EE219 Large Scale Data Mining

Winter 2018

Project 5 Popularity Prediction on Twitter

Shizhong Hao 605035020 Zhicheng Zhang 104946990

Yucong Wang 305036163 Zhengtao Zhou 005036433

Background

As one of the leading platforms of social communications and information dissemination, Twitter has become a major source of information for common Web users. Twitter, with its public discussion model, is a good platform to predict future popularity of a topic or an event. Conversations on Twitter are featured with their “burstiness”, the phenomenon that a topic of discussion suddenly gains a considerable popularity, and then quickly fades away. Such bursting topics are usually triggered by breaking news, real world events, malicious rumors, or various types of behavior cascades such as campaigns of persuasion. If knowing current and previous tweet activity for a hashtag, we can predict if it becomes more prominent and trendy in the future and if yes by how much.

Introduction

In this project, twitter data is collected by querying popular hashtags related to the Super Bowl spanning a period starting from 2 weeks before the game to a week after the game. We are required to use these data to train a regression model and then use the model to making predictions for other hashtags. The test data consists of tweets containing a hashtag in a specified time window, and we have then used our model to predict the number of tweets containing the hash-tag posted within one hour immediately following the given time window. Finally, we use the knowing data to define our problem and try to implement our idea and show how to work.

The dataset contains raw tweets information that was obtained during the 2015 Super Bowl, spanning a period starting from 2 weeks before the event to 1 week after the event. Every tweet is related to one or more of the following 6 hashtags: #GoHawks, #GoPatriots, #NFL, #Patriots, #SB49 and #SuperBowl. Each text file contains all tweets that have one of the 6 hashtags.

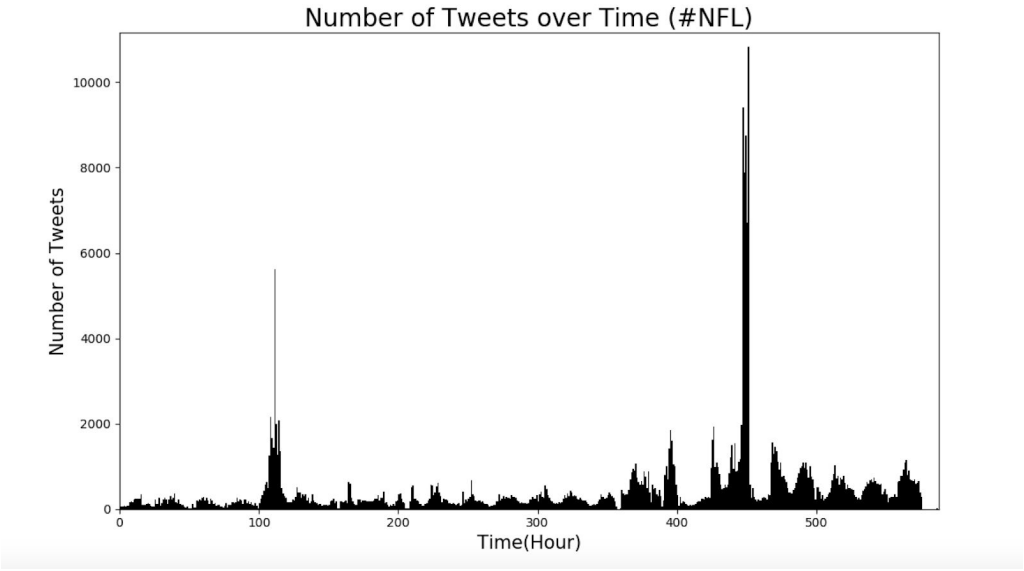
In the original text files, information is stored as JSON strings and tweets are sorted with respect to their posting time, from the earliest to the latest. Each tweet can be converted to a Python Dictionary object, which includes information such as the name of the author, the posting time of the tweet, the number of followers of the user, etc.

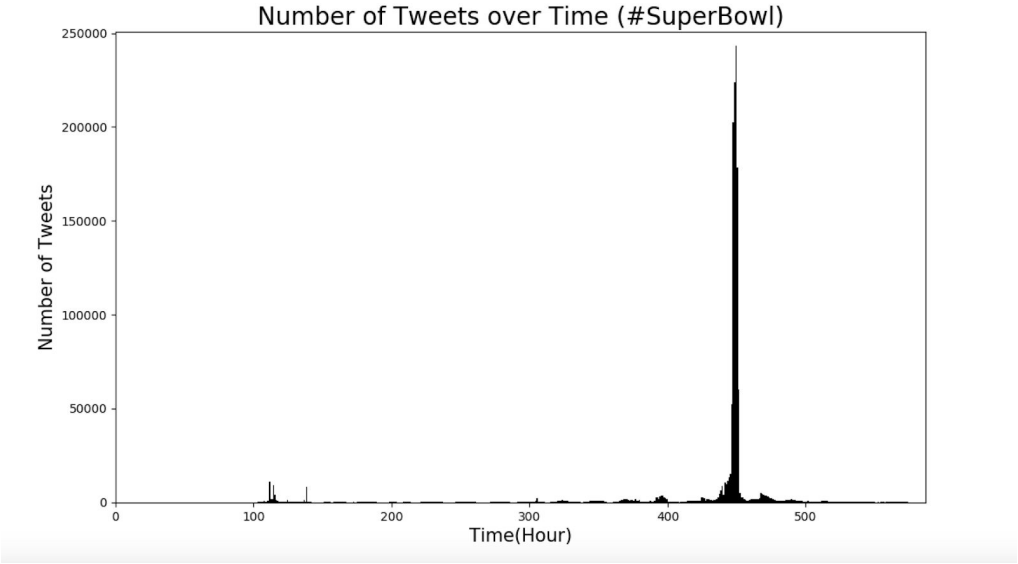
Part 1 popularity prediction

Problem 1.1

In this part, in order to get a whole sense of the pattern embedded in the datasets, first we want to find some basic statistical results, including the average number of tweets per hour, the average number of followers, and the average number of retweets.
Average number of tweets per hour: we set the hourly time window and then fill in the missing data. The first hourly time window starts at the earliest time of all the posted tweets, then subsequently add later hours after that. The posting time of each tweet can be directly accessed after the JSON string has been loaded. We calculate the quotient of the total tweets and the total hours passed as the average number of tweets.
Average number of followers for users: we should know the total number of users, because one user may post multiple tweets during the designated period, there could be multiple records of the user in the dataset. Therefore, we use a Python Dictionary to obtain the number of followers of each user. For the case that one user posted multiple tweets, the number of his followers may change over time, so it is reasonable to use the average value of all records.
Average number of retweets: we calculate the total retweets and the number of tweets to get the average value of retweets to get a sense of the average popularity of the tweets in this period. As the instruction described, we set the timestamp for a tweet by json object['citation date'], also assess the number of retweets of a tweet by json object['metrics']['citations']['total'].
Besides, the number of followers of the person tweeting can be retrieved by json object['author']['followers'].
The results for all the hashtags are shown below.

Statistics for #GoHawks
   Average number of tweets per hour: 325.371591304
   Average number of followers per user: 1588.18866293
   Average number of retweets per tweet: 2.01461708551
##
Statistics for #GoPatriots
   Average number of tweets per hour: 45.6945105736
   Average number of followers per user: 1294.46936646
   Average number of retweets per tweet: 1.40008386703
##
Statistics for #NFL
   Average number of tweets per hour: 441.323431137
   Average number of followers per user: 4221.07698787
   Average number of retweets per tweet: 1.5385331089
##
Statistics for #Patriots
   Average number of tweets per hour: 834.555509164
   Average number of followers per user: 1695.27106215
   Average number of retweets per tweet: 1.78281564917
##
We also plot the number of tweets in hour over time for #SuperBowl and #NFL.





From the two plots, we can see that the there are two peaks in the Super Bowl period. The highest peaks occurred on the Super Bowl time, which makes sense because people post much more tweets during the important event. Another peak occurs approximately 350 hours before Super Bowl, which should be another important event of NFL. The characteristic of both hashtags is that they “burst” in short periods of time.

And if we compare the 2 plots, we could notice that the maximum value of number of tweets of #SuperBowl at peak hour is much larger than that of #NFL’s, meaning that #SuperBowl’s spike is more extreme.

Problem 1.2

For this part, we are going to fit a linear regression model with 5 features to predict tweets in the next hour, from data in the previous hour. The 5 features are:

-number of tweets
-total number of retweets
-sum of the number of followers of the users
-maximum number of followers of the users
-time of the day

As for the “time of the day” feature, we used one-hot encoding method, where we introduced 24 new dimensions to the feature vector, each one representing an hour of the day, and is turned to 1 if the tweet was at the corresponding hour of the day, and 0 otherwise.

We built a DataFrame using the parameters extracted from tweet data based on hours and rewrite the useful information into a new file to shorten the processing time for future using.

For each hashtag we trained a linear regression model and got the predicted values, we calculate RMSE and R-squared measure, when the linear model has high accuracy, R squared value will be closer to 1. Therefore, the value of R-squared can be used as the representation of the accuracy of the model. We also perform t-test and calculate p value as the representation of the accuracy of the model. We use the library statsmodels.api in Python to get a statistic summary of t test.

The results are shown below.

==========

Processing hashtag "#GoHawks".....

RMSE = 910.9921996805798

R-squared = 0.532

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	1.4237	0.164	8.659	0.000	1.101	1.747
num_retweets	-19.8237	3.889	-5.097	0.000	-27.463	-12.184
sum_followers	-0.0003	8.51e-05	-3.270	0.001	-0.000	-0.000
max_followers	0.0002	0.000	1.126	0.261	-0.000	0.001
0th_hour	16.9938	187.031	0.091	0.928	-350.387	384.375
1th_hour	12.1762	186.821	0.065	0.948	-354.793	379.145
2th_hour	3.2168	186.790	0.017	0.986	-363.691	370.124
3th_hour	23.4880	190.666	0.123	0.902	-351.034	398.010
4th_hour	34.2407	190.662	0.180	0.858	-340.273	408.754
5th_hour	52.2219	190.884	0.274	0.785	-322.727	427.171
6th_hour	120.2952	190.978	0.630	0.529	-254.838	495.428
7th_hour	142.3044	190.957	0.745	0.456	-232.788	517.397
8th_hour	205.6231	193.187	1.064	0.288	-173.850	585.096
9th_hour	197.4692	191.701	1.030	0.303	-179.085	574.023
10th_hour	324.2885	191.838	1.690	0.092	-52.535	701.112
11th_hour	215.3796	191.699	1.124	0.262	-161.170	591.929
12th_hour	17.7893	192.423	0.092	0.926	-360.182	395.761
13th_hour	253.5218	192.556	1.317	0.189	-124.711	631.755
14th_hour	999.8690	193.175	5.176	0.000	620.420	1379.318
15th_hour	-243.6088	198.152	-1.229	0.219	-632.835	145.618
16th_hour	199.9072	193.022	1.036	0.301	-179.241	579.055
17th_hour	172.4385	193.656	0.890	0.374	-207.955	552.832
18th_hour	-130.1525	192.812	-0.675	0.500	-508.889	248.583
19th_hour	-16.4555	197.652	-0.083	0.934	-404.699	371.788
20th_hour	157.2451	193.179	0.814	0.416	-222.213	536.703
21th_hour	85.4406	193.508	0.442	0.659	-294.662	465.543
22th_hour	26.5137	191.449	0.138	0.890	-349.545	402.572
23th_hour	-7.8748	190.873	-0.041	0.967	-382.803	367.053

Omnibus:	854.260	Durbin-Watson:	2.193
Prob(Omnibus):	0.000	Jarque-Bera (JB):	683918.794
Skew:	7.479	Prob(JB):	0.00
Kurtosis:	170.706	Cond. No.	1.62e+07

P values:

num_tweets	5.249441e-17
num_retweets	4.744914e-07
sum_followers	1.141717e-03
max_followers	2.607242e-01
0th_hour	9.276361e-01
1th_hour	9.480579e-01
2th_hour	9.862662e-01
3th_hour	9.020025e-01
4th_hour	8.575415e-01
5th_hour	7.845103e-01
6th_hour	5.290263e-01
7th_hour	4.564584e-01
8th_hour	2.876252e-01
9th_hour	3.034199e-01
10th_hour	9.151201e-02
11th_hour	2.617010e-01
12th_hour	9.263749e-01
13th_hour	1.885149e-01
14th_hour	3.181616e-07
15th_hour	2.194461e-01
16th_hour	3.008091e-01
17th_hour	3.736187e-01
18th_hour	4.999437e-01
19th_hour	9.336788e-01
20th_hour	4.160052e-01
21th_hour	6.589981e-01
22th_hour	8.899036e-01
23th_hour	9.671063e-01

T values:

num_tweets	8.659436
num_retweets	-5.097228
sum_followers	-3.270232
max_followers	1.125838
0th_hour	0.090861
1th_hour	0.065176
2th_hour	0.017221
3th_hour	0.123189
4th_hour	0.179589
5th_hour	0.273580
6th_hour	0.629892
7th_hour	0.745217
8th_hour	1.064375
9th_hour	1.030090
10th_hour	1.690426
11th_hour	1.123532
12th_hour	0.092449
13th_hour	1.316616
14th_hour	5.175981
15th_hour	-1.229401
16th_hour	1.035673
17th_hour	0.890439
18th_hour	-0.675024
19th_hour	-0.083255
20th_hour	0.813985
21th_hour	0.441536
22th_hour	0.138490
23th_hour	-0.041257

Processing hashtag "#GoPatriots".....

