Checkpoint B: Data and Schema for the Knowledge Graph

Jake Remsza

Sean Padgett

MSDS 459: Knowledge Engineering

Abstract

The purpose of this paper is to outline plans for a knowledge base concerning public companies in the healthcare sector. This outline is provided in three main sections. These are a literature review, a section entitled methods, and a results section. In the literature review the researchers detail related published works. The methods section presents the intended project approach for the researchers covering the four questions presented in the assignment posting in addition to the eight steps outlined by Chakrabarti et al. (1999). Finally, the results section details outcomes for some of the work completed thus far.

Introduction

Traditional methods of evaluating the future performance potential of stocks were mainly computational. Over time various other methods have been proposed including analyzing sentiment of publicly available information. One of the earliest and most influential papers challenging the status quo was that by Cootner (1962). He challenged the prevailing theory of the trajectory of stock prices being purely a random walk forecastable by precise calculations. Cootner (1962), instead proposed that this random walk wavered between less precise price barriers. This wavering was driven by professional traders not utilizing the same methodology, and non-professional traders assessing value with less precise measures (Cootner, 1962).

The work by Cootner (1962) laid the foundation that has since been expanded on. This began as early as the 1970's with work by Zweig (1973) to analyze closed-end funds as an indicator of investor sentiment. In this work, the researchers present a plan to expand upon this area with the intent of analyzing sentiment related to specific stocks present in various

focused blogs to forecast whether a particular stock would make a good investment. This begins with a literature review of some of the more recent efforts of other researchers in this area followed by a presentation of the planned methods.

Literature Review

Various other researchers have investigated the merits of utilizing sentiment found in online text documents for forecasting stock performance. Some of these studies such as that by Feuerriegel and Gordon (2018) have utilized company and regulatory reports. Additionally, the works presented by Schumaker et al. (2012), Wüthrich et al. (1998), and Wu et al. (2021) have utilized news articles. Finally, some such studies have also utilized social media and comment boards. Examples of this final type are works by Deng, Sinha, and Zhao (2017), Li and Li (2013), and Bollen, Mao, and Zeng (2011). Following is a review of one study from each of these three approaches.

The research by Feuerriegel and Gordon (2018) analyzed sentiment in the text of 75,927 ad hoc regulatory announcements in both English and German. Researchers in this study utilized several different machine learning techniques to analyze this text data including “lasso, ridge regression, elastic net, gradient boosting, principal component regression, and random forest” (Feuerriegel and Gordon, 2018). They then used the results of these analyses to predict the performance of the German stock indexes DAX, CDAX, and STOXX Europe 600. The results of utilizing sentiment analysis were compared to results from a linear autoregressive model utilizing historic time series price data (Feuerriegel and Gordon, 2018). Feuerriegel and Gordon (2018) found their methods utilizing sentiment analysis were able to outperform established linear models when forecasting future stock prices.

Schumaker et al. (2012) chose to analyze sentiment in news articles for their research. The researchers in this article utilized the existing stock prediction tool AZFinText and augmented the results from that system utilizing the OpinionFinder tool. The combined results were then used as instructions for a trading engine to make a trade 20 minutes after each prediction was made. The researcher's indicated that their approach netted returns from those trades between 2.4 and 3.3 percent (Schumaker et al., 2012).

In their research Bollen, Mao, and Zeng (2011), utilized the tools OpinionFinder and Google-Profile of Mood States to measure public reactions around the 2008 election. The researchers then compared the results of this analysis to changes in the value of the Dow Jones Industrial Average (DJIA). The researchers found that simply comparing the sentiment results to trends in the DJIA did not seem to reveal much correlation. However, they then trained a Fuzzy Neural Network with their results organized in a time series along with historical DJIA values. They claimed the results of the neural network showed a higher degree of accuracy compared to their initial attempts (Bollen, Mao, and Zeng 2011).

Methods

The healthcare sector is incredibly dynamic and is responsible for human progress through the development of cutting-edge technologies. This sector is also distinctive in that it is less sensitive to broader economic fluctuations compared to most industries (Correa 2024). Given these characteristics, the healthcare sector presents a unique opportunity for focused analysis of its leading companies in the market.

