Web Crawler

Description

This Python-based web crawler scrapes data from a given URL and its sub-links up to a depth of 5 levels. It extracts titles and first paragraphs from each page and prints sub-links with depth levels to show the hierarchy of the crawled data.

Features

* Scrapes titles and first paragraphs from web pages.
* Recursively crawls sub-links up to 5 levels deep.
* Prints depth levels for each title and sub-link to show data hierarchy.
* Avoids revisiting the same URLs to prevent infinite loops and redundant requests.
* Mimics a real browser request to avoid being blocked by websites.

Prerequisites

* Python 3.x must be installed on your system.
* Install the required libraries using pip: pip install requests beautifulsoup4

Notes

* Adjust Max Depth: You can adjust the `max\_depth` variable to set the maximum depth for crawling.
* Handling Rate Limiting: The script includes a brief pause (`time.sleep(1)`) to avoid overloading the server. Adjust the duration as needed.