Project 4 Report:Popularity Prediction on Twitter *

Tushar Sudhakar Jee, Shubham Agarwal, Pulkit Aggarwal, Ishan Upadhyaya

March 19, 2016

1 Part 1

Hashtag:gopatriots
average number of tweets per hour: 38.3861
average number of followers of users posting the tweets: 1558
average number of retweets: 1.0307

Hashtag:gohawks
average number of tweets per hour: 215.47
average number of followers of users posting the tweets: 1709
average number of retweets: 45.0641

Hashtag:patriots
average number of tweets per hour: 556.5743
average number of followers of users posting the tweets: 1859
average number of retweets: 50.6097

Hashtag:sb49
average number of tweets per hour: 1419.8896
average number of followers of users posting the tweets : 2243
average number of retweets: 252.9347

Hashtag:nfl
average number of tweets per hour:294.8925
average number of followers of users posting the tweets : 4376
average number of retweets: 15.0273

Hashtag:superbowl
average number of tweets per hour: 1624.5715
average number of followers of users posting the tweets : 4221
average number of retweets: 222.1168

As the hashtag becomes more generic, like #superbowl is more generic than #gopatriots, number of tweets per hour increases.Similarly, as more users are posting tweets with generic hashtags, average number of followers also increases.

* EE 239AS ; Winter 2016

Plots for number of tweets in an hour for superbowl,nfl

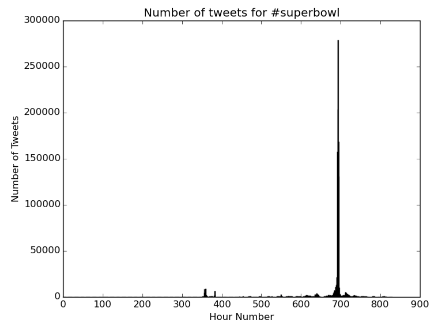


Figure 1: Plot for number of tweets in an hour for superbowl

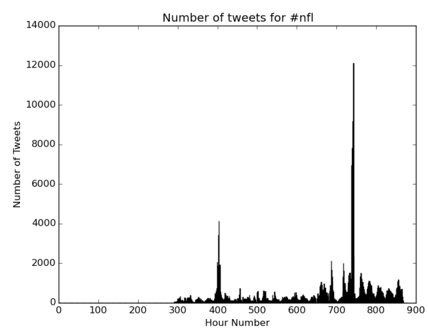


Figure 2: Plot for number of tweets in an hour for nfl

In the above graphs , we can see that number of tweets in a an hour were high near the super-bowl event as compared to days other than super-bowl event.

2 Part 2

The linear regression model was trained on the dataset described in the question to predict the number of tweets in the next hour. This model was trained on each of the hashtags. Features-
x1-number of tweets
x2-total number of retweets
x3- sum of the number of followers of the users posting the hashtag
x4-maximum number of followers of the users posting the hashtag
x5- time of the day.

Linear model for #gopatriots

```
Currently working with #gopatriots
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.614
Model:                            OLS   Adj. R-squared:                  0.610
Method:                 Least Squares   F-statistic:                     142.8
Date:                Sat, 19 Mar 2016   Prob (F-statistic):           1.95e-90
Time:                        01:56:34   Log-Likelihood:                -428.14
No. Observations:                 454   AIC:                             866.3
Df Residuals:                     449   BIC:                             886.9
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            25.9936      3.320      7.828      0.000        19.468    32.519
x2           -19.3543      1.470    -13.163      0.000       -22.244   -16.465
x3             7.6036      3.897      1.951      0.052        -0.054    15.262
x4            -3.4688      2.015     -1.721      0.086        -7.430     0.492
x5            -2.0302      0.629     -3.227      0.001        -3.267    -0.794
==============================================================================
Omnibus:                      731.674   Durbin-Watson:                   2.379
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           308193.251
Skew:                           8.991   Prob(JB):                         0.00
Kurtosis:                     129.368   Cond. No.                         15.1
==============================================================================
```

Figure 3: Figure showing p,t and R$^2 values for \#gopatriots$

Significant Features:x1,x2,x5 are significant as they have very low p value.Features which have p value less than 0.05 are considered as significant.These features also have a high t-value.This also suggests that these features are significant. Training accuracy can be seen from the value of R-squared. R-squared value is 0.614.

