Predicting Financial Markets through Media

Team

Teddy Knox, 2015
Will Potter, 2014.5
Parker Woodworth, 2013.5

Summary

Our goal is to predict the performance of financial markets using data from financial news sources. Based off of historical data, we will search for performance trends that follow articles published about a specific company, industry, country or index with respect to a particular security, index or currency. We will look to determine not only whether negative media surrounding a particular entity affects the future performance of a company, but additionally to what degree a number of bad reviews will affect the performance. Does a history of negative performance following a bad article lead to bad performance in the future following a current bad article? To what degree does a given company respond to articles regarding competitors or substitutes? To start off, we will try and gather a historical dataset of articles regarding a few companies and then tie a particular article (or set of articles) to a set of time intervals corresponding to stock performance. Therefore, we can cluster realtime articles with historical articles and then try and come up with a likely predicted market response based on the average of historical responses. Additionally, we may investigate creating a classifier based on training data, however, we could use the weighted average of market responses to formulate a composite response. That is, if there is a 30% likelihood that it rises by 2.5% and a 70% chance that it rises by 1%, then the weighted average would be .3 * 2.5 +.7 * 1 = 1.45%. After a set amount of time, our algorithm could then compare the predicted response to the actual response and then store the new response as a part of the training data to classify future articles.

Resources

Financial Information

- ystockquote.py (http://goldb.org/ystockquote.html) - A python API for Yahoo Finance market data.
- PYQL (https://code.google.com/p/pyql/) - Python API to retreive stock information from Yahoo! Finance using Yahoo Query Language.

News

- NYTimes Article Search API (http://developer.nytimes.com/docs/article_search_api/) - provides a search for headlines and abstracts.
- Yahoo! Finance Company News RSS Feed (http://developer.yahoo.com/finance/company.html) - provides an RSS feed of news relevant to a given company.

Machine Learning

- PyML (http://pyml.sourceforge.net/) - an interactive object oriented framework for machine learning written in Python. PyML focuses on SVMs and other kernel methods. It is supported on Linux and Mac OS X.
- SCIKit - http://scikit-learn.sourceforge.net/stable/
- PyBrain - http://pybrain.org/
- MLPy - http://mlpy.sourceforge.net/

Using the above resources we will attempt various grouping algorithms on our collected textual dataset, and use a boosting technique to weight the most accurate market predictions. These libraries will speed our development process by allowing us to focus on high-level algorithmic design instead of fine-tuning and debugging.

Background

Reading through a few research papers that approach similar problems was a good way to gain a background in this particular type of machine learning problem. We found these papers particularly useful in giving us initial direction in choosing which grouping algorithms to use for the task.

The research paper cited below addresses the same challenge that we approach in our project, but instead of focusing on the details of clustering implementation, uses statistics to determine if a correlation between news articles and the market’s fluctuations exists. If such a correlation were not found to exist, then a machine learning technique based off of this data would not be effective. Fortunately, the results significantly showed that connection between online media and markets.

W. Zhang and S. Skiena. *Trading strategies to exploit blog and news sentiment*. In Fourth Int. Conf. on Weblogs and Social Media (ICWSM), 2010.

The following paper addresses the difficulty of extracting relevant features from texts of a narrow domain. Narrow textual domains often have high vocabulary overlap between clusters, rendering conventional clustering difficult. This paper introduces a few textual feature extraction algorithms that are agnostic to vocabulary overlap.

Pinto, D., Rosso, P., & Jime´nez-Salazar, H. (2011). *A self-enriching methodology for clustering narrow domain short texts*. Computer Journal, 54(7), 1148–1165.