**Extract New Product Information from DOW30**

This project aims to develop an automated system that retrieves the most recent 10-K filings for companies in the Dow Jones Industrial Average (DOW30) using the U.S. Securities and Exchange Commission (SEC) API. The primary objective is to extract information about new products released by these companies and save this data in a CSV file for easy analysis and further use.

The process begins by retrieving the Central Index Key (CIK) for each company using its ticker symbol. The CIK is a unique identifier for public companies that is essential for accessing filings with the SEC. The system queries the SEC’s company\_tickers.json file, which maps ticker symbols to their corresponding CIKs. Once the CIK is identified, the system moves on to the next step of retrieving the company’s recent 10-K filings.

The 10-K filing is a comprehensive annual report filed by public companies, which includes detailed information about their financial performance, operations, and risks. Using the CIK, the system queries the SEC's data endpoint to retrieve a list of recent filings for the company. The code filters these filings to identify those of type “10-K,” ensuring that the focus is on annual reports. Up to four of the most recent filings are retrieved for each company.

Once the relevant filings are identified, the next challenge is extracting the new product information from these documents. Each filing is linked to an index page that contains URLs to different files associated with the filing. The system retrieves these files, parses the text content using BeautifulSoup, and breaks the text into manageable chunks. These chunks are then processed by OpenAI’s GPT-4 model to extract specific details about new products introduced by the company. The GPT model is prompted to identify and list the products, along with brief descriptions, and to format the output in a structured manner. The output is expected to include the company name, stock ticker, filing date, new product name, and a description of the product.

Additionally, the system handles cases where no relevant filing or product data is found by skipping the company and moving on to the next one in the DOW30 list. If no matching files are found for a particular ticker, the system logs this and proceeds with the next company.

After the product data is successfully extracted, the system saves the information in a CSV file named new\_products.csv. The file is structured with columns for the company name, stock ticker, filing date, product name, and a description of the product. Each entry corresponds to a new product introduced by a company. This CSV file serves as the final output of the process, enabling easy analysis and reporting.

The system processes filings for all the companies in the DOW30 index, which includes prominent companies like 3M, American Express, Apple, Boeing, and others. For each company, the system retrieves up to four of the most recent 10-K filings, extracts new product data, and saves the results to the CSV file. The ability to automate this process significantly reduces the time and effort required to gather and analyze such information manually.

Throughout the project, several potential issues are addressed, including rate limiting, missing data, and missing files. The rate limiting issue is handled by introducing a delay when the OpenAI API reaches its token limit, allowing the system to retry the request without failure. In cases where no relevant filings or product data are found, the system prints a message and continues processing the next company. If a file that matches the ticker symbol is not found, the system logs this and skips to the next company in the list.

In conclusion, this project offers an efficient and scalable solution for extracting valuable product information from SEC filings, specifically focusing on new products introduced by companies in the DOW30 index. By combining the SEC’s API with OpenAI’s GPT-4 model, the system ensures accurate extraction and structuring of data. The final output, stored in CSV format, is easily accessible for further analysis or reporting. This project is a valuable tool for business intelligence, financial analysis, and competitive benchmarking, providing insights into the latest product offerings of some of the world’s most influential companies.