Linear model for #gohawks

```
Currently working with #gohawks
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.522
Model:                            OLS   Adj. R-squared:                  0.518
Method:                 Least Squares   F-statistic:                     144.0
Date:                Sat, 19 Mar 2016   Prob (F-statistic):          3.75e-103
Time:                        01:56:34   Log-Likelihood:                -697.04
No. Observations:                 664   AIC:                             1404.
Df Residuals:                     659   BIC:                             1427.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            12.3965      3.198      3.877      0.000         6.117    18.676
x2            -0.0975      0.701     -0.139      0.889        -1.473     1.278
x3            10.5023      4.049      2.594      0.010         2.552    18.453
x4            -5.9060      1.595     -3.702      0.000        -9.038    -2.774
x5            -2.1650      0.747     -2.896      0.004        -3.633    -0.697
==============================================================================
Omnibus:                      953.589   Durbin-Watson:                   2.240
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           982112.218
Skew:                           7.103   Prob(JB):                         0.00
Kurtosis:                     190.873   Cond. No.                         12.6
==============================================================================
```

Figure 4: Figure showing p,t and R^2 values for #gohawks

Significant Features:x1,x4,x5,x3 are significant features with low p values. R-squared value is 0.522.

Linear model for #patriots

```
Currently working with #patriots
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.692
Model:                            OLS   Adj. R-squared:                  0.690
Method:                 Least Squares   F-statistic:                     296.4
Date:                Sat, 19 Mar 2016   Prob (F-statistic):          8.31e-166
Time:                        01:56:34   Log-Likelihood:                -549.83
No. Observations:                 663   AIC:                             1110.
Df Residuals:                     658   BIC:                             1132.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            27.4734      1.072     25.621      0.000        25.368    29.579
x2            -1.6965      0.673     -2.522      0.012        -3.017    -0.376
x3            -7.8920      1.368     -5.769      0.000       -10.578    -5.206
x4             2.1136      0.788      2.681      0.008         0.566     3.662
x5            -2.3237      0.623     -3.731      0.000        -3.547    -1.101
==============================================================================
Omnibus:                     1296.287   Durbin-Watson:                   1.799
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2006895.273
Skew:                          13.791   Prob(JB):                         0.00
Kurtosis:                     271.118   Cond. No.                         5.09
==============================================================================
```

Figure 5: Figure showing p,t and R^2 values for #patriots

Significant Features:All features are significant, with x1,x3,x5 more significant than x2,x4.Significant features also have a high t-value. R-squared value is 0.692.

Linear model for #sb49

```
Currently working with #sb49
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.793
Method:                 Least Squares   F-statistic:                     419.9
Date:                Sat, 19 Mar 2016   Prob (F-statistic):          9.27e-184
Time:                        01:56:34   Log-Likelihood:                -342.97
No. Observations:                 547   AIC:                             695.9
Df Residuals:                     542   BIC:                             717.5
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            25.3438      1.083     23.404      0.000        23.217    27.471
x2            -1.7019      0.546     -3.115      0.002        -2.775    -0.629
x3            -7.1830      1.408     -5.103      0.000        -9.948    -4.418
x4             4.8808      0.723      6.749      0.000         3.460     6.301
x5            -3.3721      0.482     -6.996      0.000        -4.319    -2.425
==============================================================================
Omnibus:                      988.887   Durbin-Watson:                   1.190
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           833300.916
Skew:                          11.504   Prob(JB):                         0.00
Kurtosis:                     192.822   Cond. No.                         6.52
==============================================================================
```

Figure 6: Figure showing p,t and R^2 values for #sb49

Significant Features: x1,x4,x5 are significant features with x1 as the most important feature.. R-squared value is 0.795.

Linear model for #nfl

```
Currently working with #nfl
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.623
Model:                            OLS   Adj. R-squared:                  0.620
Method:                 Least Squares   F-statistic:                     202.9
Date:                Sat, 19 Mar 2016   Prob (F-statistic):          2.14e-127
Time:                        01:56:34   Log-Likelihood:                -575.19
No. Observations:                 618   AIC:                             1160.
Df Residuals:                     613   BIC:                             1183.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            20.3825      2.015     10.115      0.000        16.425    24.340
x2            -9.8686      0.812    -12.155      0.000       -11.463    -8.274
x3            10.1277      2.907      3.484      0.001         4.419    15.836
x4            -5.0599      1.630     -3.104      0.002        -8.261    -1.859
x5            -8.2981      0.747    -11.112      0.000        -9.765    -6.831
==============================================================================
Omnibus:                      917.371   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           289729.932
Skew:                           8.106   Prob(JB):                         0.00
Kurtosis:                     107.828   Cond. No.                         10.8
==============================================================================
```

Figure 7: Figure showing p,t and R^2 values for #nfl

Significant Features:x1,x2,x5 are significant with high t-value and low p-value. R-squared value is 0.623.

