Principles of Programming Final Project Topic Proposal
DSCI 510 – Fall 2025 – Dr. Alexey Tregubov
University of Southern California
Madeleine Willson

Project Title

Predicting IBM Stock Movements Through Public Interest & Sentiments

Project Description

This project will develop a machine learning model to predict IBM's stock price trends using news and public interest data. It will combine stock data from Yahoo Finance, Google Trends search interest, and headlines from The New York Times. Stock data will be aggregated to show overall trends. Google Trends will capture public attention over time. Sentiment scores from New York Times articles with IBM in the header will be averaged to measure overall media tone. Together, these data sources will be analyzed to understand how market sentiment and public interest influence IBM's stock movements.

Data Sources

Data source #	Name / short description	Source URL	Type: - API - Web page - file	List of fields	Format: - json - xml - csv - sql - other	Have tried to access/collect data with Python? yes/no	Estimated data size, number of data points you plan to use
1	Yahoo Finance Data (IBM ticker)	This is the data: https://finance.yahoo.com/quote/IBM/ This is the library documentation: https://ranaroussi.github.io/yfinance/	yfinance library or file	- open - high - low - close - volume	DataFrame or csv	yes	~500
2	Google Trends Data about IBM (Interest Over Time)	https://pypi.org/project/pytrends/	pytrends library	- Date/Timestamp Index - Keyword Columns (I will probably start with just "IBM") - isPartial	DataFrame	no	~500
3	NYT News Articles Data about IBM	Source: https://developer.nytimes.com/ Documentation: https://developer.nytimes.com/docs/articlesearch-product/1/overview	Article Search API	-headline - Byline - Date - URL - abstract - section name - news desk - type of material	JSON	no	~500

Proposed Analysis

In this project I plan to begin with cleaning up the data by filtering out unnecessary columns or rows, transforming values and integrating the data into one source. Then I will need to conduct exploratory data analysis (EDA) to understand my three data sets each using a common time grid. This will help me begin to understand the relationships between IBM's stock price, public interest volume, and media sentiments over time. Then, I will experiment with several machine learning models to predict changes in IBM's stock price based on Google search volume and headline sentiment from The New York Times. I want to test linear, Random Forest and XGBoost regression models. Then I will use model explainability methods such as SHAP values and performance indicators such as f1 scores to determine the best model.

Link to GitHub Repository

https://github.com/maddwillson/dsci510_fall2025_final_project