AI WEB SCRAPER

Contents

Project Description

The AI Web Scraper is a streamlined tool designed to extract, clean, and analyze data from websites using cutting-edge AI techniques. It leverages Selenium for dynamic website scraping, BeautifulSoup for content extraction and cleaning, and the Ollama language model to parse and interpret data according to user-defined queries.

This tool is ideal for researchers, data analysts, or developers seeking to extract and interpret structured or unstructured web content efficiently.

Key Features

1. Dynamic Web Scraping:
    - Automates the extraction of HTML content from dynamic websites using Selenium.
2. Content Cleaning:
    - Removes extraneous elements like scripts and styles for a cleaner, text-focused result.
3. AI Parsing:
    - Processes and interprets scraped content using the OllamaLLM model for customized data extraction.
4. Streamlit Interface:
    - Provides a user-friendly web interface for inputting URLs, managing DOM content, and defining parsing tasks.

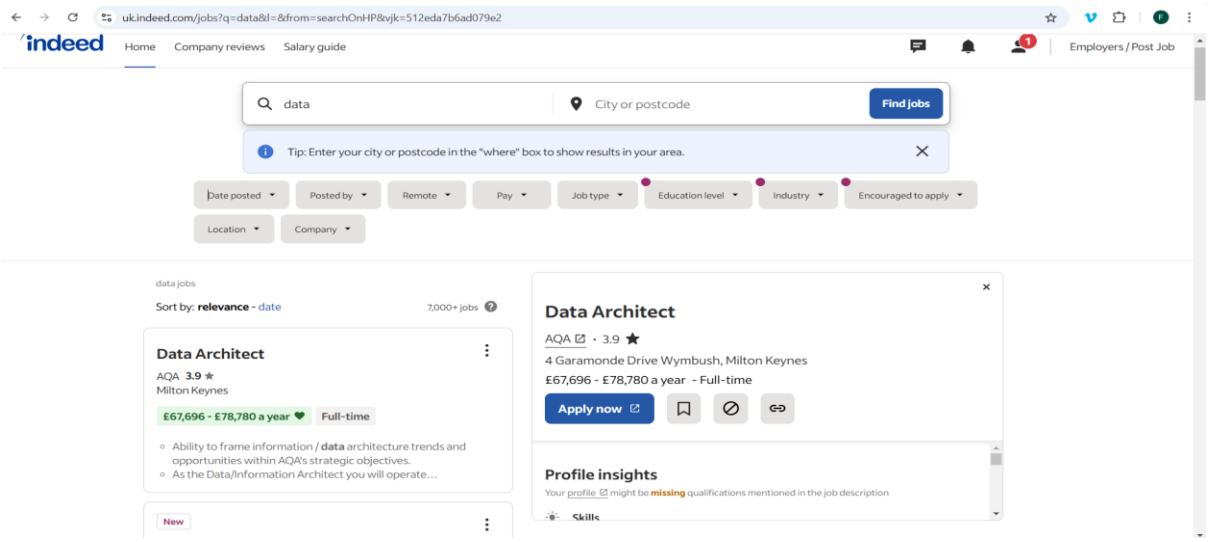
Project Workflow

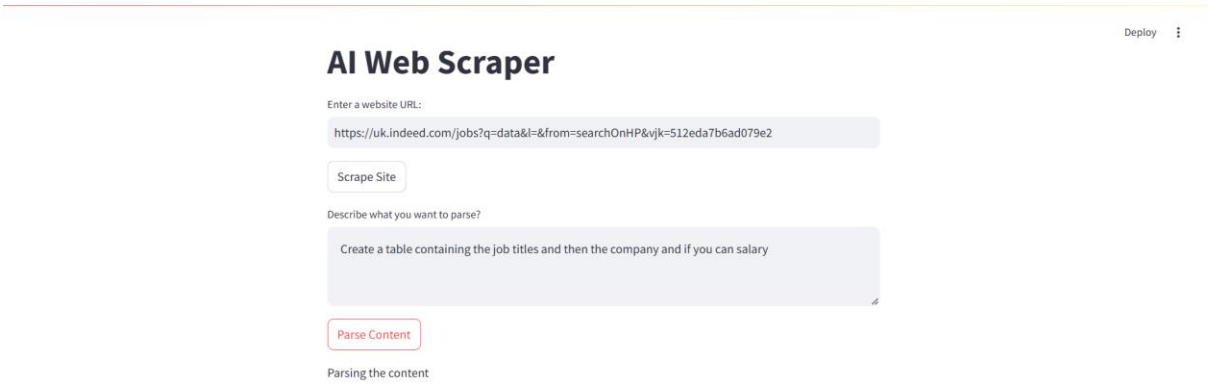
1. Input a Website URL:
    - The user enters a URL via the Streamlit interface.
2. Scrape Website:
    - The scrape_website function loads the page using Selenium and extracts its HTML source.
3. Extract and Clean Content:
    - The raw HTML is processed by BeautifulSoup to isolate the <body> content.
    - Scripts, styles, and unnecessary whitespace are removed for readability.
4. Chunk DOM Content:

- Large DOM content is split into smaller, manageable chunks for processing by the language model.

5. AI Parsing:
   - The user describes the type of information they want to extract.
   - The tool uses the OllamaLLM model to parse and extract only the relevant information, as defined by the user.
6. Display Results:
   - Parsed results are displayed directly in the Streamlit app for review and analysis.

Full example using Indeed page





Job Title	Company	Salary
Data Architect	AQA	£67,696 - £78,780 a year
AI Data Entry - Chemistry	Outlier Ai	£30 - £50 an hour
AI Data Entry - Physics	Outlier Ai	£30 - £50 an hour
Data Annotator	Heartfelt Technologies	
Junior Data Analyst	e-Careers Limited	
Data AI Specialist/Annotator for AI Models	RWS Group	
Business Intelligence & Data Manager	YeoValley	£55,000 - £65,000 a year
Admissions Data and Reporting Manager	University of Bristol	£42,632 - £47,874 a year
Junior Data Analyst	SMALL DATA ANALYTICS LIMITED	
Senior Data Cabling Engineer	Sceptre Networking Limited	£29,000 - £38,000 a year
Lead Generator/Research Assistant/Data Input	Octavian Facilities Management t/a Octavian...	From £25,000 a year

---

Code Overview

1. **main.py:**
    - **Serves as the entry point for the Streamlit application.**
    - **Manages user inputs, session states, and interactions with scraping and parsing functions.**
2. **scrape.py:**
    - **Handles website scraping and content cleaning.**
    - **Includes utility functions for splitting large content into smaller chunks.**
3. **parse.py:**
    - **Defines the AI-driven parsing logic using the OllamaLLM model.**
    - **Processes DOM chunks based on user-defined descriptions.**
4. **Dependencies:**
    - **Streamlit: Web app framework.**
    - **Selenium: For browser automation and dynamic web scraping.**
    - **BeautifulSoup: HTML parsing and content cleaning.**
    - **LangChain Ollama: AI-powered content parsing.**