CS 109a - Milestone #3

Alex Lin Melissa Yu

November 5, 2016

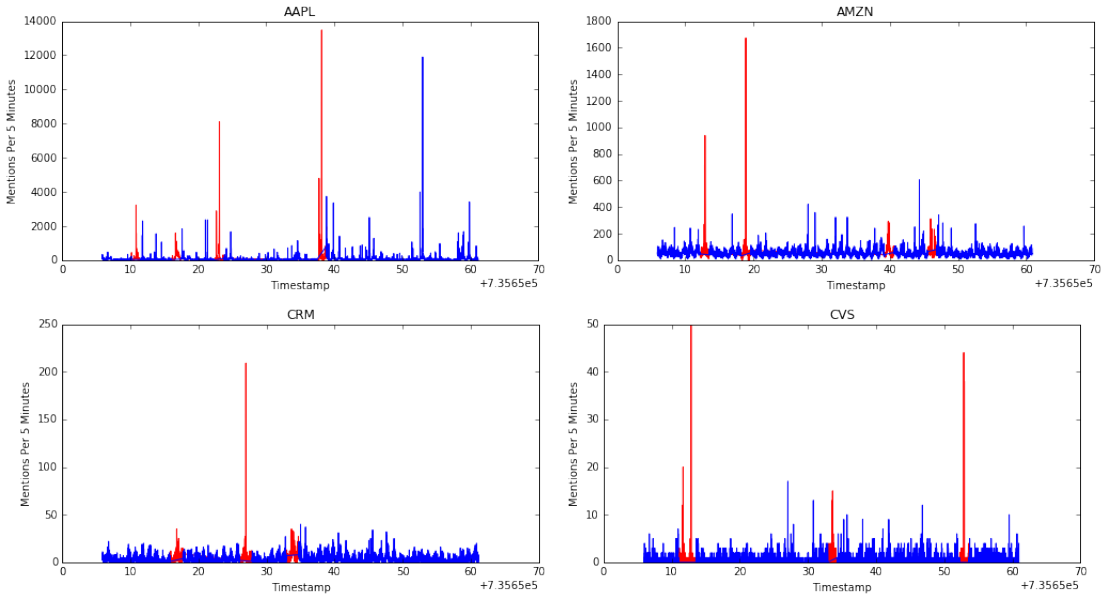
1 Datasets

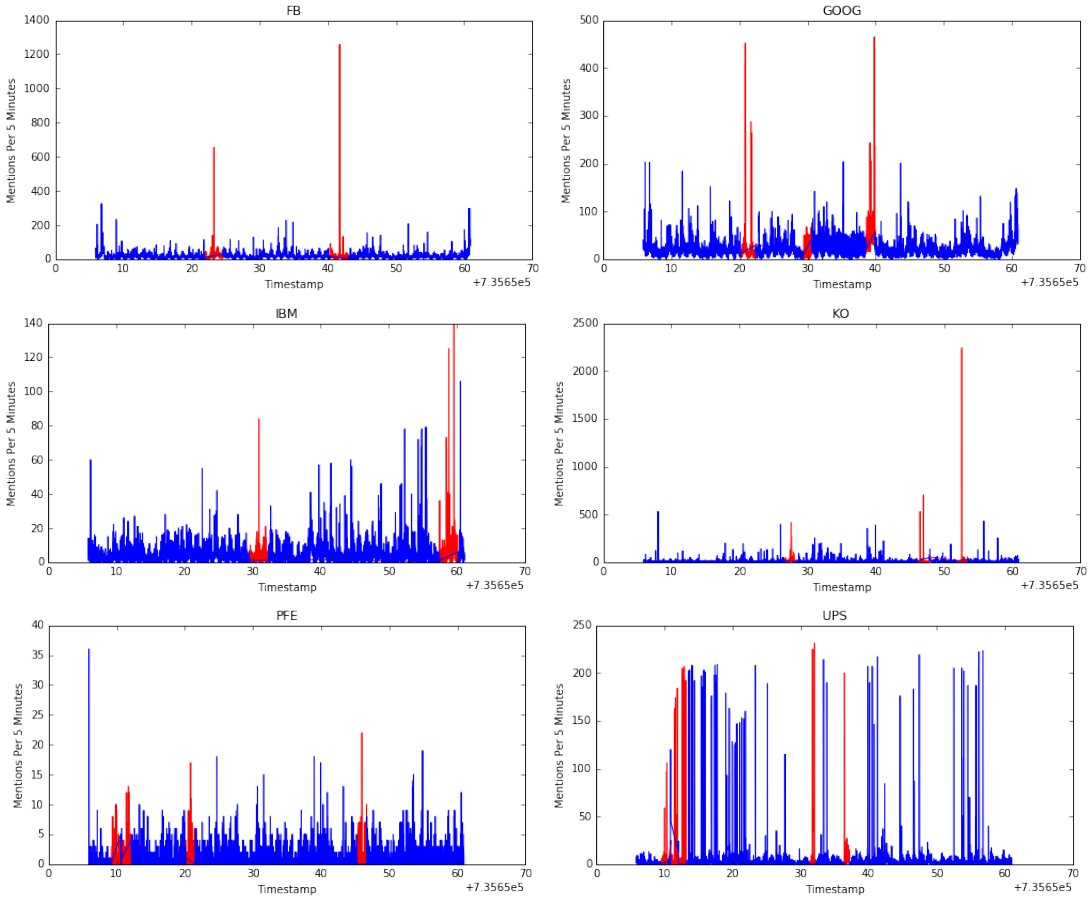
We obtained our time series data from the following link:

`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`.





Note that anomalies generally correspond to a sudden peak in Twitter mentions. However, the relationship is not always clear cut, given that there are many non-anomalous peaks and it is not always easy to tell whether or not a given rise in activity is unusual or expected. Here are some statistics giving the mean number of mentions for anomalous and non-anomalous time frames:

Company	Mean Mentions (Anomalous)	Mean Mentions (Non-Anomalous)
GOOG	29.1208100559	19.9031922276
PFE	1.30604534005	0.817659425368
IBM	5.52641509434	4.26393064392
AAPL	214.986775819	71.1928182199
KO	16.7971014493	10.7964806506
CVS	0.617955439056	0.332100230334
FB	18.1466498104	17.7740509438
AMZN	63.4063291139	52.1778120834
UPS	10.5621451104	4.88964358238
CRM	6.16886377903	3.03543224544