Topics Covered

- 1. Objective
- 2. Introduction to Web Scraping
- 3. Why is Web Scrapping Needed?
- 4. How is Web Scraping Done?
- 5. Tools for Web Scraping
- 6. Getting Started with Requests
- 7. Getting Started with Beautiful Soup
- 8. Getting Started with Pandas
- 9. How to Scrape a Website?
 - Step 1. Sending an HTTP Request: to retrieve the desired web page.
 - Step 2. Retrieving Web Page Content : usually in HTML format.
 - Step 3. Parsing the HTML: to locate specific data elements using Python libraries(e.g., BeautifulSoup, Scrapy).
 - Step 4. Extracting Data: The relevant data is extracted from the parsed HTML structure
 - Step 5. Storing Data: for further analysis or processing.

Use the links above to jump directly to any section in this notebook.

Objective

Comprehensive understanding of **web scraping techniques** using **Python**, specifically utilizing the **Beautiful Soup**(bs4) and **Requests** modules. By applying these tools, I successfully extracted faculty information from my college's website.

Introduction to Web Scraping

Web scraping is a powerful technique used to extract data from websites.

It involves fetching web pages and parsing their content to retrieve specific information, which can be stored for further analysis or use.

Web scraping is commonly used for data collection, research, and various applications across different industries.

This can be useful for various applications, such as data analysis, research, and competitive analysis.

However, it's essential to respect the website's terms of service and robots.txt file to avoid legal issues.

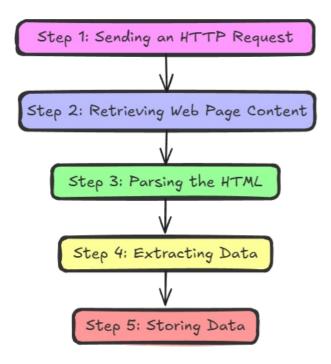
Why is Web Scraping Needed?

- 1. **Data Collection**: Web scraping allows for the automated gathering of data from multiple online sources, making it easier to collect large datasets for analysis or research.
- 2. **Dynamic Content Extraction**: Many websites display dynamic content that may be difficult or impossible to collect manually. Web scraping can efficiently extract this data.
- 3. **Time Efficiency**: Automating the data collection process saves time and reduces human error, enabling faster and more accurate results.
- 4. **Access to Unstructured Data**: A significant amount of information on the web is unstructured. Web scraping helps convert this data into structured formats, facilitating analysis and integration.
- 5. **Cost-Effectiveness**: By automating data collection, businesses and researchers can reduce the need for extensive human resources, lowering operational costs.

How is Web Scraping Done?

Web scraping typically involves the following steps:

- 1. Sending an HTTP Request: A request is sent to the target web server to retrieve the desired web page.
- 2. Retrieving Web Page Content: The server responds with the web page content, usually in HTML format.
- 3. **Parsing the HTML**: The HTML content is parsed to locate specific data elements using libraries or tools designed for this purpose (e.g., BeautifulSoup, Scrapy).
- 4. Extracting Data: The relevant data is extracted from the parsed HTML structure based on predefined rules or criteria.
- 5. **Storing Data**: The extracted data can be stored in various formats, such as CSV, JSON, or databases, for further analysis or processing.



Tools for Web Scraping

Beautiful Soup (bs4): A library for parsing HTML and XML documents, it provides Pythonic idioms for iterating, searching, and modifying the parse tree, making it easy to extract data from web pages.

Pandas: Primarily a data manipulation library, it can also be used for web scraping by reading HTML tables directly into DataFrames,

allowing for efficient data analysis and manipulation after extraction.

Requests: A simple and elegant HTTP library for Python, it allows you to send HTTP requests easily, handle responses, and manage sessions, making it essential for fetching web content before parsing it with tools like Beautiful Soup.

Getting Started with Requests

Importance of Requests in Web Scraping

The Requests library is a popular HTTP library for Python, making it easy to send HTTP requests and handle responses. Here are the key reasons why Requests is important in web scraping:

1. Simplified HTTP Requests

- User -Friendly API: Requests provides a simple interface for making GET and POST requests, allowing you to fetch web pages with minimal code.
- Session Management: It supports session management, enabling you to maintain cookies and headers across multiple requests.