Linear model for #superbowl

```
Currently working with #superbowl
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.824
Model:                            OLS   Adj. R-squared:                  0.823
Method:                 Least Squares   F-statistic:                     567.1
Date:                Sat, 19 Mar 2016   Prob (F-statistic):          1.35e-225
Time:                        01:56:34   Log-Likelihood:                -335.42
No. Observations:                 610   AIC:                             680.8
Df Residuals:                     605   BIC:                             702.9
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             9.9400      2.114      4.701      0.000         5.787    14.092
x2           -16.7365      0.675    -24.779      0.000       -18.063   -15.410
x3            25.3237      2.511     10.086      0.000        20.393    30.254
x4            -2.2855      0.852     -2.684      0.007        -3.958    -0.613
x5            -3.3723      0.489     -6.897      0.000        -4.333    -2.412
==============================================================================
Omnibus:                     1086.186   Durbin-Watson:                   2.152
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           953499.953
Skew:                          11.291   Prob(JB):                         0.00
Kurtosis:                     195.366   Cond. No.                         14.2
==============================================================================
```

Figure 8: Figure showing p,t and R^2 values for #superbowl

Significant features:x2,x3 are significant features. R-squared value is 0.824.

3 Part 3

Features used in our analysis of the twitter dataset

3.1 Network Features

Network indicates the connectivity of the users posting the tweets. The connectivity is indicative of how well the tweet can be diffused in the network. Here we take several of the connectivity features which are important for predicting the number of tweets.
Number of retweets: Sum of number of retweets in an hour(x1).
Number of max followers: Here we count a list of followers for the users who tweeted in last hour and take the maximum. This indicates the maximum extent to which a single user can affect his network(x3).
Sum of number of people following the hashtag: As people following the tweets are the likely users to tweet, we take the number of people following that hashtag as a feature(x2).
Number of mentions: Sum of number of tweets in a given hour containing '@' mentions(x5).
Number of unique users: We also take the number of unique users which posted in last hours as a feature(x6).

3.2 Time Series Features

The time series features indicate the trend of tweets in a given time interval. Since the past number of tweets values are extremely important, through these features we try to extract the tweet variation with time.
Moving Average: Averaging number of tweets in last five hours with reference to present value(x7).
Moving Standard Deviation: Standard deviation of tweets in last five hours with reference to present value(x8).
Derivative: Taking number of tweets to be a time-series,the Derivative indicates Slope value at present time(x9).
Derivative mean: Mean of past five derivative values. The derivate gives the trend for past values which is a very good indicator for prediction(x10).
Past value: We take the past five values of number of tweets. This ensures we have enough past information to predict the values in next hour. This is similar to the linear prediction model used in many cases(x10-x15).
Time of the day: Represent hours of the day with respect to a given time reference(x4).

Using the features described on the previous page, we built a Linear Regression model. The training accuracy and significant variables for each hashtag are shown below:

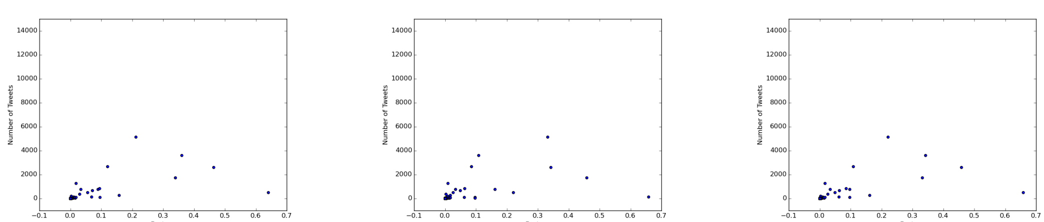
Linear model for #gopatriots

```
Currently working with #gopatriots
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.720
Model:                            OLS   Adj. R-squared:                  0.711
Method:                 Least Squares   F-statistic:                     80.91
Date:                Thu, 17 Mar 2016   Prob (F-statistic):          4.84e-112
Time:                        02:10:53   Log-Likelihood:                -355.05
No. Observations:                 454   AIC:                             738.1
Df Residuals:                     440   BIC:                             795.8
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1           -15.0933      2.398     -6.295      0.000       -19.806   -10.381
x2            23.8492      6.411      3.720      0.000        11.249    36.449
x3            -9.9803      2.903     -3.437      0.001       -15.687    -4.274
x4            -2.0515      0.571     -3.594      0.000        -3.173    -0.930
x5             0.2105      3.785      0.056      0.956        -7.228     7.649
x6            -1.3659     38.708     -0.035      0.972       -77.441    74.710
x7           -31.3866     12.013     -2.613      0.009       -54.997    -7.776
x8             7.4237      4.468      1.661      0.097        -1.358    16.205
x9            -1.0552     10.342     -0.102      0.919       -21.382    19.271
x10            0.9777      1.074      0.910      0.363        -1.133     3.088
x11            1.8936      1.621      1.169      0.243        -1.291     5.079
x12            9.3083      2.314      4.022      0.000         4.760    13.856
x13           14.1025      2.934      4.806      0.000         8.335    19.870
x14           10.0488     12.865      0.781      0.435       -15.235    35.333
x15            9.1682     21.365      0.429      0.668       -32.823    51.159
==============================================================================
Omnibus:                      577.377   Durbin-Watson:                   2.209
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           132837.969
Skew:                           5.784   Prob(JB):                         0.00
Kurtosis:                      85.997   Cond. No.                     1.51e+16
==============================================================================
```

Figure 9: Figure showing p,t and R^2 values for #gopatriots

From the values obtained, the significant features were found to be x4(time of day), x12 and x13(past values). These are sort of intuitive as well because we can expect the future value to depend on past values and the current time.

