AI Lead Scraper

By Fathia Zulfa Alfajr

Overview

This project aims to develop a streamlined lead generation scraping tool, inspired by the reference application (e.g., Cohesive AI Scraper). With a 5-hour code sprint, the focus is on building a high-impact feature that enhances lead qualification and data extraction accuracy. My approach centers on a **Quality First** strategy by improving the extraction of actionable data (emails, social links, and concise AI-generated summaries) to support informed business decisions.

Approach & Strategic Focus

- **Feature Selection:**
  I chose to enhance the AI-generated summarization and lead qualification component. By integrating an Artificial Intelligence model, the tool not only extracts raw data but also provides context and insight for each lead. It also supports customizable output fields based on either keyword search or AI-powered Q&A, enabling users to extract exactly what they need.
- **Design Rationale:**
  Clear summaries help companies quickly judge lead relevance without manual effort. It supports efficient lead generation within a focused 5-hour build.

Data Preprocessing

- **Web Scraping:**
  The tool leverages lightweight libraries (BeautifulSoup) to extract structured data (emails, social media links, text content) from target websites.
- **Cleaning & Structuring:**
  Extracted HTML is cleaned to remove extraneous formatting and noise. Text is normalized to ensure consistency for the summarization process.
- **Lead Qualification:**
  A simple rule-based system marks leads as “Qualified” when valid email addresses are detected. It ensures the prioritization of actionable contacts.

Model Selection

- **Chosen Model:**
  For summarization and intelligent content extraction, I used **Together.ai**, an inference platform for open-source LLMs like **Mixtral-8x7B**. This model is hosted by Together.ai and is optimized for efficient API use with strong performance on instruction-following and summarization tasks.
- **Why Together.ai:**
  Due to API quota constraints with OpenAI, I opted for Together.ai, which offers free access to performant large language models. This model provides sufficiently high quality for tasks like summarization, classification, and question answering; all essential to the lead qualification pipeline.
- **Integration:**
  The selected Together.ai model is accessed via API to:
  - Generate concise, context-aware summaries of each website
  - Answer user-defined customized queries (e.g., “Is this an edtech company?”)

Performance Evaluation

- **Extraction Accuracy:**
  I validated the scraping results by manually checking several websites and confirming that the extracted emails and social media links matched what was actually on the page. This helped ensure high precision in detecting useful contact data.
- **Summary Quality:**
  To check whether the AI-generated summaries were clear and accurate, I compared them with my own quick descriptions of the websites. While not formally benchmarked, I iteratively refined the prompt and input cleaning until the summaries aligned well with the actual site content.
- **Iterative Feedback:**
  Throughout the 5-hour build, I iteratively tested various websites and refined the scraping and AI logic based on real-time results. It allows me to quickly catch and fix issues like repetition and unclear summaries.