2. Handling URL Parameters

• Easy Parameter Passing: You can easily pass URL parameters and form data using dictionaries, making it straightforward to interact with web forms and APIs.

3. Response Handling

- . Accessing Content: The library allows you to access the response content easily, including HTML, JSON, and binary data.
- Status Code Checking: You can check the status code of the response to handle errors and ensure successful requests.

4. Custom Headers and Authentication

- **Custom Headers**: Requests allows you to set custom headers, which can be useful for simulating browser requests or handling APIs that require specific headers.
- Authentication: It supports various authentication methods, including Basic Auth and OAuth, making it versatile for different web services.

5. Integration with Beautiful Soup

- Fetching HTML Content: Requests is often used to fetch HTML content, which can then be parsed using Beautiful Soup for data extraction.
- Streamlined Workflow: The combination of Requests and Beautiful Soup creates a streamlined workflow for web scraping.

Getting Started with Beautiful Soup

Importance of Beautiful Soup in Web Scraping

Beautiful Soup is a powerful library in Python for parsing HTML and XML documents. It is widely used in web scraping for the following reasons:

1. HTML Parsing

- Ease of Use: Beautiful Soup provides a simple and intuitive API for navigating and searching the parse tree of HTML documents.
- Flexible Parsing: It can handle poorly formatted HTML, making it robust for real-world web scraping scenarios.

2. Data Extraction

- Element Selection: Use methods like .find(), .find_all(), and CSS selectors to locate specific HTML elements and extract data.
- Text and Attribute Retrieval: Easily extract text content and attributes from HTML tags, allowing for efficient data collection.

3. Integration with Other Libraries

- Combining with Requests: Beautiful Soup is often used in conjunction with the Requests library to fetch HTML content and then
 parse it for data extraction.
- Compatibility: It works well with other libraries like Pandas for data manipulation and analysis.

Getting Started with Pandas

Importance of Pandas in Web Scraping

Web scraping is the process of extracting data from websites. When combined with Pandas, it becomes a powerful tool for data analysis and manipulation. Here are some key reasons why Pandas is important in web scraping:

1. Data Structuring

Pandas provides two primary data structures: Series and DataFrame. These structures allow you to organize scraped data in a tabular format, making it easier to analyze and manipulate.

- DataFrame: Ideal for handling structured data, where each column can hold different data types (e.g., integers, strings).
- 2. Exporting Data

After processing the scraped data, Pandas makes it easy to export the results to various formats:

- CSV: Use df.to csv('output.csv', index=False) to save the DataFrame as a CSV file.
- Excel: Save data to Excel files using df.to excel('output.xlsx', index=False).
- 3. Integration with Other Libraries

Pandas integrates well with other libraries commonly used in web scraping, such as:

- Beautiful Soup: For parsing HTML and XML documents.
- · Requests: For making

How to Scrape a Website?

Sending an HTTP Request

```
import requests

url = 'http://gcambalacantthry.edu.in/Faculty'
response = requests.get(url)

if response.status_code == 200 :
    print(" Successfully fetched the page!")
else :
    print(f"Failed to retrieve the page ! Status Code :{response.status_code}")
```

Successfully fetched the page!

Retrieving Web Page Content

Parsing the HTML

```
In [35]: from bs4 import BeautifulSoup
    soup = BeautifulSoup(response.content, 'html.parser')
In [36]: #print(soup.prettify()) # formated Version
```

Extracting Data

```
'faculty_dept' : fdept,
'faculty_photo' : fpic })
```

Storing Data

In [39]: import pandas as pd

Creating a DataFrame with pandas

In [40]: df = pd.DataFrame(faculty_info)

faculty dent Out[40]:

