Jose Ayala

Professor Cuneyt Gurcan Akcora

CAP5619 – Artificial Intelligence for FinTech

March 23, 2025

LLM Document Analysis - SEC 8-K Filings Analysis for New Product Releases

Introduction

This report details the methodology and challenges encountered during the analysis of SEC 8-K filings for new product announcements. The goal of the assignment was to extract company names, stock names, filing times, new product names, and product descriptions from SEC 8-K filings. Named Entity Recognition (NER) and Large Language Models (LLMs) were utilized to automate the extraction and structure the data into a tabular format.

Data Preprocessing

The data was accessed via two key URLs:

- Company Information URL: A JSON file containing company ticker symbols and their associated Central Index Key (CIK) was fetched.
- 8-K Filings URL: The SEC EDGAR database was queried using the CIKs to retrieve 8-K filings. These filings were accessed in XML format, and I used BeautifulSoup to parse the response and extract URLs for individual 8-K filings.

The first challenge I encountered was extracting URLs for the 8-K filings. The SEC EDGAR site uses a non-standard format for 8-K document URLs, and I had to manually remove the "/ix?doc=" portion from the URLs to obtain the correct links.

Model Usage and Extraction Process

The extraction process relied on Named Entity Recognition (NER) to identify and categorize entities such as company names, stock names, product names, and dates from unstructured text. I used the llama3.2:1b model via Ollama to process the extracted HTML content. To identify new product information, I focused on specific keywords related to product announcements. Initially, I attempted to convert the 8-K HTML files to JSON, but the model struggled with this approach, so I switched to manually extracting sections of the HTML, converting them to strings, cleaning them, and processing them with the LLM.

Challenges

The biggest challenge encountered was the scarcity of new product information in the 8-K filings. Despite processing up to 10 filings for 9,720 companies over 25,016 seconds (about 7 hours), very few contained relevant product information. This resulted in a limited set of usable results.

Conclusion

This project demonstrated the application of NER and LLMs for extracting structured information from SEC 8-K filings. Despite successfully processing many filings, the lack of new product information in the filings was the major limitation. Nonetheless, the methodology employed was effective in extracting structured information when available. The structured output file can provide valuable insights for companies where new product data is included in the filings.

GitHub Repository - https://github.com/Jose-Ayala/LLM-Document-Analysis