CS 410: Project Proposal
Causal analysis of stock prices vs. news (Free Topic)

ritikd2

October 24, 2021

Members

Ritik Dutta - ritikd2

Topic

The goal of this project is to implement a system that explores if there exists a causal relation between news on popular financial market websites and stock prices. Relations between stock prices and news cycles are of particular interest to those that engage in equities trading. Being able to identify trends early on might allow traders to allocate capital in ways that maximizes their gains.

I plan to implement the following:

1. Fetch news data from popular financial websites. I plan to collect this by scraping news aggregators like Zerodha Pulse
2. Fetch closing prices of stocks from the National Stock Exchange of India
3. Implement the Granger causality test to determine if there exists a causal relation between news data and the corresponding stock price
4. Explore building a predictive model should a correlation between the two exist

The project I'm proposing is similar to this project from the previous iteration of the course but differs in the following major ways:

1. The previous team used a fixed 5-day lag to check to test their hypothesis. A fixed time-lag can be restrictive, since different kinds of news can affect prices at different speeds. I plan to explore a much broader window of time lag for any possible correlations
2. The previous team used tweets as the data source. However, parsing for good tweets can be difficult, and they have a limited reach compared to news websites, which can thus influence a much larger audience. I plan to rely on headlines and content scraped from financial news websites instead.
3. Equities might often show correlations amongst each other, which was not explored by the previous team. Furthermore, sometimes general news can cause overall changes in financial markets. I plan to explore this as well.

The project will be implemented using Python.

Since I'm working on this project alone, the expected workload has to be at least 20 hours. I expect the following distribution across tasks:

1. Data scraping - 5 hours
2. Data cleaning, formatting - 4 hours
3. Check for causal relations - 15 hours
4. Evaluate results, prepare report - 5 hours

All updates will be pushed to this Github repo.