With this project, I analyzed the SEC 8-K filings to extract information about new product releases by implementing a two-step process. The first part involves fetching and storing the filings, while the second part used an LLM-based approach to extract the relevant information. The reason this had to be split into two separate parts is because the SEC filings had rate limiting when hitting the endpoints and was causing the code to break and fail. After running the first part and generating a filings_data.json, I could run the second part to return relevant data to be exported to a csv file.

The initial step involves gathering S&P 500 company tickers from the Wikipedia page. This is then followed by using the pre-provided JSON file, `company_tickers.json`, to map the tickers to their respective CIK (Central Index Key) codes. These CIK codes are essential to query the SEC EDGAR database and to fetch up to 20 of the most recent 8-K filings per company. I had initially started with 10 but scaled up to retrieve more data. I also realized that the SEC EDGAR database only does it in increments of 10. Using the URL endpoints, the program requested the filings in XML format. If a filing was identified, the program extracts the filing's URL from the entry data and then parses the content using BeautifulSoup. During this step, safeguards such as timeouts and exception handling are implemented to avoid connection issues and handle errors gracefully. All extracted data was stored in a JSON file named `filing_data.json`.

The second part of the project focused on extracting specific entities from the collected filings using Ollama, a local LLM. I used Ollama 3.2 for this. A structured prompt was created to instruct the LLM to extract only relevant information about new product announcements. This prompt strictly excluded financial results, acquisitions, dividends, or executive changes. I had to tweak the prompt multiple times over the course of running the extraction. Adding in the restrictions on financial information was tricky as a large number of filings that were returned were related to stock data changing versus being a filing for a new product. The model was asked to output structured JSON with fields such as "Company Name," "Stock Name," "Filing Time," "New Product," and "Product Description." The extracted data was validated to ensure the required fields were populated with meaningful information. Invalid or incomplete responses were filtered out.

To enhance efficiency, a `ThreadPoolExecutor` was used to perform concurrent processing of the filings, enabling faster extraction. The extraction was something I was able to run in parallel, but the 8-K filings from EDGAR from before were impossible. Because of the rate limiting and only being able to hit the URL every so often, I was running into timeouts frequently. Each successful extraction was written to a CSV file named `extracted_entities.csv` using a pipe-delimited format. The program also included

error-handling mechanisms to capture and log any issues encountered during the extraction process.

This two-step methodology ensured a streamlined pipeline for fetching, storing, and analyzing SEC 8-K filings. The use of the LLM for entity extraction provided a flexible and efficient approach to identifying new product announcements. The final output in CSV format is well-structured for further analysis, providing actionable insights into product trends and corporate activities.

GitHub url: https://github.com/riyagharat/llm-document-analysis

I have left this repository as private until 12:00AM March 27th.