RMSE = 188.2116004251095

R-squared = 0.627

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	0.2511	0.203	1.236	0.217	-0.148	0.650
num_retweets	-10.9993	3.836	-2.867	0.004	-18.535	-3.464
sum_followers	0.0006	0.000	3.230	0.001	0.000	0.001
max_followers	-0.0007	0.000	-3.548	0.000	-0.001	-0.000
0th_hour	3.1339	39.392	0.080	0.937	-74.245	80.512
1th_hour	3.7415	39.391	0.095	0.924	-73.634	81.117
2th_hour	6.1516	39.394	0.156	0.876	-71.231	83.535
3th_hour	9.5014	39.413	0.241	0.810	-67.918	86.921
4th_hour	10.9079	39.429	0.277	0.782	-66.543	88.359
5th_hour	11.4425	39.409	0.290	0.772	-65.970	88.855
6th_hour	9.7617	39.410	0.248	0.804	-67.651	87.174
7th_hour	20.9183	39.428	0.531	0.596	-56.531	98.368
8th_hour	15.4460	39.403	0.392	0.695	-61.953	92.845
9th_hour	23.8400	39.557	0.603	0.547	-53.862	101.542
10th_hour	14.1391	39.437	0.359	0.720	-63.328	91.606
11th_hour	33.6158	40.157	0.837	0.403	-45.265	112.497
12th_hour	119.8666	39.450	3.038	0.002	42.374	197.359
13th_hour	49.6551	39.945	1.243	0.214	-28.810	128.120
14th_hour	-22.2959	40.144	-0.555	0.579	-101.151	56.559
15th_hour	112.0844	39.919	2.808	0.005	33.671	190.497
16th_hour	80.0945	39.711	2.017	0.044	2.089	158.100
17th_hour	-122.2721	39.740	-3.077	0.002	-200.335	-44.210
18th_hour	13.7698	39.708	0.347	0.729	-64.229	91.769
19th_hour	7.9766	39.415	0.202	0.840	-69.447	85.400
20th_hour	7.8640	39.419	0.199	0.842	-69.567	85.295
21th_hour	6.7570	39.440	0.171	0.864	-70.715	84.229
22th_hour	2.7549	39.393	0.070	0.944	-74.625	80.135
23th_hour	2.9905	40.239	0.074	0.941	-76.051	82.032

Omnibus:	272.512	Durbin-Watson:	2.136
Prob(Omnibus):	0.000	Jarque-Bera (JB):	358447.958
Skew:	-0.297	Prob(JB):	0.00
Kurtosis:	125.315	Cond. No.	2.16e+06

P values:

num_tweets	0.217032
num_retweets	0.004301
sum_followers	0.001312
max_followers	0.000421
0th_hour	0.936619
1th_hour	0.924361
2th_hour	0.875970
3th_hour	0.809590
4th_hour	0.782157
5th_hour	0.771659
6th_hour	0.804461
7th_hour	0.595951
8th_hour	0.695210
9th_hour	0.546974
10th_hour	0.720091
11th_hour	0.402898
12th_hour	0.002492
13th_hour	0.214371
14th_hour	0.578848
15th_hour	0.005166
16th_hour	0.044192
17th_hour	0.002197
18th_hour	0.728895
19th_hour	0.839700
20th_hour	0.841948
21th_hour	0.864031
22th_hour	0.944272
23th_hour	0.940784

T values:

num_tweets	1.235883
num_retweets	-2.867217
sum_followers	3.230065
max_followers	-3.548034
0th_hour	0.079557
1th_hour	0.094986
2th_hour	0.156153
3th_hour	0.241072
4th_hour	0.276646
5th_hour	0.290350
6th_hour	0.247698
7th_hour	0.530542
8th_hour	0.392001
9th_hour	0.602676
10th_hour	0.358521
11th_hour	0.837107
12th_hour	3.038409
13th_hour	1.243081
14th_hour	-0.555401
15th_hour	2.807810
16th_hour	2.016915
17th_hour	-3.076767
18th_hour	0.346774
19th_hour	0.202373
20th_hour	0.199497
21th_hour	0.171325
22th_hour	0.069934
23th_hour	0.074318

Processing hashtag "#NFL".....

RMSE = 567.1052814025201

R-squared = 0.585

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	0.4752	0.103	4.601	0.000	0.272	0.678
num_retweets	-1.5832	2.807	-0.564	0.573	-7.096	3.930
sum_followers	7.315e-05	2.65e-05	2.756	0.006	2.1e-05	0.000
max_followers	-9.432e-05	3.7e-05	-2.547	0.011	-0.000	-2.16e-05
0th_hour	56.3083	116.806	0.482	0.630	-173.124	285.741
1th_hour	67.4735	116.873	0.577	0.564	-162.091	297.038
2th_hour	76.3124	116.595	0.655	0.513	-152.705	305.330
3th_hour	92.3850	116.643	0.792	0.429	-136.727	321.497
4th_hour	83.8884	116.783	0.718	0.473	-145.500	313.276
5th_hour	169.5487	116.801	1.452	0.147	-59.874	398.971
6th_hour	196.7151	117.237	1.678	0.094	-33.563	426.993
7th_hour	149.7344	117.061	1.279	0.201	-80.199	379.668
8th_hour	166.4218	117.790	1.413	0.158	-64.943	397.787
9th_hour	120.6821	117.350	1.028	0.304	-109.818	351.182
10th_hour	237.9457	117.583	2.024	0.043	6.986	468.905
11th_hour	244.0414	120.708	2.022	0.044	6.945	481.138
12th_hour	83.3442	120.386	0.692	0.489	-153.121	319.809
13th_hour	220.2633	121.927	1.807	0.071	-19.229	459.755
14th_hour	634.0510	121.109	5.235	0.000	396.167	871.934
15th_hour	15.7214	123.964	0.127	0.899	-227.771	259.213
16th_hour	269.0316	122.193	2.202	0.028	29.018	509.045
17th_hour	218.6963	125.524	1.742	0.082	-27.860	465.253
18th_hour	283.3371	122.493	2.313	0.021	42.734	523.940
19th_hour	-198.0234	122.052	-1.622	0.105	-437.760	41.713
20th_hour	86.8392	120.082	0.723	0.470	-149.027	322.706
21th_hour	81.4536	119.888	0.679	0.497	-154.032	316.939
22th_hour	33.0973	119.432	0.277	0.782	-201.494	267.688
23th_hour	63.9559	119.234	0.536	0.592	-170.245	298.156

Omnibus:	581.889	Durbin-Watson:	2.364
Prob(Omnibus):	0.000	Jarque-Bera (JB):	286567.909
Skew:	3.503	Prob(JB):	0.00
Kurtosis:	111.016	Cond. No.	2.97e+07

P values:

num_tweets	5.202953e-06
num_retweets	5.729409e-01
sum_followers	6.040466e-03
max_followers	1.113151e-02
0th_hour	6.299479e-01
1th_hour	5.639550e-01
2th_hour	5.130521e-01
3th_hour	4.286772e-01
4th_hour	4.728576e-01
5th_hour	1.471726e-01
6th_hour	9.391927e-02
7th_hour	2.013874e-01
8th_hour	1.582512e-01
9th_hour	3.042081e-01
10th_hour	4.348326e-02
11th_hour	4.367808e-02
12th_hour	4.890328e-01
13th_hour	7.137615e-02
14th_hour	2.333895e-07
15th_hour	8.991269e-01
16th_hour	2.809471e-02
17th_hour	8.201196e-02
18th_hour	2.107983e-02
19th_hour	1.052705e-01
20th_hour	4.698793e-01
21th_hour	4.971567e-01
22th_hour	7.817891e-01
23th_hour	5.919009e-01

T values:

num_tweets	4.601158
num_retweets	-0.564058
sum_followers	2.756098
max_followers	-2.547021
0th_hour	0.482066
1th_hour	0.577321
2th_hour	0.654511
3th_hour	0.792033
4th_hour	0.718324
5th_hour	1.451604
6th_hour	1.677932
7th_hour	1.279115
8th_hour	1.412868
9th_hour	1.028397
10th_hour	2.023632
11th_hour	2.021750
12th_hour	0.692306
13th_hour	1.806511
14th_hour	5.235393
15th_hour	0.126822
16th_hour	2.201694
17th_hour	1.742266
18th_hour	2.313089
19th_hour	-1.622452
20th_hour	0.723168
21th_hour	0.679414
22th_hour	0.277122
23th_hour	0.536392

Processing hashtag "#Patriots".....

RMSE = 2302.118322202846

R-squared = 0.723

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	0.8230	0.039	21.051	0.000	0.746	0.900
num_retweets	-19.7359	3.548	-5.562	0.000	-26.706	-12.766
sum_followers	3.505e-05	2.54e-05	1.382	0.168	-1.48e-05	8.49e-05
max_followers	0.0001	9.69e-05	1.402	0.161	-5.45e-05	0.000
0th_hour	59.0273	472.067	0.125	0.901	-868.216	986.270
1th_hour	47.4428	471.927	0.101	0.920	-879.524	974.409
2th_hour	50.7466	471.983	0.108	0.914	-876.330	977.823
3th_hour	77.6890	472.291	0.164	0.869	-849.993	1005.371
4th_hour	79.4322	473.004	0.168	0.867	-849.650	1008.515
5th_hour	76.4594	474.558	0.161	0.872	-855.676	1008.594
6th_hour	70.7354	473.728	0.149	0.881	-859.769	1001.240
7th_hour	21.2432	476.106	0.045	0.964	-913.932	956.419
8th_hour	96.7734	473.981	0.204	0.838	-834.228	1027.774
9th_hour	500.0589	474.386	1.054	0.292	-431.738	1431.856
10th_hour	1805.4470	474.871	3.802	0.000	872.696	2738.198
11th_hour	-44.4154	487.337	-0.091	0.927	-1001.651	912.820
12th_hour	-306.7086	489.203	-0.627	0.531	-1267.609	654.192
13th_hour	31.0503	488.518	0.064	0.949	-928.505	990.606
14th_hour	435.8040	491.708	0.886	0.376	-530.017	1401.625
15th_hour	-78.9151	492.324	-0.160	0.873	-1045.946	888.115
16th_hour	-463.0496	498.382	-0.929	0.353	-1441.979	515.880
17th_hour	1203.5731	486.460	2.474	0.014	248.061	2159.085
18th_hour	-92.8105	490.764	-0.189	0.850	-1056.778	871.157
19th_hour	-841.7744	484.449	-1.738	0.083	-1793.337	109.788
20th_hour	271.5677	491.369	0.553	0.581	-693.587	1236.722
21th_hour	68.8244	483.349	0.142	0.887	-880.578	1018.226
22th_hour	95.4799	482.402	0.198	0.843	-852.062	1043.022
23th_hour	28.7491	482.046	0.060	0.952	-918.094	975.593

Omnibus:	974.392	Durbin-Watson:	1.943
Prob(Omnibus):	0.000	Jarque-Bera (JB):	835105.074
Skew:	9.619	Prob(JB):	0.00
Kurtosis:	186.777	Cond. No.	5.27e+07

P values:

num_tweets	6.991804e-73
num_retweets	4.140126e-08
sum_followers	1.676426e-01
max_followers	1.614308e-01
0th_hour	9.005368e-01
1th_hour	9.199596e-01
2th_hour	9.144167e-01
3th_hour	8.694018e-01
4th_hour	8.666980e-01
5th_hour	8.720595e-01
6th_hour	8.813577e-01
7th_hour	9.644272e-01
8th_hour	8.382938e-01
9th_hour	2.922844e-01
10th_hour	1.593389e-04
11th_hour	9.274148e-01
12th_hour	5.309443e-01
13th_hour	9.493431e-01
14th_hour	3.758333e-01
15th_hour	8.727096e-01
16th_hour	3.532349e-01
17th_hour	1.365125e-02
18th_hour	8.500720e-01
19th_hour	8.283357e-02
20th_hour	5.807061e-01
21th_hour	8.868227e-01
22th_hour	8.431750e-01
23th_hour	9.524639e-01

T values:

num_tweets	21.051240
num_retweets	-5.561936
sum_followers	1.381613
max_followers	1.402132
0th_hour	0.125040
1th_hour	0.100530
2th_hour	0.107518
3th_hour	0.164494
4th_hour	0.167931
5th_hour	0.161117
6th_hour	0.149317
7th_hour	0.044619
8th_hour	0.204172
9th_hour	1.054118
10th_hour	3.801970
11th_hour	-0.091139
12th_hour	-0.626956
13th_hour	0.063560
14th_hour	0.886307
15th_hour	-0.160291
16th_hour	-0.929107
17th_hour	2.474147
18th_hour	-0.189114
19th_hour	-1.737591
20th_hour	0.552676
21th_hour	0.142391
22th_hour	0.197926
23th_hour	0.059640

Processing hashtag "#SB49".....

