JUNAID GIRKAR | 60004190057 | BE COMPS A2 | WEB INTELLIGENCE

EXPERIMENT - 2

AIM: Design a crawler to gather web information

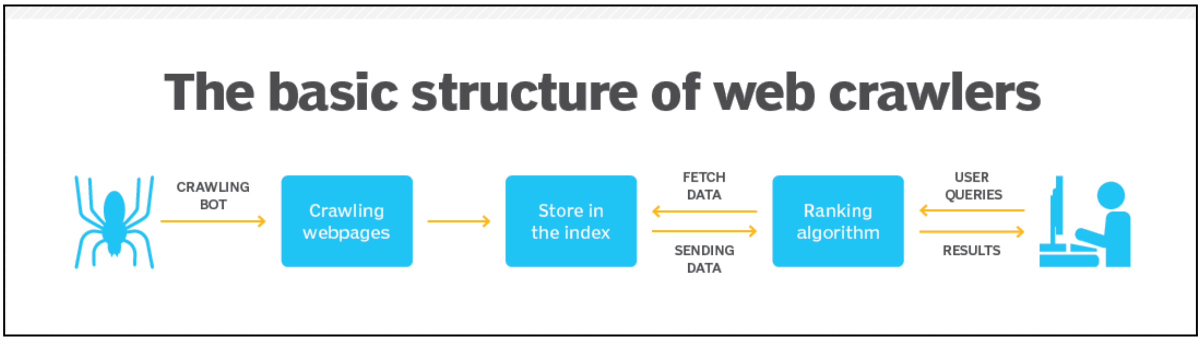
THEORY:

A Web crawler, sometimes called a spider or spiderbot and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web and that is typically operated by search engines for the purpose of Web indexing (web spidering). Web crawlers copy pages for processing by a search engine, which indexes the downloaded pages so that users can search more efficiently.

The number of Internet pages is extremely large; even the largest crawlers fall short of making a complete index. For this reason, search engines struggled to give relevant search results in the early years of the World Wide Web, before 2000. Today, relevant results are given almost instantly.

Web scraping is data scraping used for extracting data from websites. The web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a page). Therefore, web crawling is a main component of web scraping to fetch pages for later processing. Once fetched, extraction can take place. The content of a page may be parsed, searched, and reformatted, and its data copied into a spreadsheet or loaded into a database.



Reference: https://www.techtarget.com/whatis/definition/crawler

CODE & OUTPUT:

```
import requests
from bs4 import BeautifulSoup

URL = "https://en.wikipedia.org/wiki/Dwarkadas_J._Sanghvi_College_of_Engineering"
page = requests.get(URL)

soup = BeautifulSoup(page.content, "html.parser")
```

```
# Find element by ID
results = soup.find(id="content")
print(results.prettify())
```

```
<main class="mw-body" id="content" role="main">\n <header class="mw-body-header
vector-page-titlebar">\n  <label aria-controls="toc-toggle-list" class="mw-ui-button
mw-ui-quiet mw-ui-icon mw-ui-icon-flush-left mw-ui-icon-element
mw-ui-icon-wikimedia-listBullet mw-checkbox-hack-button"
data-event-name="vector.toc-toggle-list" for="vector-toc-collapsed-checkbox"
id="vector-toc-collapsed-button" role="button" tabindex="0" title="Table of Contents">\n
Toggle the table of contents\n  </label>\n...
```

```
# Find elements by tag and class name
page_titles = results.find_all("span", class_="mw-headline")
for title in page_titles:
    print(title)
```

```
<span class="mw-headline" id="History">History</span>
<span class="mw-headline" id="Departments">Departments</span>
<span class="mw-headline" id="Library">Library</span>
<span class="mw-headline" id="Activities">Activities</span>
<span class="mw-headline" id="Festivals">Festivals</span>
<span class="mw-headline" id="See_also">See also</span>
<span class="mw-headline" id="References">References</span>
<span class="mw-headline" id="External_links">External links</span>
```

```
# Extract text from HTML elements
for title in page_titles:
    print(title.text.strip())
```

```
History
Departments
Library
Activities
Festivals
See also
References
External links
```

```
URL2 = "https://quotes.toscrape.com/"
page2 = requests.get(URL2)
soup2 = BeautifulSoup(page2.content, "html.parser")

# Find element by Class
results2 = soup2.find_all(class_="quote")
print(results2)
```

```
<span class="text" itemprop="text">“Try not to become a man of success. Rather become a man of value.”</span>
<span>by <small class="author" itemprop="author">Albert Einstein</small>
<a href="/author/Albert-Einstein">(about)</a>
</span>
<div class="tags">
            Tags:
            <meta class="keywords" content="adulthood,success,value" itemprop="keywords"/>
<a class="tag" href="/tag/adulthood/page/1/">adulthood</a>
```

```
<a class="tag" href="/tag/success/…
```

```
for quote in results2:
  quote = quote.find(class_="text")
  print(quote.text.strip())
```

```
“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”
“It is our choices, Harry, that show what we truly are, far more than our abilities.”
“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”
“The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”
“Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring.”
“Try not to become a man of success. Rather become a man of value.”
“It is better to be hated for what you are than to be loved for what you are not.”
“I have not failed. I've just found 10,000 ways that won't work.”
“A woman is like a tea bag; you never know how strong it is until it's in hot water.”
“A day without sunshine is like, you know, night.”
```

CONCLUSION: Web crawling is the functionality that is responsible for connecting all the websites over the internet. The core feature of search engines like Google, Bing, and DuckDuckGo is to index all the websites using web crawlers so that they can quickly generate and send results to the user when asked. In this experiment, we implemented web scraping, an important attribute of web crawling using the BeautifulSoup library available in python.