FIFTH Program

__/__/____

Web scrapping is also known as web crawling. It is the process of extracting data from websites. Web crawlers (spiders or bots) automatically navigate the web by following links from one page to another.

The main goal is to discover new pages and revisit existing ones to see if they have been updated. Crawlers download the content of web pages for further processing.

Indexing is the process of processing and storing the data collected by web crawlers into a structure format: it is easy and fast to retrieve the data. ex: sports, movies, politics.

Sitemap is a file written in XML. It list all the URLs of a website. Site owner submits sitemap data. If you want to design a website. your website is able to shown in the search engine you should give crawling and indexing.

Send HTTP Requests:- Using program language like python requests library to send request to web servers and retrieve the HTML content of web pages.

Parse HTML:- Use parsing libraries like BeautifulSoup extract specific data from HTML (Heading, paragraph, links, tables, images).

Extract Data:- with the parsed HTML. you can locate and extract the relevant data using HTML Tags.

__/__/____

Cleaning and Structuring:- After extracting you need to clean & structure it.

Finally Store Data:- Save the extracted data in a structured format like CSV, JSON (or) database for further analysis.

Robots.txt file:- It tells search engine which pages to access and which are not access.

User-agent : * All

Command disallow → Folder/Pages.

User-agent : /directory//

The 'robots.txt' file must be placed in the root directory of the website.

robots.txt file consists of one or more blocks of directives.

each block starts with user-agent line to specify which crawlers the directives apply follow by one or more Disallow (or) Allow line

ex:- User-agent *

Disallow : /private/

Disallow : /tmp/

Allow : /public/

User-agent : Googlebot

Disallow : /no-google.

Block all webcrawlers from accessing the private directory.

User-agent : *

Disallow : /private/

__/__/____

Allow all crawlers to access the public directory
Allow: /public/
prevent the Google web crawler from accessing the /no-google directory.
User-agent: Googlebot
Disallow: /no-google/
Provide the location of sitemap
Sitemap: https://www.sample.com/sitemap.xml.

```
Import requests
response = requests.get("http://en.wikipedia.org/robots.txt")
test = response.text
print("robots.txt for http://www.wikipedia.org//")
print(test)
```

(or)

```
Import requests
# make a get request to the robots.txt file.
response = request.get("_____")
# check if the request was successfull
if response.status_code==200:
# Store the content of the robots.txt file
    test = response.text
# print the contents
    print(test)
else:
    print('Failed to retrieve the robots.txt file')
```