RMSE = 3799.8433322883743

R-squared = 0.857

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	0.9414	0.030	31.904	0.000	0.883	0.999
num_retweets	-75.1845	10.554	-7.124	0.000	-95.916	-54.453
sum_followers	3.355e-06	4.49e-06	0.748	0.455	-5.46e-06	1.22e-05
max_followers	0.0002	4.15e-05	4.185	0.000	9.23e-05	0.000
0th_hour	-70.8379	780.105	-0.091	0.928	-1603.158	1461.482
1th_hour	-102.1442	782.115	-0.131	0.896	-1638.412	1434.124
2th_hour	-164.2185	782.418	-0.210	0.834	-1701.082	1372.645
3th_hour	-197.2752	782.395	-0.252	0.801	-1734.093	1339.542
4th_hour	-256.9267	786.731	-0.327	0.744	-1802.261	1288.407
5th_hour	2494.5636	785.829	3.174	0.002	951.000	4038.127
6th_hour	1319.3125	785.055	1.681	0.093	-222.729	2861.354
7th_hour	-1295.7993	807.914	-1.604	0.109	-2882.742	291.144
8th_hour	-1039.8350	803.523	-1.294	0.196	-2618.153	538.483
9th_hour	30.3122	819.470	0.037	0.971	-1579.329	1639.953
10th_hour	-182.7876	809.576	-0.226	0.821	-1772.996	1407.421
11th_hour	205.5069	804.837	0.255	0.799	-1375.392	1786.406
12th_hour	-430.8907	817.501	-0.527	0.598	-2036.664	1174.883
13th_hour	-1031.0163	804.730	-1.281	0.201	-2611.704	549.672
14th_hour	-1397.9380	797.865	-1.752	0.080	-2965.143	169.267
15th_hour	-65.5140	799.750	-0.082	0.935	-1636.421	1505.393
16th_hour	-50.2012	799.358	-0.063	0.950	-1620.339	1519.936
17th_hour	-292.5488	803.164	-0.364	0.716	-1870.161	1285.063
18th_hour	-59.0724	795.882	-0.074	0.941	-1622.381	1504.237
19th_hour	-174.8924	797.157	-0.219	0.826	-1740.706	1390.921
20th_hour	-136.8856	795.658	-0.172	0.863	-1699.754	1425.983
21th_hour	-6.9382	795.166	-0.009	0.993	-1568.841	1554.965
22th_hour	40.7550	795.142	0.051	0.959	-1521.100	1602.610
23th_hour	41.0264	795.259	0.052	0.959	-1521.059	1603.112

Omnibus:	945.198	Durbin-Watson:	1.327
Prob(Omnibus):	0.000	Jarque-Bera (JB):	749151.776
Skew:	9.166	Prob(JB):	0.00
Kurtosis:	177.653	Cond. No.	4.82e+08

P values:

num_tweets	1.210368e-127
num_retweets	3.279663e-12
sum_followers	4.550319e-01
max_followers	3.319350e-05
0th_hour	9.276799e-01
1th_hour	8.961391e-01
2th_hour	8.338338e-01
3th_hour	8.010241e-01
4th_hour	7.441124e-01
5th_hour	1.584506e-03
6th_hour	9.341582e-02
7th_hour	1.093088e-01
8th_hour	1.961708e-01
9th_hour	9.705063e-01
10th_hour	8.214542e-01
11th_hour	7.985552e-01
12th_hour	5.983466e-01
13th_hour	2.006599e-01
14th_hour	8.030923e-02
15th_hour	9.347413e-01
16th_hour	9.499469e-01
17th_hour	7.158134e-01
18th_hour	9.408601e-01
19th_hour	8.264228e-01
20th_hour	8.634681e-01
21th_hour	9.930413e-01
22th_hour	9.591407e-01
23th_hour	9.588750e-01

T values:

num_tweets	31.904034
num_retweets	-7.123533
sum_followers	0.747577
max_followers	4.184694
0th_hour	-0.090806
1th_hour	-0.130600
2th_hour	-0.209886
3th_hour	-0.252143
4th_hour	-0.326575
5th_hour	3.174434
6th_hour	1.680536
7th_hour	-1.603883
8th_hour	-1.294095
9th_hour	0.036990
10th_hour	-0.225782
11th_hour	0.255340
12th_hour	-0.527083
13th_hour	-1.281196
14th_hour	-1.752098
15th_hour	-0.081918
16th_hour	-0.062802
17th_hour	-0.364246
18th_hour	-0.074223
19th_hour	-0.219395
20th_hour	-0.172041
21th_hour	-0.008725
22th_hour	0.051255
23th_hour	0.051589

Processing hashtag "#SuperBowl".....

RMSE = 6433.372471300713

R-squared = 0.870

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	2.2978	0.084	27.221	0.000	2.132	2.464
num_retweets	-3.9616	1.741	-2.275	0.023	-7.381	-0.542
sum_followers	-0.0002	1.2e-05	-19.193	0.000	-0.000	-0.000
max_followers	0.0011	0.000	9.488	0.000	0.001	0.001
0th_hour	-1152.9725	1329.978	-0.867	0.386	-3765.347	1459.402
1th_hour	-523.1336	1320.899	-0.396	0.692	-3117.676	2071.408
2th_hour	-320.5041	1321.701	-0.242	0.808	-2916.622	2275.614
3th_hour	-754.8762	1327.757	-0.569	0.570	-3362.888	1853.136
4th_hour	-325.6118	1322.931	-0.246	0.806	-2924.144	2272.921
5th_hour	-656.1878	1335.849	-0.491	0.623	-3280.096	1967.720
6th_hour	-210.0972	1330.078	-0.158	0.875	-2822.669	2402.474
7th_hour	-707.7970	1337.028	-0.529	0.597	-3334.019	1918.425
8th_hour	-50.0254	1340.333	-0.037	0.970	-2682.741	2582.690
9th_hour	-659.1034	1337.665	-0.493	0.622	-3286.578	1968.371
10th_hour	-954.4856	1376.758	-0.693	0.488	-3658.747	1749.776
11th_hour	-460.5540	1373.900	-0.335	0.738	-3159.202	2238.094
12th_hour	-182.9047	1365.986	-0.134	0.894	-2866.008	2500.198
13th_hour	981.7322	1375.887	0.714	0.476	-1720.818	3684.283
14th_hour	5249.6969	1390.505	3.775	0.000	2518.433	7980.961
15th_hour	-2006.0965	1395.997	-1.437	0.151	-4748.148	735.955
16th_hour	-1450.4399	1388.073	-1.045	0.297	-4176.926	1276.047
17th_hour	24.4359	1389.325	0.018	0.986	-2704.509	2753.381
18th_hour	-2880.8218	1370.936	-2.101	0.036	-5573.647	-187.997
19th_hour	-3352.1905	1369.294	-2.448	0.015	-6041.791	-662.590
20th_hour	-1327.1132	1361.904	-0.974	0.330	-4002.199	1347.972
21th_hour	-1105.3829	1355.967	-0.815	0.415	-3768.806	1558.040
22th_hour	-799.4634	1351.176	-0.592	0.554	-3453.476	1854.549
23th_hour	-1423.8574	1358.102	-1.048	0.295	-4091.475	1243.760

Omnibus:	1121.631	Durbin-Watson:	1.721
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1763400.781
Skew:	12.869	Prob(JB):	0.00
Kurtosis:	270.505	Cond. No.	7.60e+08

P values:

num_tweets	1.848781e-104
num_retweets	2.325511e-02
sum_followers	2.015395e-63
max_followers	6.763354e-20
0th_hour	3.863633e-01
1th_hour	6.922242e-01
2th_hour	8.084868e-01
3th_hour	5.699004e-01
4th_hour	8.056728e-01
5th_hour	6.234683e-01
6th_hour	8.745466e-01
7th_hour	5.967517e-01
8th_hour	9.702408e-01
9th_hour	6.223995e-01
10th_hour	4.884192e-01
11th_hour	7.375877e-01
12th_hour	8.935304e-01
13th_hour	4.758181e-01
14th_hour	1.768838e-04
15th_hour	1.512686e-01
16th_hour	2.965074e-01
17th_hour	9.859736e-01
18th_hour	3.605768e-02
19th_hour	1.466744e-02
20th_hour	3.302533e-01
21th_hour	4.153061e-01
22th_hour	5.543047e-01
23th_hour	2.949004e-01

T values:

num_tweets	27.220695
num_retweets	-2.275486
sum_followers	-19.193210
max_followers	9.487883
0th_hour	-0.866911
1th_hour	-0.396044
2th_hour	-0.242494
3th_hour	-0.568535
4th_hour	-0.246129
5th_hour	-0.491214
6th_hour	-0.157959
7th_hour	-0.529381
8th_hour	-0.037323
9th_hour	-0.492727
10th_hour	-0.693285
11th_hour	-0.335217
12th_hour	-0.133899
13th_hour	0.713527
14th_hour	3.775389
15th_hour	-1.437035
16th_hour	-1.044931
17th_hour	0.017588
18th_hour	-2.101355
19th_hour	-2.448116
20th_hour	-0.974454
21th_hour	-0.815199
22th_hour	-0.591680
23th_hour	-1.048417

When the p-value of each feature turned out to be small, which means it is a significant feature in the feature vector, since a smaller p-value means we need a larger significance value to reject the null hypothesis.

Here we list the accuracy results of all hashtags.

	#GoHawks	#GoPatriots	#NFL	#Patriots	#SB49	#SuperBowl
R-squared	0.532	0.627	0.585	0.723	0.857	0.870
RMSE	910	188	567	2302	3799	6433

Comparing between each hashtag, we can see that larger training datasets yield better accuracy. For example, the dataset for “#GoPatriots” contains about 26000 tweets, and the R-squared value is as low as 0.627. In comparison, the dataset for “#SuperBowl” contains about 1340000 tweets, and the corresponding R-squared value is as high as 0.870.

Basically, for each hashtag the P-values of attributes “number of tweets”, “total number of retweets”, “sum of number of followers” and “max number of followers” are very close to zero, which means that these four attributes are critical and they are important during the regression analysis. While the attribute “time of the day”, which has been split into 24 sub-attributes by one hot encoding, is much less important because the P-values of this attribute is very large and close to 1. Therefore, it is reasonable to exclude “time of the day” without impairing the accuracy of the model.

Besides, we could observe that the RMSE value is very high which means the prediction is not quite accurate, which is reasonable since the characteristic of all hashtags is that they “burst” in a short period of time rather than gradually increasing.

Problem 1.3

For this part, we are required to design a new regression model using features from the papers we find useful for this problem.

We applied the meme features and hashtag features mentioned by the paper, which indicate popularity on the Internet.

-Tweet count. We use the number of tweets containing a hashtag to represent current popularity of the hashtag, instead of using the appearance count of the hashtag. This is because some tweets may use the same hashtag multiple times.

-Author count. Besides tweet count for a hashtag, we also consider the unique number of authors who posted tweets containing the hashtag. This feature can be used to recognize those hashtags automatically posted by some fake accounts.

-Retweet count. Retweeting is the typical way of information diffusion in Twitter. Interesting information can spread quickly and broadly through retweets. If a user retweeted a tweet, that means the content of the tweet successfully attracted the attention of this user and motivated him to share it. Besides indicating the interestingness of messages, the retweeting behavior of a user may also affect his followers.

-Mention count. Mention is a directional sharing behavior in Twitter. Messages can be shared to a designated user using @ as the prefix of the user's name. If a user was mentioned in a tweet with a hashtag, he probably took part in the topic, especially when this mention came from his friends.

-Url ratio. A url in Twitter can be a link of a picture, a song, a video, or a piece of news. High ratio of tweets with urls may indicate a topic about a good song, an interesting picture or video, or a piece of breaking news.

Actually we used "url count" instead of "url ratio", since we found extreme cases where there is only 1 tweet of a certain hashtag during a certain hour and that tweet contains a url in it, which makes the url ratio of that hour 1, and therefore skewed the meaning it is supposed to represent.

-Number of hashtags. Sometimes, some hashtags are not used individually, but are used together with other hashtags, e.g. #boston#explosion. It's reasonable to guess the number of hashtag in tweets are critical to indicate the popularity of the topic.

-Ranking score. Ranking scores are listed in each tweet to show its scores intuitively, which shows its spread ability.

Considering the features we already have in problem 1.2, now the modified training attributes are:

num_tweets: total number of tweets within an hour
num_retweets: total number of retweets within an hour
sum_followers: sum of the number of followers of the users
max_followers: maximum number of followers of the users
time_of_day: one of all 24 hours of the day
num_URLs: total number of URLs included by the tweets
num_authors: total number of authors involved in an hour
num_mentions: total number of mentions in the tweets

ranking_score: sum of ranking scores of all tweets in an hour
num_hashtags: total number of hashtags in the tweets
After adding attributes and extracting data from files, we apply the same linear regression model on the dataset, just as that in the previous question. Then we also perform t-test and use P-values to determine the importance of the attributes. The results are listed below.

##
Processing hashtag "#GoHawks".....
RMSE = 700.117466230265
R-squared = 0.724

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	-67.2442	4.190	-16.047	0.000	-75.476	-59.013
num_retweets	13.4966	3.547	3.805	0.000	6.529	20.464
sum_followers	-0.0003	6.86e-05	-4.390	0.000	-0.000	-0.000
max_followers	0.0002	0.000	1.234	0.218	-9.95e-05	0.000
num_URLs	7.5628	1.508	5.016	0.000	4.601	10.525
num_authors	4.5277	0.745	6.079	0.000	3.065	5.991
num_mensions	2.8669	0.468	6.130	0.000	1.948	3.786
ranking_score	13.5304	0.838	16.141	0.000	11.884	15.177
num_hashtags	0.3588	0.329	1.091	0.276	-0.287	1.005
0th_hour	54.3914	144.659	0.376	0.707	-229.764	338.547
1th_hour	-18.1857	144.375	-0.126	0.900	-301.784	265.413
2th_hour	16.6854	144.484	0.115	0.908	-267.128	300.499
3th_hour	14.8236	147.270	0.101	0.920	-274.461	304.108
4th_hour	-21.3342	147.266	-0.145	0.885	-310.611	267.943
5th_hour	-4.0843	147.593	-0.028	0.978	-294.005	285.836
6th_hour	-90.3962	148.223	-0.610	0.542	-381.552	200.760
7th_hour	-131.8647	149.518	-0.882	0.378	-425.565	161.835
8th_hour	-180.9324	152.608	-1.186	0.236	-480.702	118.837
9th_hour	-242.7575	152.420	-1.593	0.112	-542.158	56.643
10th_hour	-276.5702	154.535	-1.790	0.074	-580.126	26.986
11th_hour	-475.3761	155.993	-3.047	0.002	-781.795	-168.957
12th_hour	-452.4758	153.927	-2.940	0.003	-754.837	-150.115
13th_hour	-175.3495	152.959	-1.146	0.252	-475.809	125.110
14th_hour	575.9658	153.188	3.760	0.000	275.056	876.876
15th_hour	-289.4265	154.147	-1.878	0.061	-592.219	13.366
16th_hour	-67.9202	151.560	-0.448	0.654	-365.632	229.791
17th_hour	163.8076	152.011	1.078	0.282	-134.791	462.406
18th_hour	-156.5957	149.677	-1.046	0.296	-450.608	137.417
19th_hour	12.6608	155.421	0.081	0.935	-292.636	317.957
20th_hour	-2.0205	151.888	-0.013	0.989	-300.378	296.337
21th_hour	-52.0524	150.618	-0.346	0.730	-347.913	243.808
22th_hour	7.2817	148.972	0.049	0.961	-285.346	299.910
23th_hour	-2.1436	147.790	-0.015	0.988	-292.451	288.164

