Questions:

1. How does rumors (speculations) and news affect stock prices?
2. How does changes in one firm affect stock prices of other firms in the same industry?
3. Discovering relationship between different industries e.g how real-estate industry impacts manufacturing firms?

Data Sources:

- Intrino API (www.intrino.com) - for extracting historical stock prices and other metrics for different companies.
- Yahoo Finance() - for extracting historical stock prices and other metrics
- News API - for extracting text documents

The extracted data from Intrino API and Yahoo Finance API is in JSON format. Matching entities between two data sources is the company name.

Since our data consists of multiple files about 400 KB each, we could not add all files to Github which limits the number of files in each directory to 1000 files.

Description of how we extracted structured data from data sources:

We wrote scripts to make API calls to the data sources and collect data:
getCompanyData.py - extracts historical data from Intrino API
getCompanyDataYahoo2.py - extracts stock prices data of last 10/ 5 /3 /2 /1 years (whichever is greater) from Yahoo Finance
getTickers.py - extracts Nasdaq symbols for each company

Text Documents:

We will be using text documents for sentiment analysis.

Open Source Tools:

We did not use any open source tool yet in our project. We are just using API calls for web crawling.

Links to data and scripts:

Intrino data and Yahoo data (compressed data):-
https://uwmadison.box.com/s/f3awltz1od9hlv1scrswhqhfqkaqdtg3
Since our data set is huge, approximately 3.7 GB with thousands of files, we could not host it on Github.

News articles:
https://github.com/guneetsinghmehta/financial-modeling/tree/master/Data/news_articles

Scripts:
https://github.com/guneetsinghmehta/financial-modeling/tree/master/Data_Collection_Scripts