EE239AS, Project 3
Popularity Prediction on Twitter

Cheyun Xia 504422348 Shengzhi Jiang 704514808
Fuxing Liu 804516755 Yining Li 204516697

1

We download the training tweet data, calculate the corresponding statistics for each hashtag and list them in the following table.

Hashtag	Avg. tweets per hour	Avg. followers	Avg. retweets
#superbowl	1399	10136	2.388
#nfl	279	4865	1.539
#gohawks	193	2477	2.015
#gopatriots	38	1619	1.400
#patriots	499	3760	1.783
#sb49	1418	10496	2.511

Table 1: Statistics for each hashtag

Specifically, we plot " number of tweets in hour " over time for #SuperBowl and #NFL as follows.

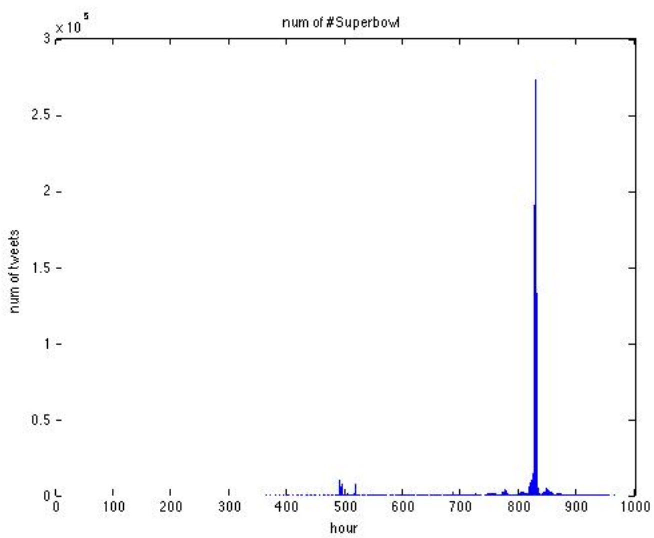


Figure 1: number of tweets in hour for SuperBowl

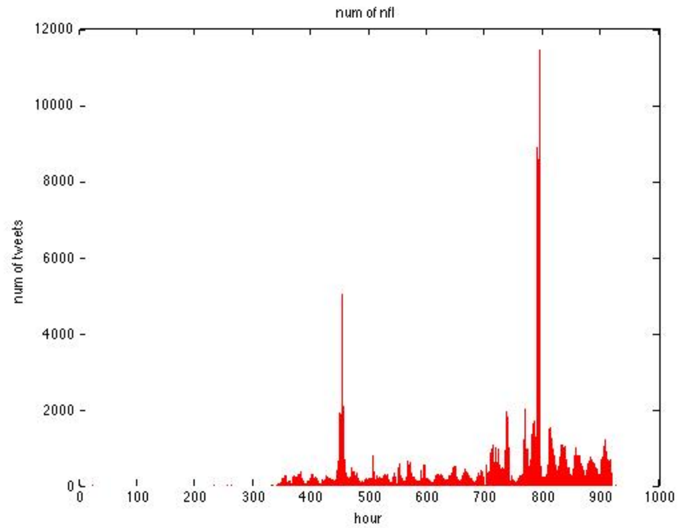


Figure 2: number of tweets in hour for NFL

2

In this part, we want to fit a linear regression model using 5 features to predict numbers of tweets in the next hour, with features extracted from tweet data in the previous hour. Figure 3 and 4 show the linear regression result of the #SuperBowl and #NFL with the use of OLS (Ordinary Least Square). Here we use average error to interpret the accuracy of our prediction.

$$E_{error} = \frac{|N_{real} - N_{prediction}|}{|N_{real}|}$$

N_{real} represents the real number of tweets, and $N_{prediction}$ is the number of tweets calculated in our prediction model.

2.1 Superbowl

Dep. Variable:	y	R-squared:	0.815
Model:	OLS	Adj. R-squared:	0.806
Method:	Least Squares	F-statistic:	83.09
Date:	Mon, 16 Mar 2015	Prob (F-statistic):	5.95e-33
Time:	16:50:14	Log-Likelihood:	-716.86
No. Observations:	100	AIC:	1446.
Df Residuals:	94	BIC:	1461.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
const	222.7326	74.791	2.978	0.004	74.233	371.233
x1	0.7770	0.116	6.704	0.000	0.547	1.007
x2	0.0198	0.044	0.446	0.657	-0.068	0.108
x3	4.819e-06	4.29e-06	1.123	0.264	-3.7e-06	1.33e-05
x4	-1.392e-05	1.42e-05	-0.982	0.329	-4.21e-05	1.42e-05
x5	-11.5429	4.871	-2.370	0.020	-21.215	-1.871

Omnibus:	55.530	Durbin-Watson:	1.548
Prob(Omnibus):	0.000	Jarque-Bera (JB):	309.767
Skew:	1.684	Prob(JB):	5.43e-68
Kurtosis:	10.937	Cond. No.	7.15e+07

Figure 3: OLS of #SuperBowl

So the model for number of tweets with hashtag #SuperBowl is :

$$y = 222.7326 + 0.7770x_1 + 0.0198x_2 + 4.819 \times 10^{-6}x_3 - 1.392 \times 10^{-5}x_4 + -11.5429x_5$$

Here x_1 denotes the number of tweets, x_2 denotes the total number of retweets, x_3 denotes the sum of the number of followers, x_4 denotes the maximum number of followers and x_5 denotes

the time of the day.
The average error of #SuperBowl is $E_{error} = 0.4146$.

