WEB SCRAPING

WORKSHEET – 1

In Q1 to Q9, only one option is correct, Choose the correct option:

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

*B) Web scraping*

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

*C) Requests*

**3. Selenium tests \_\_\_\_\_\_\_\_\_\_?**

*A) Browser based applications*

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

*B) Crawler ; C) Boat ; D) Spider*

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

*B) tag.name*

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

*A) html.parser*

**7. In selenium the webdriver is used to?**

*D) to download any content from a webpage*

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

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

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

*D) ‘a’ number of pixels vertically*

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

*A) <a> ; B) <b>*

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

* *web crawlers usually focus on indexing the web while web scrapers extract or scrape data from webpages.*
* *Web Crawling basically visits a website and reads web pages for the purpose of building entries for search engine index. Web Scraping is a technique used to extract a large amount of data from websites and then saving it to the local machine in the form of XML, excel or SQL.*
* *The tools that are used for web crawling are known as web crawlers or spiders. The tools used for web scraping are known as web scrapers.*

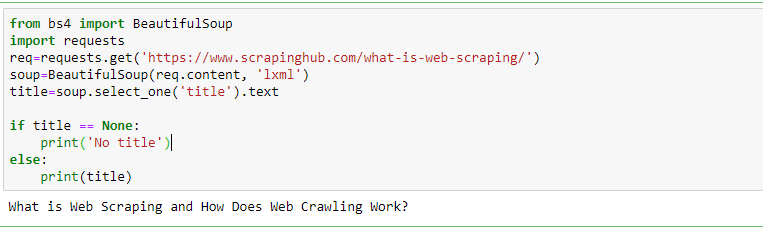
**12. What is ‘robots.txt’ file? What is the use of ‘robots.txt’ file?**

* *A robots.txt file tells search engine crawlers which pages or files the crawler can or can't request from your site.*
* *When a crawler arrives at a website, it first reads the robots.txt file to determine which parts of the website it should crawl and which parts it should ignore, according to the so-called Robots Exclusion Standard Protocol.*
* *robots.txt is used mainly to avoid overloading a site with requests.*

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

* *"Static" means unchanged or constant, while"dynamic" means changing or lively. Therefore, static* [*Web pages*](https://pc.net/glossary/definition/webpage) *contain the same prebuilt content each time the page is loaded, while the content of dynamic Web pages can be generated on-the-fly.*
* *Static Web pages are very simple. It is written in languages such as HTML, JavaScript, CSS, etc.Dynamic Web Pages are written in languages such as CGI, AJAX, ASP, ASP.NET, etc.*
* *For static web pages when a server receives a request for a web page, then the server sends the response to the client without doing any additional process. Pages will remain the same until someone changes it manually. In dynamic web pages, the Content of pages is different for different visitors. These pages contain "server-side" code, which allows the server to generate unique content each time the page is loaded.*

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

****

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

****