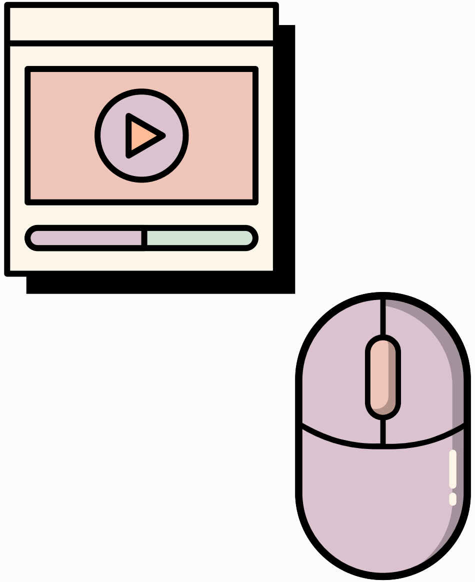
#WEB CRAWLER

with Multithreading & Multiprocessing



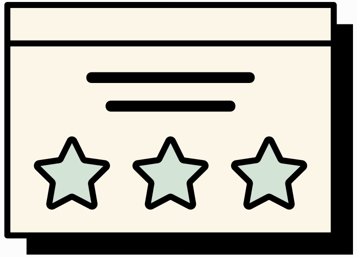
What Is Web Crawler?

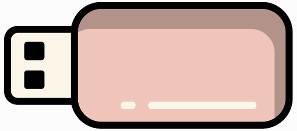
A web crawler, often referred to as a spider or bot, is an automated program designed to **browse the internet and systematically navigate through websites to collect information.** Web crawlers are an essential component of web search engines, as they help index and catalog the vast amount of content available on the internet.

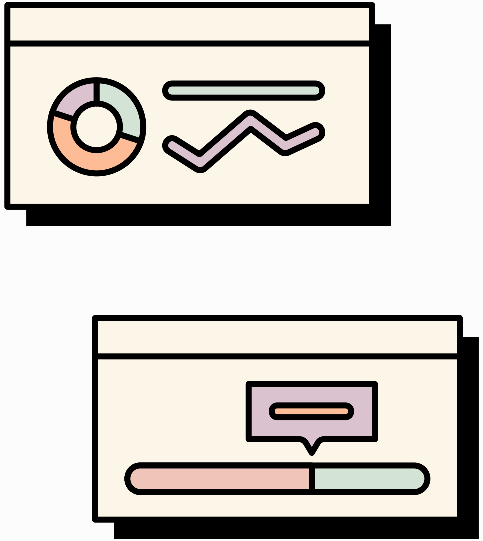
It is a Python-based application that allows you to crawl and analyze web pages, extracting information such as **page titles, meta descriptions, images, and links.**

Features

- - Crawl multiple websites simultaneously using multithreading.
- - Extract and display page titles, meta descriptions, images, and links.
- - Utilize multiprocessing for efficient data extraction and analysis.
- - Interactive web interface for easy usage.







Multithread

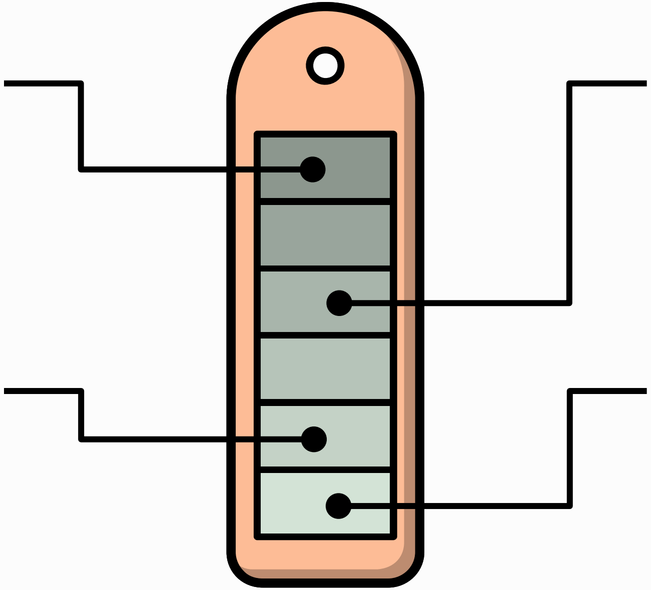
Multithreading is used in this web crawler to enable concurrent execution of multiple crawling tasks.

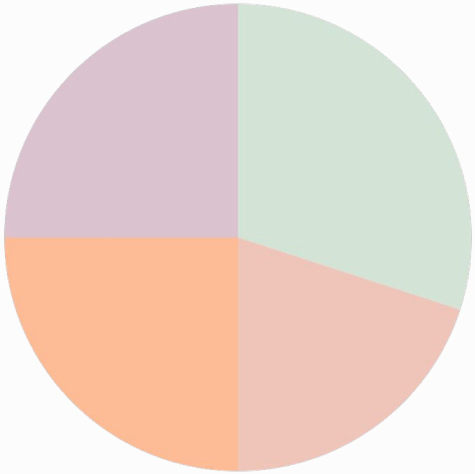
Each website to be crawled is assigned to a separate thread, allowing multiple websites to be processed simultaneously. This improves the overall crawling speed and efficiency.