2.2NFL

```
==============================================================================
Dep. Variable:                      y   R-squared:                       0.708
Model:                            OLS   Adj. R-squared:                  0.693
Method:                 Least Squares   F-statistic:                     45.67
Date:                Mon, 16 Mar 2015   Prob (F-statistic):           1.07e-23
Time:                        16:46:04   Log-Likelihood:                -830.39
No. Observations:                 100   AIC:                             1673.
Df Residuals:                      94   BIC:                             1688.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        284.6759    200.938      1.417      0.160      -114.292   683.643
x1            -1.2598      0.602     -2.092      0.039        -2.455    -0.064
x2             0.6902      0.320      2.156      0.034         0.054     1.326
x3             0.0002   5.79e-05      4.070      0.000         0.000     0.000
x4            -0.0004      0.000     -3.556      0.001        -0.001    -0.000
x5            -0.5690     15.005     -0.038      0.970       -30.361    29.223
==============================================================================
Omnibus:                       93.857   Durbin-Watson:                   2.119
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1406.433
Skew:                           2.850   Prob(JB):                    3.95e-306
Kurtosis:                      20.466   Cond. No.                     2.09e+07
==============================================================================
```

Figure 4: OLS of #NFL

So the model for number of tweets with hashtag #NFL is :

$$y = 284.6759 - 1.2598x_1 + 0.6902x_2 + 0.0002x_3 - 0.0004x_4 - 0.5690x_5$$

Similarly, x_1 denotes the number of tweets, x_2 denotes the total number of retweets, x_3 denotes the sum of the number of followers, x_4 denotes the maximum number of followers and x_5 denotes the time of the day.
The average error of #NFL is $E_{error} = 0.8092$. One possible reason for the larger error than the #superbowl may be the smaller size of the data.

As we can see from Figure 3 and Figure 4, the value of t for x_4 is quite small, so we discard the feature of maximum number of followers and replace it with the other 3 features in part 3.

3

In this part, we introduced three new features, **ranking score**, **user mentions** and **number of authors**, to fit the linear regression model. Ranking score shows the presence of query keywords and recency of one tweet. User mention measures the popularity of tweet, that is to say, the more times people are mentioned, the more popular this tweet is. Number of authors is also an index on popularity. Besides, we deleted the feature **maximum number of followers** in this part, for the result in part 2 shows that it is rather irrelavent to the prediction of tweet numbers. Figure 5 is the scatter plots for two models, where we choose retweet number, ranking score and user mention as the outstanding features.

3.1 Superbowl

After fitting the linear regression model for #superbowl, we have Equation (1) to predict the tweet number for this hashtag:

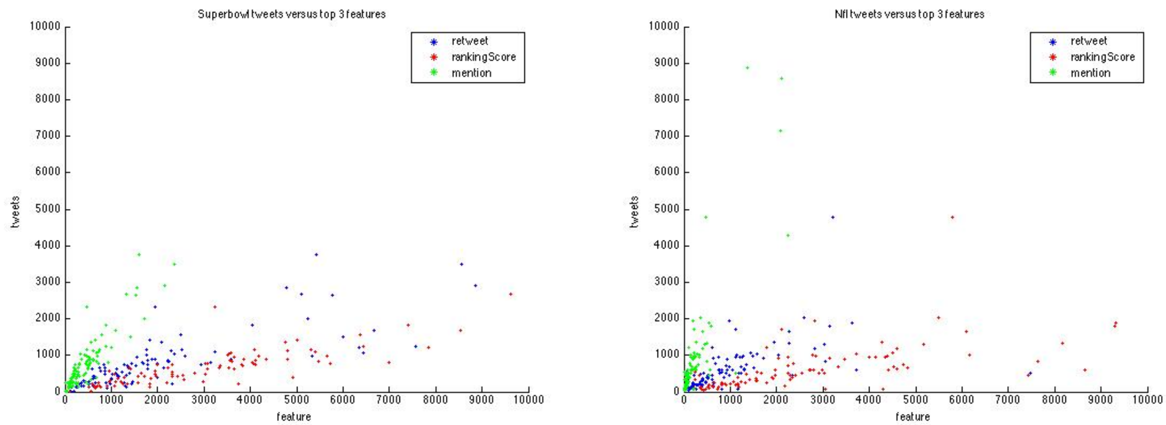
$$y = 285 - 11.3x_1 + 0.0315x_2 - 2.24 \times 10^{-6}x_3 - 10.1x_4 + 2.62x_5 + 0.900x_6 - 0.786x_7, \quad (1)$$

where x_1-x_7 represents number of tweets, total number of retweets, sum of number of followers, time of the day, sum of ranking scores, sum of user mentions, number of authors for current hour separately, and y denotes the number of tweets for next hour.

