**WEB SCRAPING**

## **WORKSHEET – 1 ANSWERS**

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

B) Web scraping

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

A) selenium

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

A) Browser based applications

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

B) Crawler

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

B) tag.name

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

C) lxml

1. In selenium the webdriver is used to?

C) execute tests on HtmlUnit browser

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

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

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

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

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

A) <a> B) <b>

D) <href>

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

Web crawling is perform by search engines. It is all about viewing a page as a whole and indexing it. When a bot crawls a website, it goes through every page and every link, until the last line of the website, looking for each information. Search engines use web crawling for the purpose of SEO, SMO. It crawls all data that are present over a web. The crawler also known as spider or bot.

Web scrapping involves specific data extraction on a targeted webpage, for instance, extract data about sales leads, real estate listing and product pricing. It is more focused and intentional and usually is directed toward a website to scrap. It’s the programmatic analysis of a web page to download information from it. By using various scrapping tools we can scrap images, videos, text etc, from any site.

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

Answer: A robots.txt is a file with set of instructions for bots or crawlers. This file is included in the source file of websites. It is a text file that tells which parts of website can be crawled and which part cannot be, it also includes a link to sitemap (can consider as directory of website) of the website.

By using a robots.txt individual files in a directories, complete directories, sub directories can be excluded from crawling. The robots.txt data is stored in the root of the domain. It is the first document that is accessed by a bot when it visits a website. The bots of the biggest search engines such as Google and Bing follow the instructions. Otherwise there is no guarantee that a bot will adhere to the robots.txt requirements. In short, Robots.txt helps to control the crawling of search engine bots. Robots.txt is very helpful for SEO purpose.

13. What are static and dynamic web pages?

Static website:

Static website is using pure HTML, CSS, JS without relying on the content, or parts of the content, being pulled from various databases. This means each page is constructed on its own, coded, and loaded individually, no matter who or how the visitor arrives on the page. In Static website content will not change until someone change it manually. Static website is very easy to manage and have the low maintenance cost.

Example

In a Static website we would code each element of this page individually. We repeat the process for as many pages that you have on your website. If we make a change in the navigation key (e.g. add a new button), or Recent Posts (add a new blog), or want to change our telephone number, we have to make that change on every single page. So if you have 100 pages, that is 100 changes!

Website example: User portfolio, some institute’s website.

Dynamic website:

A dynamic website pulls different content into a webpage from the repository of database within the website for different user.  A dynamic website is more functional. It allows users to interact with the information that is listed on the page.

This enables the website to present varying content in a page based on the visitor’s choices, as well as past visit history using cookies, and other variables. It is the combination of server-side and client-side scripting. Client-side scripting refers to code that is executed by the browser, usually with JavaScript. Meanwhile, server-side scripting refers to code that is executed by the server such as PHP, Python, Java etc.

Example of dynamic website: Facebook, Live stock website, YouTube etc.