MultiProcess

Multiprocessing is employed for efficient data extraction and analysis.

While crawling, the application utilizes multiple processes to extract information from web pages in parallel. This further enhances the performance and allows for faster data retrieval.





GitHub

- The code of our project can be found on our public GitHub Repository: [GitHub]
- The code has been deployed successfully on render: [Web-Crawler]

Code

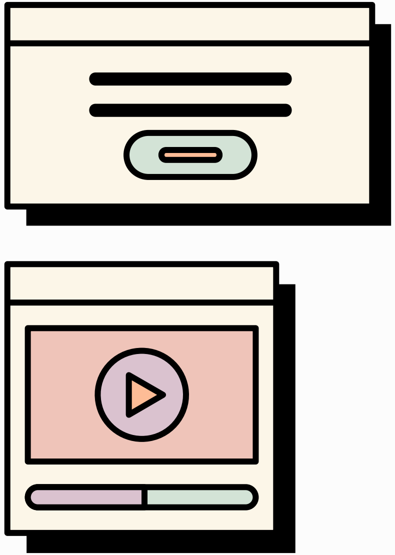
crawl_website(url, max_depth): Initiates crawling for a given URL, up to a specified depth.

extract_info(soup): Parses HTML content, extracts page data (title, meta description, images, links).

get_images(soup): Extracts image URLs from HTML content.

get_links(soup): Extracts hyperlinks from HTML content.

process_urls(urls): Concurrently crawls and extracts data from multiple URLs.



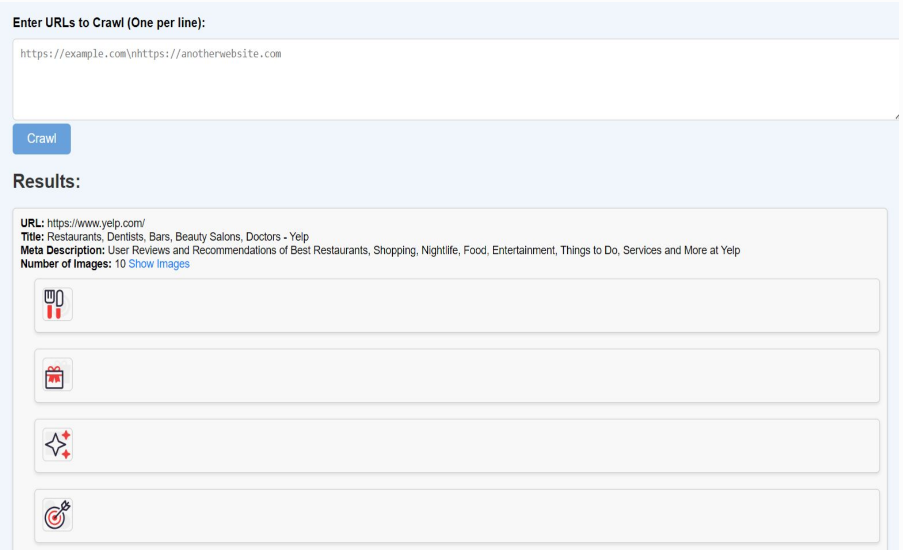
Web Crawler

Enter URLs to Crawl (One per line):

https://example.com\nhttps://anotherwebsite.com

Results:

ScreenShot



Enter URLs to Crawl (One per line):

```
https://example.com\nhttps://anotherwebsite.com
```

Results:

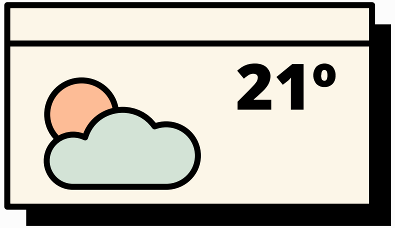
URL: https://www.yelp.com/
Title: Restaurants, Dentists, Bars, Beauty Salons, Doctors - Yelp
Meta Description: User Reviews and Recommendations of Best Restaurants, Shopping, Nightlife, Food, Entertainment, Things to Do, Services and More at Yelp
Number of Images: 10 Show Images

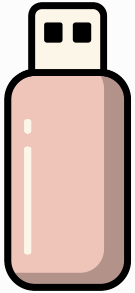
Team

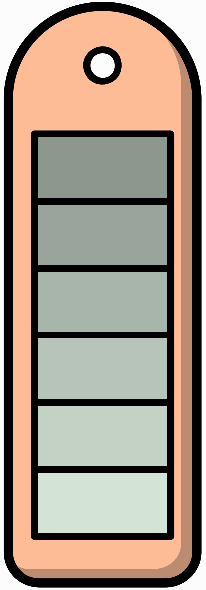
Akanksha Rathore - RA2211003010396

Manya Mehrotra - RA2211003010399

Tuhina Tripathi - RA2211003010423







Thanks!