3.2 NFL

Similarly we can derive Equation (2) for #NFL:

$$y = 127 + 4.46x_1 - 0.497x_2 + 3.22 \times 10^{-5}x_3 - 11.0x_4 - 1.22x_5 + 9.11x_6 + 0.490x_7 \tag{2}$$



(a) #superbowl (b) #nfl

Figure 5: Scatter plots of predictants VS features

As we could see from the figures, each of our three features and the number of tweets are linear related hence our prediction model based on OLS makes sense.

4

For #Superbowl and #NFL, we train regression models for three time periods, each period using cross-fold validation.
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.

4.1 Superbowl

The average errors are calculated as:

$$\text{Total Error}: \begin{cases} 320.0580 & : \text{Before Feb. 1, 8:00 a.m.} \\ 54052.2986 & : \text{Between Feb. 1, 8:00 a.m. and 8:00 p.m.} \\ 1673.2612 & : \text{After Feb. 1, 8:00 p.m.} \end{cases}$$

The error in the second period is apparently larger than the others, the reason may be the amount of time in the second period is much less than others, making the prediction more difficult.

For each period, the best model can be expressed as follows:

$y = 1.2705 + 0.0735x_1 + 0.3280x_2 - 1.1359x_3$
$y = 7.5027 \times 10^3 + 6.7699x_1 - 0.3633x_2 - 43.8255x_3$
$y = 2.2359 \times 10^2 - 6.4481 \times 10^{-3}x_1 - 0.1230x_2 + 2.4670x_3$

Here x_1 denotes the retweet feature, x_2 denotes the rankingScore feature and x_3 denotes the mention feature. These features can be used to make better predictions.

4.2 NFL

The average errors are calculated as:

$$\text{Total Error}: \begin{cases} 166.5032 & : \text{Before Feb. 1, 8:00 a.m.} \\ 2029.9597 & : \text{Between Feb. 1, 8:00 a.m. and 8:00 p.m.} \\ 199.2946 & : \text{After Feb. 1, 8:00 p.m.} \end{cases}$$

Also, the error in the second period is apparently larger than the others due to the relatively shorter time period, making the prediction more difficult.

For each period, the best model can be expressed as follows:

$y = 1.2796^2 + 8.5339 \times 10^{-2}x_1 + 2.9560 \times 10^{-2}x_2 - 1.9004x_3$
$y = -1.8196 \times 10^3 + 1.8079 \times 10^{-1}x_1 - 6.6142 \times 10^{-1}x_2 + 16.240x_3$
$y = 3.0550 \times 10^2 - 1.8087 \times 10^{-1}x_1 + 1.8647 \times 10^{-1}x_2 - 0.6667x_3$

Similarly, x_1 denotes the retweet feature, x_2 denotes the rankingScore feature and x_3 denotes the mention feature. These features can be used to make better predictions.

5

In this part, we use the best model calculated in problem 4 to predict the 10 testing samples, we applied both NFL and Superbowl model to implement our predictions. The predicted numbers of tweets are shown as following:

5.1 Superbowl

Hour	2	3	4	5	6
Sample1	3.32	3.09	3.82	39.58	53.87
Sample2	52199	63988	52329	90861	163550
Sample3	75.52	11.94	70.18	33.67	12.23
Sample4	206.16	25.45	55.09	74.01	56.64
Sample5	51.86	116.79	5.18	16.97	8.94
Sample6	13151	749490	4642082	3967817	2957552
Sample7	166.72	141.08	173.56	141.18	173.12
Sample8	16.30	24.17	19.78	9.48	
Sample9	8074	9356	8223	6998	22169
Sample10	136.93	165.74	137.11	140.52	146.84

5.2 NFL

Hour	2	3	4	5	6
Sample1	175.97	152.62	144.45	207.99	203.33
Sample2	24211	34855	33230	44789	31581
Sample3	197.30	103.37	318.83	499.95	299.49
Sample4	274.99	223.69	132.10	105.61	136.60
Sample5	326.79	416.36	211.24	178.96	199.04
Sample6	149.79	265573	1639646	1390517	1036286
Sample7	207.39	240.62	266.12	282.08	274.47
Sample8	101.86	83.06	87.32	108.80	
Sample9	3677	4434	2935	3067	3427
Sample10	275.38	275.52	271.99	275.36	276.99