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?
   B) get_element_by_name()
   C) find_element_by_name()
2. Which of the following functions can be used when you want to locate an element by tag name?
   A) get_elements_by_tagid() B) get_element_by_tagsid()

3. In what type of Waits, a WebDriver waits for a certain condition to occur before proceeding further with execution.
   C) Both of them
4. Which of the following is an expected condition in selenium (python)?
   B) visibility_of
   C) staleness_of
5. Which of the following is a disadvantage of html5lib parser in beautiful soup?
   B) Very Slow
   C) External Pyhton Dependency
6. What are the advantages of using Scrapy over Selenium for web-scraping?
   A) For large data Scrapy is faster than selenium
   B) It supports javascript better than Selenium
   C) Scrapy is better than Selenium for simple projects
   D) All of the above
7. Which of the following is (are) true regarding Scrapy?
   A) spiders are classes which define how a certain site will be scrapped.
   B) spiders are the place where you define the custom behaviour for crawling.

   C) both A & B 8. Full form of HTML:
   A) Hyper Text Markup Link B) Hyper Text Mark language
   C) Hyper Text Markup Language ++++++++++ D) Hyper Text Mining Link

9. Which among the following is the correct syntax for parsing a html page?
   A) soup=BeautifulSoup(html_doc, html) B) soup=BeautifulSoup(html_doc,'html.parser')

10. Which among the following is not a valid parser in BeautifulSoup?
    A) "lxml" B) "html.parser" C) "lxml-xml"

11. Which of the following functions is used to go to the next element in the page?
    B) Find_all()
    C) find_next()
12. Which of the following functions are used to iterate over an element's siblings that precede it in the tree?
    B) Get_prev_sibs()
    C) get_siblings()

FLIP ROBO

13. 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.
    A) stop_at B) stop_before
    B) limit

14. How would you set the recursive argument in find_all() if you want Beautiful Soup to consider only the direct children.
    A) recursive = True
    C) recursive argument has no effect

Q15 is subjective answer type question, Answer it briefly.

15. What is the difference between find() and find_all() in Beautiful Soup?

find() :

Gets the first tag of the incoming HTML object that satisfies the condition and returns. A function of a label group or a single label.

Prototype: find(tag, atributes, recursive, text, keywords)

Parameter tag: Indicates the type of tag you need to find. There can be multiple tags.

Parameter attributes: The tag search is performed by using the attribute value corresponding to several attributes in the tag, and the attribute value may be multiple.

note:

You can search for CSS class, this string using soup.find("tagName", { "class" : "cssClass" })**The default value is the parameter value of the CSS class.**。

The parameter recursive: Whether to use a recursive method to traverse each sub-tab, the default is on, True. If set to False, findAll() only looks for the first-level label of the document. In general use, there is no need to change the default value of this parameter.

Parameter text: Find the list of tags based on the text content of the tag, usually in conjunction with regular expressions.

Parameter keywords:A list of tags for the specified attribute within the tag. Similar to the attribute parameter, there is an exception when the class attribute is used to find the label. Direct findAll(class='green') will report an error because class is a reserved word.

findAll()：

Get all the conditions of the incoming HTML object and return it.

Prototype: findAll(tag, atributes, recursive, text, limit, keywords)

Parameter tag: Indicates the type of tag you need to find. There can be multiple tags.

Parameter attributes: The tag search is performed by using the attribute value corresponding to several attributes in the tag, and the attribute value may be multiple.

note:

You can search for CSS class, this string using soup.find("tagName", { "class" : "cssClass" })**The default value is the parameter value of the CSS class.**。

The parameter recursive: Whether to use a recursive method to traverse each sub-tab, the default is on, True. If set to False, findAll() only looks for the first-level label of the document. In general use, there is no need to change the default value of this parameter.

Parameter text: Find the list of tags based on the text content of the tag, usually in conjunction with regular expressions.

Parameter limit: The range limit parameter can obviously only be used for the findAll() function. Limit the number of returns, such as how many tag information to extract as a sample.

Parameter keywords:A list of tags for the specified attribute within the tag. Similar to the attribute parameter, there is an exception when the class attribute is used to find the label. Direct findAll(class=‘green’) will report an error because class is a reserved word.