Omnibus:	933.001	Durbin-Watson:	2.021
Prob(Omnibus):	0.000	Jarque-Bera (JB):	675766.335
Skew:	9.082	Prob(JB):	0.00
Kurtosis:	169.376	Cond. No.	1.92e+07

P values:

num_tweets	9.329546e-48
num_retweets	1.577650e-04
sum_followers	1.363767e-05
max_followers	2.177873e-01
num_URLs	7.158163e-07
num_authors	2.269332e-09
num_mensions	1.687794e-09
ranking_score	3.314710e-48
num_hashtags	2.757178e-01
0th_hour	7.070643e-01
1th_hour	8.998089e-01
2th_hour	9.081050e-01
3th_hour	9.198603e-01
4th_hour	8.848681e-01
5th_hour	9.779332e-01
6th_hour	5.422027e-01
7th_hour	3.782004e-01
8th_hour	2.362939e-01
9th_hour	1.118080e-01
10th_hour	7.405751e-02
11th_hour	2.419756e-03
12th_hour	3.425957e-03
13th_hour	2.521388e-01
14th_hour	1.883448e-04
15th_hour	6.096794e-02
16th_hour	6.542287e-01
17th_hour	2.816873e-01
18th_hour	2.959198e-01
19th_hour	9.351051e-01
20th_hour	9.893913e-01
21th_hour	7.297813e-01
22th_hour	9.610333e-01
23th_hour	9.884331e-01

T values:

num_tweets	-16.047053
num_retweets	3.805114
sum_followers	-4.389606
max_followers	1.233856
num_URLs	5.015660
num_authors	6.079238
num_mensions	6.129760
ranking_score	16.141492
num_hashtags	1.091082
0th_hour	0.375998
1th_hour	-0.125961
2th_hour	0.115483
3th_hour	0.100656
4th_hour	-0.144869
5th_hour	-0.027673
6th_hour	-0.609868
7th_hour	-0.881934
8th_hour	-1.185605
9th_hour	-1.592692
10th_hour	-1.789692
11th_hour	-3.047427
12th_hour	-2.939555
13th_hour	-1.146383
14th_hour	3.759865
15th_hour	-1.877606
16th_hour	-0.448142
17th_hour	1.077602
18th_hour	-1.046226
19th_hour	0.081461
20th_hour	-0.013303
21th_hour	-0.345594
22th_hour	0.048879
23th_hour	-0.014504

The most 3 important features for #GoHawks are …

features	p-value
ranking_score	3.314710e-48
num_tweets	9.329546e-48
num_mensions	1.687794e-09

##
Processing hashtag "#GoPatriots".....
RMSE = 100.15289056780102

R-squared = 0.894

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	-8.0301	2.727	-2.944	0.003	-13.388	-2.672
num_retweets	-43.2781	2.608	-16.592	0.000	-48.402	-38.154
sum_followers	-0.0020	0.000	-9.179	0.000	-0.002	-0.002
max_followers	0.0019	0.000	8.756	0.000	0.001	0.002
num_URLs	13.5502	0.853	15.892	0.000	11.875	15.225
num_authors	-6.6753	0.593	-11.248	0.000	-7.841	-5.510
num_mensions	3.1152	0.402	7.740	0.000	2.325	3.906
ranking_score	2.2835	0.472	4.836	0.000	1.356	3.211
num_hashtags	1.8346	0.315	5.823	0.000	1.216	2.453
0th_hour	-13.4494	21.067	-0.638	0.523	-54.832	27.934
1th_hour	-0.9546	21.060	-0.045	0.964	-42.325	40.415
2th_hour	-11.1825	21.073	-0.531	0.596	-52.578	30.213
3th_hour	1.2885	21.079	0.061	0.951	-40.119	42.696
4th_hour	-11.0857	21.099	-0.525	0.600	-52.532	30.360
5th_hour	-12.2316	21.084	-0.580	0.562	-53.649	29.186
6th_hour	-30.7719	21.122	-1.457	0.146	-72.263	10.720
7th_hour	-17.9052	21.133	-0.847	0.397	-59.417	23.607
8th_hour	-23.6883	21.136	-1.121	0.263	-65.206	17.829
9th_hour	-28.8476	21.277	-1.356	0.176	-70.642	12.947
10th_hour	-50.5863	21.388	-2.365	0.018	-92.600	-8.573
11th_hour	-15.3564	21.620	-0.710	0.478	-57.826	27.114
12th_hour	66.9282	21.341	3.136	0.002	25.006	108.850
13th_hour	11.5140	21.887	0.526	0.599	-31.480	54.508
14th_hour	-39.9123	21.823	-1.829	0.068	-82.781	2.956
15th_hour	37.0422	21.613	1.714	0.087	-5.414	79.498
16th_hour	-6.6506	21.945	-0.303	0.762	-49.759	36.458
17th_hour	-5.8325	21.765	-0.268	0.789	-48.586	36.921
18th_hour	-2.3024	21.274	-0.108	0.914	-44.093	39.488
19th_hour	-4.3360	21.109	-0.205	0.837	-45.801	37.129
20th_hour	6.1273	21.077	0.291	0.771	-35.276	47.531
21th_hour	15.0085	21.090	0.712	0.477	-26.419	56.436
22th_hour	-5.5452	21.062	-0.263	0.792	-46.918	35.828
23th_hour	-5.2516	21.517	-0.244	0.807	-47.519	37.016

Omnibus:	606.578	Durbin-Watson:	2.125
Prob(Omnibus):	0.000	Jarque-Bera (JB):	119995.116
Skew:	4.265	Prob(JB):	0.00
Kurtosis:	73.255	Cond. No.	2.26e+06

P values:

num_tweets	3.377311e-03
num_retweets	2.681923e-50
sum_followers	9.143321e-19
max_followers	2.577612e-17
num_URLs	5.793266e-47
num_authors	1.587829e-26
num_mensions	4.884546e-14
ranking_score	1.729867e-06
num_hashtags	9.906028e-09
0th_hour	5.234749e-01
1th_hour	9.638620e-01
2th_hour	5.958776e-01
3th_hour	9.512834e-01
4th_hour	5.995134e-01
5th_hour	5.620720e-01
6th_hour	1.457388e-01
7th_hour	3.972135e-01
8th_hour	2.628787e-01
9th_hour	1.757175e-01
10th_hour	1.837252e-02
11th_hour	4.778401e-01
12th_hour	1.805205e-03
13th_hour	5.990620e-01
14th_hour	6.796447e-02
15th_hour	8.712699e-02
16th_hour	7.619635e-01
17th_hour	7.888177e-01
18th_hour	9.138596e-01
19th_hour	8.373272e-01
20th_hour	7.713871e-01
21th_hour	4.769845e-01
22th_hour	7.924348e-01
23th_hour	8.072737e-01

T values:

num_tweets	-2.944151
num_retweets	-16.592051
sum_followers	-9.178608
max_followers	8.756102
num_URLs	15.891813
num_authors	-11.248076
num_mensions	7.740396
ranking_score	4.835841
num_hashtags	5.823074
0th_hour	-0.638412
1th_hour	-0.045329
2th_hour	-0.530652
3th_hour	0.061124
4th_hour	-0.525409
5th_hour	-0.580124
6th_hour	-1.456843
7th_hour	-0.847279
8th_hour	-1.120778
9th_hour	-1.355831
10th_hour	-2.365180
11th_hour	-0.710273
12th_hour	3.136065
13th_hour	0.526060
14th_hour	-1.828896
15th_hour	1.713856
16th_hour	-0.303056
17th_hour	-0.267979
18th_hour	-0.108222
19th_hour	-0.205412
20th_hour	0.290706
21th_hour	0.711655
22th_hour	-0.263280
23th_hour	-0.244064

The most 3 important features for #GoPatriots are ...

features	p-value
num_retweets	2.681923e-50
num_URLs	5.793266e-47
num_authors	1.587829e-26

##

Processing hashtag "#NFL".....

RMSE = 426.97393700371094

R-squared = 0.765

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	-1.1317	1.496	-0.757	0.450	-4.070	1.806
num_retweets	-9.9461	2.517	-3.951	0.000	-14.891	-5.002
sum_followers	-2.376e-05	2.42e-05	-0.980	0.327	-7.14e-05	2.38e-05
max_followers	1.302e-05	3.21e-05	0.405	0.686	-5.01e-05	7.62e-05
num_URLs	-0.4606	0.174	-2.648	0.008	-0.802	-0.119
num_authors	-4.4403	0.343	-12.962	0.000	-5.113	-3.767
num_mensions	2.9480	0.644	4.574	0.000	1.682	4.214
ranking_score	0.2665	0.307	0.867	0.386	-0.337	0.870
num_hashtags	1.1542	0.083	13.882	0.000	0.991	1.318
0th_hour	-85.0332	91.043	-0.934	0.351	-263.866	93.799
1th_hour	-76.5635	90.406	-0.847	0.397	-254.144	101.017
2th_hour	-68.3849	89.549	-0.764	0.445	-244.283	107.513
3th_hour	20.6908	89.562	0.231	0.817	-155.233	196.614
4th_hour	67.9376	90.042	0.755	0.451	-108.929	244.804
5th_hour	172.6458	89.976	1.919	0.056	-4.089	349.381
6th_hour	236.0467	91.980	2.566	0.011	55.375	416.719
7th_hour	180.6493	93.876	1.924	0.055	-3.748	365.047
8th_hour	163.9958	96.150	1.706	0.089	-24.867	352.859
9th_hour	120.3897	95.002	1.267	0.206	-66.219	306.998
10th_hour	187.8101	94.409	1.989	0.047	2.366	373.254
11th_hour	144.3960	99.966	1.444	0.149	-51.963	340.755
12th_hour	95.9948	99.206	0.968	0.334	-98.871	290.860
13th_hour	92.6347	97.977	0.945	0.345	-99.817	285.086
14th_hour	530.2282	98.621	5.376	0.000	336.512	723.945
15th_hour	50.4429	99.130	0.509	0.611	-144.275	245.160
16th_hour	18.9676	96.550	0.196	0.844	-170.681	208.616
17th_hour	189.4597	99.700	1.900	0.058	-6.376	385.295
18th_hour	119.1781	97.274	1.225	0.221	-71.892	310.248
19th_hour	-59.1401	96.239	-0.615	0.539	-248.179	129.898
20th_hour	-50.2045	95.079	-0.528	0.598	-236.963	136.554
21th_hour	-64.2906	95.030	-0.677	0.499	-250.955	122.373
22th_hour	-167.6198	93.279	-1.797	0.073	-350.844	15.605
23th_hour	-131.1583	93.127	-1.408	0.160	-314.083	51.766

Omnibus:	691.527	Durbin-Watson:	2.155
Prob(Omnibus):	0.000	Jarque-Bera (JB):	89935.999
Skew:	5.454	Prob(JB):	0.00
Kurtosis:	62.650	Cond. No.	4.42e+07

P values:

num_tweets	4.495591e-01
num_retweets	8.783207e-05
sum_followers	3.272762e-01
max_followers	6.855078e-01
num_URLs	8.331178e-03
num_authors	9.487487e-34
num_mensions	5.904038e-06
ranking_score	3.861195e-01
num_hashtags	8.148776e-38
0th_hour	3.507183e-01
1th_hour	3.974251e-01
2th_hour	4.453973e-01
3th_hour	8.173839e-01
4th_hour	4.508653e-01
5th_hour	5.552243e-02
6th_hour	1.054108e-02
7th_hour	5.482516e-02
8th_hour	8.863818e-02
9th_hour	2.056048e-01
10th_hour	4.715747e-02
11th_hour	1.491789e-01
12th_hour	3.336500e-01
13th_hour	3.448283e-01
14th_hour	1.121638e-07
15th_hour	6.110576e-01
16th_hour	8.443268e-01
17th_hour	5.791203e-02
18th_hour	2.210260e-01
19th_hour	5.391302e-01
20th_hour	5.976883e-01
21th_hour	4.989885e-01
22th_hour	7.288552e-02
23th_hour	1.595777e-01

T values:

num_tweets	-0.756687
num_retweets	-3.951143
sum_followers	-0.980482
max_followers	0.405172
num_URLs	-2.647855
num_authors	-12.962474
num_mensions	4.574162
ranking_score	0.867362
num_hashtags	13.881983
0th_hour	-0.933985
1th_hour	-0.846885
2th_hour	-0.763656
3th_hour	0.231021
4th_hour	0.754507
5th_hour	1.918806
6th_hour	2.566287
7th_hour	1.924330
8th_hour	1.705626
9th_hour	1.267232
10th_hour	1.989317
11th_hour	1.444446
12th_hour	0.967633
13th_hour	0.945476
14th_hour	5.376427
15th_hour	0.508853
16th_hour	0.196454
17th_hour	1.900305
18th_hour	1.225185
19th_hour	-0.614511
20th_hour	-0.528032
21th_hour	-0.676527
22th_hour	-1.796966
23th_hour	-1.408386

The most 3 important features for #NFL are ...

features	p-value
num_hashtags	8.148776e-38
num_authors	9.487487e-34
num_mensions	5.904038e-06

##

Processing hashtag "#Patriots".....

