README.md

# 📚 API Doc Scraper 🕷️

## Description

API Doc Scraper is a powerful document scraper that helps you gather API documentation effortlessly! 🚀 Provide it with the URL of an API's documentation, and it will crawl through the pages, creating a comprehensive text file ready for use with your favourite LLM. 🤖

## Features

- 🌐 Asynchronous web scraping for fast performance

- 🧠 Intelligent URL filtering using GPT-4

- 🔄 Automatic retry mechanism for resilient scraping

- 📁 Organized output in a single text file

- 🛡️ Respects robots.txt through Jina API integration

## Prerequisites

- Python 3.7+

- OpenAI API key

- Jina API key

## Installation

1. Clone the repository:

```

git clone https://github.com/ali-abassi/API-Documentation-Scraper-for-LLMs.git

cd API-Documentation-Scraper-for-LLMs

```

2. Install the required packages:

```

pip install -r requirements.txt

```

3. Create a `.env` file in the project root and add your API keys:

```

OPENAI\_API\_KEY=your\_openai\_api\_key\_here

JINA\_API\_KEY=your\_jina\_api\_key\_here

```

## Usage

1. Run the script:

```

python API-Doc-Scraper.py

```

2. Enter the URL of the API documentation when prompted like example: https://platform.openai.com/docs/overview

3. Sit back and watch as the crawler does its magic! 🪄

4. Find your scraped documentation in the generated `{domain}\_docs.txt` file.

5. Add that to your conversations with an LLM to help you write code using the api.

## Output

The script will create a text file named after the domain of the provided URL (e.g., `example\_docs.txt`). This file will contain the scraped content from all relevant pages, neatly organized into sections.

## Contributing

Contributions are welcome! Please feel free to submit a Pull Request. 🤝

## License

This project is licensed under the MIT License. See the `LICENSE` file for details.

## Disclaimer

Please use this tool responsibly and in accordance with the terms of service of the websites you're scraping. 🙏

Happy scraping! 🎉