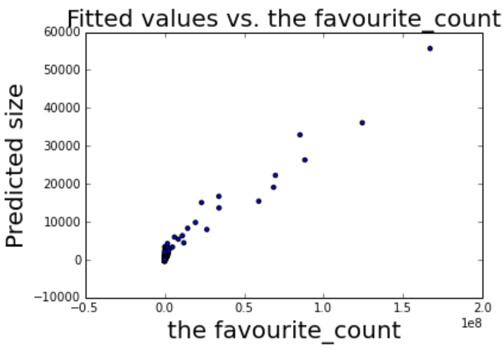
2.3.4 Patriots

For Patriots tweets, we use linear regression and get the following statistics.

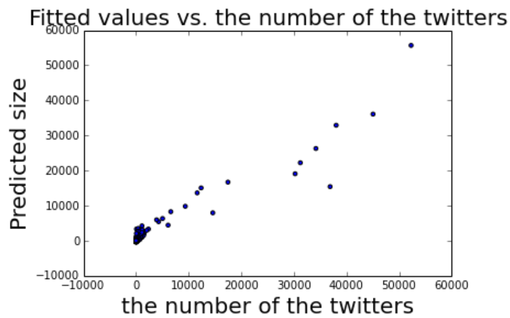
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.733
Model:                            OLS   Adj. R-squared:                  0.731
Method:                 Least Squares   F-statistic:                     319.3
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          4.82e-164
Time:                        19:29:40   Log-Likelihood:                -5370.3
No. Observations:                 586   AIC:                         1.075e+04
Df Residuals:                     581   BIC:                         1.077e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0002   2.45e-05      7.939      0.000       0.000       0.000
x2             0.0006   5.78e-05     10.661      0.000       0.001       0.001
x3            -0.1432      0.323     -0.443      0.658      -0.778       0.492
x4             0.0256      0.007      3.719      0.000       0.012       0.039
x5            -3.0608      0.596     -5.136      0.000      -4.231      -1.890
==============================================================================
Omnibus:                      979.093   Durbin-Watson:                   2.040
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           611846.691
Skew:                           9.924   Prob(JB):                         0.00
Kurtosis:                     160.050   Cond. No.                     8.64e+04
==============================================================================
```

Figure 21: Regression Results

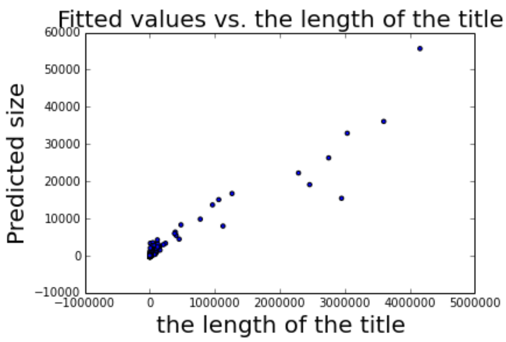
The R-squared is 0.72507. And the top three features are favorite count, length of the tile and the number of twitters. And the scatter plot of predict versus value of the features are given below.



(a)



(b)



(c)

Figure 22: scatter plot of predict versus value of the features

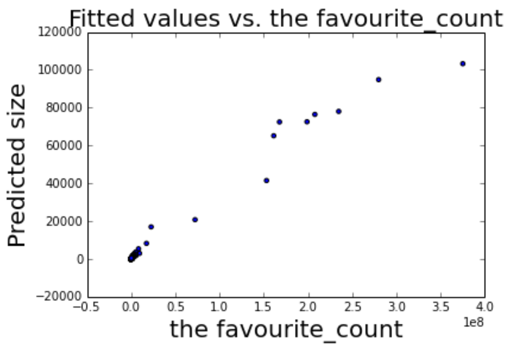
2.3.5 SB49

For SB49 tweets, we use linear regression and get the following statistics.

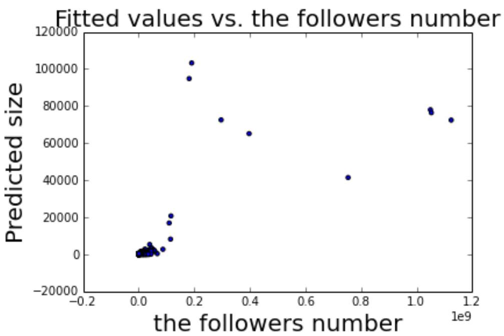
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.807
Model:                            OLS   Adj. R-squared:                  0.805
Method:                 Least Squares   F-statistic:                     482.5
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          2.12e-203
Time:                        19:31:18   Log-Likelihood:                -5720.2
No. Observations:                 582   AIC:                         1.145e+04
Df Residuals:                     577   BIC:                         1.147e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1         -1.812e-05    6.5e-06     -2.789      0.005   -3.09e-05   -5.36e-06
x2            -0.0001   6.66e-05     -1.823      0.069      -0.000     9.4e-06
x3             0.4691      0.621      0.755      0.450      -0.751       1.689
x4             0.0177      0.009      1.886      0.060      -0.001       0.036
x5            -0.0368      0.732     -0.050      0.960      -1.474       1.400
==============================================================================
Omnibus:                     1168.510   Durbin-Watson:                   1.563
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2110184.460
Skew:                          14.290   Prob(JB):                         0.00
Kurtosis:                     296.600   Cond. No.                     3.90e+05
==============================================================================
```