The aim of this study is to build a knowledge base that is used for market sentiment analysis within the healthcare sector. The intended user of this tool is likely a financial

professional such as a fund manager or financial analyst. However, it is not limited to only professionals, but anyone looking to invest, such as retail investors. Successful implementation of the knowledge base relies on prerequisite steps that include the identification of appropriate web sources, the development of multiple website specific web crawlers using the Python programming language to perform web scraping with the scrapy package, and the development of a database schema where the information will be housed and built with the graph-relational database Gel.

The knowledge domain must encapsulate the healthcare sector for effective sentiment analysis. The healthcare sector is vast, not only in the range of company sizes, but also in the diversity of its subsectors. For this reason, the top ten healthcare companies by market capitalization are the focus of the crawler, as they are the primary drivers of the market (Kavout 2025). Web sources chosen for the analysis are centered around these ten companies.

As outlined by Chakrabarti et al. (1999), the researchers for this project have been researching possible sources who's websites can be crawled for relevant information. This includes both manual inspection of sample web pages and web crawler testing. The goal is to verify data availability, assess page structure, and identify potential challenges such as dynamic content or potential pay wall issues.

The team's original intention was to incorporate text data from social media postings relating specifically to the 10 target corporations. In performing further research as to the crawlability of various social media and news websites for potential text data, multiple websites appeared to be actively restricting the activities of website crawlers. While some news websites appeared to restrict crawler activity, the social media websites were even more stringent. We

therefore refocused our efforts toward news articles. These efforts have resulted in three successful web crawlers and one API thus far. The first of these is for Wikipedia. This first source provides background information on the ten companies such as name, ticker symbol, market capitalization, location and description. Two web crawlers have also been successfully created for news articles from CNBC and Investopedia. The CNBC web crawler obtains text data from a list of news articles relating to the healthcare sector. The Investopedia web crawler works in two parts. First, it gathers links to news article search results for each of the 10 selected companies. The webcrawler then proceeds to pull text data from all of the identified news articles.

Chakrabarti et al. (1999) discusses the importance of incorporating a distiller and classifier into a full-blown focused crawler. So far, for the purposes of this project the researchers are utilizing the news site's native search function as the means of distillation. Three approaches have been utilized for this. In the first method, links from the search results are gathered manually and then incorporated as an iterative list in the web crawler instructions. The second approach automates the process to gather links for news articles. In this approach the researchers first identify the html class tag relating to links for news articles in the search results. The crawler is then programed to identify this class in the target page's html and gather the associated links which it then uses an iterative list for data extraction. The third method utilizes an API to fetch data from top new sources that match defined tags such as "healthcare".

The process of defining a schema for the database is necessary for laying a foundation for data storage once text data is obtained through the crawler activities. The type of database used for this study is a graph-relational database built using Gel. This schema combines

relational tables (objects) with graph-like relationships (links), making it ideal for modeling our healthcare sector effectively.

Once the data is properly parsed and cleaned it will be fed into the database. Nodes that will be used include company, sentiment source, and financial metrics. Properties stored within the company nodes include name, ticker, market cap, and description. This will serve as the central node, connecting to sentiment sources, financial metrics, and sectors for sentiment and financial analysis. Sentiment source nodes will have properties such as source type, content, publish date, sentiment score, and url. These nodes will connect companies to capture sentiment data, assisting with the analysis of public perceptions. Financial metric nodes will have properties such as metric type, metric value and date. These nodes will link to companies to track financial trends over time.

Edges will link the relationships between nodes that will resemble: Company -to- Sentiment Source (Multi-Link), Sentiment Source -to- Company (Multi-Link, Inverse), Company -to- Financial Metrics (Multi-Link), Financial Metrics -to- Company (Single Link, Inverse).

With this initial structure in place, the final application will serve to build up a knowledge base to expose users to the sentiment within the healthcare sector. The intended purpose of this application is for information retrieval within the knowledge base. Additionally, this market sentiment application aids in the decision-making process for trading stocks by offering the user recommendations of buy or sell points in the market. The completion of these methods outlined will ensure a complete foundation for the subsequent analysis, enabling an informed interpretation of the sector's underlying forces and market drivers.

