James Thornton and Dylan Hurwitz
Artificial Intelligence
Total Time: 20 Hours Each
Time Since Last Status Report: 12 Hours Each

Stock Analysis Status Report 2

Summary

Since our last status report, we've quadrupled the size of our data set, reworked the DataReader file, laid the groundwork for several new concrete features, and obtained some preliminary evaluations of the program. Previously, our data set contained 10 quarterly reports from Microsoft and 10 from Apple. We've added another 20 from Microsoft, 20 from Google, 10 from HP, and 10 from Amazon. In the process, we added several new attributes to each document. Where previously a document only had one attribute, the label, each document now has a price attribute, a date attribute, a company attribute, and a risklength attribute. The risklength attribute is the word count of the section in which a company is legally required to disclose all risks, and is a concrete feature we intend to use. In order to incorporate this new functionality into our program, we needed to alter the DataReader file so that it would gather and store all of these attributes. It now returns all of them when "next" is called, which leaves room for several experimental paths. In particular, we now have the potential to consider concrete features including risklength, time of year, scale of stockprice (i.e. Apple's is several hundred while Microsoft's below fifty). Over the next week, we will be brainstorming other features to investigate. We also intend to investigate within company data sets in comparison to cross-company data sets. When classifying, we will also consider assigning strength weights to our classifications. These will be derived from the difference between p(positive) and p(negative), and we can develop a more sophisticated evaluation metric to determine whether or not these weights correlate to success rates.

Results:

These are results using the current form of our final evaluation metric for the program. The program labels a document buy or sell depending on when the stock price has risen or fallen since the previous report. It is a strict buy/sell for positive/negative price change. To evaluate, we use the evaluation metric from the Naive Bayes assignment. The table below shows an average set of results, although we still need a larger data set for success rates to stabilize. In the result set below, 18 of 28 Documents have been classified correctly, representing 62% accuracy. 14 correct out of 16 led to 88% precision for "Buy." 4 correct out of 12 led to 33% precision for Sell.

Table A

Results*	**All**	**Buy**	**Sell**
Correct	**18**	**14**	**4**

Total	**28**	**16**	**12**
Accuracy	**0.62**	**.875**	**0.3333**

*Data being used is our full data set of stock history of multiple tech oriented companies

Table B

Results*	**All**	**Buy**	**Sell**
Correct	**3**	**2**	**1**
Total	**7**	**4**	**3**
Accuracy	**.375**	**.5**	**.3333**

*Data being used is only Microsoft Stock History, predicting Microsoft Quarterly's

Our current results are good for our large data set of multiple companies with an accuracy of 62%. If the program could maintain 62% accuracy, it would be profitable. A portfolio consisting of companies chosen with 62% accuracy would see a steady increase in value from year to year. The second table represents the results of using only data from Microsoft. The results for the single company predictions are significantly worse with an accuracy of 37.5%, which makes us believe that the classifier might need a lot of data to do well, because the single company data set is much smaller than the combined data set.

Problems:

At the current moment we have not encountered any problems that might prevent the completion of our project.