The scatter plot of Predictant values versust the significant features is shown below



(a) (b) (c)

Figure 10: Scatter plot of significant variable with Predictant for #gopatriots

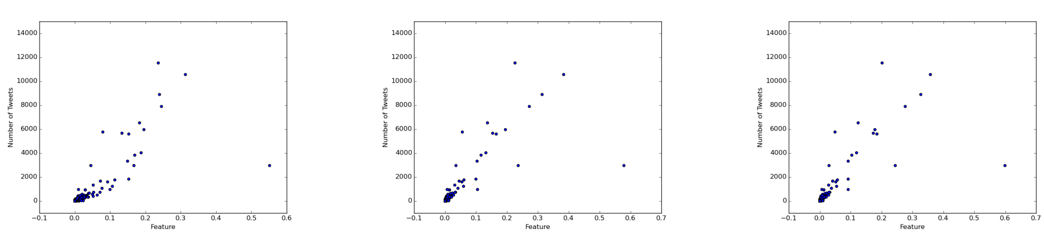
Linear model for #gohawks

```
Currently working with #gohawks
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.564
Model:                            OLS   Adj. R-squared:                  0.554
Method:                 Least Squares   F-statistic:                     59.99
Date:                Thu, 17 Mar 2016   Prob (F-statistic):          4.82e-107
Time:                        02:10:53   Log-Likelihood:                -666.80
No. Observations:                 664   AIC:                             1362.
Df Residuals:                     650   BIC:                             1425.
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.0119      0.803      0.015      0.988        -1.564     1.588
x2             2.6839      5.164      0.520      0.603        -7.457    12.825
x3            -3.8270      1.841     -2.079      0.038        -7.441    -0.213
x4            -2.6586      0.794     -3.349      0.001        -4.217    -1.100
x5            -0.5538      4.039     -0.137      0.891        -8.486     7.378
x6            41.9631     16.698      2.513      0.012         9.175    74.752
x7            -0.4951      8.745     -0.057      0.955       -17.666    16.676
x8            -0.4563      2.665     -0.171      0.864        -5.690     4.778
x9            -9.1293      3.549     -2.573      0.010       -16.098    -2.161
x10           -0.1185      1.321     -0.090      0.929        -2.712     2.475
x11           -4.8186      2.017     -2.389      0.017        -8.780    -0.857
x12            0.9974      2.017      0.494      0.621        -2.963     4.958
x13            3.4266      2.089      1.641      0.101        -0.675     7.528
x14           -7.2658      4.905     -1.481      0.139       -16.898     2.366
x15          -13.8771      7.305     -1.900      0.058       -28.220     0.466
==============================================================================
Omnibus:                     1105.691   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1282077.957
Skew:                           9.710   Prob(JB):                         0.00
Kurtosis:                     217.390   Cond. No.                     7.83e+15
==============================================================================
```

Figure 11: Figure showing p,t and R^2 values for #gohawks

The significant variable obtained for this hashtag were x4(time), x6(of unique users) and x14(most recent value). The reasoning for time and most recent value is same as that we saw above, while number of unique users is also intuitive to understand as this indicates to some extent the reach of the current networ.

The scatter plot of Predictant values versus the significant features is shown below



(a) (b) (c)

