**WEB SCRAPING**

## **Yogendra Verma (1814)**

**Solutions for we scraping Questions ---**

1. Which of the following extracts information from user generated content?

A) Java script tagging B) Web scraping

C) A/B testing D) MROCs

Ans. (B)

1. Which of the following is not a web scraping library in python?

A) selenium B) Beautiful soup

C) Requests C) scrapy

Ans. (C)

1. Selenium tests \_\_\_\_\_\_\_\_\_\_?

A) Browser based applications B) DOS applications

C) GUI applications D) All of the above

Ans. (A)

1. Task of crawling is performed by a complex software which is known as:

A) Scraper B) Crawler

C) Boat D) Spider

Ans. (B) (I think Boat and Spider also the right option)

1. Which of the following commands is used to access name of a tag in Beautiful Soup?

A) tag.attrs B) tag.name

C) tag,id C) tag[‘id’]

Ans. (B)

1. Which of the following is the default parser in Beautiful Soup?

A) html.parser B) html5lib

C) lxml D) lxml-xml

Ans. (C)

1. In selenium the webdriver is used to?

A) design a test using selenese

B) test a web application on firefox only

C) execute tests on HtmlUnit browser

D) to download any content from a webpage

Ans. (C)

1. In selenium, driver**.find\_elements\_by\_xpath(‘given xpath’)** returns:

A) the first webelement associated with the ‘given xpath’

B) the url of first webelement associated with the ‘given xpath’

C) the list of all webelements associated with the ‘given xpath’

D) all the attributes of the first webelement associated with the ‘given xpath’

Ans. (C)

1. The script **‘window.scrollBy(0,a)** scrolls the webpage by?

A) **‘a’** number of horizontal spaces

B) **‘a’** number of lines

C) **‘a’** number of pixels horizontally

D) **‘a’** number of pixels vertically

Ans. (D) (scrolled down by a pixels)

**In Q10, more than one options are correct, Choose all the correct options:**

1. Which of the following is(are) tags of HTML?

A) <a> B) <b>

C) <image> D) <href>

Ans. (A), (B)

**Q11. What is the main difference between a web scraper and a web crawler?**

**Ans.**

**Web Crawler: -**

* A web crawler sometimes called a **“spider”** is a standalone bot that systematically scans the Internet for indexing and searching for content, following internal links on web pages.
* In general, the term **“crawler”** means the ability of a program to navigate web pages on its own, possibly even without a clearly defined end goal or goal, endlessly exploring what a site or network can offer.
* Web crawlers are actively used by search engines such as Google, Bing and others to extract content for a URL, check this page for other links, and get URLs for these links and so on.
* Web crawling creates a copy of what’s there
* Data crawling involves certain degree of scraping, like saving all the keywords, the images and the URLs of the web page.

**Web Scraper:-**

* Web Scraper is a process of extracting specific data.
* A web scraper searches for specific information on specific websites or pages.
* Web scraping extracts specific data for analysis, or to create something new. However, in order to conduct web scraping you would first have to do some sort of web crawling to find the information you need.
* Web scraping is essentially targeted at specific websites for specific data, e.g. for stock market data, business leads, supplier product scraping.

**Q12. What is ‘robots.txt’ file? What is the use of ‘robots.txt’ file?**

**Ans.**

**Robots.txt**:-

* A robots.txt file tells search engines where they can and can’t go on your site. Primarily, it lists all the content you want to lock away from search engines like Google. You can also tell some search engines (not Google) *how* they can crawl allowed content.
* Most search engines are obedient. They aren’t in the habit of breaking an entry. That said, some aren’t shy about picking a few metaphorical locks. Google isn’t one of those search engines. They obey the instructions in a robots.txt file.
* Each search engine identifies itself with a different user-agent. You can set custom instructions for each of these in your robots.txt file. There are [hundreds of user-agents](http://www.user-agents.org/), but here are some useful ones for SEO:
* Google: Googlebot
* Google Images: Googlebot-Image
* Bing: Bingbot
* Yahoo: Slurp
* Baidu: Baiduspider
* DuckDuckGo: DuckDuckBot

Here’s the basic format of a robots.txt file:

* Sitemap: [URL location of sitemap]
* User-agent: [bot identifier]
* [directive 1]
* [directive 2]
* [directive ...]
* User-agent: [another bot identifier]
* [directive 1]
* [directive 2]
* [directive ...]