Figure 23: Regression Results

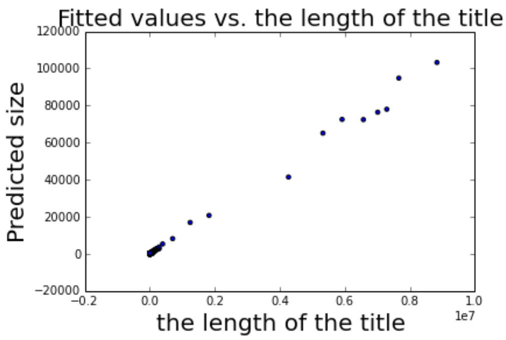
The R-squared is 0.80318. And the top three features are favorite count, length of the tile and the number of twitters. And the scatter plot of predict versus value of the features are given below.



(a)



(b)



(c)

Figure 24: scatter plot of predict versus value of the features

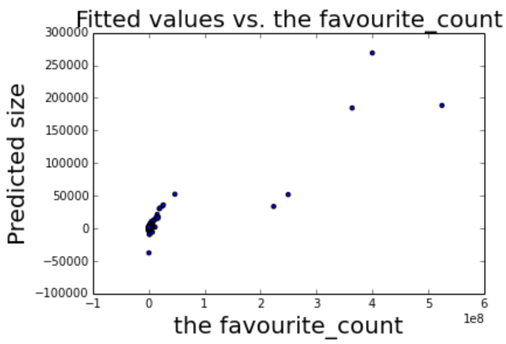
2.3.6 Superbowl

For Superbowl tweets, we use linear regression and get the following statistics.