In the near term, we plan to continue our work to build out the database. This will require efforts over multiple facets. First, we are working toward incorporating historical prices in our database. For this, the researchers are actively working on building out an API. Second, the researchers will construct a Natural Language Processing pipeline to convert results from our web crawlers into data digestible by Gel. Once that is complete, we can begin to add data into our database and experiment with the level of success we are able to obtain in asking various questions of the data via queries. This will inform any iterative database updates, additional updates that may be required, and any adjustment to our current plans for analyzing sentiment.

Results

The results demonstrate that it is viable to collect healthcare sector resources from the web using a crawler, though not without challenges. To date, crawlers for Wikipedia, CNBC, Investopedia, and NewsAPI have successfully retrieved healthcare sector data and stored it as JSON Lines files. Figure 1 illustrates a news article that was scraped from CNBC. The figure shows that the URL, source, title, tags and article content were successfully obtained. While success in scraping was obtained using Wikipedia and CNBC, Yahoo Finance was met with challenges. When the crawler was set up to scrape company news data from Yahoo Finance a 429 restricted error was observed. This error is typical of too many requests to the server, but adjustments made to the request delay did not correct the error. Further review suggested that Yahoo Finance has changed policies with how web crawlers are allowed to interact with the content. This is likely the cause for the difficulties observed with the process.

The NewsAPI approach was successful. With this approach the data obtained is from various reputable news sources writing about the healthcare sector, but the content is limited to 100 words. Given that many articles contain extraneous content not required for sentiment

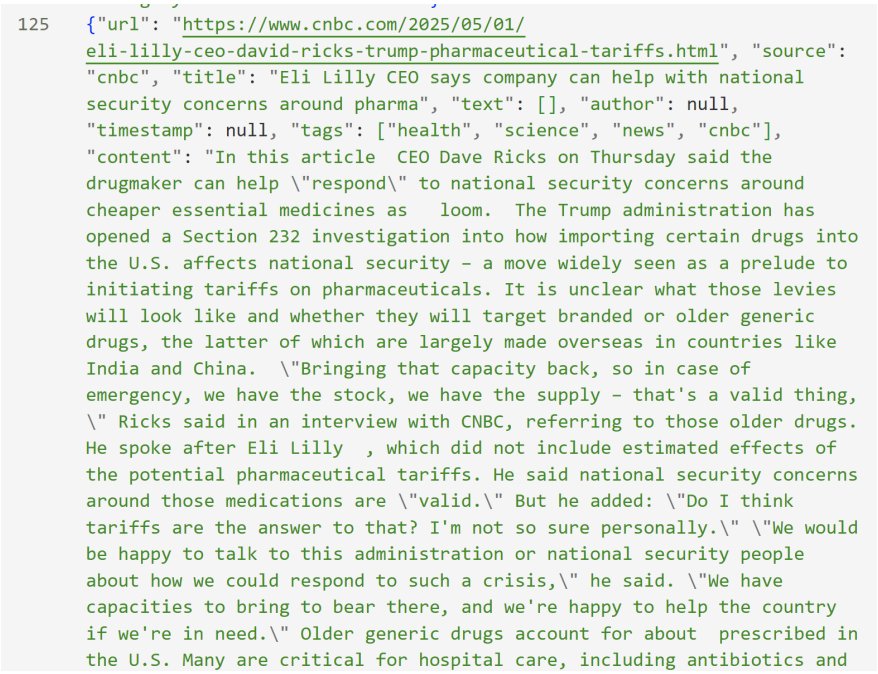
125 {"url": "https://www.cnbc.com/2025/05/01/eli-lilly-ceo-david-ricks-trump-pharmaceutical-tariffs.html", "source": "cnbc", "title": "Eli Lilly CEO says company can help with national security concerns around pharma", "text": [], "author": null, "timestamp": null, "tags": ["health", "science", "news", "cnbc"], "content": "In this article CEO Dave Ricks on Thursday said the drugmaker can help \"respond\" to national security concerns around cheaper essential medicines as loom. The Trump administration has opened a Section 232 investigation into how importing certain drugs into the U.S. affects national security – a move widely seen as a prelude to initiating tariffs on pharmaceuticals. It is unclear what those levies will look like and whether they will target branded or older generic drugs, the latter of which are largely made overseas in countries like India and China. \"Bringing that capacity back, so in case of emergency, we have the stock, we have the supply – that's a valid thing, \" Ricks said in an interview with CNBC, referring to those older drugs. He spoke after Eli Lilly , which did not include estimated effects of the potential pharmaceutical tariffs. He said national security concerns around those medications are \"valid.\" But he added: \"Do I think tariffs are the answer to that? I'm not so sure personally.\" \"We would be happy to talk to this administration or national security people about how we could respond to such a crisis,\" he said. \"We have capacities to bring to bear there, and we're happy to help the country if we're in need.\" Older generic drugs account for about prescribed in the U.S. Many are critical for hospital care, including antibiotics and

