Semester	B.E. Semester VIII – Computer Engineering
Subject	Data Science Honors
Subject Professor In-charge	Prof. Amit Alyani
Academic Year	2024-25
Student Name	Deep Salunkhe
Roll Number	21102A0014
Assignment	10

Introduction

Web mining is the process of extracting useful information from web pages. It involves techniques such as web scraping, web crawling, and data analysis. One of the most popular Python libraries for web scraping is **BeautifulSoup**, which allows easy parsing and extraction of HTML and XML data.

Why Use BeautifulSoup?

BeautifulSoup provides:

- Easy navigation and searching of HTML/XML parse trees.
- Support for parsing different types of HTML structures.
- Simple methods to extract text, attributes, and table data.

```
pip install beautifulsoup4 requests

Requirement already satisfied: beautifulsoup4 in
/usr/local/lib/python3.11/dist-packages (4.13.3)
Requirement already satisfied: requests in
/usr/local/lib/python3.11/dist-packages (2.32.3)
Requirement already satisfied: soupsieve>1.2 in
/usr/local/lib/python3.11/dist-packages (from beautifulsoup4) (2.6)
Requirement already satisfied: typing-extensions>=4.0.0 in
/usr/local/lib/python3.11/dist-packages (from beautifulsoup4) (4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.11/dist-packages (from requests) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests) (2025.1.31)
```

Basic Workflow

The general process of extracting data using BeautifulSoup involves:

1. **Sending a Request**: Use the `requests` module to fetch a webpage.

2. **Parsing HTML**: Convert raw HTML into a structured format.
3. **Navigating the DOM**: Extract required data using BeautifulSoup functions.
4. **Saving the Data:** Store the extracted data for further analysis.

```python
import requests
from bs4 import BeautifulSoup

url = "http://books.toscrape.com/"
response = requests.get(url)
soup = BeautifulSoup(response.text, "html.parser")

for book in soup.find_all("article", class_="product_pod"):
    title = book.h3.a["title"]
    price = book.find("p", class_="price_color").text
    print(f"Title: {title}, Price: {price}")
```

```
Title: A Light in the Attic, Price: Â£51.77
Title: Tipping the Velvet, Price: Â£53.74
Title: Soumission, Price: Â£50.10
Title: Sharp Objects, Price: Â£47.82
Title: Sapiens: A Brief History of Humankind, Price: Â£54.23
Title: The Requiem Red, Price: Â£22.65
Title: The Dirty Little Secrets of Getting Your Dream Job, Price:
Â£33.34
Title: The Coming Woman: A Novel Based on the Life of the Infamous
Feminist, Victoria Woodhull, Price: Â£17.93
Title: The Boys in the Boat: Nine Americans and Their Epic Quest for
Gold at the 1936 Berlin Olympics, Price: Â£22.60
Title: The Black Maria, Price: Â£52.15
Title: Starving Hearts (Triangular Trade Trilogy, #1), Price: Â£13.99
Title: Shakespeare's Sonnets, Price: Â£20.66
Title: Set Me Free, Price: Â£17.46
Title: Scott Pilgrim's Precious Little Life (Scott Pilgrim #1), Price:
Â£52.29
Title: Rip it Up and Start Again, Price: Â£35.02
Title: Our Band Could Be Your Life: Scenes from the American Indie
Underground, 1981-1991, Price: Â£57.25
Title: Olio, Price: Â£23.88
Title: Mesaerion: The Best Science Fiction Stories 1800-1849, Price:
Â£37.59
Title: Libertarianism for Beginners, Price: Â£51.33
Title: It's Only the Himalayas, Price: Â£45.17
```

Conclusion

BeautifulSoup is a powerful and easy-to-use library for web scraping. By following the methods outlined in this document, you can extract and analyze data efficiently. If working with dynamic content, integrating Selenium can enhance scraping capabilities.