**Analysis of SEC Form 8-K Filings Using LLM and NER**

**Introduction**

This report outlines the development of an automated pipeline using Python and Named Entity Recognition (NER) to extract structured information from 8-K filings efficiently. The primary objective was to automate the identification of company names, stock symbols, filing dates, new product mentions, and product descriptions, and to compile this data into a structured format for further analysis.

**Methodology**

The data extraction process was designed to leverage publicly available APIs from the U.S. Securities and Exchange Commission (SEC). The first step involved retrieving a list of S&P500 companies and their corresponding Central Index Key (CIK) codes from the SEC’s JSON endpoint. This allowed the system to map stock symbols to their respective company names and CIK codes. Once this mapping was established, the extraction process focused on retrieving 8-K filings from the SEC’s Atom feed using the CIK codes.

Python's requests library was employed to handle HTTP requests to the SEC API. The response data was structured in XML format, which was parsed using the ElementTree library. One challenge in working with XML data was the presence of embedded HTML tags within the filing summaries. To address this, the BeautifulSoup library was used to clean and extract plain text from the HTML content.

To identify product mentions within the filing summaries, spaCy was employed for Named Entity Recognition (NER). spaCy is a robust NLP library that can identify entities such as products, organizations, and dates. The NER model (en\_core\_web\_sm) was pre-trained on general language patterns, which allowed it to identify product mentions with reasonable accuracy. The extracted data was formatted into a structured CSV file with five key fields: company\_name, stock\_name, filing\_time, new\_product, and product\_description. The product\_description field was truncated to a maximum of 180 characters to maintain consistency and improve readability.

Error handling was a crucial aspect of the project. Parsing errors in XML data were managed using try-except blocks to prevent program crashes. Similarly, a timeout of 10 seconds was implemented for HTTP requests to avoid hanging connections and to improve overall system stability.

**Results**

The system successfully processed filings from five S&P500 companies: Apple Inc., NVIDIA Corporation, Microsoft Corporation, Amazon.com Inc., and Alphabet Inc. (Google). A total of 492 records were extracted, covering a diverse range of corporate activities. Approximately 80% of the extracted filings included meaningful product-related information. The data extraction process was completed without any major errors, and the average runtime per extraction session was approximately 30 seconds.

The output CSV file was cleanly structured, with company names and product descriptions correctly populated. Although the new\_product field was frequently left blank due to the NER model's limitations in recognizing product names, the overall extraction quality was high. The use of BeautifulSoup ensured that HTML tags were successfully removed from the summaries, enhancing the readability of the extracted text.

**Challenges and Solutions**

Several challenges were encountered during the development and execution of the pipeline. First, the SEC API imposed request limitations that resulted in blocked requests. This was resolved by setting a custom User-Agent header in the HTTP request, which aligned with best practices for accessing the SEC API.

XML parsing errors also presented difficulties due to inconsistencies in the SEC’s data formatting. This issue was addressed through improved exception handling using try-except blocks. Another challenge was the inconsistent performance of the NER model in identifying product names. This was mitigated by cleaning the input text and adjusting the NER processing parameters. Finally, the SEC’s Atom feed occasionally responded slowly, which was managed by setting a timeout to prevent prolonged connection issues.

**Conclusion**

The automated extraction of SEC 8-K filings using Python and NER was highly successful. The pipeline demonstrated stability and accuracy in identifying key information from corporate filings. The cleaned and structured data is now available for further analysis, providing valuable insights into corporate activity and product launches. The use of BeautifulSoup and spaCy ensured that the extracted data was both accurate and well-structured. Overall, the project met its objective of automating the extraction process while maintaining high data quality.

The extraction process was executed successfully, and the output met the expected quality standards. The structured CSV file is ready for further analysis, and the underlying code is stable and adaptable for future expansions. This project demonstrates the potential of combining NLP techniques with financial data analysis to generate actionable insights from large-scale unstructured data sources.