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

1. FINAL PROJECT GOAL

The final project for the course tasks each team with building a knowledge base centered on a specific industry sector or a defined group of firms within that sector. The primary goal is to create a structured repository that enables analysis of companies or emerging trends, with a particular focus on integrating media coverage, financial data, and other relevant sources to generate meaningful insights. Our team requested—and was approved to work within—the technology sector.

2. INITIAL PROJECT STRATEGY (Checkpoint A)

In Checkpoint A, we outlined our initial strategy to build a knowledge graph and web-crawling framework aimed at supporting competitive intelligence in the Generative AI space. The project addressed the challenge of tracking emerging competitors across fragmented data sources like Crunchbase, SEC filings, Wired, and TechCrunch. Our goal was to create a unified graph-based system linking entities such as companies, products, patents, and media coverage to enable semantic queries and time-series analysis. By the end of Phase 1, we had finalized our topic, designed the draft schema, identified 24 relevant data sources, and prepared to launch web scraping using Python tools like Scrapy and BeautifulSoup.

2. PROFESSOR FEEDBACK

The professor’s feedback acknowledged the strengths of our initial efforts—particularly our work on web scraping, data source identification, and schema development—but raised important concerns about the project's alignment with the course objectives. While our technical foundation was sound, our original focus on Generative AI was deemed misaligned with the primary goal of constructing a knowledge base to support media-based time series measures for predicting market movements. The professor noted several limitations with the Generative AI theme, including insufficient data history, a lack of public pure-play companies, and the challenge of isolating relevant media coverage. Additionally, our objective of tracking emerging competitors was seen as too abstract. The feedback urged us to refocus the project on specific public companies with at least two years of stock price data and align our approach toward building a testable, prediction-oriented model grounded in measurable media indicators.

3. REVISED APPROACH (Checkpoint B)

In response to feedback and early findings, we have refined our project scope to focus on a more targeted and analytically tractable domain. Our goal is to construct a robust and insightful knowledge base that captures the dynamic interplay between public news sentiment and stock market performance across four major technology companies—Apple, Amazon, Google, and Microsoft. To do this, we integrated a diverse array of structured and unstructured data sources, including news headlines, natural language processing outputs, and daily stock prices. The resulting multidimensional dataset supports both analytical depth and knowledge graph modeling. The following sections outline our data sources, describe the proposed schema, present implementation and visualization strategies, and detail our roadmap for completing the knowledge base in the final phase of the project.

List the data sources you have identified, describing the kinds of information being gleaned from these sources.

We identified and integrated multiple data sources to construct a comprehensive knowledge base that enables financial and sentiment-driven analysis. These sources include digital news articles focused on four major technology firms—Apple, Amazon, Google, and Microsoft—sourced from leading outlets such as TechCrunch, Bloomberg, and Reuters.

Each article was enriched through NLP techniques, generating features like VADER-based sentiment scores, emotion classifications (e.g., joy, fear), topic tags (e.g., acquisition, regulation), headline length, readability metrics, and novelty measures. To connect this qualitative data to market dynamics, we merged daily stock price data (close values) corresponding to each article's publication date. Finally, we labeled future price movements over 1-day, 3-day, and 7-day intervals as "up," "down," or "flat," allowing for predictive modeling. Together, these sources provide a multidimensional foundation for analyzing how corporate announcements and public sentiment influence financial outcomes.

Describe the information you have collected so far, providing examples of individual documents or tables.

We collected and integrated both structured and unstructured data into a unified dataset titled nlp_news_with_price_labels-FINAL.csv, which serves as the foundation for our analytics and knowledge graph construction. This file includes news metadata such as article IDs, company names, and publication dates, as well as NLP-enriched features like compound sentiment scores, emotion confidence levels (e.g., emotion_fear, emotion_joy), and binary topic indicators (e.g., topic_earnings, topic_acquisition). Additionally, it contains financial data including the close_price on the article's publication date and price movement labels for future timeframes: movement_next, movement_3d, and movement_7d.

For example, an article titled "Apple announces major acquisition"may have a high anticipation emotion score (emotion_anticipation: 0.72), a topic flag for acquisition (topic_acquisition: 1), a close_price of 177.82, and a three-day price movement label of "up." This richly structured dataset enables multi-layered insight generation and supports graph-based exploration of business events and market dynamics.

Provide a detailed description of the schema that you are proposing for the knowledge base. Code and visualizations from the database you have selected for the term project. If you have selected Gel/EdgeDB, a figure showing the graph nodes and links makes sense.

The knowledge graph schema captures key relationships between companies, news articles, events, emotions, and market signals to support reasoning and analysis. Core node types include Company (e.g., Apple, Google, Amazon, Microsoft), NewsArticle (e.g., News_00001), Date (in YYYY-MM-DD format), Sentiment (Positive, Negative, Neutral), Emotion(Joy, Fear, Surprise, etc.), EventTopic (such as Acquisition, Earnings, or Regulation), and Movement (up, down, flat).

These entities are linked by directed edges or predicates that define their relationships, including: (NewsArticle) —[mentions]→ (Company), (NewsArticle) —[published_on]→ (Date), (NewsArticle) —[has_sentiment]→ (Sentiment), (NewsArticle) —[has_emotion]→ (Emotion), (NewsArticle) —[about_event]→ (EventTopic), (Company) —[has_stock_price_on]→ (Date), and (NewsArticle) —[price_movement_3d]→ (Movement). Together, these triples form the structure of the knowledge_graph_edge_list.csv file, which contains thousands of such relationships and serves as the foundation for graph-based querying and pattern discovery.

Plans for completing the knowledge base, as needed for the third assignment.

To finalize the knowledge base for Assignment 3, we plan to import the edge list into NetworkX for local graph analysis and into Neo4j to enable advanced querying and interactive visualization. We will create a schema visualization using Matplotlib and NetworkX to display a sample of nodes and edges, clearly illustrating the relationships among entities such as companies, news articles, sentiments, and events. To enhance analytical capabilities, we will design Cypher queries that identify sentiment patterns preceding stock price drops, detect recurring news-event associations, and cluster emotional tone by company.

The knowledge graph will also incorporate temporal reasoning by modeling future-oriented price movements, such as price_movement_7d and price_movement_next, enabling predictive analyses. As optional enhancements, we may introduce additional node types like journalists or media sources and use LLMs to enrich event detection from full article texts. Together, these steps will ensure the knowledge graph is dynamic, insightful, and aligned with both business relevance and academic rigor.

4. 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/finance

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