🌐 Web Scraping Quiz

1. What is Web Scraping?

   A. The process of building a website using HTML.

   B. The automated process of extracting data from websites.

   C. The technique used to display web pages in a browser.

   D. The design layout of a webpage.

   Answer: B

2. Before scraping any website, what is the most important initial step regarding ethical considerations?

   A. Sending a legal notice to the website owner.

   B. Checking the website's robots.txt file.

   C. Immediately running the scraper at high speed.

   D. Downloading all the website's images.

   Answer: B

3. What is the primary legal consideration when web scraping?

   A. Ensuring the scraper runs quickly.

   B. Respecting Copyright and Terms of Service of the website.

   C. Using the latest version of Python.

   D. Saving the data in a CSV file.

   Answer: B

4. Which Python library is commonly used to make HTTP requests and retrieve the content of a web page?

   A. Pandas

   B. BeautifulSoup

   C. Requests

   D. NumPy

   Answer: C

5. When setting up the environment for web scraping in Python, which library is specifically designed for parsing HTML and XML documents?

   A. Requests

   B. BeautifulSoup

   C. Selenium

   D. Matplotlib

   Answer: B

6. In the context of making web requests, what does the HTTP status code 200 generally indicate?

   A. Access is forbidden (Permission Denied).

   B. The page was successfully retrieved.

   C. The page was not found (Error).

   D. The request was too large.

   Answer: B

7. In the context of making web requests, what does the HTTP status code 404 generally indicate?

   A. Server is overloaded.

   B. Request was redirected.

   C. The requested resource was Not Found.

   D. Successful connection.

   Answer: C

8. When parsing HTML with BeautifulSoup, which object represents the entire structured document?

   A. HTMLParser

   B. soup object

   C. request object

   D. web_page object

   Answer: B

9. Which method in BeautifulSoup is used to find only the first matching tag (element) in the HTML document?

   A. find_all()

   B. select()

   C. find_one()

   D. find()

   Answer: D

10. Which method in BeautifulSoup is used to find all matching tags (elements) in the HTML document and returns them as a list?

    A. find_all()

    B. get_all()

    C. select_all()

    D. search()

    Answer: A

11. To extract the text content inside an HTML tag (e.g., extracting "Hello" from <h1>Hello</h1>), you would use which attribute or property on the tag object?

    A. .content

    B. .text or .get_text()

    C. .string

    D. .value

    Answer: B

12. If you are extracting the URL from an anchor tag (<a href="URL">...</a>), which attribute must you access on the BeautifulSoup tag object?

    A. ['link']

    B. .url

    C. ['href']

    D. .source

    Answer: C

13. Which CSS selector, when used with BeautifulSoup's select() method, targets an element based on its ID attribute (e.g., id="main")?

    A. [main]

    B. .main

    C. main#

    D. #main

    Answer: D

14. Which CSS selector, when used with BeautifulSoup's select() method, targets all elements based on their class attribute (e.g., class="product-title")?

    A. [product-title]

    B. .product-title

    C. #product-title

    D. ::product-title

    Answer: B

15. What is Web Crawling?

    A. Extracting data from a single webpage.

    B. The process of following hyperlinks from one page to another to discover and index content.

    C. The process of designing the layout of a website.

    D. Running JavaScript on a webpage.

    Answer: B

16. What is the key difference between Web Scraping and Web Crawling?

    A. Scraping is for static content; Crawling is for dynamic content.

    B. Scraping is about extracting data; Crawling is about discovering URLs and pages.

    C. Scraping is done by search engines; Crawling is done by users.

    D. Scraping is always legal; Crawling is always illegal.

    Answer: B

17. To avoid overloading a website's server and adhere to ethical guidelines, what should a scraper implement?

    A. Faster processing speed.

    B. Request delays (sleeps) between requests.

    C. More complex parsing logic.

    D. Use of multiple simultaneous IP addresses.

    Answer: B

18. What is the primary file format used by websites to signal which parts of the site should not be crawled or scraped?

    A. sitemap.xml

    B. webpages.html

    C. robots.txt

    D. license.txt

    Answer: C

19. Why is it important to check the User-Agent header when making web requests?

    A. It tells the website what programming language you are using.

    B. It prevents the website from redirecting you.

    C. It identifies the client making the request, and some sites block requests from known bot User-Agents.

    D. It saves the extracted data faster.

    Answer: C

20. What is a "Headless Browser" used for in advanced web scraping?

    A. Rendering HTML without CSS.

    B. Scraping only static content.

    C. Simulating a full user browser experience (running JavaScript) without a graphical interface.

    D. Making simple HTTP requests only.

    Answer: C

21. When a scraped data element looks like ['\n', 'Product Name', '\n'], what needs to be done during the data extraction step?

    A. Convert it to an integer.

    B. Use the .strip() method to clean up extra whitespace/newlines.

    C. Convert it to a List.

    D. Raise a ValueError.

    Answer: B

22. In Python, how do you install the requests library?

    A. install requests

    B. pip install requests

    C. load requests

    D. requests.install()

    Answer: B

23. Which term refers to the nested structure of HTML elements that BeautifulSoup relies on for parsing?

    A. HTTP Protocol

    B. Object-Oriented Programming

    C. DOM (Document Object Model) structure

    D. JSON format

    Answer: C

24. If a website loads its primary data using JavaScript after the initial page load, which tool might be necessary instead of just requests and BeautifulSoup?

    A. Pandas

    B. Selenium (or a similar headless browser)

    C. NumPy

    D. The time module

    Answer: B

25. What is the primary purpose of using Exceptions Handling (like try...except) in a web scraping script?

A. To increase the speed of the script.

B. To manage and recover from common errors like 404 responses or network connection failures.

C. To print data to the console.

D. To define the output file format.

Answer: B