**Q13. What are static and dynamic web pages?**

**Ans.**

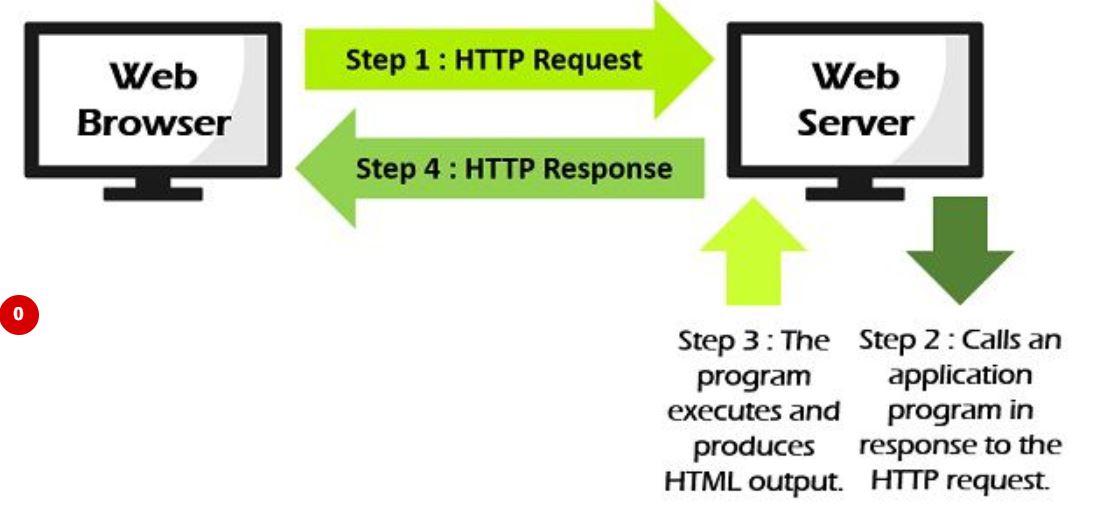
**Static web pages**: - are simple and written in the HTML language and stored in web server. Whenever server receives a request regarding a web page, it sends a response along with the requested web page to the client without performing any additional processing. It just locates that page on its hard disk and add HTTP headers, and reply back an HTTP response.

The peculiar thing in a static web page is that the content in these types of the web page does not change depending on the request. They are always the same unless the content is changed physically on the server’s hard disk. That is the reason these web pages are known as static web pages.



**Dynamic web pages: -**

* It Provide a solution for the static web pages. The dynamic web page content can vary depending on the number of parameters. It not just simply send HTML page in response. The web server calls a program located on the hard disk which can access a database, perform transaction procedure, etcetera.
* If the application program produces HTML output, which is used to construct an HTTP response by the web server. The web server sends the HTTP response thus created, back to the web browser.
* The dynamic web pages are employed where the information changes very often such as stock prices, weather information, news and sports updates.
* Let’s assume a person has to physically change the Web page every 10 seconds to show the latest update of the stock prices which is impractical to physically alter the HTML pages very often, so in this case, a dynamic web page can be used.

****

**Q14. Write a python program to check whether a webpage contains a title or not.**

**Ans.**

import selenium

from selenium import webdriver as wb

webD=wb.Chrome('chromedriver.exe')

driver = wb.Chrome()

url = "https://www.fliprobo.com/"

driver.get(url)

# Getting current URL Title

get\_title = driver.title

# Printing the title of this URL

if get\_title =={}:

print("NO Title")

else:

print(get\_title)

**Q15. Write a python program to access the search bar and search button on images.google.com.**

**Ans.**

import selenium

from selenium import webdriver as wb

webD=wb.Chrome('chromedriver.exe')

driver = wb.Chrome()

driver.get("https://images.google.com/")

que=driver.find\_element\_by\_xpath("//input[@name='q']")

que.send\_keys("Data Science")

que.send\_keys(Keys.RETURN)