Pulse - Module Extraction AI Agent

It is an AI-powered documentation analysis tool. It automatically crawls technical documentation websites, filters the content, and uses Large Language Models (LLMs) to extract a structured hierarchy of **Modules** and **Submodules** with detailed descriptions.

Features

AI-Powered Extraction: Uses LLMs (Groq Llama 3 or OpenAI GPT) to understand context and generate descriptions.

Intelligent Crawling: Recursively crawls documentation pages while handling filtering of non-content elements.

Structured Output: Generates clean, nested JSON output suitable for integration into other systems.

Streamlit UI: User-friendly web interface for inputting URLs and visualizing results.

Bypass Protections: Includes browser-mimicking headers to scrape sites protected against basic bots.

Tech Stack

Language: Python 3.x

Interface: Streamlit

Scraping: Requests, BeautifulSoup4

AI Integration: OpenAI Client (Compatible with Groq)

Data Processing: JSON, urllib3

Project Structure

📁 pulse-module-extractor

```
├── main.py          # Entry point for the Streamlit Application
├── crawler.py       # Logic to crawl websites and handle recursion
├── extractor.py     # Orchestrates the data flow from crawler to AI
├── utils.py         # Handles API communication with Groq/OpenAI
├── parser.py        # Cleans HTML to remove navigation/footers
├── requirements.txt  # List of dependencies
└── README.md        # Project documentation
```

⚙️ **Installation**

1. **Clone the repository** (or download files):

Bash

```
git clone <your-repo-url>
cd pulse-module-extractor
```

2. **Install Dependencies:**

Bash

```
pip install -r requirements.txt
```

🏃 **Usage**

1. **Run the Application:**

Bash

```
streamlit run main.py
```

2. **Get a Free API Key:**
    - Go to Groq Console.
    - Create a free API Key.
3. **In the App:**
    - Paste your **Groq API Key** in the sidebar.
    - Enter a documentation URL (e.g., https://requests.readthedocs.io/en/latest/).

- Click **Start Extraction**.

🧪 Testing

The tool has been tested on the following documentation structures:

- **Requests Docs:** https://requests.readthedocs.io/en/latest/ (Stable)
- **WordPress Docs:** https://wordpress.org/documentation/
- **Chargebee:** https://www.chargebee.com/docs/2.0/

🤝 Troubleshooting

- **Error 401 (Invalid API Key):** Ensure you are using a valid Groq key (starts with gsk_) and that the base_url in utils.py is set to https://api.groq.com/openai/v1.
- **Error 400 (Model Decommissioned):** Check utils.py and update the model parameter to a current version like llama-3.3-70b-versatile.
- **No Content Found:** The website might be blocking the crawler. Try a different, more developer-friendly documentation URL