# WEB SCRAPING – WORKSHEET 4

## In Q1 to Q14 have one or more than one correct options, Choose all the correct options:

1. Which of the following functions can be used to get an element from webpage when we know the Name attribute of the element?
   1. get\_by\_name() B) get\_element\_by\_name()

C) find\_element\_by\_name() D) None of the above

**Answer:** C) find\_element\_by\_name()

1. Which of the following functions can be used when you want to locate an element by tag name?
   1. get\_elements\_by\_tagid() B) get\_element\_by\_tagsid()

C) find\_element\_by\_tag\_name() D) All of the above

**Answer:** C) find\_element\_by\_tag\_name()

1. In what type of Waits, a WebDriver waits for a certain condition to occur before proceeding further with execution.
   1. Implicit wait B) Explicit wait

C) Both of them D) None of them

**Answer:** D) Explicit wait

1. Which of the following is an expected condition in selenium (python)?
   1. title\_is B) visibility\_of

C) staleness\_of D) All of the above

**Answer:** D) All of the above

1. Which of the following is a disadvantage of html5lib parser in beautiful soup?
   1. External C dependency B) Very Slow

C) External Pyhton Dependency D) all of the above

**Answer:** C) External Pyhton dependency

1. What are the advantages of using Scrapy over Selenium for web-scraping?
   1. For large data Scrapy is faster than selenium
   2. It supports javascript better than Selenium
   3. Scrapy is better than Selenium for simple projects
   4. All of the above

**Answer**: A) For large data Scrapy is faster than selenium

1. Which of the following is (are) true regarding Scrapy?
   1. spiders are classes which define how a certain site will be scrapped.
   2. spiders are the place where you define the custom behaviour for crawling.
   3. None of them
   4. both A & B

**Answer:** D) both A &B

1. Full form of HTML:
   1. Hyper Text Markup Link B) Hyper Text Mark language

C) Hyper Text Markup Language D) Hyper Text Mining Link

**Answer:** C) Hyper Text Markup Language

1. Which among the following is the correct syntax for parsing a html page?
   1. soup=BeautifulSoup(html\_doc, html)
   2. soup=BeautifulSoup(html\_doc,’html.parser’)
   3. soup=BeautifulSoup(html\_doc,’html\_parser’)
   4. none of the above

**Answer:** B) soup = BeautifulSoup(html\_doc, ‘html.parser’)

1. Which among the following is not a valid parser in BeautifulSoup?
   1. “lxml” B) “html.parser”

C) “lxml-xml" D) “html-xml”

**Answer:** D) “html-xml”

1. Which of the following functions is used to go to the next element in the page?
   1. findNext\_all() B) Find\_all()

C) find\_next() D) None of the above

**Answer:** C) find\_next()

1. Which of the following functions are used to iterate over an element’s siblings that precede it in the tree?
   1. find\_previous\_siblings() B) Get\_prev\_sibs()

C) get\_siblings() D) None of the above

**Answer:** A) find\_previous\_siblings()

1. Which of the following is an argument used in find\_all() which tells Beautiful Soup to stop gathering results after it’s found a certain number.
   1. stop\_at B) stop\_before

C) limit D) None of the above

**Answer:** C) limit

1. How would you set the recursive argument in find\_all() if you want Beautiful Soup to consider only the direct children.
   1. recursive = True B) recursive = False

C) recursive argument has no effect D) None of the above

**Answer:** B) recursive =False

## Q15 is subjective answer type question, Answer it briefly.

1. What is the difference between find () and find\_all() in Beautiful Soup?

**Answer:** find () Gets the first tag of the incoming HTML element that satisfies the condition and returns a function of a label group or a single label. While find\_all() gets all the conditions of the incoming HTML element and return it.

**Beautiful Soup** is a **Python** library for getting data out of HTML, XML, and other markup languages. Say you've found some WebPages that display data relevant to your research, such as date or address information, but that do not provide any way of downloading the data directly

