Principles of Programming: Final Project Progress Report
DSCI 510 – Fall 2025 – Dr. Alexey Tregubov
University of Southern California
Madeleine Willson

Project Scope Update

This project aims to build a machine learning model that predicts IBM's stock price using Google search interest and sentiment from New York Times (NYT) articles. Since beginning the project, the overall scope has remained the same, but the timeline for data collection has somewhat expanded.

I successfully gathered initial IBM stock data using the yfinance library, and article data from the New York Times Article Search API. After reviewing the stock data frequency and the limited number of NYT articles returned for two years (227 data points), I determined that the timeframe must be expanded to ensure sufficient observations for modeling. Because yfinance only supports specific historical intervals, my next viable window is five years. I expect this will provide enough stock and article data for alignment, but I am still validating this assumption.

Data Sources

I collected historical IBM stock data using the Ticker.history() method from the yfinance library and saved the resulting DataFrame as a CSV file. Also, I obtained article data on IBM from the NYT Article Search API, which returns detailed JSON documents. A typical entry includes fields such as abstract, headline, keywords, pub_date, and web_url, along with relevant metadata about the article's content and publication context.

The New York Times Article Search API allows data to be searched for by keywords and filtered using parameters such as date ranges. It returns structured JSON objects for each article, which include metadata, headlines, abstracts, and publication details.

Issues / Difficulties

Several challenges arose during early development.

First, I expected the .history() output to automatically index by datetime, but it did not, so I added the index manually.

This project was my first time editing in VS Code which unfortunately introduced version control complications. Before enabling autosave, I committed unintended files (they were thankfully empty) because my .gitignore had not yet been saved.

This is my first project fully using Git, and at one point I edited files without realizing I had forgotten to push previous changes. Because I was the sole contributor, I was able to resolve the issue by force pushing.

Also, learning to work with the NYT API was initially difficult. When using a short time window, the API returned zero results, which made it seem like the request was failing.

Looking ahead, I expect data cleaning and overall dataset alignment to be tedious for many reasons. For example, datetime values across sources will likely not match up perfectly. This means, I will need to figure out a system to aggregate or average sentiment and interest measures to align them with stock price intervals.