:	faculty_name	faculty_post	faculty_dept	faculty_photo
0	Anju Tanwar	Assistant Professor	Botany	https://mis.highereduhry.ac.in/EmpImages/11150
1	Dr Ravinder Kaur	Extension Lecturer	Botany	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
2	Rohini Singh	Assistant Professor	Botany	https://mis.highereduhry.ac.in/EmpImages/10196
3	Shilpa	Extension Lecturer	Botany	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
4	Abha Chaudhary	Assistant Professor	Chemistry	https://mis.highereduhry.ac.in/EmpImages/13052
5	Deepak Sharma	Associate Professor	Chemistry	https://mis.highereduhry.ac.in/EmpImages/10725
6	Dharam Singh	Assistant Professor	Chemistry	https://mis.highereduhry.ac.in/EmpImages/12915
7	Dr Avtar Singh Rahi	Associate Professor	Chemistry	https://mis.highereduhry.ac.in/EmpImages/10599
8	Monika Rani	Assistant Professor	Chemistry	https://mis.highereduhry.ac.in/EmpImages/13012
9	Ritu Mittal	Extension Lecturer	Chemistry	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
10	Aarzoo Katyal	Assistant Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/15766
11	Anil Kumar	Associate Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/10820
12	Anuradha	Assistant Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/12447
13	Dr Anjna Rani	Extension Lecturer	Commerce	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
14	Gaytri Devi	Associate Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/12011
15	Lucky Verma	Assistant Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/15661
16	Mona Saini	Extension Lecturer	Commerce	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
17	Monika Sharma	Assistant Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/10628
18	Ms Rekha Rani	Extension Lecturer	Commerce	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
19	Naib Singh	Associate Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/11065
20	Navita Rani	Extension Lecturer	Commerce	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
21	Nisha	Extension Lecturer	Commerce	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
22	Pardeep Kumar	Assistant Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/10617
23	Parmil Kumar	Associate Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/11158
24	Poonam Rajora	Associate Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/10828
25	Prabjot Kaur	Extension Lecturer	Commerce	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
26	Ravi Kant	Assistant Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/10677
27	Richa Lomas	Extension Lecturer	Commerce	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
28	Shagun Ahuja	Associate Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/12013
29	Shilpa Rani	Extension Lecturer	Commerce	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
30	Shweta Garg	Extension Lecturer	Commerce	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
31	Sujata Sasan	Assistant Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/11991
32	Usha Rani Chahal	Assistant Professor	Commerce	https://mis.highereduhry.ac.in/EmpImages/10607
33	Ashish Jolly	Assistant Professor	Computer Science	https://mis.highereduhry.ac.in/EmpImages/11746
34	Ashok	Assistant Professor	Computer Science	https://mis.highereduhry.ac.in/EmpImages/13663
35	Dimple Rani	Assistant Professor	Computer Science	https://mis.highereduhry.ac.in/EmpImages/13664
36	Dr Ruchi Gupta	Guest Lecturer	Computer Science	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
37	Dr. Vikas Sharma	Guest Lecturer	Computer Science	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
38	Dr.Shruti Sharma	Guest Lecturer	Computer Science	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
39	Indresh Aggarwal	Guest Lecturer	Computer Science	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
40	Karmjit Kaur	Assistant Professor	Computer Science	https://mis.highereduhry.ac.in/EmpImages/13528
41	Malika Rhivana	Accietant Professor	Computer Science	https://mis-highereduhry-ac-in/Emplmages/117//2

