Checkpoint B: Data and Schema for the Knowledge Graph

MSDS 459

Technology Sector Team

Nick Butler, Michael Rivera, and Richard Pereira

May 11, 2025

Northwestern University

INTRODUCTION

The final project for the course tasks each team with developing a knowledge base for a specific industry or set of companies, emphasizing integration of media coverage, financial data, and structured information. Our team focused on the Technology sector, aiming to build a framework that enables analysis of how public sentiment and news events correlate with financial performance (particularly, stock prices). Originally, we targeted the emerging Generative AI landscape, intending to track new players and innovations. However, early feedback from Professor Miller indicated a need for a more data-rich, measurable, and company-specific focus. In response, we pivoted our efforts toward analyzing well-established public companies with sufficient historical data to provide a meaningful time-series for further analysis.

Embedded within the standard structure for this paper, we will also cover the data sources identified, the information collected thus far, schemas/code/visualizations, and our planned next steps for reaching the next checkpoint.

LITERATURE REVIEW

The project is informed by several streams of existing work. Prior studies have shown that sentiment analysis of news and social media can predict short-term financial market trends. Techniques such as VADER sentiment scoring, topic modeling, and emotion classification have long been used to extract signals from unstructured text. Additionally, knowledge graphs have emerged as a powerful method for linking various entities (companies, products, events, etc.) and supporting semantic querying.

Our revised approach builds on this foundation by merging sentiment-labeled media coverage data with financial time-series data, creating a structure that better supports predictive analysis to forecast stock prices.

METHODS

Our revised project approach now focuses on building a knowledge base linking unstructured or textual data, such as news sentiment, with objective financial outcomes (stock prices) for Apple, Amazon, Google, and Microsoft. We used a combination of:

- **Data Collection:** Web scraping and API-based ingestion of news articles from TechCrunch, Bloomberg, and Reuters, and more.

- **NLP Processing:** VADER sentiment scoring, emotion classification (e.g., joy, fear), topic tagging (e.g., acquisition, regulation), and novelty detection.

- **Financial Data:** Stock closing prices for each company on the publication date of relevant news articles, with future movement labels over 1-day, 3-day, and 7-day intervals.

- **Integration:** Combined structured (stock prices) and unstructured (news) data into a single dataset nlp_news_with_price_labels-FINAL.csv.

- **Schema Design:** We created a knowledge graph schema including node types (Company, NewsArticle, Sentiment, Emotion, EventTopic, Date, Movement) and edges (e.g., mentions, published_on, has_sentiment, price_movement_3d).

- **Tools:** Python (Scrapy, BeautifulSoup, NetworkX), Neo4j for visualization and querying, and Matplotlib for schema representation.

This set of methods was chosen to support both downstream analytics and extensibility (for example, to other companies). It allows us to bridge qualitative content with quantitative market indicators, such as the inclusion of more companies, or the application of machine learning models. With this foundation in place, we're now well-positioned to begin populating the graph and analyzing it accordingly.

RESULTS

We successfully created a labeled dataset containing thousands of records linking news events to financial outcomes. Each row in the dataset includes:

- Article metadata (e.g., title, publication date)
- NLP features (e.g., sentiment score, emotions, topics)
- Financial metrics (e.g., close price, price movement labels)

Example: An article titled *"Apple announces major acquisition"* might include a high anticipation score (0.72), a topic flag for acquisition, a stock close price of $177.82, and a movement label of "up" over the next 3 days.

We generated a knowledge graph (knowledge_graph_edge_list.csv) visualizing these relationships and enabling graph-based queries. The graph includes thousands of nodes and edges and supports exploration of trends such as:

- Emotional tone by company over time
- Event types most associated with stock movements
- Sentiment patterns preceding price drops

Together, the structured dataset and knowledge graph provide a fairly flexible foundation for answering both descriptive and predictive questions about how public discourse affects financial performance. This sets the stage for the final phase, where we will operationalize the knowledge base for advanced querying and predictive analysis.

CONCLUSIONS

The project demonstrates the potential of knowledge graphs as a tool for connecting qualitative and quantitative indicators of market performance. By structuring the interplay between public sentiment and financial data, we created a resource that supports both exploratory analysis and predictive modeling.

Next steps include:

- Importing the edge list into Neo4j for advanced queries
- Building Cypher queries for pattern detection
- Enhancing the model with temporal dynamics and potential LLM integration for deeper event understanding

This knowledge base not only fulfills the academic goals of the course but also offers practical applications for financial analysts, competitive intelligence professionals, and data scientists exploring market signals from public news.

REFERENCES

Chakrabarti, Soumen. *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco: Morgan Kaufmann, 2003.

Chakrabarti, Soumen, Martin van den Berg, and Byron Dom. “Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery.” *Computer Networks* 31, no. 11–16 (1999): 1623–1640. https://www.cse.iitb.ac.in/~soumen/doc/www1999f/pdf/www1999f.pdf.

Hajba, Gábor László. *Website Scraping with Python: Using Beautiful Soup and Scrapy*. New York: Apress, 2018.

Mitchell, Ryan. *Web Scraping with Python: Collecting More Data from the Modern Web*. 2nd ed. Sebastopol, CA: O’Reilly Media, 2018.

Miller, Thomas W. *Web and Network Data Science: Modeling Techniques in Predictive Analytics*. Upper Saddle River, NJ: Pearson FT Press, 2015.

Nair, Vineeth G. *Getting Started with Beautiful Soup: Build Your Own Web Scraper and Learn All About Web Scraping with Beautiful Soup*. Birmingham, UK: Packt Publishing, 2014.

Olston, Christopher, and Marc Najork. “Web Crawling.” *Foundations and Trends in Information Retrieval* 4, no. 3 (2010): 175–246.

Patel, Jay M. *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale*. New York: Apress, 2020.

Smith, Vincent. *Go Web Scraping Quick Start Guide: Implement the Power of Go to Scrape and Crawl Data from the Web*. Birmingham, UK: Packt Publishing, 2019.

APPENDIX A – Data Sources Table

Source	Content Type	Details	URL
Google News Results	News & Headlines	Scraped headlines and previews mentioning Apple, Amazon, Google, and Microsoft	https://news.google.com
TechCrunch	News & Blogs	Unstructured tech news articles referenced in NLP processing	https://techcrunch.com
The Verge	News & Blogs	Supplementary tech coverage included in scraped headlines	https://www.theverge.com/tech
Yahoo Finance	Financial Data	Used to obtain daily closing prices for each company	https://finance.yahoo.com
Google Finance	Financial Data	Alternative source for stock prices (CSV download into Google Sheets)	https://www.google.com/fina

APPENDIX B – Knowledge Graph Data and Code

Dataset: https://drive.google.com/file/d/14hLJDeHOt4IQ_VJAn_fMpniFFVL8MzC4/view?usp=share_link

Folder: https://drive.google.com/drive/folders/1sp-hmw51uD1QifgU_8qkgRRE9gxVn2NO?usp=sharing