Figure 12: Scatter plot of significant variable with Predictant for #gohawks

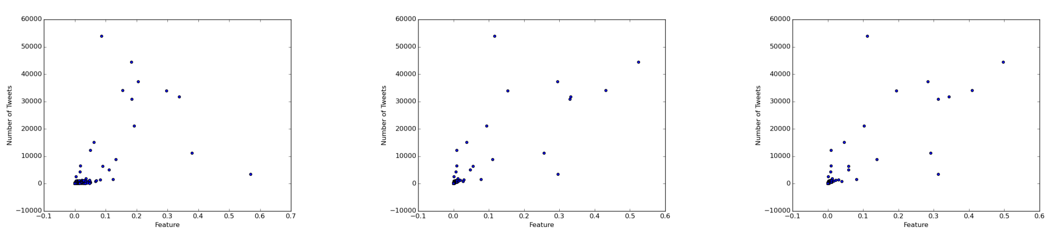
Linear model for #patriots

```
Currently working with #patriots
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.746
Model:                            OLS   Adj. R-squared:                  0.741
Method:                 Least Squares   F-statistic:                     136.4
Date:                Thu, 17 Mar 2016   Prob (F-statistic):          1.33e-182
Time:                        02:10:53   Log-Likelihood:                -485.94
No. Observations:                 663   AIC:                             999.9
Df Residuals:                     649   BIC:                             1063.
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            -1.0777      0.654     -1.647      0.100        -2.362     0.207
x2            15.3767      5.330      2.885      0.004         4.911    25.843
x3            -3.0994      1.331     -2.329      0.020        -5.713    -0.486
x4            -3.0439      0.581     -5.236      0.000        -4.185    -1.902
x5            42.1066      7.950      5.296      0.000        26.496    57.718
x6          -125.6345     24.876     -5.050      0.000      -174.481   -76.788
x7           -39.5285      8.008     -4.936      0.000       -55.253   -23.804
x8             1.7800      1.714      1.038      0.300        -1.586     5.146
x9            27.3657      4.122      6.639      0.000        19.271    35.460
x10           -5.6527      1.235     -4.576      0.000        -8.078    -3.227
x11            8.8564      2.518      3.518      0.000         3.913    13.800
x12            7.1597      1.724      4.153      0.000         3.774    10.545
x13           14.6535      2.119      6.917      0.000        10.493    18.814
x14           47.9428      8.137      5.892      0.000        31.965    63.921
x15           63.9687     10.495      6.095      0.000        43.360    84.578
==============================================================================
Omnibus:                     1287.941   Durbin-Watson:                   1.940
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2090556.493
Skew:                          13.557   Prob(JB):                         0.00
Kurtosis:                     276.754   Cond. No.                     9.19e+15
==============================================================================
```

Figure 13: Figure showing p,t and R^2 values for #patriots

Significant Features are:

The significant variable for this hashtag were the x15,x14 and x13. These correspond to previous values, which we have seen are intuitive to understand. The scatter plot of Predictant values versus the significant features is shown below



(a) (b) (c)

Figure 14: Scatter plot of significant variable with Predictant for #patriots

Linear model for #sb49

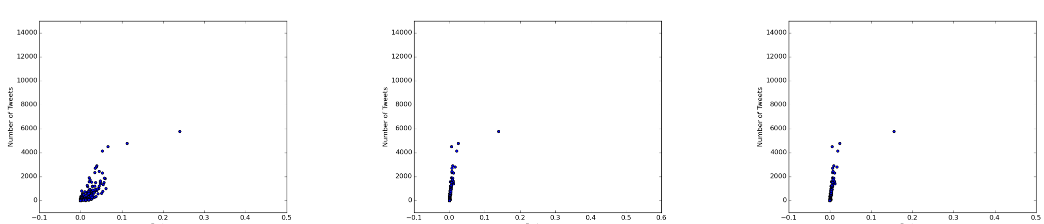
```
Currently working with #sb49
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.922
Model:                            OLS   Adj. R-squared:                  0.919
Method:                 Least Squares   F-statistic:                     417.5
Date:                Thu, 17 Mar 2016   Prob (F-statistic):          1.08e-282
Time:                        02:10:54   Log-Likelihood:                -79.482
No. Observations:                 547   AIC:                             189.0
Df Residuals:                     532   BIC:                             253.5
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
                coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1           -1.9093      0.429     -4.448      0.000        -2.752    -1.066
x2           30.9628      2.169     14.273      0.000        26.701    35.224
x3           -9.0643      0.861    -10.528      0.000       -10.755    -7.373
x4           -2.6180      0.307     -8.528      0.000        -3.221    -2.015
x5           86.3467      7.759     11.128      0.000        71.104   101.590
x6          -86.2569     15.921     -5.418      0.000      -117.532   -54.982
x7           77.9401     16.107      4.839      0.000        46.299   109.581
x8          -12.4929      2.918     -4.282      0.000       -18.224    -6.761
x9        -1.062e+04   2.93e+04     -0.362      0.717     -6.82e+04  4.69e+04
x10           1.3592      1.438      0.945      0.345        -1.465     4.184
x11         -23.2565      4.169     -5.579      0.000       -31.446   -15.067
x12          -9.2562      3.346     -2.766      0.006       -15.829    -2.683
x13          -2.4810      2.814     -0.882      0.378        -8.009     3.047
x14         -2.52e+04   6.94e+04    -0.363      0.717     -1.62e+05  1.11e+05
x15         2.517e+04   6.94e+04     0.362      0.717     -1.11e+05  1.62e+05
==============================================================================
Omnibus:                     1020.076   Durbin-Watson:                   1.658
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1235742.075
Skew:                          12.126   Prob(JB):                         0.00
Kurtosis:                     234.583   Cond. No.                     1.07e+06
==============================================================================
```

Figure 15: sFigure showing p,t and R^2 values for #sb49

Significant Features:

The significant features for this variable were found to be x6(moving average), x4(time of tweet) and x10(past value). The presence of x6 makes sense as the moving average denotes the current mean and is a strong indicator of the range we can expect the future values to lie in.