41	ווום ווועם ווועם ווועם ווועם	Assistant Fidicasul	Computer Science	Intpo.//Inio.ingneredurity.ac.in/Linpiniages/11/42
42	Manisha Saini	Assistant Professor	Computer Science	https://mis.highereduhry.ac.in/EmpImages/13661
43	Pallavi Rani	Extension Lecturer	Computer Science	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
44	Poonam Dhiman	Assistant Professor	Computer Science	https://mis.highereduhry.ac.in/EmpImages/13672
45	Sangharsh Saini	Extension Lecturer	Computer Science	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
46	Sarbjit Kaur	Assistant Professor	Computer Science	https://mis.highereduhry.ac.in/EmpImages/13660
47	Shakti Kumar	Extension Lecturer	Computer Science	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
48	Shalin Bhola	Extension Lecturer	Computer Science	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
49	Naresh Kumar	Assistant Professor	Economics	https://mis.highereduhry.ac.in/EmpImages/11483
50	Ombir	Assistant Professor	Economics	https://mis.highereduhry.ac.in/EmpImages/12978
51	Rajesh Kumar	Assistant Professor	Economics	https://mis.highereduhry.ac.in/EmpImages/10361
52	Sharmila Ahlawat	Assistant Professor	Economics	https://mis.highereduhry.ac.in/EmpImages/10218
53	Uma Shashi	Assistant Professor	Economics	https://mis.highereduhry.ac.in/EmpImages/16917
54	Anju Jagpal	Associate Professor	English	https://mis.highereduhry.ac.in/EmpImages/11888
55	Barjinder Singh	Assistant Professor	English	https://mis.highereduhry.ac.in/EmpImages/15913
56	Gurvinder Singh	Associate Professor	English	https://mis.highereduhry.ac.in/EmpImages/10700
57	Kusum Bhatia	Assistant Professor	English	https://mis.highereduhry.ac.in/EmpImages/15905
58	Munish Trighatia	Assistant Professor	English	https://mis.highereduhry.ac.in/EmpImages/10516
59	Pooja Goyal	Assistant Professor	English	https://mis.highereduhry.ac.in/EmpImages/11030
60	Rekha Batta	Assistant Professor	English	https://mis.highereduhry.ac.in/EmpImages/15906
61	Roopa Mangal	Associate Professor	English	https://mis.highereduhry.ac.in/EmpImages/11181
62	Saroj Bala	Extension Lecturer	English	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
63	Saroj Duhan	Extension Lecturer	English	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
64	Shardha Sharma	Assistant Professor	English	https://mis.highereduhry.ac.in/EmpImages/10091
65	Sunil Kumar	Assistant Professor	English	https://mis.highereduhry.ac.in/EmpImages/10562
66	Vinod Kumar	Associate Professor	English	https://mis.highereduhry.ac.in/EmpImages/11161
67	Randeep Singh	Assistant Professor	Fine Arts	https://mis.highereduhry.ac.in/EmpImages/11174
68	Ajay Chauhan	Associate Professor	Geography	https://mis.highereduhry.ac.in/EmpImages/10574
69	Dolan Kaler	Extension Lecturer	Geography	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
70	Dr Vandana Kumari Chauhan	Extension Lecturer	Geography	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
71	Pawan Kumar	Assistant Professor	Geography	https://mis.highereduhry.ac.in/EmpImages/10625
72	Sanjeet Singh	Assistant Professor	Geography	https://mis.highereduhry.ac.in/EmpImages/17194
73	Dr Manjeet Kaur	Extension Lecturer	Hindi	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
74	Dr Reetu Rani	Extension Lecturer	Hindi	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
75	Dr Rita Tandon	Extension Lecturer	Hindi	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
76	Gulab Singh	Assistant Professor	Hindi	https://mis.highereduhry.ac.in/EmpImages/15932
77	Neeru Bhatia	Assistant Professor	Hindi	https://mis.highereduhry.ac.in/EmpImages/10899
78	Suresh Kumar	Assistant Professor	Hindi	https://mis.highereduhry.ac.in/EmpImages/15646
79	Atul Yadav	Associate Professor	History	https://mis.highereduhry.ac.in/EmpImages/11171
80	Dr Manjesh Kumar	Assistant Professor	History	https://mis.highereduhry.ac.in/EmpImages/16904
81	Ram Kumar	Associate Professor	History	https://mis.highereduhry.ac.in/EmpImages/11871
82	Sanjay Kumar	Extension Lecturer	History	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
83	Baljinder Kaur	Assistant Professor	Mass Communication	https://mis.highereduhry.ac.in/EmpImages/11147
84	Dharmvir Singh	Extension Lecturer	Mass Communication	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
85	Dr Tanjum Kamboj	Extension Lecturer	Mass Communication	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
86	Hitesh Bhardwaj	Extension Lecturer	Mass Communication	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
87	Rakesh Sharma	Assistant Professor	Mass Communication	https://mis.highereduhry.ac.in/EmpImages/10982
88	Sandeep Kumar	Extension Lecturer	Mass Communication	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
89	Deepak	Assistant Professor	Maths	https://mis.highereduhry.ac.in/EmpImages/13226
90	Dr Pooja Nagpal	Extension Lecturer	Maths	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
91	Hardish Kaur	Assistant Professor	Maths	https://mis.highereduhry.ac.in/EmpImages/15647