Figure 1: Sample news article obtained from CNBC and stored in JSON Lines format.

classification, we are confident that the available material is sufficient to accurately determine the sentiment. Figure 2 provides a sample of some of the news that api is fetching. Additionally, efforts made to obtain news information from Investopedia have been conducted successfully. With this appraoch we scraped 160 recent news articles to date. This collection includes articles for each of the top 10 companies. A sample of one of the articles obtained with the Investopedia crawler is shown in figure 3.

{"source": {"id": null, "name": "Yahoo Entertainment"}, "author": "Michael Erman", "title": "Bristol Myers posts better-than-expected quarterly revenue on strong cancer drug sales", "description": "(Reuters) -Bristol Myers Squibb reported better-than-expected first-quarter revenue on Thursday and raised its full-year forecast due to growth from its...", "url": "https://ca.finance.yahoo.com/news/bristol-myers-posts-better-expected-110305143.html", "urlToImage": "https://media.zenfs.com/en/reuters.com/506e70cbb31e951dd227ed295c47ce7e", "publishedAt": "2025-04-24T11:03:05Z", "content": "By Michael Erman\r\n(Reuters) -Bristol Myers Squibb reported better-than-expected first-quarter revenue on Thursday and raised its full-year forecast due to growth from its portfolio of drugs that spur… [+1980 chars]", "tags": ["healthcare"]}

Figure 2: A sample of healthcare news obtained using NewsAPI.org

2 {"url": "https://www.investopedia.com/what-to-expect-in-the-markets-this-week-april-28-2025-11721321", "title": "What To Expect in the Markets This Week", "text": "Key TakeawaysPresident Donald Trump's 100th day in office comes on Wednesday.Apple, Amazon, Microsoft, Meta Platforms, ExxonMobil, Coca-Cola and McDonald\u2019s are among the firms scheduled to release quarterly results in a packed week of corporate earnings.The Federal Reserve will get the April jobs report and key inflation data this week as Trump has reiterated his calls for the central bank to cut interest rates.Investors will also be watching out for first-quarter GDP data, the latest consumer confidence report, a trade balance update and housing market reports. Major corporate earnings, April jobs data and the latest inflation report are on tap for investors this week.

Figure 3: A sample section of a news article obtained from Investopedia.

With confirmation of viable data availability for the database, the focus now shifts to defining the schema that will dictate its data structure. Appendix A illustrates the schema definition language (SDL) code for Gel. This demonstrates the approach that will be used to build the database. Wikipedia sources will fill in the company background data such as location, founding year, products produced, and acquisitions etc. CNBC, NewsAPI and Investopedia sources provide current happenings in the market and will be the bulk of text used for sentiment analysis. This data will be housed in the SentimentSource type outlined in the schema.

Conclusions

The healthcare sector is vast, so it was decided to stick to the top 10 companies for the scope. The articles collected thus far have related either to the healthcare market as a whole or to the 10 target companies specifically. The intent of this is that questions could be asked of the database to help determine whether or not an investor should buy one of those 10 stocks. Data on the healthcare sector in general as well as data specific to those 10 companies are both potentially valuable in that regard. The future prospects of a corporation are affected by both internal and external factors outside of its control.

