Enhancing Data Extraction from Restaurant Websites Using OpenAI API and Web Scraping Technologies

May 31, 2024

Abstract

This thesis explores the integration of advanced web scraping technologies and AI-driven text analysis to extract and analyze information from restaurant websites and review platforms like Yelp. It addresses challenges such as rate limits, dynamic content, and data summarization, offering a comprehensive system for effective data gathering, processing, and presentation.

1 Introduction

This thesis explores the application of language model-driven prompts, particularly utilizing OpenAI's LLMs, to enhance the effectiveness of data scraping. The goal is to harness the advanced capabilities of LLMs to navigate and interpret web content dynamically, overcoming traditional challenges associated with web scraping such as handling complex site structures, rate limits, and dynamically loaded content. This approach aims to not only improve the accuracy and efficiency of data extraction but also to streamline the process by integrating natural language processing directly into the scraping workflow.

2 Problem Statement

- Rate Limits and Data Fragmentation: Initial scraping efforts faced challenges due to API rate limits, necessitating the development of advanced chunking mechanisms.
- Dynamically Loaded Content: Conventional static scraping tools fail to capture content loaded dynamically via JavaScript.
- Data Accuracy and Summarization: Extracting precise data like dietary options and summarizing customer reviews from diverse formats posed significant challenges.

3 Methodology

- Structure-Aware Chunking: Developed a method to preserve HTML structure during data extraction, ensuring important data is not split across chunks.
- Selenium for Dynamic Content: Implemented Selenium to fully render pages before scraping, capturing dynamically loaded elements effectively.
- OpenAI-Powered Data Extraction and Summarization: Utilized OpenAI's API to extract and summarize data, enhancing the depth of analysis for dietary information and customer reviews.

4 Implementation

Integration of Python with libraries such as BeautifulSoup for HTML parsing and Selenium for handling JavaScript is detailed. OpenAI's GPT models were employed to enhance text analysis and summarization capabilities.

5 Conclusion

The system developed in this thesis effectively managed rate limits, captured dynamically loaded content, and enhanced data accuracy through AI-powered analysis, demonstrating its practical utility in fields such as digital marketing and customer insights. This research highlights the significant potential of integrating AI with traditional web scraping techniques to transcend common data extraction barriers. The effective combination of these technologies has not only proven successful in navigating complex web environments but also in enriching the data extraction process, thereby providing more actionable insights.

6 Future Work

Future enhancements will explore real-time data scraping and the integration of additional AI models to broaden the scope and applicability to other dynamic web domains. Additionally, considering scalable and compliant alternatives such as using Yelp or Google Reviews APIs could mitigate potential legal issues and service interruptions commonly associated with direct web scraping methods. This approach would ensure sustainability and scalability, maintaining compliance with web service terms and avoiding potential IP bans.