The scatter plot of Predictant values versus the significant features is shown below



(a) (b) (c)

Figure 16: Scatter plot of significant variable with Predictant for #patriots

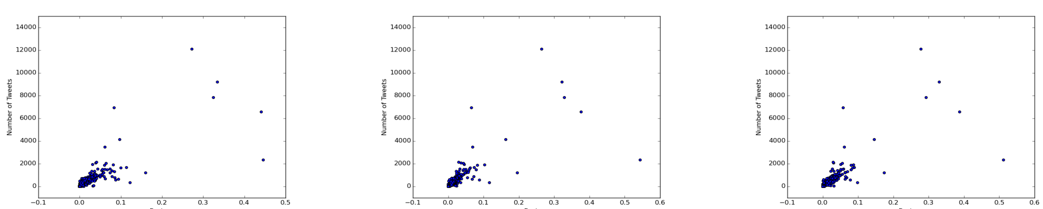
Linear model for #nfl

```
Currently working with #nfl
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.726
Model:                            OLS   Adj. R-squared:                  0.720
Method:                 Least Squares   F-statistic:                     114.5
Date:                Thu, 17 Mar 2016   Prob (F-statistic):          2.00e-159
Time:                        02:10:53   Log-Likelihood:                -476.58
No. Observations:                 618   AIC:                             981.2
Df Residuals:                     604   BIC:                             1043.
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            -9.6759      0.856    -11.306      0.000       -11.357    -7.995
x2            10.8040      2.934      3.682      0.000         5.042    16.566
x3            -4.8498      1.569     -3.091      0.002        -7.932    -1.768
x4            -5.9471      0.745     -7.985      0.000        -7.410    -4.484
x5             5.2639      3.760      1.400      0.162        -2.120    12.647
x6             3.7172      5.341      0.696      0.487        -6.772    14.207
x7           -44.2920      7.910     -5.599      0.000       -59.827   -28.757
x8            11.1935      1.834      6.104      0.000         7.592    14.795
x9             0.1203      1.423      0.085      0.933        -2.674     2.914
x10           -2.3161      1.198     -1.933      0.054        -4.669     0.037
x11            0.6009      2.066      0.291      0.771        -3.457     4.659
x12           10.1402      2.066      4.907      0.000         6.082    14.198
x13           11.2701      1.949      5.784      0.000         7.443    15.097
x14           12.9365      2.960      4.371      0.000         7.124    18.749
x15           13.0088      3.628      3.585      0.000         5.883    20.134
==============================================================================
Omnibus:                      713.044   Durbin-Watson:                   1.889
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           114761.781
Skew:                           5.185   Prob(JB):                         0.00
Kurtosis:                      68.949   Cond. No.                     2.15e+16
==============================================================================
```

Figure 17: Figure showing p,t and R^2 values for #nfl

Significant Features for this hashtag are x13,x14 and x15. We've seen the reasoning behind these features.

The scatter plot of Predictant values versus the significant features is shown below



(a) (b) (c)

Figure 18: Scatter plot of significant variable with Predictant for #nfl

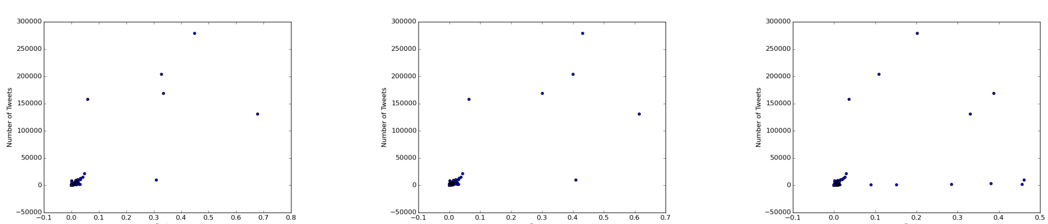
Linear model for #superbowl

```
Currently working with #superbowl
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.886
Model:                            OLS   Adj. R-squared:                  0.884
Method:                 Least Squares   F-statistic:                     331.6
Date:                Thu, 17 Mar 2016   Prob (F-statistic):          2.62e-270
Time:                        02:10:54   Log-Likelihood:                -202.65
No. Observations:                 610   AIC:                             433.3
Df Residuals:                     596   BIC:                             495.1
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1           -16.6300      0.893    -18.614      0.000       -18.385   -14.875
x2            19.4817      2.960      6.582      0.000        13.669    25.294
x3            -3.2324      0.808     -3.999      0.000        -4.820    -1.645
x4            -2.6742      0.408     -6.559      0.000        -3.475    -1.873
x5           -20.6658      5.041     -4.100      0.000       -30.565   -10.766
x6            35.7385      8.980      3.980      0.000        18.102    53.375
x7            58.4107     15.113      3.865      0.000        28.730    88.091
x8           -15.4531      5.076     -3.045      0.002       -25.421    -5.485
x9            -5.7352      2.492     -2.302      0.022       -10.629    -0.841
x10            5.1579      1.013      5.091      0.000         3.168     7.147
x11           -1.6313      2.796     -0.584      0.560        -7.122     3.859
x12          -20.8473      4.476     -4.657      0.000       -29.639   -12.056
x13          -22.1994      2.719     -8.165      0.000       -27.539   -16.860
x14           -1.2402      3.941     -0.315      0.753        -8.980     6.500
x15           -4.8355      5.365     -0.901      0.368       -15.373     5.702
==============================================================================
Omnibus:                     1090.469   Durbin-Watson:                   1.898
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1037195.335
Skew:                          11.345   Prob(JB):                         0.00
Kurtosis:                     203.731   Cond. No.                     9.01e+15
==============================================================================
```

