**Introduction**

My summary report illustrates the methodology I utilized for my LLM extraction assignment. The main goal was to engineer a python script that autonomously extracts and analyzes SEC 8-K Filings to triangulate potential investment opportunities based public sentiment of new product/software related releases, company updates, or management related announcements. The process encompassed company ticker retrieval, accessing recent 2024-2025 filings, parsing through the identified documents, and then integrating keyword focused filtering to detect relevant data/information for extraction. I downloaded the following Python libraries needed for our analysis:

A screen shot of a computer program

AI-generated content may be incorrect.

**Data Collection**

I first began by retrieving the list of publicly traded companies’ ticker data from the SEC’s public filings dataset. This public dataset included their respective company ticker symbols as well as their individual CIK numbers. These two data metrics were vital in querying the necessary SEC datasets to produce our intended query results. I then connected to the dataset directly via API request at <https://www.sec.gov/files/company_tickers.json> and then converted the data into a dictionary for more efficient lookup process for our algorithm.

**Retrieving Filings**

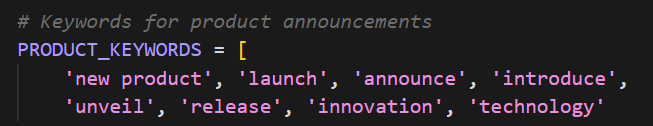
My algorithm fetched the latest 8-K filings for the companies I selected at the bottom of the script by leveraging the public SEC EDGAR API. The JSON response prompted from [https://data.sec.gov/submissions/CIK{cik}.json](https://data.sec.gov/submissions/CIK%7bcik%7d.json) contained the following data points that I queried for further data analysis: *form type, filing date, accession #, and direct document URL*. Moreover, my algorithm prioritized 8-K filings principally and limited the number of companies and number of filings to 5 to reduce efficiency issues and pulling too many API requests. For example, at first I was pulling data without limiting the date range, but experienced significant loading issues. Thus, I had to adjust the algorithm to limit the requests to a max of 10 per second and introduced a more conservative delay of 0.2 second between request pulls while filtering through the 2024-2025 latest filings first to avoid any extra noise in the datasets from 2023 and prior.

**Parsing Filings**

I accessed each retrieved document through their respective index pages collected from SEC EDGAR. The documents URLs were extracted through web scraping techniques by parsing through the company’s index page HTML content and then extracting their relevant “filing details” with BeautifulSoup and Pandas. The extracted filings details were then saved as an XML file in our project folder under the following: *“AMZN\_filings\_response.xml”, “AAPL\_filings\_response.xml”, “GOOGL\_filings\_response.xml”, “META\_filings\_response.xml”, and “MSFT\_filings\_response.XML*” file names. The filings responses included all 5 reports per each of the 5 companies mentioned above with all the data points we queried in the previous steps.

**Keyword-Based Filtering**

Next, I identified a pre-set of keywords to focus the LLM’s extraction process more efficiently on the 8-K document releases that mattered to our data analysis process. I focused the extraction on the following keywords depicted below:



Once one of the keywords are found in our query, the algorithm then pulls 300 characters of content for further analysis. Consequently, the script also extracts supplemental product names, CEO names, and any other pertinent news utilizing heuristic patterns related to quotation marks or capitalized phrases directly following one of our keywords.

**Data Output**

The algorithm output results were then saved as *"sec\_8k\_product\_filings.csv"*. The output results were structured in our indicated format with the relevant supplementary files saved in our project folder. Sample 8K data and fallback data were also included prior, directing the formatting of our intended results to ensure correct results formats prior to exporting.

**Conclusion**

Needless to say, my algorithm creates an automated process for detecting, analyzing, and transcribing new product announcements in publicly traded companies’ SEC 8-K filings while leveraging public APIS, web scraping techniques, and keyword detection queries.

**CSV Output - Screenshot:**

