CS 109a - Milestone #3

Alex Lin Melissa Yu

November 5, 2016

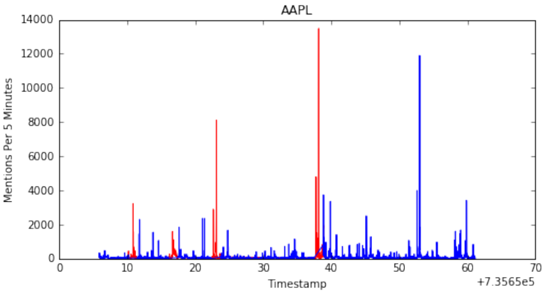
1 Datasets

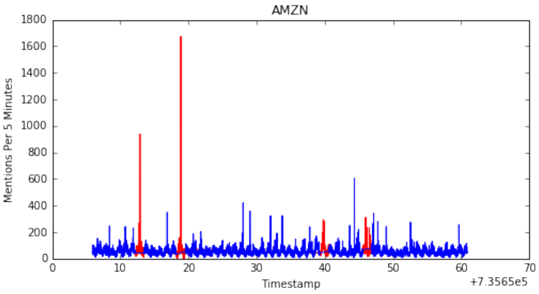
We obtained our time series data from the following link:

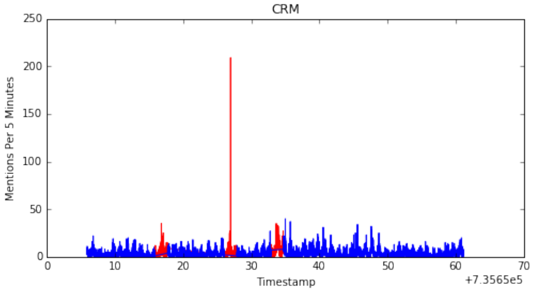
`https://github.com/numenta/NAB/tree/master/data/realTweets`

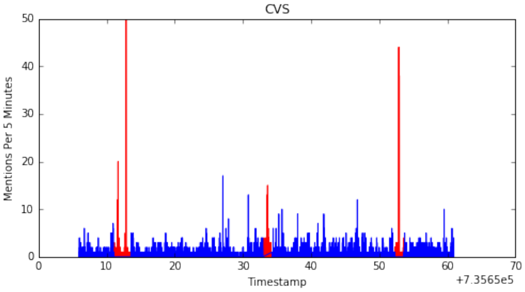
Each dataset corresponds to a count of the number of Twitter mentions of large, publicly-traded companies over an interval of 5 minutes. Each row in a dataset contains a timestamp for when the 5-minute interval starts and the corresponding counts for the associated company. There are 10 companies/datasets in total - AAPL, AMZN, CRM, CVS, FB, GOOG, IBM, KO, PFE, UPS. These can be found in our zip file under the folder `data`. We plan to use these 10 time series to test the three algorithms we read about in the theoretical papers for Milestone # 2.

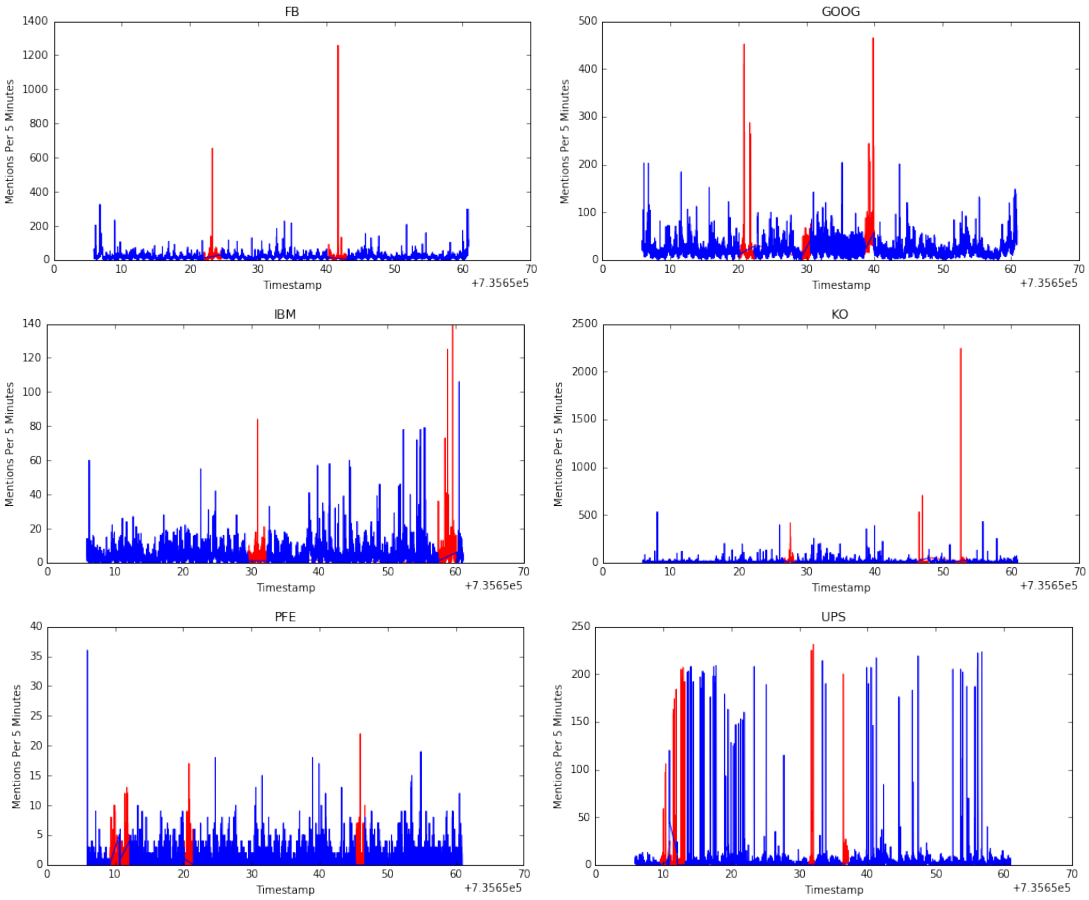
We created a Python script (`visualizations.ipynb`) to process each dataset and visualize the associated time series. Each dataset is stored as a Pandas dataframe with three columns - timestamp, value, and label. Timestamp and value come straight from the raw data files; label denotes whether or not the specific instance is an anomaly. The time intervals of anomalies are stored in another file, `anomalies.json`. Our Python script processes this file and assigns each instance the label 'True' if it lies within an anomalous time interval and 'False' otherwise. Then, we plot the entire time series, using blue for the regular data and red for the anomalies. The graphs of the 10 time series are below. They can also be found in the folder `img`.










FB
Mentions Per 5 Minutes
1400
1200
1000
800
600
400
200
0
0
10
20
30
40
50
60
70
Timestamp
+7.3565e5
GOOG
Mentions Per 5 Minutes
500
400
300
200
100
0
0
10
20
30
40
50
60
70
Timestamp
+7.3565e5
IBM
Mentions Per 5 Minutes
140
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
Timestamp
+7.3565e5
KO
Mentions Per 5 Minutes
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60
70
Timestamp
+7.3565e5
PFE
Mentions Per 5 Minutes
40
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
Timestamp
+7.3565e5
UPS
Mentions Per 5 Minutes
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Timestamp
+7.3565e5