Figure 19: Figure showing p,t and R^2 values for #superbowl

Significant Features for this hashtag are x4,x6 and x10. We've seen the reasoning behind these features.

The scatter plot of Predictant values versus the significant features is shown below



(a) (b) (c)

Figure 20: Scatter plot of significant variable with Predictant for #superbowl

In general, the scatter plots indicate a linear relation between the significant variables and predictant values. This behavior is what we expected and also follows the theoretical explanation of p and t values.

4 Part 4

The average prediction error obtained from 10 fold cross validation over full dataset is:

hashtag	Average Prediction Error
#gohawks	49.83
#gopatriots	192.83
#nfl	648.23
#patriots	1491.11
#sb49	3103.92
#superbowl	6722.28

Table 1: Table shows the average prediction error for different hashtags

Now we divide the dataset for each hashtag into three periods:

Period 1: Before Feb 1, 8:00

Period 2: Between Feb 1, 8:00 AM and 8:00 PM

Period 3: After Feb 1, 8:00 PM

The cross validation results were calculated for each hashtag in these three periods. The error obtained from cross validation is shown in the tables below.

Period	Cross Validation Error
Period 1	34.60
Period 2	10732.80
Period 3	141.65

Table 2: Cross Validation Error for #gopatriots

Period	Cross Validation Error
Period 1	769.17
Period 2	3190.07
Period 3	349.09

Table 3: Cross Validation Error for #gohawks

Period	Cross Validation Error
Period 1	263.32
Period 2	29426.67
Period 3	179.91

Table 4: Cross Validation Error for #patriots

Period	Cross Validation Error
Period 1	47.42
Period 2	143924.43
Period 3	258.24

Table 5: Cross Validation Error for #sb49

Period	Cross Validation Error
Period 1	121.73
Period 2	12275.64
Period 3	147.26

Table 6: Cross Validation Error for #nfl

Period	Cross Validation Error
Period 1	258.96
Period 2	881809.54
Period 3	679.02

Table 7: Cross Validation Error for #superbowl

5 Part 5

We read the period from the file and computed the hashtag as the one which occurs maximum number of times in a given test file. The files along with their obtained hashtag and corresponding tweet prediction for the next hour are shown in table.

File	Computed hashtag	Window	Number of tweets predicted in next hour
Sample1_period1.txt	#nfl	Period 1	7.35e+03
Sample2_period2.txt	#patriots	Period 2	2.98e+6
Sample3_period3.txt	#patriots	Period 3	1.18e+3
Sample4_period1.txt	#nfl	Period 1	9.83e+4
Sample5_period1.txt	#nfl	Period 1	4.04e+3
Sample6_period2.txt	#superbowl	Period 2	1.26e+8
Sample7_period3.txt	#superbowl	Period 3	2.42e+5
Sample8_period1.txt	#nfl	Period 1	8.83e+4
Sample9_period2.txt	#superbowl	Period 2	1.50e+9
Sample10_period3.txt	#nfl	Period 3	3.87e+02

Table 8: Table shows the prediction for number of tweets in next hour for the test files

6 Part 6

Problem Statement:Stock Prediction of selected advertising companies using Sentiment Analysis

The task at hand is to predict the stock market price of a particular company on the next day, given the closing Stock price on the previous day and the sentiment score for that company on the previous day.

The idea is to study the effect of a famous event like SuperBowl on the stock prices of the given companies using sentimental analysis.

Eight companies were taken for analysis from the twitter dataset. The hashtags for those eight companies are. 'doritos' 'makeithappy' 'budweiser' 'mcdonalds' 'mercedes' 'lexus' 'audi' 'minion'

Note:'makeithappy' hashtag represents coca-cola

We did the analysis for a period close to the SuperBowl event.Period from 28 January 2015 to 7 February 2015 was chosen. For each company, the sentiments score on each day was calculated.

Approach for Sentiment Analysis

There are multiple options to do sentimental analysis and classify the tweet into different categories. One way to explore sentiments is to use a list of keywords with tagged sentiment information (e.g. “happy” or “awesome' might have high sentiment scores whereas “terrible” or “awful” might have very low sentiment score.) Then we count the occurrence of these tagged keywords to get a sense of how people feel about that particular company at hand.

