Report - Methodology for SEC 8-K Filings Extraction and Product Identification

The goal of the LLM Document Analysis student project was to extract product information from SEC 8-K filings. It requires focusing on identifying new products and services mentioned by companies. I leveraged web scraping, natural language processing (NLP), and the OpenAI API in order to accomplish this and ultimately satisfy and fulfill the requirements of the project. This process involved overcoming a number of challenges. Challenges faced involved parsing inconsistent filing formats and dealing with technical issues such as HTTP errors, including 403 Forbidden errors as these errors recognize the user as a bot and will deny access or extraction of data pertaining to the company’s information like the CIK string, Ticker, and Title of the organization.

A major challenge that was experienced often was data extraction from the SEC 8-K filings. These filings were often inconsistent in structure, with relevant product information scattered across legal and financial language typically in XML format. The variations in formatting made it very difficult to automate the process of extracting specific product details like names, products, and descriptions. I utilized Python libraries such as BeautifulSoup for XML parsing to handle this allowing me to navigate and extract text from the filings. However, the unstructured nature of the content required more advanced techniques to identify useful information even after parsing.

Another challenge was the inconsistent nature of product mentions in the filings. While some companies provided clear and detailed descriptions of new products, others only briefly mentioned new offerings without sufficient context. In addition, other companies returned a fixed and generic New Product name and description for all companies returning a static trend instead of a more dynamic and unique New Product name and description, so this needed to be fixed also. This variability made it difficult to consistently extract and categorize product-related data. In order to fix and address this properly, I designed prompts for the OpenAI API that guided the model to identify key information, such as product names and descriptions, from the extracted text. Although the model was finally able to parse complex sentences at the end that uniquely identified each company and their new product, there were times when the filings contained so much irrelevant or technical language that the model's accuracy could have been potentially compromised if not handled correctly.

Additionally, I came across challenges related to API rate limits and errors throughout the project. I dealt with issues related to OpenAI’s rate limits as I processed a large number of filings, which restricted how many requests could be made in a given time frame. I implemented error-handling mechanisms that would pause and resume the process in order to manage this. This ensured that I could continue the analysis without losing data. Another significant issue I encountered was the 403 Forbidden error when trying to access some of the SEC's filings. This error typically occurred due to permission issues or access restrictions and required adjustments to my request headers or the use of alternate access methods like adding a User Agent that mimicked the one of the browser’s URL. Furthermore, I had to ensure that all filings were accessible and that my code could handle these issues accurately; thus, resolving these technical errors was not easy.

The methodology behind this project involved several key steps. First, I scraped SEC 8-K filings using the EDGAR database and identified relevant company ticker symbols. For instance, companies’ ticker symbol is in the format shown as “320193” which is the one used for Apple Inc. I used an Atom feed to compile the associated 8-K filings once the symbols were retrieved. After parsing the filings using BeautifulSoup, I utilized the OpenAI API to analyze the text. This requires focusing on extracting product information such as the new product name, product description, and filing date. Afterwards when running the code in Jupyter Notebook, the extracted data was then structured into a CSV format for analysis in a table like format consisting of 5 columns including “Company Name”, “Stock Name”, “Filing Time”, “New Product”, and “Product Description” along with rows for each company’s information using the extracted data.

I made sure that the OpenAI API was appropriately guided by using a detailed prompt that helped the model identify and return relevant details throughout the project because it was essential for converting unstructured text into usable data. Plus, I incorporated error-handling mechanisms to deal with any inconsistent or missing data. When the model returned incomplete results, I logged these instances for review to make sure that the final dataset was as accurate as possible.

I analyzed filings from major S&P500 companies. Companies included industry leaders like Apple, Microsoft, Tesla, Amazon, Coca-Cola, and more. These companies were selected because of their impact on the market and the likelihood of having new product launches mentioned in their SEC 8-K filings. For example, filings from Apple mentioned the launch of the MacBook Pro with M2 chip, while Tesla included details about the Cybertruck. Other companies like Meta Platforms and NVIDIA also provided valuable insights into their latest product offerings that could be used for the project analysis, such as the Meta Quest Pro and GeForce RTX 4090. Moreover, the data set created from this analysis provided an overview of current product launches and highlighted the different ways in which companies disclose new products in their filings.

All in all, the project demonstrated the combination of web scraping, Natural Language Processing, and the OpenAI API usages to extract and structure important company product information from SEC 8-K filings. While the process involved overcoming several technical challenges including inconsistent filing structures, managing API rate limits, and handling errors like 403 Forbidden, they were addressed and handled appropriately through careful implementation of error-handling mechanisms. Thus, the result was a comprehensive dataset of new products launched by leading companies, which can most definitely serve as a useful resource for market analysis and business strategy.

For further information, please visit the official website for my code on the LLM Document Analysis student project: <https://github.com/MatteoMorange/LLM-Document-Analysis.git>