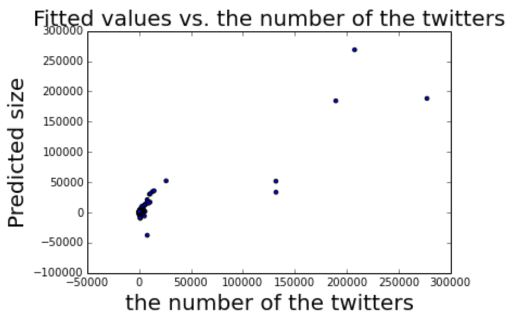
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.817
Model:                            OLS   Adj. R-squared:                  0.815
Method:                 Least Squares   F-statistic:                     517.8
Date:                Mon, 12 Mar 2018   Prob (F-statistic):         2.36e-211
Time:                        19:33:46   Log-Likelihood:                -6080.6
No. Observations:                 586   AIC:                         1.217e+04
Df Residuals:                     581   BIC:                         1.219e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0002   1.27e-05    -16.291      0.000      -0.000      -0.000
x2             0.0027      0.000     10.433      0.000       0.002       0.003
x3             1.7479      1.129      1.548      0.122      -0.470       3.966
x4             0.1114      0.011      9.846      0.000       0.089       0.134
x5           -11.6534      1.056    -11.033      0.000     -13.728      -9.579
==============================================================================
Omnibus:                     1023.353   Durbin-Watson:                   2.150
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           895687.803
Skew:                          10.745   Prob(JB):                         0.00
Kurtosis:                     193.320   Cond. No.                     4.98e+05
==============================================================================
```

Figure 25: Regression Results

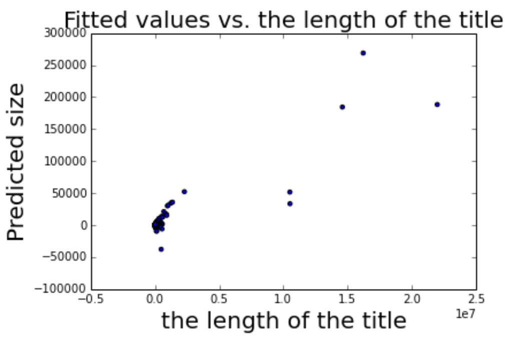
The R-squared is 0.81380. And the top three features are favorite count, length of the tile and the number of twitters. And the scatter plot of predict versus value of the features are given below.



(a)



(b)



(c)

Figure 26: scatter plot of predict versus value of the features

2.3.7 Conclusion

From the above scatter plots of predict versus value of the features, we can get a relatively linear relationship between the predict values and the selected feature, from which we can conclude that the linear regression is quite successful.

2.4 Problem 1.4

In this section, we use k-fold cross validation to get a more accurate judgments to our model. And we use 90-10% splitting for each model. Apart from this, we also separate the data by three time period, and get the more precise model.

Also, in this part, we use three different models to fit the data. And the models we use in this section is linear model, neutral network and svm.

For all the model, we train the data using these three different models and the final results are shown in the following table.

Table 3: Before Feb. 1, 8:00 a.m

tags	linear model	neutral network	svm
gohawks	303.63	227.77	197.76
gopatriots	15.09	13.97	45.57
sb49	37.15	93.91	1919.62
superbowl	301.88	412.34	2459.91
nfl	120.04	189.98	231.58
patriots	229.14	287.25	227.44

Table 4: Between Feb. 1, 8:00 a.m. and 8:00 p.m

tags	linear model	neutral network	svm
gohawks	7971.91	5452.35	4284.13
gopatriots	813.88	1471.83	819.53
sb49	75383.89	64287.71	70395.90
superbowl	148049.23	94192.73	81612.61
nfl	2029.75	4637.33	1214.34
patriots	23030.77	26419.31	57055.42

Table 5: After Feb. 1, 8:00 p.m.

tags	linear model	neutral network	svm
gohawks	35.44	37.67	113.85
gopatriots	3.87	3.56	5.66
sb49	172.56	360.77	285.48
superbowl	183.69	696.34	677.95
nfl	109.20	466.04	125.82
patriots	88.95	146.23	278.87

From the table we can see that we can have a relatively accurate prediction during the time period 1 and 3, but in period 2, namely between Feb. 1, 8:00 a.m. and 8:00 p.m, we have a less accurate answer.

Finally, we aggregate all the data and get our final result, which is displayed below.

Table 6: Aggregated data

	Time Period 1	Time Period 2	Time Period 3
linear regression	253.27	115980.00	175.92
neutral network	411.36	87689.34	711.74
svm	1080.54	85851.11	385.96

2.5 Problem 1.5

In this part, we will use the all of the hash tag as the training sets. And we will use the training sets to produce a trained model. The number of tweets in the next hour we predict by using with features from previous 5 hour window. The test files have content with different hash tags, then we need to combine six hash tag sets to be one set. And use this set to train the model. After we train the model, we use the trained model to predict the number of tweets in last hour.

The following graph shows the predicted value vs actual value in the test set. The graph shown in order from sample 1 to sample 10.