We use the AFINN Sentiment Dictionary for our keyword list. Link here:
$http : //www2.imm.dtu.dk/pubdb/views/publication_{d}etails.php?id = 6010$

Using this approach of sentiment analysis , we calculated an average sentiment score for each day for each company.

Sentiment Score(SS):
0: Neutral
large positive value : high positive sentiment
large negative value : high negative sentiment

We plotted the sentiment scores for the 8 companies on the day of SuperBowl event and saw that #minion has the highest positive sentiment and #mercedes was neutral.It is also seen that , no company had negative sentiment scores.

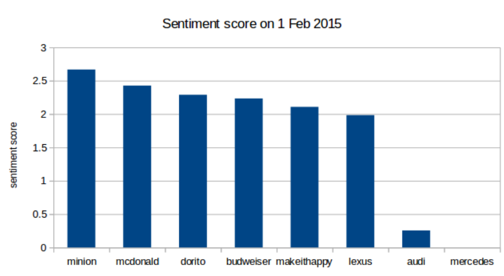


Figure 21: Sentiment scores on February 1, 2015

How Sentiment Analysis is done ?

Take #dorito, for this company we calculate the sentiment score for each day for the given range of days . For calculating the score , we took only those tweet which contain the hashtag related to that particular company. If the company hashtag was present in the tweet text, then using the AFINN Sentiment Dictionary , we calculated the score.

Closing Stock Prices

Other challenge for this task was to collect the closing stock prices for each day for the given range of days.This data was collected from the yahoo stock quote python library. (Library Name:'ystockquote 0.2.4') Once this data of closing stock prices was collected , it was used as a feature in the training dataset. We plotted graphs depicting the variation of closing stock with each day for three comapnies.

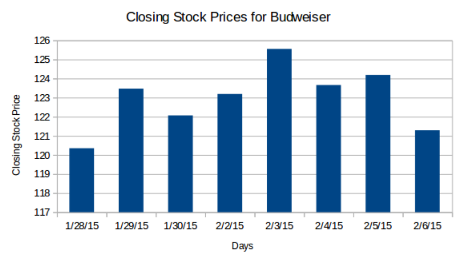


Figure 22: Closing stock prices for budweiser

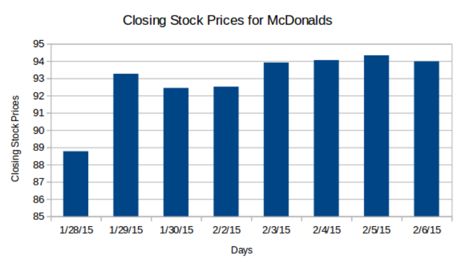


Figure 23: Closing stock prices for McDonalds

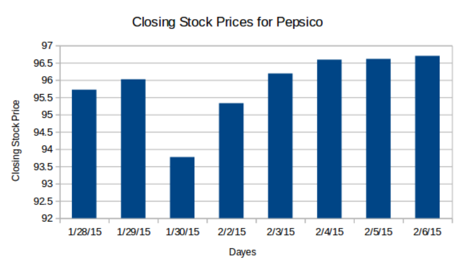


Figure 24: Closing stock prices for Pepsico

It is clear from the graph that closing stock price changes everyday for the three companies. Finally, the training dataset is made for each company.

How is the training data made? Training data consists of days from 01/28/2015 to 02/06/2015 as rows. Each row represent a single day. Features used for the matrix are sentiment score for a day and the closing stock price on that day. The training label was the closing stock price on the next day. If closing stock price for the next day was not available(as in case of Saturday and Sunday) , we used the stock price on the next available day.

Further work to be done:
The training dataset will be used to make a linear regression model. Once the model is built, we will check the cross validation accuracy of the model.Then make a testing dataset from the days after 02/06/2016 and predict the stock prices for the next days and find the testing RMSE. Since, advertising companies are more talked about near the SuperBowl event, it is likely that the stock prices of these advertising companies will increase.So if we predict a decrease in the stock prices for a particular company after the event, then company can change its marketing strategies for handling with the decrease in the stock prices.But how public sentiment effects the stock price has to studied further.

Other approaches:
Sentimental analysis done in this task can be modified.Instead of the sentiment score , we can find score for different categories of the mood and then use each category as a feature. There are few online tools available for this task like OpinionFiner. Apart from this , we can make a language model of the tweets and combine it with the sentiment analysis features.

References for Part 6

1.Mittal, Anshul, and Arpit Goel. "Stock prediction using twitter sentiment analysis." Standford University, CS229 (2012).
2.Chung, Sang, and Sandy Liu. "Predicting stock market fluctuations from twitter." Berkeley, California (2011).
3.Au, Benjamin, Qian Zhang, and Wanlu Zhang. "Learning Dow Jones From Twitter Sentiment."