

ASSINGMENT

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) Scrapy
3. Selenium tests _____?
A) Browser based applications
4. Task of crawling is performed by a complex software which is known as:
D) Scraper
5. Which of the following commands is used to access name of a tag in BeautifulSoup?
B) tag.name
6. Which of the following is the default parser in BeautifulSoup?
C) lxml
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 web elements associated with the 'given xpath'
9. The script '**window.scrollTo(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)
D) <href>

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

Web Scraper is a tool used for Web Scraping, which is a technique to extract huge amount of data from any websites within a fraction of time and then saving those data to the local machine in the form SQL, Excel and XML. It is used for downloading information. It need not to visit all the pages of website for information. Web Scraper doesn't obey robots.txt in most cases. Data de-duplication is not a necessary part of web scraping.

Web Crawler or Spider is a tool which is used for web crawling. It basically visits a website and reads web pages for the purpose of building entries for search engine index. It is used for indexing of web pages. It visits all the pages, until the last line of information. It always obeys robots.txt. Data de-duplication is an integral part for web crawling.

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

A robots.txt file tells search engines where they can and can't go on a website. The primary purpose of robots.txt files is to exclude certain pages from search engine indexing. This text file contains "directives" which helps to dictate to search engines which pages are to 'Allow' and 'Disallow' search engine access.

13. What are static and dynamic web pages?

In the case of Static Web Pages, pages will remain same until someone changes it manually. But in Dynamic Web Pages, content of pages are different for different visitors.

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

```
from selenium import webdriver
import time
```

```
def title(url):
    driver=webdriver.Chrome(executable_path=r"C:\Users\hp\cdriver.exe")
    driver.get(url)
    time.sleep(5)
    title=driver.title
    if title==None:
        print('No Title found',driver.current_url)
    else:
        print('\nTitle:\t',title)
```

```
url=input('Input URL: ')
title(url=url)
```

Input URL: <https://www.hsbc.co.in/>

Title: HSBC India - Credit Cards, NRI Services, Saving and Deposit

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

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
import time
```

```
driver=webdriver.Chrome(executable_path=r"C:\Users\hp\cdriver.exe")
url="https://images.google.com/"
driver.get(url)
```

```
search_bar=driver.find_element(By.NAME,"q")
```

```
search_button=driver.find_element_by_xpath("//div[@class='FAuhyb']")
```

```
search=input("What do u want to search?: ")
search_bar.send_keys(search)
search_button.click()
```

What do u want to search?: cake

```
driver.quit()
```