Document

Script Overview

The script is designed to scrape event data from a target website using an HTTP request that closely mimics a real browser session. It integrates the use of proxies, headers, and cookies to reduce the risk of detection and blocking. The core workflow ensures robustness by checking for expected HTML structures and falling back to raw parsing if needed.

Detailed Workflow

1. Import Libraries

Loads all necessary libraries such as requests, lxml, and json.

2. Initialize Output

Prepares an empty list or structure to store the scraped event data.

3**. Configure Request**

Proxies: Use proxies to anonymize the request.

Headers: Sets up headers to simulate a browser request (e.g., user-agent).

Cookies: Includes session or tracking cookies to bypass bot detection.

4. HTML Content Validation

Checks if the response contains the expected HTML elements:

If found: Proceeds to parse and extract data from this structured content.

If not found: Falls back to parsing the raw HTML, possibly rendered by JavaScript.

5. Data Extraction

Uses XPath expressions to locate and extract specific data points such as:

a. Event name

b. Date/time

c. Location

d. Image_url

6.Data Transformation

Formats the extracted data into a JSON structure, ready for storage or further processing.