Web Scraper & RAG Assistant - Project Report

Generated on: 2025-04-25 14:27:02

This application is a Streamlit-based tool that allows users to either scrape data from an online bookstore or ask questions based on the content of a web page using a Retrieval-Augmented Generation (RAG) method powered by a LLM (LLaMA 4 - via Groq).

Key Features:

- Web scraping of book listings from an online catalog.
- Filtering books by maximum price and number of pages.
- Displaying scraped data in a table.
- Generating and downloading CSV files of results.
- Visualizing price distribution of books.
- Using LLM to summarize book data (top 10 entries) based on filters.
- LLM-powered question answering system using web page content.

Scope:

- Designed for educational and demonstration purposes.
- Suitable for small to medium-scale data extraction tasks.
- Can be extended to other sites with similar HTML structures.
- Can support more LLM use cases such as summarization, classification, etc.

Technologies Used:

- Python
- Streamlit for frontend UI

- Requests & BeautifulSoup for scraping

- pandas for data manipulation

- matplotlib for visualizations

- fpdf for PDF report generation

- LangChain + Groq API for LLM integration

Modules Overview:

1. Web Scraper:

   - Inputs: URL, max price, number of pages

   - Outputs: DataFrame of books with title, price, availability

2. RAG Assistant:

   - Inputs: Web URL and user question

   - Outputs: LLM-generated answer based on content of page

3. Book Data Analysis:

   - Uses scraped data (CSV format)

   - Provides LLM-based analysis summary for top 10 entries

4. Session Management:

   - Uses Streamlit session_state to persist scraped data across interactions

Deployment:

- Can be deployed locally or via cloud (e.g., Streamlit Cloud, Heroku, etc.)

Future Enhancements:

- Add support for login/authentication

- Extend scraper to handle pagination automatically
- Support for multiple websites
- Save user query history
- More robust HTML parsing and error handling