Project Overview: Intelligent Web Data Extraction and Visualization

1. Approach

Our tool makes it easy to pull data from websites. You just type in a company name or website link, and it finds the right information for you. Then, you can choose what specific details you want to collect. Using a Large Language Model (LLM)-powered scraper (ScrapegraphAI), it grabs the data in an organized way and shows it to you on a simple web page. You can interact with the data using features like dark mode or loading more data as you go. It is designed to be user-friendly and flexible.

---

2. Model Selection

We use ScrapegraphAI, a smart tool built for pulling structured and semi-structured data from websites. It is great at understanding what you need and grabbing the right information from web pages.

Why ScrapegraphAI?

- It works well with messy or partially organized web data.
- You can ask for multiple pieces of information at once using simple language.
- It's flexible and works with different types of content without needing a lot of adjustments.

---

3. Data Preprocessing

1. **Input Handling:**
   - Accepts company name or URL.
   - Resolves missing information using search-based functions.
2. **Column Specification:**
   - Users define columns and provide contextual descriptions for targeted scraping.
3. **Data Extraction:**
   - ScrapegraphAI processes the webpage and extracts relevant content.
4. **Storage:**
   - Data is stored in a structured JSON format for ease of access and visualization.

Citation:

We used **ScrapegraphAI** for intelligent content extraction. For more details on the model's architecture and performance, refer to its official documentation.