Appendix A: Schema Definition Language Outlined for the Proposed Gel Database

The following Schema Definition Language (SDL) code defines the structure of the Gel database designed for a knowledge base focused on the healthcare sector.

```
module healthcare {
    type Company {
        required property name -> str;
        property headquarters -> str;
        property founding_year -> int32;
        property revenue -> float64;
        property leadership -> int32;
        property stock_ticker -> str;
    }
    type FinancialMetrics {
        required property date -> cal::local_date;
        required property close -> float64;
        property open -> float64;
        link company -> Company;
      }
    type Product {
        required property name -> str;
        property type -> str;
        property approval_date -> cal::local_date;
        property revenue -> float64;
        link manufacturer -> Company;
      }
    type Acquisition {
        property date -> cal::local_date;
        property amount -> float64;
```

```
        link acquirer -> Company;
        link acquired -> Company;
    }
type SentimentSource {
        required property title -> str;
        property publication_date -> cal::local_date;
        property source -> str;
        property url -> str;
        property summary -> str;
        property content -> str;
        multi link mentions_companies -> Company;
        multi link mentions_products -> Product;
       }
}
```

References

Bollen, Johan, Huina Mao, and Xiaojun Zeng. “Twitter Mood Predicts the Stock Market.” *Journal of Computational Science* 2, no. 1 (2011): 1–8. https://doi.org/10.1016/j.jocs.2010.12.007.

Chakrabarti, Soumen, Martin van den Berg, and Byron Dom. 1999, May 17. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks: The International Journal of Computer and Telecommunications Networking,* 31(11-16): 1623–1640.

Cootner, Paul H. 1962. "Stock Prices: Random Vs. Systematic Changes." Industrial Management Review (Pre-1986) 3 (2) (Spring): 24. http://turing.library.northwestern.edu/login?url=https://www.proquest.com/scholarly-journals/stock-prices-ramdom-vs-systematic-changes/docview/214030918/se-2.

Correa, Monica L. 2024. "Health Care Is the Most Negatively Correlated Sector with Economic Cycles." *Seeking Alpha*, May 13, 2024. https://seekingalpha.com/news/4105399-health-care-is-the-most-negatively-correlated-sector-with-economic-cycles-bofa.

Deng, Shuyuan, Atish P Sinha, and Huimin Zhao. “Adapting Sentiment Lexicons to Domain-Specific Social Media Texts.” Decision Support Systems 94 (2017): 65–76. https://doi.org/10.1016/j.dss.2016.11.001.

Feuerriegel, Stefan, and Julius Gordon. “Long-Term Stock Index Forecasting Based on Text Mining of Regulatory Disclosures.” *Decision Support Systems* 112 (2018): 88–97. https://doi.org/10.1016/j.dss.2018.06.008.

Kavout. 2025. "Healthcare Stocks on the Rise: Top Picks for Smart Investors in 2025." *Sector Insights*. March 7, 2025. https://www.kavout.com/market-lens/healthcare-stocks-on-the-rise-top-picks-for-smart-investors-in-2025.

Li, Yung-Ming, and Tsung-Ying Li. “Deriving Market Intelligence from Microblogs.” Decision Support Systems 55, no. 1 (2013): 206–17. https://doi.org/10.1016/j.dss.2013.01.023.

Schumaker, Robert P, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. “Evaluating Sentiment in Financial News Articles.” Decision Support Systems 53, no. 3 (2012): 458–64. https://doi.org/10.1016/j.dss.2012.03.001.

Wu, Jheng-Long, Min-Tzu Huang, Chi-Sheng Yang, and Kai-Hsuan Liu. “Sentiment Analysis of Stock Markets Using a Novel Dimensional Valence–Arousal Approach.” Soft Computing (Berlin, Germany) 25, no. 6 (2021): 4433–50. https://doi.org/10.1007/s00500-020-05454-x.

wüthrich, B, D Permunetilleke, Steven Leung, W Lam, Vincent Cho, and J Zhang. “Daily Prediction of Major Stock Indices from Textual WWW Data.” Transactions (Hong Kong Institution of Engineers) 5, no. 3 (1998): 151–56. https://doi.org/10.1080/1023697X.1998.10667783.

Zweig, Martin E. “AN INVESTOR EXPECTATIONS STOCK PRICE PREDICTIVE MODEL USING CLOSED-END FUND PREMIUMS.” The Journal of Finance (New York) 28, no. 1 (1973): 67–78. https://doi.org/10.1111/j.1540-6261.1973.tb01346.x