92	Neha Rani	Assistant Professor	Maths	https://mis.highereduhry.ac.in/EmpImages/15654
93	Preeti	Extension Lecturer	Maths	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
94	Rajani Saini	Associate Professor	Maths	https://mis.highereduhry.ac.in/EmpImages/11169
95	Sanjay Nanda	Assistant Professor	Music (I)	https://mis.highereduhry.ac.in/EmpImages/10611
96	Anju Bhardwaj	Assistant Professor	Music (V)	https://mis.highereduhry.ac.in/EmpImages/11069
97	Alka	Associate Professor	Physics	https://mis.highereduhry.ac.in/EmpImages/11764
98	Dr Samiksha Kumari	Extension Lecturer	Physics	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
99	Dr Sushma Devi	Extension Lecturer	Physics	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
100	Harneet Kaur	Assistant Professor	Physics	https://mis.highereduhry.ac.in/EmpImages/12912
101	Priyanka	Assistant Professor	Physics	https://mis.highereduhry.ac.in/EmpImages/13113
102	Reeta Rani	Extension Lecturer	Physics	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
103	Mamta Kumari	Extension Lecturer	Political Science	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
104	Ravinder Kumar	Assistant Professor	Political Science	https://mis.highereduhry.ac.in/EmpImages/11009
105	Dr Ajay Gauri	Assistant Professor	Psychology	https://mis.highereduhry.ac.in/EmpImages/11091
106	Dr Ravi Aggarwal	Extension Lecturer	Psychology	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
107	Dr Neelam Rani	Extension Lecturer	Punjabi	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
108	Reeta Devi	Assistant Professor	Sanskrit	https://mis.highereduhry.ac.in/EmpImages/10621
109	Daleep Kumar	Assistant Professor	Tourism	https://mis.highereduhry.ac.in/EmpImages/12191
110	Rajeev Sharma	Assistant Professor	Tourism	https://mis.highereduhry.ac.in/EmpImages/12042
111	Shikha Bhatt	Assistant Professor	Tourism	https://mis.highereduhry.ac.in/EmpImages/11173
112	Dr Ratesh Kumar	Extension Lecturer	Zoology	https://mis.highereduhry.ac.in/forms_MIS/EL/Em
113	Pooja Devi	Assistant Professor	Zoology	https://mis.highereduhry.ac.in/EmpImages/11052
114	Tarsem Kumar	Assistant Professor	Zoology	https://mis.highereduhry.ac.in/EmpImages/11056

Saving the Data

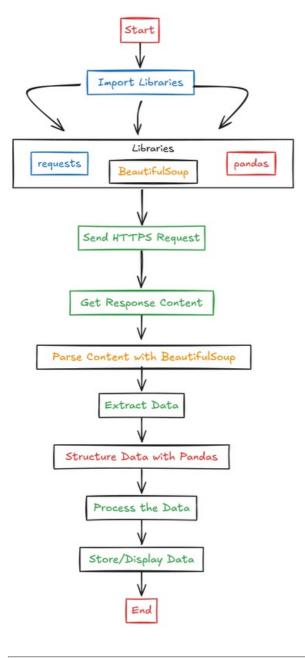
```
In [41]: MyData = df.to_csv("faculty_info")
print("Data saved successsfully as CSV format!!") #In the same folder
```

Data saved successsfully as CSV format!!

Display Entire Dataset

```
In [43]: pd.set_option('display.max_rows', None) # Show all rows
```

SUMMARYY



THANKS