

MOCK EXAM 1

Q1. What is the main role of HTTP in web scraping?

- A. Store HTML locally
- B. Define how clients and servers exchange data
- C. Execute JavaScript in the browser
- D. Parse HTML into a tree

Answer: B

Q2. What does `requests.get(url)` return?

- A. Raw HTML string
- B. A BeautifulSoup object
- C. A Response object
- D. A dictionary

Answer: C

Q3. Where is the textual HTML content stored in a requests response?

- A. `response.content`
- B. `response.text`
- C. `response.html`
- D. `response.body`

Answer: B

Q4. What does HTTP status code **404** usually mean?

- A. Server error
- B. Redirection
- C. Resource not found
- D. Request succeeded

Answer: C

Q5. Which HTTP status code indicates a successful request?

- A. 200
- B. 301
- C. 404
- D. 500

Answer: A

BeautifulSoup — Core Concepts (5 questions)

Q6. What is required to create a BeautifulSoup object?

- A. A URL only
- B. Raw HTML and a parser
- C. A Response object only
- D. A CSS file

Answer: B

Q7. What does find() return?

- A. A list of matching elements
- B. The first matching element
- C. All elements in the DOM
- D. A dictionary

Answer: B

Q8. What is returned by find_all()?

- A. One element
- B. A string
- C. A list of elements
- D. A generator

Answer: C

Q9. How do you extract the value of an attribute (e.g. href) from a tag?

- A. tag.href
- B. tag.get("href")
- C. tag["text"]
- D. tag.attr("href")

Answer: B

Q10. Which method extracts only the visible text inside an HTML tag?

- A. .string
- B. .text()

- C. .get_text()
- D. .innerHTML()

Answer: C

Regular Expressions (5 questions)

Q11. What is the main purpose of regex in scraping workflows?

- A. Download web pages
- B. Execute JavaScript
- C. Extract or standardize text patterns
- D. Parse HTML structure

Answer: C

Q12. When should regex typically be applied?

- A. Before sending the HTTP request
- B. While rendering JavaScript
- C. After extracting text
- D. While creating a DataFrame

Answer: C

Q13. Conceptually, what does a regex “group” allow you to do?

- A. Match multiple pages
- B. Capture a specific part of a pattern
- C. Replace text automatically
- D. Speed up scraping

Answer: B

Q14. What is the difference between retrieving one match vs multiple matches?

- A. One uses regex, the other doesn't
- B. One returns a string, the other returns several matches
- C. One cleans data, the other extracts
- D. There is no difference

Answer: B

Q15. Which of these is a *reasonable* expected skill for students?

- A. Writing complex regex from memory
- B. Extracting emails or dates using examples
- C. Optimizing regex engines
- D. Building parsers from scratch

Answer: B

Pandas Advanced (5 questions)

Q16. What does an **inner join** return?

- A. All rows from both tables
- B. Only rows with matching keys in both tables
- C. All rows from the left table
- D. All rows from the right table

Answer: B

Q17. Which Pandas function is used to join two DataFrames?

- A. `pd.join()`
- B. `pd.merge()`
- C. `pd.concat()`
- D. `df.append()`

Answer: B

Q18. Which parameter defines the type of join in `pd.merge()`?

- A. `on`
- B. `axis`
- C. `how`
- D. `index`

Answer: C

Q19. When would `pd.concat()` be preferred over `pd.merge()`?

- A. When joining on a key
- B. When stacking DataFrames vertically or horizontally
- C. When filtering rows
- D. When cleaning text

Answer: B

Q20. A business question requires keeping **all rows from the left table**.

Which join should be used?

- A. Inner
- B. Right
- C. Left
- D. Outer

Answer: C