الجامعة اللبنانية
كلية الاعلام
الفرع الاول

Université Libanaise
Faculté d'Information
Br. I

امتحانات: الفصل الثاني
من العام الجامعي 2025/2024

المرحلة: الاجازة	الدورة الاولى	المادة: Data Scraping
السنة المنهجية: الثانية		المدة: ساعتان
الاختصاص: علم البيانات		اسم الاستاذ: د. ليندا محمودي

Part I: Choose the correct answer (one incorrect answer will cancel out a correct one) **(15 points)**

1. Which python library is commonly used for web scraping?
   a) BeautifulSoup
   b) Selenium
   c) Requests
   d) All of the above

1. What is the purpose of using regular expressions in web scraping?
   a) To generate HTML code
   b) To scrape data from databases
   c) To extract patterns from text
   d) To render JavaScript code

2. What is Python IDLE?
   a) An Integrated Development and Learning Environment
   b) Perfect tool for a beginning programmer.
   c) Create and edit Python files and interact with the Python shell
   d) all of the above

3. What is the primary purpose of using proxies in web scraping?
   a) To increase the speed of scraping
   b) To hide the scraper's IP address
   c) To parse HTML content
   d) To handle JAVASCRIPT rendering

4. How can you handle dynamic content loaded by JAVASCRIPT in web scraping?
   a) By using the request library
   b) By manually entering the data
   c) By using a headless browser
   d) By disabling JAVASCRIPT in the browser

5. What is XPath primarily used for in web scarping?
   a) Parsing JSON
   b) Selecting elements in XML/HTML documents
   c) Scraping JAVASCRIPT
   d) Sending HTTP requests

6. Which of the following can cause issues when scraping websites?
   a) Anti-bot mechanisms
   b) JAVASCRIPT- heavy content
   c) Frequent requests
   d) all of the above

7. How can you handle rate-limiting in web scraping?
   a) Use proxies
   b) Add delays between requests
   c) Follow website's terms of service
   d) all of the above

8. What is a common method to store scraped data?
   a) Printing data to the console
   b) Sending data via email
   c) Storing data in CSV or JSON files
   d) Uploading data to social media

9. Which HTTP method is used for retrieving data in web scraping?
   a) POST
   b) PUT
   c) GET
   d) DELETE

10. What is the purpose of the requests library in web scraping?
   a) To render web pages in a browser.
   b) To handle HTTP requests and responses.
   c) To create visualizations of scraped data.
   d) To automate from submissions

11. What does a 401 unauthorized HTTP response mean?
   a) Resource not found
   b) Access is forbidden
   c) Authentication is required and has failed
   d) Request was successful

12. What does API stand for?
   a) Application programming interface
   b) Advanced programming interface
   c) Automated program interaction

13. Why is using an API generally better than scraping HTML from a website?
   a) APIs are always free
   b) APIs provide structured data and follow usage rules
   c) Scraping only during specific hours

14. What is web scraping?
   a) Designing web templates
   b) Creating websites
   c) APIs work without internet
   d) APIs return data in CSV format only

15. What is an alternative to Selenium for handling JAVASCRIPT rendering?
    a) Scrapy
    b) puppeteer
    c) numpy
    d) Flask

Part II: Answer by TRUE or FALSE with explanation (one incorrect answer will cancel out a correct one) **(30 points)**

1. Web scraping is illegal in all countries.
2. Selenium can be used to interact with dynamic elements on a web page like buttons and drop-downs list.
3. Web scarping can be used for data analysis, price monitoring, and lead generation.
4. Web scraping tools can only be written in Python.
5. Scrappy is a framework used for building web crawlers in Python.
6. The status code 200 means the API request was successful.
7. CAPTCHA is an anti-scraping measure used to verify if a user is human.
8. JSON is never used as a data format in web scraping
9. APIs cannot be rate-limited like websites
10. When using a public API, you don't need to worry about authentication or authorization.
11. The purpose of Pandas library in web scraping is to handling HTTP request.
12. "re" is a python library that can be used to interact with HTML Form.
13. The headless browser mode in Mechanicalsoup work runs a browser with a visible user interface and extra features.
14. This code line finds all links tag: all_links = soup.find ('link').
15. Beautifulsoup is a python library that can be used to interact with HTML Form.

Part III: (55 points)

1- What is data scarping? Explain with an example.
2- List and briefly describe any three Python libraries commonly used for web scraping.
3- What are some common ethical issues involved in web scraping?
4- What is the difference between static and dynamic scraping?
5- Mention two methods to avoid getting blocked while scraping a website.
6- What is the role of the robots.txt file in web scraping?
7- Describe how proxies are used in large-scale web scraping projects.
8- What is a headless browser, and why is it used in web scraping?
9- How does the user-Agent header help in web scraping?
10- Describe a real-life application where web scraping is used.