



امتحانات: الفصل الرابع
من العام الجامعي 2026/2025

المادة: Data Scraping

المدة: ساعتان

اسم الاستاذ: د. محمد أبوطعام

الدورة الثانية

المرحلة: الاجازة

السنة المنهجية: الثانية

الاختصاص: علم البيانات

Part I: Choose the correct answer (30 points)

- What is a python library that can be used to send and receive data over HTTP?
 - header
 - Port
 - ☒ urllib
 - HTTP
- What is the process by which search engines retrieve webpages and build a search index called?
 - ☒ Scrape
 - Parse
 - Beautiful Soup
 - Spider
- What does the following block of code do?


```
url=http://www.nytimes.com
html = urllib.request.urlopen(url).read()
soup = BeautifulSoup(html, 'html.parser')
```

 - Retrieves and displays the webpage
 - ☒ Parses the html content of the "http://www.nytimes.com" webpage
 - Downloads the webpage
- What is the role of re module in web scraping?
 - Handling HTTP requests
 - Parsing JSON responses
 - ☒ Working with regular expressions
 - Creating web forms

How can handle websites that load data through JavaScript in web scraping?

- By scraping only static content
- ☒ By using a headless browser

- c) By ignoring dynamic content
- d) By avoiding such websites

6. What protocol can be used to retrieve web pages using python?

- ☒ a) Hyper Text Markup Language
- b) Hyper Tech Markup Language
- c) Home Tool Markup Language
- d) Hyperlinks and Text Markup Language

7. What does API stand for?

- ☒ a) Application programming interface
- b) Advanced programming interface
- c) Automated program interaction

8. 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

9. What is web scraping?

- a) Designing web templates
- b) Creating websites
- ☒ c) APIs work without internet
- d) APIs return data in CSV format only

10. 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

Part II: (30 points)

→ It is an automated process by which HTTP request and the recieved content is parsed into structured form

1- What is web scarping?

2- Explain web scraping procedure, step by step.

3- Can you scrape data behind a login page?

4- Which Python libraries are used for web scraping?

5- What is web scraping used for?

6- What are the challenges of Data Scrapping.

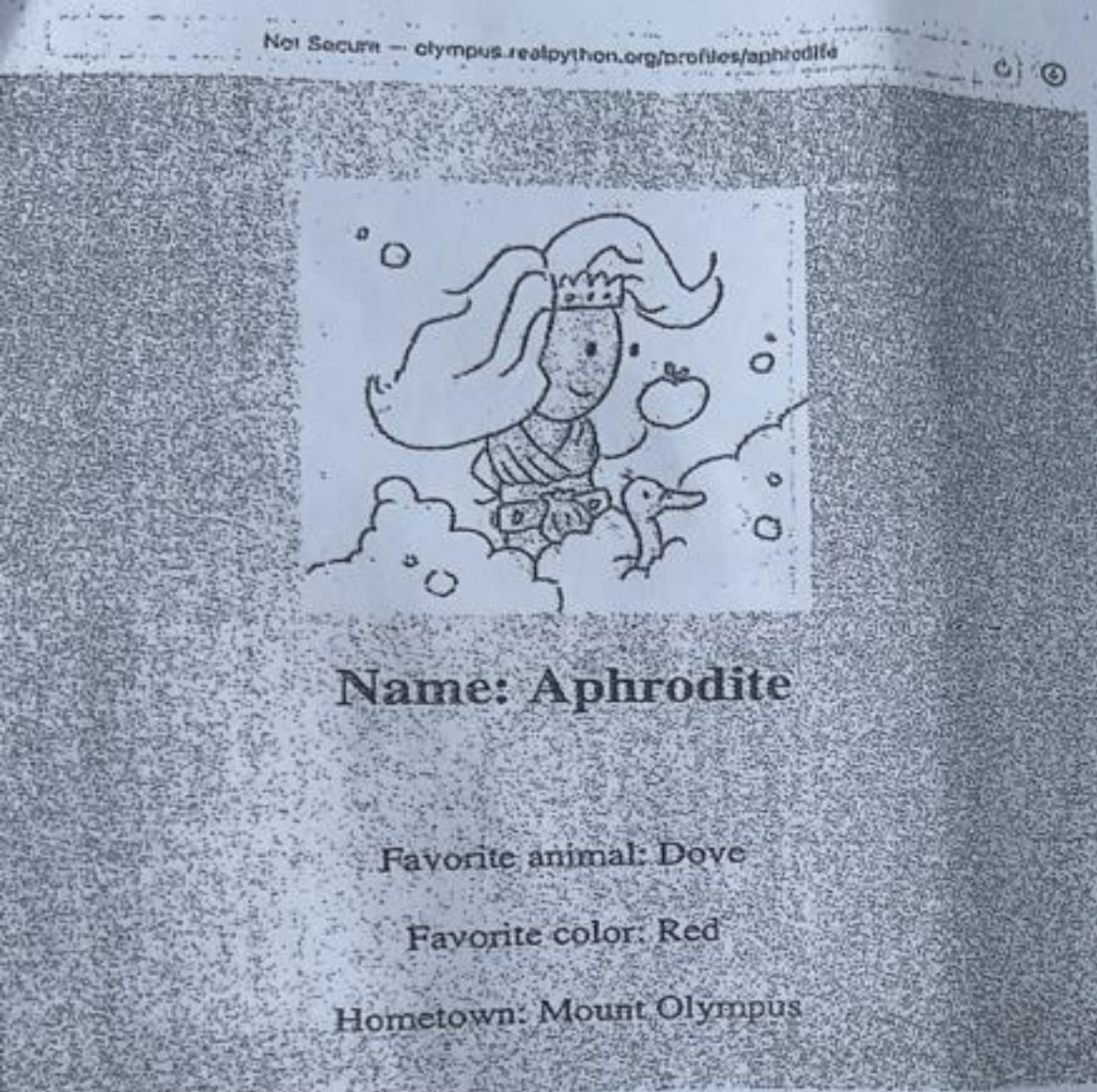
- False data

- legal issues

- firing a lot

Part III: (40 points)

Write a program that grabs the full HTML from the page at the URL <http://olympus.realpython.org/profiles/aphrodite>.



```
<html>
<head>
<title>Profile: Aphrodite</title>
</head>
<body bgcolor="yellow">
<center>
<br><br>

<h2>Name: Aphrodite</h2>
<br><br>
Favorite animal: Dove
<br><br>
Favorite color: Red
<br><br>
Hometown: Mount Olympus
</center>
</body>
</html>
```

Using string method (first solution), and using Python library (second solution),

- a- To display text following profile, Name, favorite animal, favorite color, hometown.
- b- Then storing scraped data to CSV file.

The final output should look like this:

Profile: Dionysus

Name: Dionysus

Hometown: Mount Olympus

Favorite animal: Leopard

Favorite Color: Wine

```
from bs4 import BeautifulSoup
from urllib import urlopen()
url = "http://olympus.realpython.org/"
page = urlopen(url)
```