RMSE = 1840.1375511287067

R-squared = 0.823

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	-63.0835	4.992	-12.636	0.000	-72.889	-53.278
num_retweets	-27.0591	2.945	-9.188	0.000	-32.844	-21.274
sum_followers	0.0004	5.7e-05	6.938	0.000	0.000	0.001
max_followers	-0.0006	0.000	-5.621	0.000	-0.001	-0.000
num_URLs	-5.0346	1.739	-2.895	0.004	-8.450	-1.619
num_authors	2.2108	0.994	2.224	0.027	0.258	4.163
num_mensions	7.0053	0.955	7.339	0.000	5.130	8.880
ranking_score	11.0880	0.949	11.687	0.000	9.224	12.952
num_hashtags	3.6577	0.413	8.852	0.000	2.846	4.469
0th_hour	197.8735	381.113	0.519	0.604	-550.729	946.476
1th_hour	-65.4957	382.060	-0.171	0.864	-815.958	684.967
2th_hour	-36.3362	380.504	-0.095	0.924	-783.744	711.072
3th_hour	-156.8004	380.059	-0.413	0.680	-903.333	589.732
4th_hour	-114.6024	381.599	-0.300	0.764	-864.160	634.955
5th_hour	-164.1821	383.748	-0.428	0.669	-917.962	589.597
6th_hour	-331.3956	388.221	-0.854	0.394	-1093.961	431.169
7th_hour	-669.8187	392.247	-1.708	0.088	-1440.291	100.654
8th_hour	-756.3629	394.680	-1.916	0.056	-1531.615	18.889
9th_hour	-313.3698	405.646	-0.773	0.440	-1110.163	483.423
10th_hour	972.2628	391.872	2.481	0.013	202.526	1741.999
11th_hour	-843.7053	401.738	-2.100	0.036	-1632.821	-54.590
12th_hour	-711.3069	403.419	-1.763	0.078	-1503.725	81.111
13th_hour	-581.7998	403.619	-1.441	0.150	-1374.610	211.010
14th_hour	-313.4766	407.543	-0.769	0.442	-1113.994	487.041
15th_hour	-562.1471	405.120	-1.388	0.166	-1357.906	233.612
16th_hour	-76.3642	404.922	-0.189	0.850	-871.734	719.006
17th_hour	140.2104	401.606	0.349	0.727	-648.646	929.067
18th_hour	345.9975	409.073	0.846	0.398	-457.527	1149.522
19th_hour	-362.4667	400.155	-0.906	0.365	-1148.473	423.540
20th_hour	330.5370	402.736	0.821	0.412	-460.540	1121.614
21th_hour	-309.1792	397.389	-0.778	0.437	-1089.753	471.395
22th_hour	151.0046	393.973	0.383	0.702	-622.860	924.869
23th_hour	225.2162	392.846	0.573	0.567	-546.433	996.866

Omnibus:	1036.548	Durbin-Watson:	1.905
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1014783.857
Skew:	10.956	Prob(JB):	0.00
Kurtosis:	205.510	Cond. No.	6.69e+07

P values:

num_tweets	2.423382e-32
num_retweets	7.979424e-19
sum_followers	1.112168e-11
max_followers	3.007636e-08
num_URLs	3.937402e-03
num_authors	2.655707e-02
num_mensions	7.729120e-13
ranking_score	2.331112e-28
num_hashtags	1.153019e-17
0th_hour	6.038292e-01
1th_hour	8.639499e-01
2th_hour	9.239564e-01
3th_hour	6.800821e-01
4th_hour	7.640443e-01
5th_hour	6.689352e-01
6th_hour	3.936810e-01
7th_hour	8.826225e-02
8th_hour	5.582903e-02
9th_hour	4.401361e-01
10th_hour	1.339442e-02
11th_hour	3.616809e-02
12th_hour	7.841852e-02
13th_hour	1.500201e-01
14th_hour	4.421100e-01
15th_hour	1.658146e-01
16th_hour	8.504834e-01
17th_hour	7.271287e-01
18th_hour	3.980248e-01
19th_hour	3.654271e-01
20th_hour	4.121543e-01
21th_hour	4.368857e-01
22th_hour	7.016547e-01
23th_hour	5.666782e-01

T values:

num_tweets	-12.636398
num_retweets	-9.188057
sum_followers	6.938247
max_followers	-5.621251
num_URLs	-2.895297
num_authors	2.223907
num_mensions	7.338616
ranking_score	11.687026
num_hashtags	8.852499
0th_hour	0.519199
1th_hour	-0.171428
2th_hour	-0.095495
3th_hour	-0.412569
4th_hour	-0.300322
5th_hour	-0.427838
6th_hour	-0.853626
7th_hour	-1.707646
8th_hour	-1.916395
9th_hour	-0.772520
10th_hour	2.481073
11th_hour	-2.100140
12th_hour	-1.763197
13th_hour	-1.441459
14th_hour	-0.769187
15th_hour	-1.387607
16th_hour	-0.188590
17th_hour	0.349124
18th_hour	0.845808
19th_hour	-0.905816
20th_hour	0.820728
21th_hour	-0.778026
22th_hour	0.383286
23th_hour	0.573294

The most 3 important features for #Patriots are ...

features	p-value
num_tweets	2.423382e-32
ranking_score	2.331112e-28
num_retweets	7.979424e-19

##

Processing hashtag "#SB49".....

RMSE = 3143.1473539568765

R-squared = 0.902

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	-62.3620	7.510	-8.304	0.000	-77.114	-47.610
num_retweets	38.2823	16.974	2.255	0.025	4.940	71.624
sum_followers	0.0002	2.08e-05	9.917	0.000	0.000	0.000
max_followers	-0.0004	6.51e-05	-6.221	0.000	-0.001	-0.000
num_URLs	5.7079	1.908	2.992	0.003	1.960	9.456
num_authors	-2.3755	0.971	-2.446	0.015	-4.283	-0.468
num_mensions	3.0465	1.066	2.857	0.004	0.952	5.141
ranking_score	11.9501	1.512	7.906	0.000	8.981	14.919
num_hashtags	2.4295	0.546	4.447	0.000	1.356	3.503
0th_hour	-165.0774	648.509	-0.255	0.799	-1438.935	1108.780
1th_hour	48.1939	650.880	0.074	0.941	-1230.320	1326.708
2th_hour	-685.9166	652.544	-1.051	0.294	-1967.700	595.867
3th_hour	-831.8146	655.672	-1.269	0.205	-2119.742	456.113
4th_hour	-880.0948	661.754	-1.330	0.184	-2179.969	419.779
5th_hour	1536.0553	658.248	2.334	0.020	243.068	2829.042
6th_hour	234.3266	669.661	0.350	0.727	-1081.079	1549.732
7th_hour	-1771.5668	674.098	-2.628	0.009	-3095.688	-447.446
8th_hour	-1310.1479	668.042	-1.961	0.050	-2622.373	2.077
9th_hour	-885.5791	683.618	-1.295	0.196	-2228.400	457.242
10th_hour	-713.8312	678.722	-1.052	0.293	-2047.035	619.373
11th_hour	-334.9490	675.902	-0.496	0.620	-1662.615	992.717
12th_hour	-418.1016	682.315	-0.613	0.540	-1758.363	922.160
13th_hour	-797.8293	670.214	-1.190	0.234	-2114.322	518.663
14th_hour	-289.9327	676.372	-0.429	0.668	-1618.521	1038.655
15th_hour	-252.1275	667.650	-0.378	0.706	-1563.584	1059.329
16th_hour	296.4913	666.397	0.445	0.657	-1012.503	1605.486
17th_hour	396.5588	670.308	0.592	0.554	-920.118	1713.236
18th_hour	139.2750	662.873	0.210	0.834	-1162.797	1441.347
19th_hour	276.7323	663.295	0.417	0.677	-1026.170	1579.635
20th_hour	221.8280	661.866	0.335	0.738	-1078.266	1521.922
21th_hour	78.6283	660.838	0.119	0.905	-1219.446	1376.703
22th_hour	33.6887	661.351	0.051	0.959	-1265.395	1332.773
23th_hour	-60.6663	661.336	-0.092	0.927	-1359.720	1238.387

Omnibus:	991.551	Durbin-Watson:	1.577
Prob(Omnibus):	0.000	Jarque-Bera (JB):	850934.405
Skew:	10.136	Prob(JB):	0.00
Kurtosis:	189.062	Cond. No.	4.91e+08

P values:

num_tweets	7.879917e-16
num_retweets	2.450427e-02
sum_followers	1.952924e-21
max_followers	9.790209e-10
num_URLs	2.899248e-03
num_authors	1.475423e-02
num_mensions	4.434342e-03
ranking_score	1.463012e-14
num_hashtags	1.055765e-05
0th_hour	7.991664e-01
1th_hour	9.410021e-01
2th_hour	2.936549e-01
3th_hour	2.051045e-01
4th_hour	1.840883e-01
5th_hour	1.997858e-02
6th_hour	7.265340e-01
7th_hour	8.827277e-03
8th_hour	5.036247e-02
9th_hour	1.957152e-01
10th_hour	2.933860e-01
11th_hour	6.204040e-01
12th_hour	5.402821e-01
13th_hour	2.343989e-01
14th_hour	6.683394e-01
15th_hour	7.058482e-01
16th_hour	6.565548e-01
17th_hour	5.543569e-01
18th_hour	8.336611e-01
19th_hour	6.766889e-01
20th_hour	7.376355e-01
21th_hour	9.053324e-01
22th_hour	9.593925e-01
23th_hour	9.269437e-01

T values:

num_tweets	-8.303612
num_retweets	2.255340
sum_followers	9.916675
max_followers	-6.221009
num_URLs	2.991640
num_authors	-2.446077
num_mensions	2.857299
ranking_score	7.905957
num_hashtags	4.446627
0th_hour	-0.254549
1th_hour	0.074044
2th_hour	-1.051142
3th_hour	-1.268645
4th_hour	-1.329943
5th_hour	2.333552
6th_hour	0.349918
7th_hour	-2.628056
8th_hour	-1.961177
9th_hour	-1.295430
10th_hour	-1.051729
11th_hour	-0.495558
12th_hour	-0.612769
13th_hour	-1.190410
14th_hour	-0.428659
15th_hour	-0.377634
16th_hour	0.444917
17th_hour	0.591607
18th_hour	0.210108
19th_hour	0.417208
20th_hour	0.335156
21th_hour	0.118983
22th_hour	0.050939
23th_hour	-0.091733

The most 3 important features for #SB49 are ...

features	p-value
sum_followers	1.952924e-21
num_tweets	7.879917e-16
ranking_score	1.463012e-14

##

Processing hashtag "#SuperBowl".....

RMSE = 4260.876213460498

R-squared = 0.943

	coef	std err	t	P>|t|	[0.025	0.975]
num_tweets	-111.7023	4.908	-22.761	0.000	-121.342	-102.063
num_retweets	48.3817	3.474	13.925	0.000	41.557	55.207
sum_followers	-0.0001	1.1e-05	-10.690	0.000	-0.000	-9.61e-05
max_followers	-7.19e-06	9.41e-05	-0.076	0.939	-0.000	0.000
num_URLs	2.4611	0.608	4.045	0.000	1.266	3.656
num_authors	15.7667	0.785	20.073	0.000	14.224	17.310
num_mensions	-13.8255	0.833	-16.606	0.000	-15.461	-12.190
ranking_score	22.3118	0.990	22.547	0.000	20.368	24.256
num_hashtags	1.7899	0.250	7.147	0.000	1.298	2.282
0th_hour	650.9831	887.874	0.733	0.464	-1093.035	2395.001
1th_hour	1004.6848	881.765	1.139	0.255	-727.333	2736.703
2th_hour	172.0183	879.944	0.195	0.845	-1556.422	1900.459
3th_hour	112.2364	884.318	0.127	0.899	-1624.796	1849.269
4th_hour	663.0132	882.862	0.751	0.453	-1071.159	2397.186
5th_hour	423.9448	890.243	0.476	0.634	-1324.727	2172.616
6th_hour	-346.4221	887.821	-0.390	0.697	-2090.336	1397.492
7th_hour	-1179.6425	895.216	-1.318	0.188	-2938.082	578.797
8th_hour	-833.1994	897.610	-0.928	0.354	-2596.341	929.942
9th_hour	-708.9208	897.564	-0.790	0.430	-2471.973	1054.131
10th_hour	-865.3510	921.721	-0.939	0.348	-2675.854	945.151
11th_hour	-158.8165	922.615	-0.172	0.863	-1971.075	1653.442
12th_hour	-27.3412	917.905	-0.030	0.976	-1830.348	1775.666
13th_hour	803.5519	918.861	0.875	0.382	-1001.332	2608.436
14th_hour	3179.5406	929.410	3.421	0.001	1353.934	5005.147
15th_hour	-239.1008	938.458	-0.255	0.799	-2082.480	1604.279
16th_hour	1537.7769	934.802	1.645	0.101	-298.419	3373.973
17th_hour	1407.0925	932.097	1.510	0.132	-423.792	3237.977
18th_hour	-330.7317	926.100	-0.357	0.721	-2149.836	1488.373
19th_hour	48.6008	933.891	0.052	0.959	-1785.807	1883.009
20th_hour	-350.2424	908.766	-0.385	0.700	-2135.299	1434.814
21th_hour	566.5183	906.024	0.625	0.532	-1213.151	2346.188
22th_hour	574.2600	901.899	0.637	0.525	-1197.307	2345.827
23th_hour	564.9300	907.577	0.622	0.534	-1217.791	2347.651

Omnibus:	430.488	Durbin-Watson:	1.368
Prob(Omnibus):	0.000	Jarque-Bera (JB):	75438.479
Skew:	2.253	Prob(JB):	0.00
Kurtosis:	58.402	Cond. No.	7.85e+08

P values:

num_tweets	2.017493e-81
num_retweets	5.337485e-38
sum_followers	2.258016e-24
max_followers	9.391348e-01
num_URLs	5.977705e-05
num_authors	9.846646e-68
num_mensions	1.528295e-50
ranking_score	2.509496e-80
num_hashtags	2.812801e-12
0th_hour	4.637514e-01
1th_hour	2.550288e-01
2th_hour	8.450829e-01
3th_hour	8.990509e-01
4th_hour	4.529829e-01
5th_hour	6.341111e-01
6th_hour	6.965436e-01
7th_hour	1.881436e-01
8th_hour	3.536867e-01
9th_hour	4.299671e-01
10th_hour	3.482214e-01
11th_hour	8.633925e-01
12th_hour	9.762481e-01
13th_hour	3.822207e-01
14th_hour	6.699281e-04
15th_hour	7.989874e-01
16th_hour	1.005319e-01
17th_hour	1.317171e-01
18th_hour	7.211360e-01
19th_hour	9.585147e-01
20th_hour	7.000862e-01
21th_hour	5.320455e-01
22th_hour	5.245687e-01
23th_hour	5.338964e-01

T values:

num_tweets	-22.761408
num_retweets	13.924823
sum_followers	-10.689776
max_followers	-0.076392
num_URLs	4.044949
num_authors	20.072980
num_mensions	-16.606335
ranking_score	22.547280
num_hashtags	7.147183
0th_hour	0.733193
1th_hour	1.139402
2th_hour	0.195488
3th_hour	0.126919
4th_hour	0.750982
5th_hour	0.476212
6th_hour	-0.390194
7th_hour	-1.317718
8th_hour	-0.928243
9th_hour	-0.789827
10th_hour	-0.938843
11th_hour	-0.172137
12th_hour	-0.029786
13th_hour	0.874509
14th_hour	3.421030
15th_hour	-0.254780
16th_hour	1.645030
17th_hour	1.509598
18th_hour	-0.357123
19th_hour	0.052041
20th_hour	-0.385404
21th_hour	0.625280
22th_hour	0.636723
23th_hour	0.622459

The most 3 important features for #SuperBowl are ...

features	p-value
num_tweets	2.017493e-81
ranking_score	2.509496e-80
num_authors	9.846646e-68

Like what we did before, we used R-squared and RMSE to measure the fitting accuracy of our model, and p-values to evaluate the significance of each feature, as they are in the output above.