```
predict value:954.6026690346104
actual value:595

predict value:6268.288017288751
actual value:204746

predict value:2474.261899333979
actual value:3188

predict value:410.6939920559822
actual value:1228

predict value:1367.9340393605019
actual value:1718

predict value:111947.98601000413
actual value:204599

predict value:65.41886829541272
actual value:403

predict value:203.0319643273405
actual value:180

predict value:6626.48115244231
actual value:9582

predict value:239.79483169725566
actual value:303
```

Figure 27: predicted value with actual value of 10 samples

From the predicted number of tweets and actual number of tweets, the accuracy of predicted tweets by five hours window improved compared with one hour window. And the training model is good, only when the number is very large, there will be a large difference. However, it is make sense, if we do the ratio of error, it is still good. This is because five hours window includes more information of past activity.

3 problem2

For this problem, we want to use the text as the feature to predict the location of the author of the tweet. In this part, we only use 'super bowl.txt' as the dataset. We use ['highlight'] as the feature. It is simply the content of the tweet. Then we create a list of text of tweet and location. After that, we use some characteristic words to identify the location of the tweet. We split the dataset into two sets. One is Washington set, the other one is Massachusetts set. At the same time, we clean the text and just remain the words. And according to the content of tweets and location we get to train a classifier. And we will try three different algorithms in this part.

3.1 SVM

In this part, we use SVM algorithm. After we got the words, we convert the content to tf-idf matrix. And we got the accuracy, recall and precision to prove the model works. The reason why we use recall and precision is the datasets are not balanced. The figure below is our result.

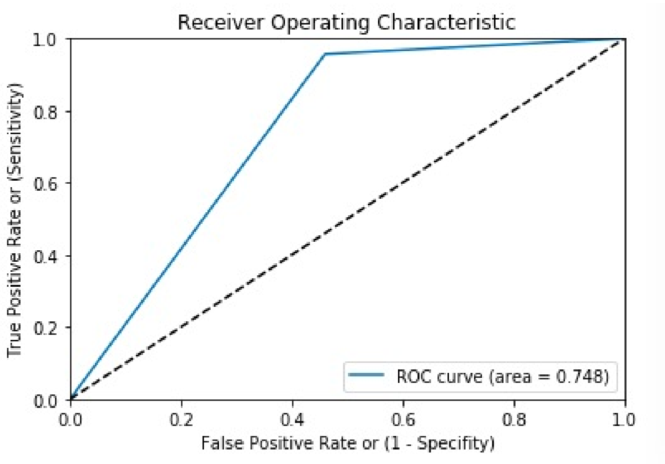


Figure 28: Roc of SVM

```
accuracy:0.759587462664
recall:0.748326648751
precision:0.810059970222
```

Figure 29: accuracy, recall, precision of SVM

```
confusion matrix:/n [[ 6258  5328]
 [  561 12349]]
```

Figure 30: matrix of SVM

AS we can see from the result, the accuracy of SVM is relatively high enough to prove that we could use the content of tweets to predict the location. And the recall and precision is also high enough to prove that. The ROC curve area is 0.748 which is also relatively high enough to prove the model is work.

3.2 Logistical Regression

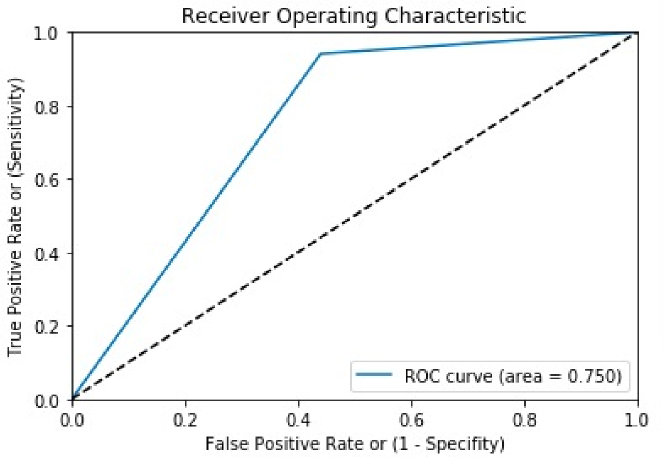


Figure 31: Roc of Logistical Regression

```
accuracy:0.760771785995
recall:0.750449909782
precision:0.805701840284
```

Figure 32: accuracy, recall, precision of Logistical Regression

