Text Mining

Carl Moser

February 2015

1 Project Overview

For my project, I decided to download all of the Olinsider blog articles and get the sentiment of them. I used Pattern to download the sourcecode and the sentiment function to get the sentiment of the posts. From doing this, I wanted to learn what the general sentiment of the Olinsider blogs was and to see how if the sentiment has changed over time.

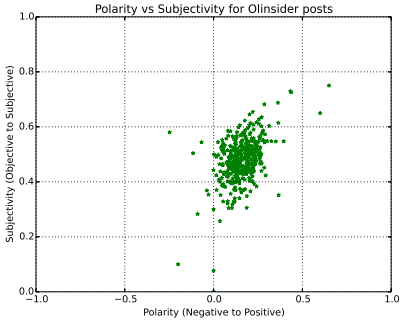
2 Implementation

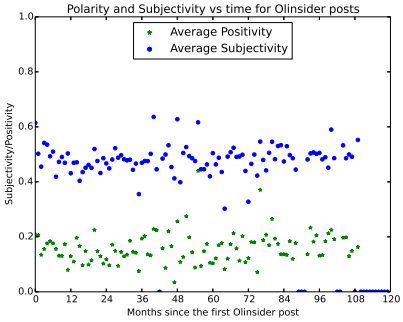
One of the major components of my implementation is the way that it downloads and saves data. Downloading the sourcecode for the articles and the intermediate webpages takes a lot of time and that can cause pattern to time out, so saving the data once it was downloaded was necessary. It also increases the amount of time needed to run the program so I implemented an algorithm that would check to see if the sourcecode file was in the Data directory and if it was, it read in from that. Because of the number of articles, close to 400, it takes a long time to parse through the sourcecode to get the article so saving the data once it was parsed was an important step in my program. One of the interesting effects of my algorithm is that it will detect if another article has been added and will update the parsed file, otherwise it will load the previously parsed data.

To store the data, I decided to use a list of lists. The inner list contained the date the article was published and the article itself. The reason that I chose this method over a dictionary where the key was the date and the article was the value was because some articles were published on the same day. Since I decided to analyze the data based on the time it was published, wanted to have easy access to the date that the article was written. I also did not want to have the date inside of the article's string for this reason. This choice made it easy to detect when a new article was published, as the length of the list would change. Another design decision that I ended up making was removing the day the article was published because the data was not dense enough to make sense to plot it vs plotting it by month.

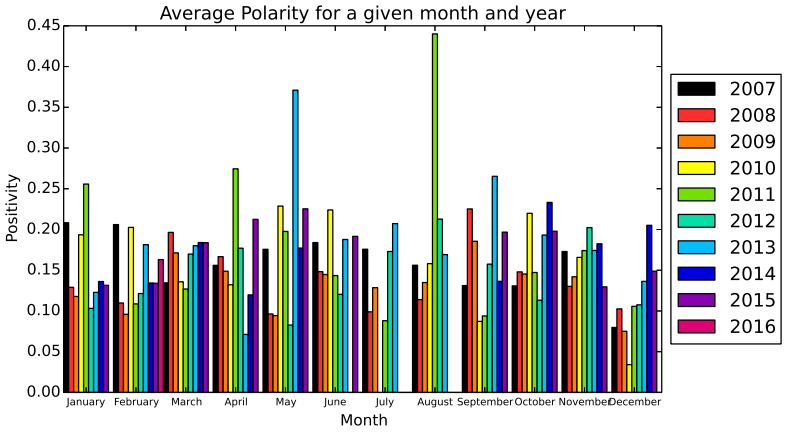
3 Results

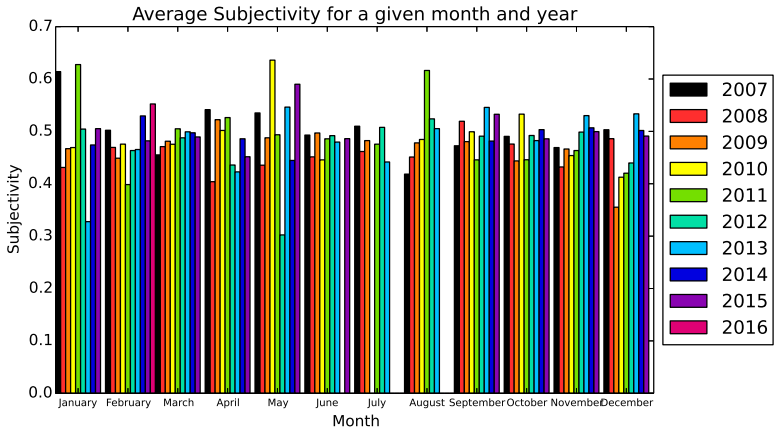
From the output that I got from my program, I found out that most of the articles on Olinsider are almost nuetral when it comes to subjectivity and objectivity and are also somewhat positive. I also discovered that the average sentiment over time has stayed about the same. Most of the articles fell directly between subjective and objective which can be seen in the graph below. Almost all of the articles were also slightly positive.





To see if there were any trends based on the time of year, I made bar graphs based on the month where the bars were different years. I had expected to see some sort of positive trend around the holidays, however there was not a trend based on holidays. There does seem to be a higher positive sentiment around May which is right before summer break. From the graph of average subjectivity for a given month and year, there seems to be a trend towards a more neutral sentiment over the years, which can be seen more clearly between August and December.





4 Reflection

The algorithms that I used to store and get the data worked very well. They allowed my program to run fairly quickly and allowed for future articles to be added to the data set without any modification. I could have improved one of the saving algorithms so that it would load the previous data and append the new processed article instead of having to re-process all of the articles. I also could have played around more with using a dictionary to store the date-article pairs, as I did not need the date in the end. Overall, I feel that I chose a project that was appropriately scoped.