Compared to the results in previous question, we can see that the accuracy of the model is substantially improved comparing with results in problem 1.2, as indicated in the table below.

	#GoHawks	#GoPatriots	#NFL	#Patriots	#SB49	#SuperBowl
1.2 R-squared	0.532	0.627	0.585	0.723	0.857	0.870

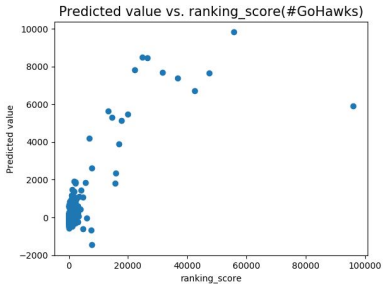
1.3 R-squared	0.724	0.894	0.765	0.823	0.902	0.943
1.2 RMSE	910	188	567	2302	3799	6433
1.3 RMSE	700	100	426	1840	3143	4260

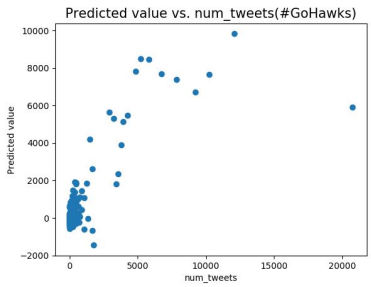
As seen from the above results we have a significant improvement in the accuracy of the model for each of the hashtag, comparing to results in problem 1.2, R squared value increased and RMSE decreased. This can be attributed to features that are not sparse and have a well-defined distribution throughout the period of the Super Bowl time. The model we trained is more fitting to the data and it proves that the new features we've applied are correct and helpful to train the linear regression model. Metrics employed in the tweet data have been used to model the importance of the tweet for a given window frame thereby increasing the accuracy.
The improvement in accuracy is easy to interpret, adding more attributes provides more information for the regression model. Since the dimension of the training model increases, the target value can be predicted from more aspects. As a result, the regression model fit better and performs more accurately.

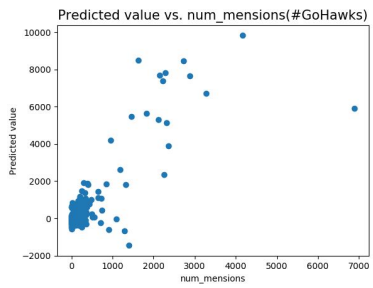
Besides, we identified the importance of each feature by its p-value. And we noticed that for different hashtags, the rankings of the significance of the features are different. According to the report above, the top 3 most important variables of each hashtag are.

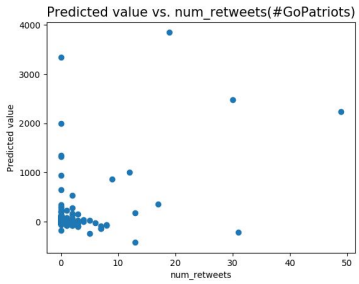
Hashtags	1st important feature	2nd important feature	3rd important feature
#GoHawks	ranking score	num tweets	num mentions
#GoPatriots	num retweets	num urls	num authors
#NFL	num hashtags	num authors	num mentions
#Patriots	num tweets	ranking score	num retweets
#SB49	sum followers	num tweets	ranking scores
#SuperBowl	num tweets	ranking score	num authors

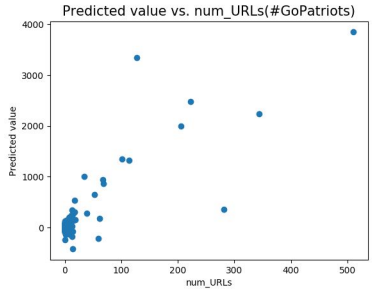
In order to better visualize the contribution of the features in the model a scatter plot was created of the Top 3 features for hashtags. The plots of predictants versus feature values are shown below.

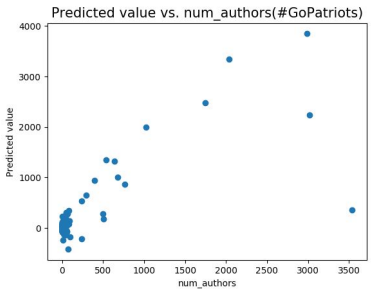

Predicted value vs. ranking_score(#GoHawks)
Predicted value
ranking_score

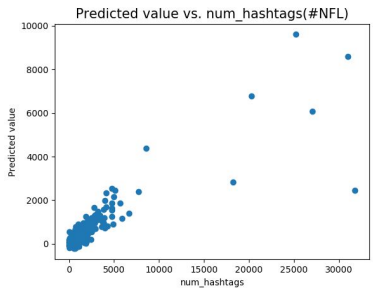

Predicted value vs. num_tweets(#GoHawks)
Predicted value
num_tweets

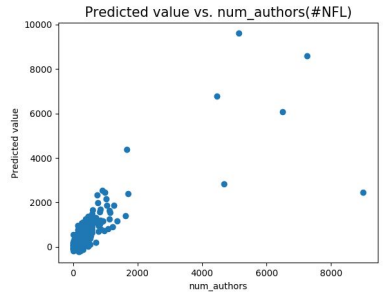

Predicted value vs. num_mensions(#GoHawks)
Predicted value
num_mensions

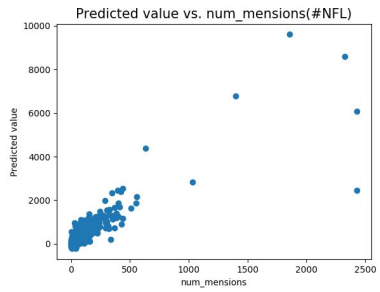

Predicted value vs. num_retweets(#GoPatriots)
Predicted value
num_retweets

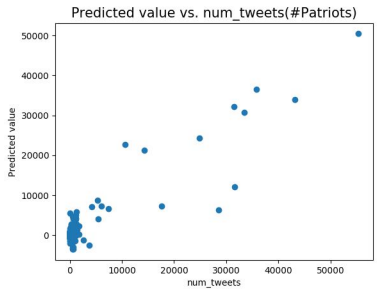

Predicted value vs. num_URLs(#GoPatriots)
Predicted value
num_URLs

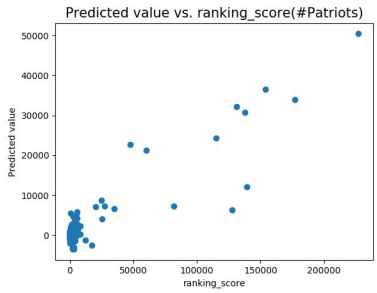

Predicted value vs. num_authors(#GoPatriots)
Predicted value
num_authors

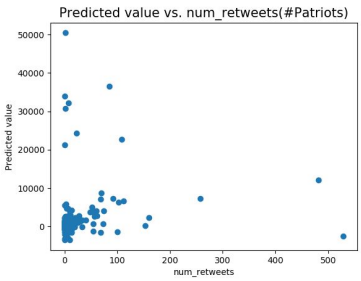

Predicted value vs. num_hashtags(#NFL)
Predicted value
num_hashtags

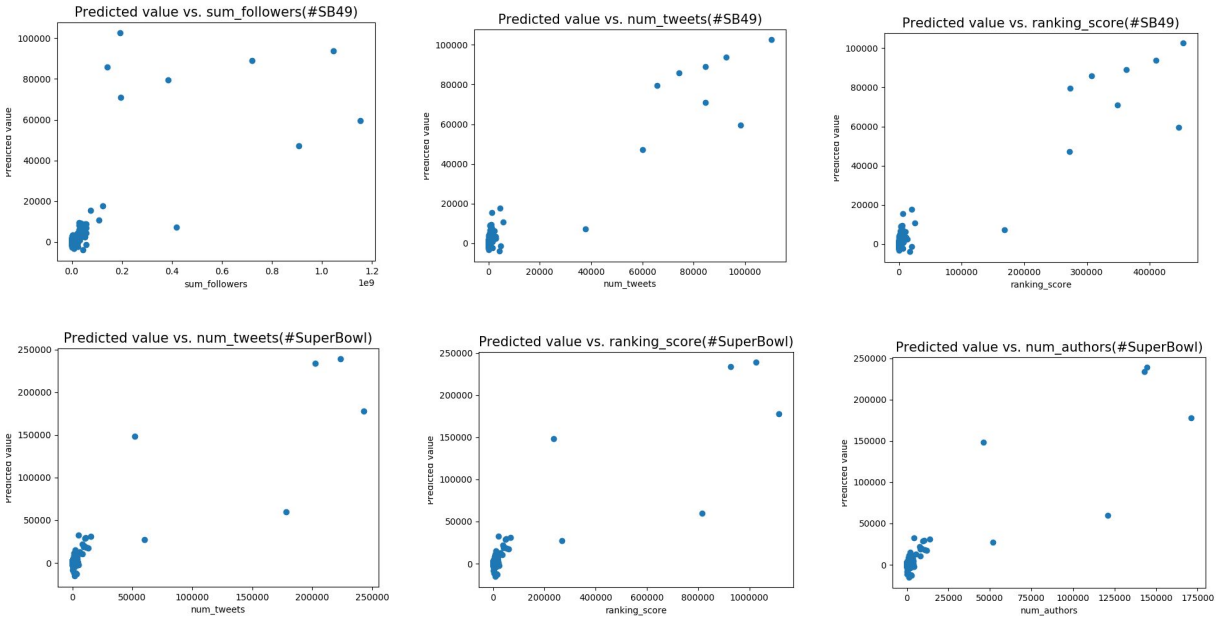

Predicted value vs. num_authors(#NFL)
Predicted value
num_authors


Predicted value vs. num_mensions(#NFL)
Predicted value
num_mensions


Predicted value vs. num_tweets(#Patriots)
Predicted value
num_tweets


Predicted value vs. ranking_score(#Patriots)
Predicted value
ranking_score


Predicted value vs. num_retweets(#Patriots)
Predicted value
num_retweets



Since the initial hours have less number of tweets, all of the graphs exhibit clustering of values near low of tweets/hour.
Additionally, the patterns have some sort of linearity, with respect to each attribute. We also observed a relatively linear relationship between the 3 top features and target value, although there are some of the extreme values beyond.
Generally, the plots could prove that the application of linear regression is quite reasonable and suitable.

Problem 1.4

For this part, we are going to use the previous linear regression model along with other two non-linear mode to conduct 10-fold cross validation of the best three features selected from part 1.3. Moreover, we are going to analyze the model within three distinct timeframes. These are before Feb 1st 8am, between Feb 1st 8am to 8 pm, after Feb 1st 8pm. For each timeframes, we will start with predicting the number of tweet next hour for each of the 6 hashtags we are interested in. That means we are going to conduct 6 by 3 by 3, in total of 54 models to compare with their root mean square error (RMSE).

As our data was manually processed with time interval in previous part, we will first export the timestamp from json file to find the corresponding time_window we are using for Feb 1st 8am and Feb 1st 8pm. Then we will conduct 10-fold cross validation for each classifier. The classifiers we chose are Linear Regression, SVC(kernel='rbf')and RandomForest Classifier.