```
confusion matrix:/n [[ 6484  5102]
 [   758 12152]]
```

Figure 33: matrix of Logistical Regression

AS we can see from the result, the accuracy of Logistical Regression is relatively high enough to prove that we could use the content of tweets to predict the location. And the recall and precision is also high enough to prove that. The ROC curve area is 0.75 which is also relatively high enough to prove the model is work.We can conclude that the Logistical Regression works well.

3.3 GaussionNB

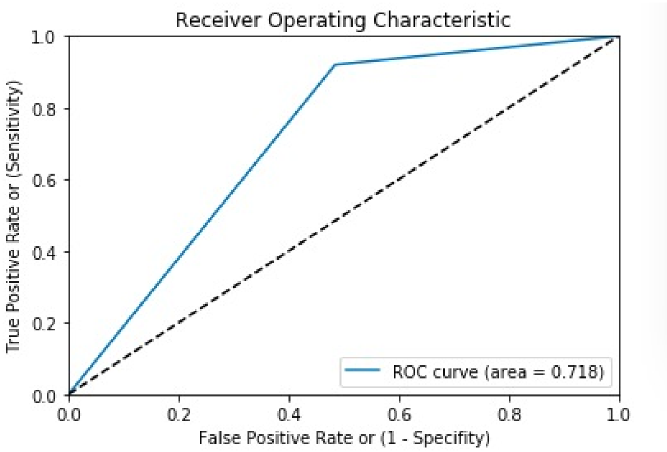


Figure 34: Roc of GaussionNB

```
accuracy:0.728684137954
recall:0.717741092794
precision:0.768110643738
```

Figure 35: accuracy, recall, precision of GaussionNB

```
confusion matrix:/n [[ 5972  5614]
 [ 1032 11878]]
```

Figure 36: matrix of GaussionNB

AS we can see from the result, the accuracy of GaussionNB is relatively high enough to prove that we could use the content of tweets to predict the location. And the recall and precision is also high enough to prove that. The ROC curve area is 0.718 which is also relatively high enough to prove the model is work.We can conclude that the GaussionNB works well.

From all of the above result, we could find all of the three algrithoms have the simlair result. That means these three algrithoms could deal with the prediction of location by the content of tweets. However, there is still one have the a liitle bit better effect. We think the Logistical Regression is the best algrithom in these three algrithoms. It has the largest ROC curve area. And the accuarcy, recall and preicsion also relatively higher than other two algrithoms.

4 Problem 3

4.1 Problem proposal:

For part 3, data analysis is that we could learn from the data to know something we do not know. For this part, we focus on the sentiment of the tweets. We could infer many things from the sentiment of the tweets. firstly, we could analyze the sentiment of the tweets and find out the positive ratio of all tweets from Wahsington and Massachusetts to predict which team won the game. It's obviouly that people living in winner state will have larger ratio of positive tweets. Secondly, we could predict the location of the author by the sentiment of the tweets, most people who tweeted positive tweets would be from winner state.

4.2 Solution:

Firstly, we choose tweets from the Washington area and Massachusetts area. The reason is the sentiment of tweet from two teams' area will be more relevant to the game result.

Secondly, We choose the time of data set to be after the game. Because only after the game, the sentiment attribute is more relevant to the result of the game.

Thirdly, We use textblob function to calculate the polarity. We clean the text and remain the words only. If we do not clean the text, it will be many neutral result which is not good for our training.

Fourthly, we find out the number of tweets which are positive from Washington and Massachusetts. The reason is only author from these two area will have strong sentiment feature.

Fifthly, we use polarity to train the model, and then predict the location of the author. After that, calculate the accuracy, recall, and precision of the model. The data in two sets is not balanced. Therefore, we need recall and precision to prove the model works.

At last, we could calculate the ratio of positive tweets and all tweets after the game from Washington and Massachusetts. We could infer from the ratio to know which team won the game.

4.3 Result:

The following graph is the accuracy, recall and precision of the trained model to predict the location of author by sentiment of the tweets.

```
accuracy:0.9018841800551611
recall:0.9320803842450701
precision:0.8293157977598848
```

Figure 37: Accuarcy of trained model

As we can see from the result, the accuracy is good. At the same time, the recall and precision is also reasonable high enough to prove that model is work. In this part, we know the Massachusetts team lost game, so we know the sentiment of tweets from Massachusetts should be negative. We could use sentiment of the tweet to know where the author from.

```
ratio_w_postive:0.843939393939
```

Figure 38: ratio of positive tweets in Washington

```
ratio_m_postive:0.364052287582
```

Figure 39: ratio of positive tweets in Massachusetts

From the two graph above, we could know that the ratio of positive tweets in Washington is much higher than the ratio in Massachusetts. Therefore, we could conclude that the Washington team probably win the game. And the result of the 2015 super bowl also prove that analysis result is correct. That solution works.