Here are the result for each of the classifier:

SVC:
#GoHawks

:Before
Feb 01 8am 1151.00210587
:Between
Feb 01 8am and 8pm 6801.76329344
:After
Feb 01 8pm 106.532539871
#GoPatriots
:Before
Feb 01 8am 75.5719554765
:Between
Feb 01 8am and 8pm 2593.27277778
:After
Feb 01 8pm 11.4504506685
#NFL
:Before
Feb 01 8am 424.446730783
:Between
Feb 01 8am and 8pm 7360.9224354
:After
Feb 01 8pm 625.638078241
#Patriots
:Before
Feb 01 8am 1077.86711
:Between
Feb 01 8am and 8pm 29464.7804633
:After
Feb 01 8pm 339.874487984
#SB49
:Before
Feb 01 8am 7529.45298719
:Between
Feb 01 8am and 8pm 77076.5862724
:After
Feb 01 8pm 628.9649923
#SuperBowl
:Before
Feb 01 8am 1190.68848304
:Between
Feb 01 8am and 8pm 170665.884925
:After
Feb 01 8pm 1092.06542568
RF:
#GoHawks

:Before
Feb 01 8am 1285.09149657
:Between
Feb 01 8am and 8pm 3467.65252008
:After
Feb 01 8pm 64.7364380412
#GoPatriots
:Before
Feb 01 8am 82.7815992949
:Between
Feb 01 8am and 8pm 1400.32021338
:After
Feb 01 8pm 3.79456426335
#NFL
:Before
Feb 01 8am 384.412092445
:Between
Feb 01 8am and 8pm 2832.9008013
:After
Feb 01 8pm 206.687720333
#Patriots
:Before
Feb 01 8am 1279.98066747
:Between
Feb 01 8am and 8pm 21197.1310594
:After
Feb 01 8pm 162.072558707
#SB49
:Before
Feb 01 8am 3892.29804316
:Between
Feb 01 8am and 8pm 35906.4602126
:After
Feb 01 8pm 204.820615699
#SuperBowl
:Before
Feb 01 8am 906.273649221
:Between
Feb 01 8am and 8pm 73670.4326253
:After
Feb 01 8pm 430.420877593

Linear Regression:

#GoHawks
:Before
Feb 01 8am 1613.17656309
:Between
Feb 01 8am and 8pm 2728.16090599
:After
Feb 01 8pm 59.3402801925
#GoPatriots
:Before
Feb 01 8am 68.6604406102
:Between
Feb 01 8am and 8pm 2266.29707247
:After
Feb 01 8pm 3.40641278939
#NFL
:Before
Feb 01 8am 386.626734623
:Between
Feb 01 8am and 8pm 4194.10358909
:After
Feb 01 8pm 149.406698246
#Patriots
:Before
Feb 01 8am 1181.25473185
:Between
Feb 01 8am and 8pm 19188.1781898
:After
Feb 01 8pm 122.993564679
#SB49
:Before
Feb 01 8am 3667.21963817
:Between
Feb 01 8am and 8pm 51640.3969227
:After
Feb 01 8pm 180.52176819
#SuperBowl
:Before
Feb 01 8am 920.655961
:Between
Feb 01 8am and 8pm 60849.2113094
:After
Feb 01 8pm 304.932339327

Each of these RMSE is the mean of 10-fold cross validation.

From the result, we can see that all of the models are giving bad prediction against the second timeframe, which is Feb 1st 8am to 8pm. Due to the high volume and changes due to either the game or pregame activities. This should be reasonable. For example, the exact time when a touchdown is made may generate countless number of tweet at the same time. The other timeframe works relatively good with the model.

After that, we aggregate all the tweet data from all 6 hashtags and perform a prediction on the next hour tweet using all of our three models.
SVC
Before Feb 01 8am 1190.40383713
Between Feb 01 8am and 8pm 178440.866133
After Feb 01 8pm 1132.5055418
RandomForestClassifier

Before Feb 01 8am 1136.11594019
Between Feb 01 8am and 8pm 104435.144132
After Feb 01 8pm 467.335489066
LinearRegression

Before Feb 01 8am 927.693492311
Between Feb 01 8am and 8pm 127215.493528
After Feb 01 8pm 310.429223464
From the result, we can see that Linear Regression gives the smallest RMSE to the aggregate data. Therefore, we choose it as our best model.

Problem 1.5

In this part, we first train a model on all hashtags. The best model found in part 1.4 is linear model, so linear regression is used to predict the number of tweets. Different from previous sections, the training takes features from previous 5 hours, and the number of tweets in sixth hour is the value to predict. The new features have suffixes like '_0h', '_1h', '_2h', which mean the data at current hour, last hour, and second last hour, respectively. Theoretically, taking a input window larger than one hour means the model has memory. The value is predicted not only on current state, but also on states happened in the past. If there is some trend-related relationships hidden in the features, this model should be able to capture these time-variant features. Below is the coefficients of linear regression model trained on aggregate data.

num_of_tweets_0h	num_of_tweets_1h	num_of_tweets_2h	num_of_tweets_3h	num_of_tweets_4h	total_num_of_retweets_0h	total_num_of_retweets_1h
1.5438	-0.1516	1.0929	1.0022	-0.4706	-0.3448	-0.1437

total_num_of_retweets_2h	total_num_of_retweets_3h	total_num_of_retweets_4h	sum_of_followers_0h	sum_of_followers_1h	sum_of_followers_2h	sum_of_followers_3h
-0.8631	0.1649	0.1665	0.0000	0.0001	0.0001	-0.0002
sum_of_followers_4h	max_num_of_followers_0h	max_num_of_followers_1h	max_num_of_followers_2h	max_num_of_followers_3h	max_num_of_followers_4h	time_of_the_day_0h
0.0000	-0.0001	0.0005	-0.0002	0.0003	0.0001	-72.8951
time_of_the_day_1h	time_of_the_day_2h	time_of_the_day_3h	time_of_the_day_4h			
82.3071	-94.8854	-9.5603	-32.4833			

It seems the model has great weight on time of the day, and ignores other features. Since linear model solely tries to minimize the error when fitting the data, it is possible to get weights like these if the relationship between time of the day and number of tweets are more significant. When evaluating results, we use features from first 5 hours in the test files as inputs, and the number of tweets in last hour as ground-truth. Note that the test files are crawled on 'first_post_date', so the time windows used in this part is also based on 'first_post_date' to keep it consistent. 'sample8_period1' only has data of 5 hours, so it is padded to 6 hours to use the same model. The predicted results are listed below.

File Name	sample2_period2	sample7_period3	sample4_period1	sample8_period1	sample1_period1	sample6_period2	sample3_period3	sample10_period3	sample9_period2	sample5_period1
Target	82923	120	201	11	178	37307	523	61	2790	213
Predicted	59251	-2374	-1125	-1539	304	51589	2242	-1143	509	3236

From the results we can see that while the predictor can get the magnitude right, for example, predict a large number for a large target value (e.g. 59251 & 82923), and a small number for a small target value (e.g. 304 and 178), the error is very significant. The model also predicts some unrealistic negative values. Like seen in the previous project, the linear model does not have idea of what the valid range of output should be, it just tries to make a prediction based on fitted linear relationship. In this case, the weights of time_of_the_day are overwhelming than other features. This makes the situation worse since intuitively the hour in a day is not sufficient to predict the number of tweets. Since adding features from previous makes the result worse for linear model, we also tried using random forest model to predict the number of tweets. Below is the result using random forest regressor.

File Name	sample2_period2	sample7_period3	sample4_period1	sample8_period1	sample1_period1	sample6_period2	sample3_period3	sample10_period3	sample9_period2	sample5_period1
Target	82923	120	201	11	178	37307	523	61	2790	213
Predicted	163950	95	347	101	260	18409	941	84	1940	283

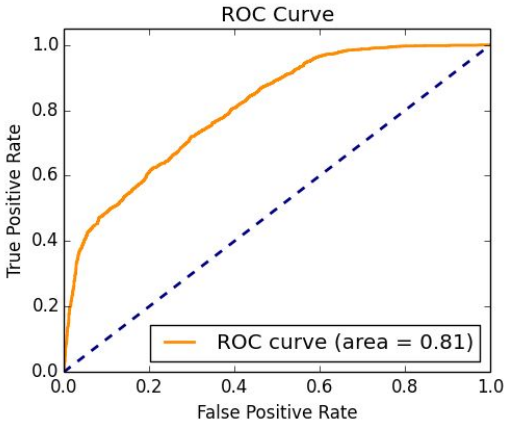
Except the error of ‘sample2_period2’ is greater than linear model, the predicted values are much more accurate than linear model. However, if we calculate the RMSE of both models, the RMSE of linear model is **8907.76**, and the RMSE of random forest model is **18922.40**. The RMSE of random forest model is still much greater than that of linear model, which means linear model is still the ‘best’ model in terms of RMSE. This is because the predicted value of ‘sample2_period2’ is too far from the true values. But overall, the random forest model seems to be more useful in real application. Even though it miscalculate the extreme value, but when we are only interest the magnitude to make further decisions, it is still valid. Moreover, random forest model predicts more accurately for small values. The error of small values does not contribute much to the RMSE. It also only produces valid positive values. In conclusion, random forest model predicts the number of tweets better in most of the time, but this advantage is not reflected in RMSE measurement.

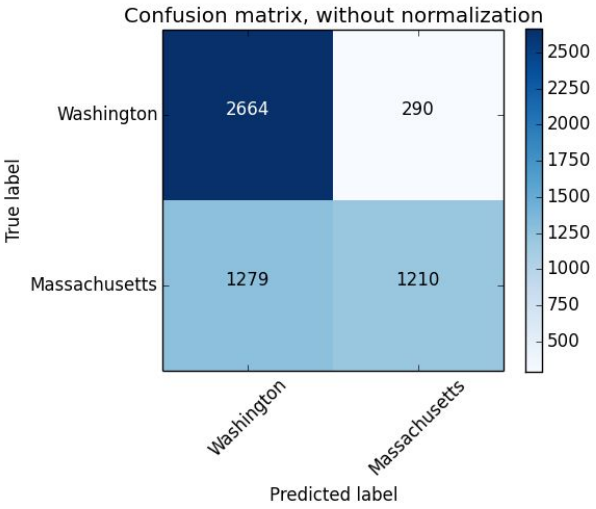
Problem 2

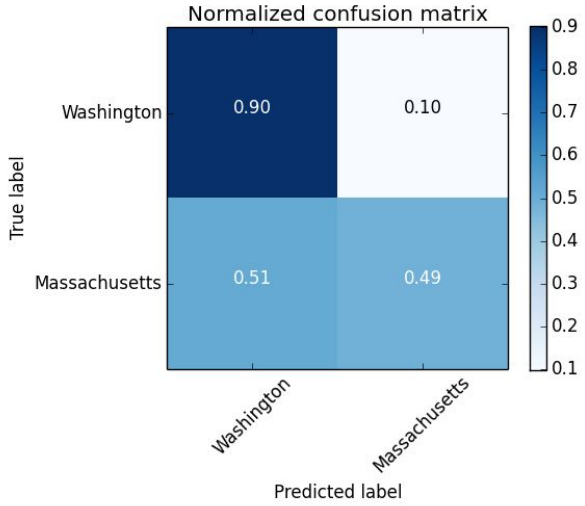
In this part, we try to use the contents of tweets to predict whether the user is located in Massachusetts or Washington state. The tweets for this section is from ‘tweets_#superbowl.txt’. To get the ground-truth labels, all tweets with words like 'Seattle', 'Washington', ‘WA' and ‘Kirkland' contained in its location attribute are considered as in Washington, except these also contain words like ‘dc’, ‘d.c.’, since Washington D.C. is not a part of Washington state. These tweets with words like ‘Massachusetts', 'Boston', 'MA', 'Springfield', ‘Mass.' contained in its location attribute are considered from Massachusetts. After that, all tweets from Washington are labeled 0 and all tweets from Massachusetts are labeled 1. Their complication is split into train set and test set.

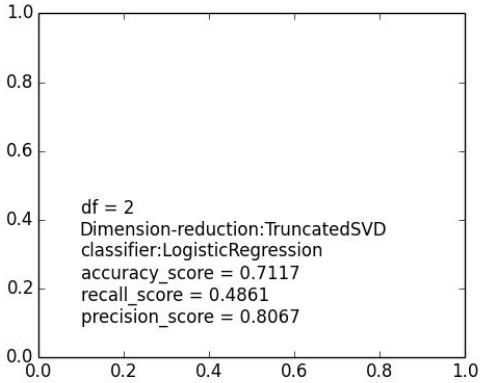
The next step is very similar to project 1. All tweets in train set are vectorized, and SnowballStemmer from nltk is used to translate words into stemmed version. Note that these tweets contain a lot of unicode characters like emojis, which are not supported by the tokenizer, so these characters are filtered out. Then TFiDF transformation and SVD dimension-reduction with n_components=50 are applied to the word-count matrix. We use three classifiers to learn from the dimension-reduced matrix, and below is the results.

Classifier: Logistic Regression





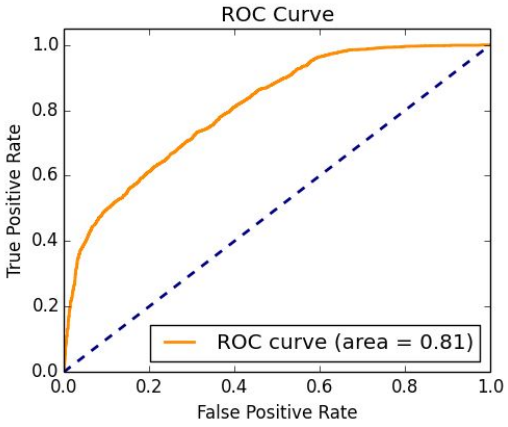


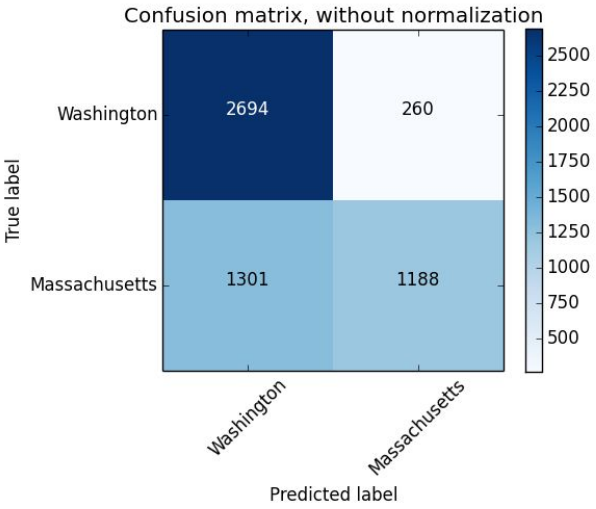


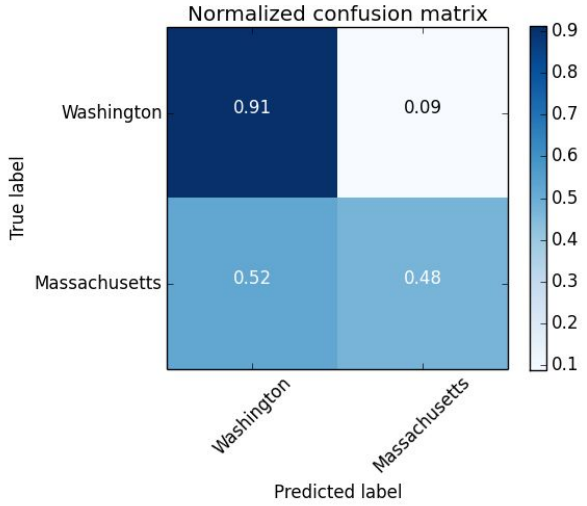
Area under ROC	Accuracy Score	Recall Score	Precision Score
0.81	0.7117	0.4861	0.8067

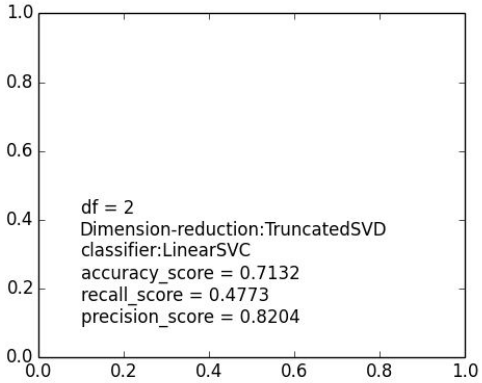
Table. Scores of Logistic Regression Classifier

Classifier: SVM





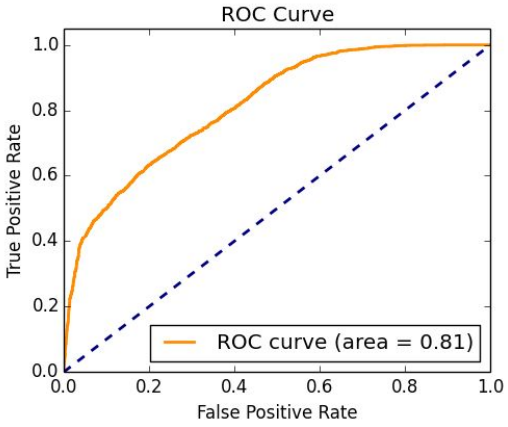


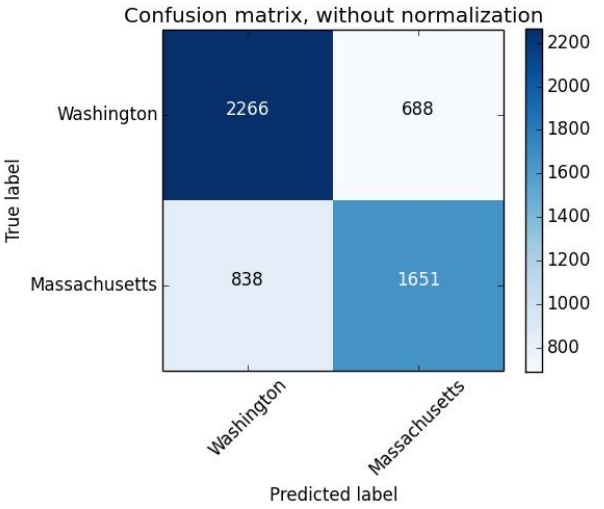


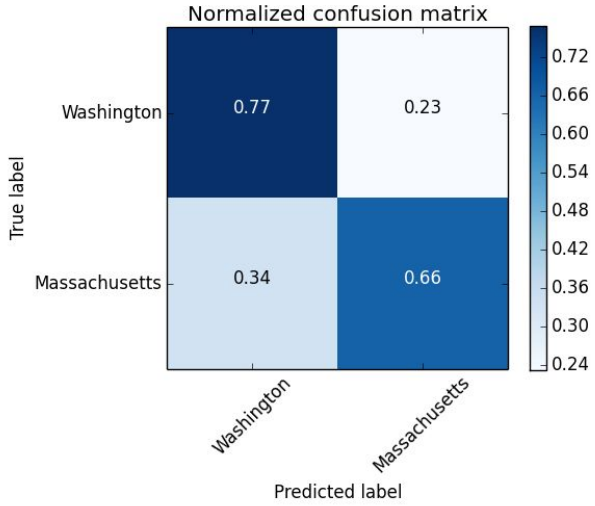
Area under ROC	Accuracy Score	Recall Score	Precision Score
0.81	0.7132	0.4773	0.8204

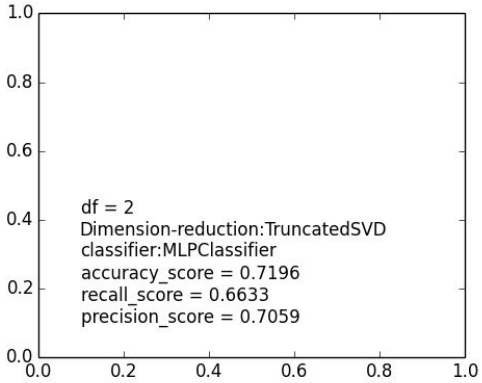
Table. Scores of SVM Classifier

Classifier: MLP







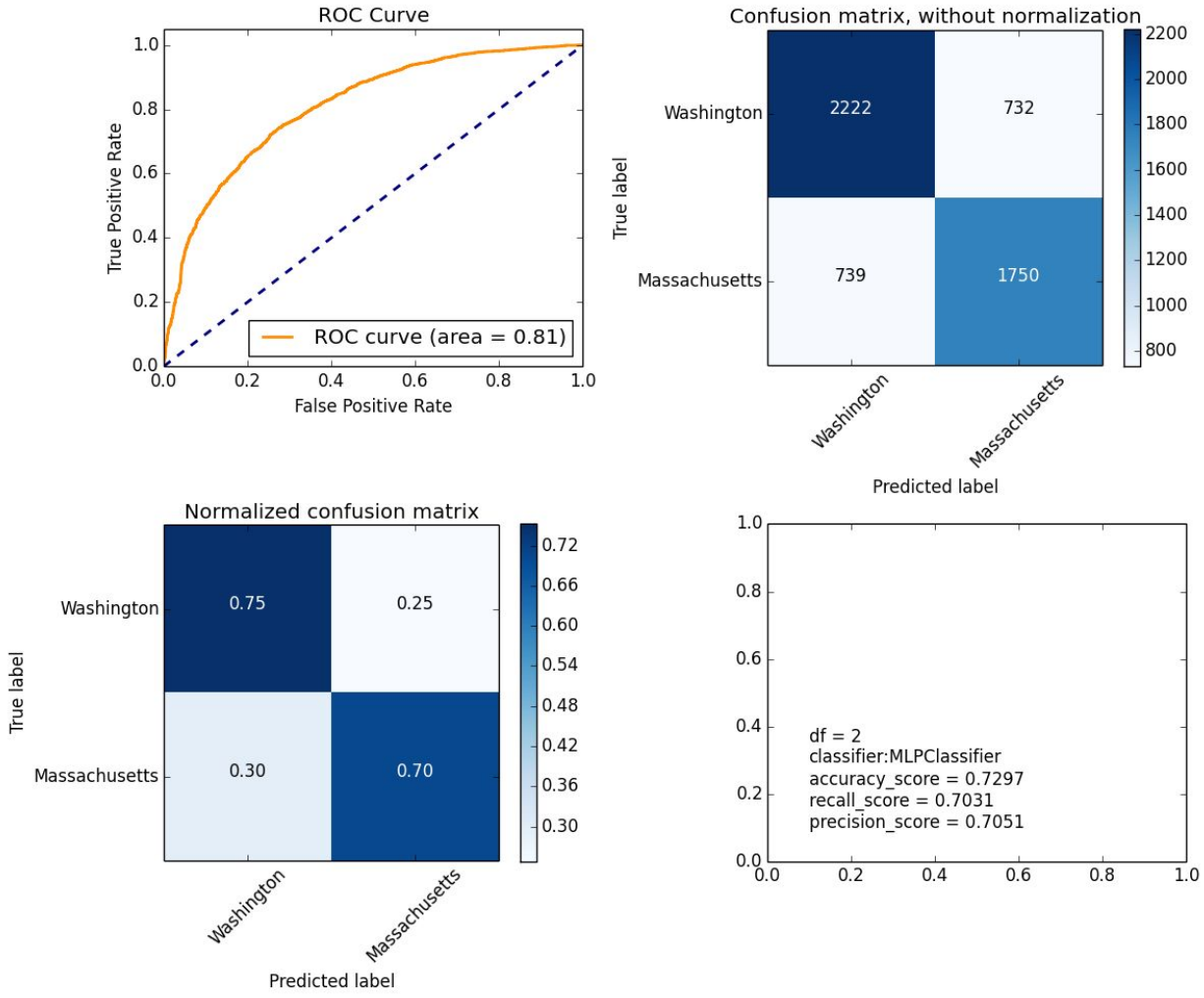


Area under ROC	Accuracy Score	Recall Score	Precision Score
0.81	0.7196	0.6633	0.7059

Table. Scores of MLP Classifier

For all three of these classifiers, their ROC curves are above the reference line, which means they are able to tell the difference between tweets from two states. However, these ROC curves are still far away from the ideal upper-left corner, and this means it's not an easy task to distinguish this difference; one has to trade off between true-negative rate and false positive rate. The precision scores of all three classifiers are around 0.7. However, the recall scores of logistic regression classifier and SVM classifier are less than 0.5. The confusion matrix shows the issue more clearly. While these two classifiers can correctly identify these tweets from Washington, their performance on predicting tweets from Massachusetts is no better than randomly guess. More than half of tweets from Massachusetts are misclassified into Washington. A classifier like this is not useful in real world since one of its prediction is totally

unreliable. Among these three classifiers, MLP classifier performs the best. Although its area under ROC curve is same as rest (0.81), it achieves accuracy score of 0.7196 and recall score of 0.6633. These scores are still not impressive, but the recall score is much better than the other two classifiers. This observation shows how difficult it is to prediction the location of a tweet solely based on its content. Intuitively, looking at a tweet and tell which state it is from is a very tough task even for human beings. This also shows algorithm based classifiers can outperform human beings at some tasks. So far, the best result we got is using MPL classifier. What if we skip the dimension-reduction and perform training on the TFiDF matrix directly. Intuitively, SVD dimension-reduction tries to capture these features that vary the most among all samples. However, it is possible that the key feature that differentiate the location is not that obvious. So we tried to train a MLP classifier on the TFiDF matrix directly. Below is the result.



Area under ROC	Accuracy Score	Recall Score	Precision Score
0.81	0.7297	0.7031	0.7051

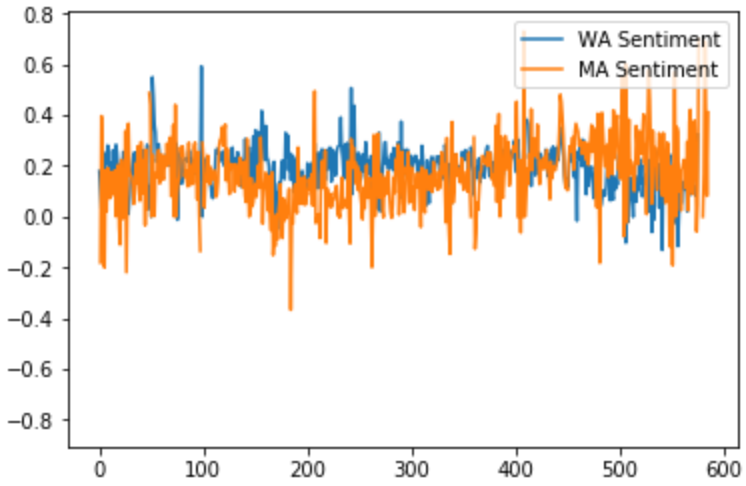
Table. Scores of MLP Classifier trained on TFiDF directly

The area under ROC does not change (0.81). The accuracy score slightly improved to 0.7297, and is the best among all. The recall score achieves 0.7031. The confusion matrix also shows this classifier is not biased to any category. So MLP classifier on TFiDF should be considered the more reliable, although the training time takes much longer.

Problem3
Sentiment analysis on tweets from Massachusetts and Washington over time

Since Super Bowl XLIX is between New England Patriots and Seattle Seahawks, then the sentiment of Super Bowl related tweets in these two states (Massachusetts and Washington) should change as the progress of game change. One important assumption we made here is people in each state support the team from their state. For example, if a tweet is sent from Massachusetts, then its user is assumed as a supporter of New England Patriots. While this assumption is not always true, but should be valid for most of the time.

We reuse the code in part 2 to get ground-truth location labels for tweets with all hashtags. After that, SentimentIntensityAnalyzer from nltk package is used to analyze the sentiment index of each tweet. The compound value is used as the sentiment index. A positive compound value means this tweet has positive sentiment, like happiness and excitement, and vice versa. Then average sentiment index of each hour is stored in a list for both states. The time is with reference to the hour when the first tweet in all dataset was posted. The sentiment index changing over time for both states is plotted below.



The plot contains a lot of noise, which makes it a little difficult to decipher. Overall, we can divide the whole plot into 3 periods. From 0 to about 100 hours, the sentiment index of both states are very close. Form about 100 to about 400 hours, the Washington got the upper hand. After about 400 hours, the sentiment index of Massachusetts overwhelms that of Washington for most of

the time. Combined with the result of that game: New England Patriots won the game on Feb 1, 2015. The trend of plot shows that users in Massachusetts went excited for winning the game for many days. This phenomenon also implies the assumption that each state support the team from their state is valid. The game ended on 452th hour, when two solid peaks appeared, and the peak of Massachusetts is higher than that of Washington. This implies people from both states enjoyed the match. Although people in Massachusetts as winner are more excited, but the sentiment in Washington is still ‘positive’ when the game ends.