






Mastering Web Scraping: Understanding Web Page Structures

Presented by:

Ch. Sindhura & S Jai Prakash (JP)

Workshop Guidelines

-  Raise your hand anytime you need help
-  Ask questions freely - no question is too basic!
-  Help your fellow participants
-  Keep devices in silent mode
-  Follow along with the live demonstrations

Quick Poll: Let's Know Our Audience

Raise your hand if you:

- Have used HTML before
- Know what CSS is
- Have used Chrome Developer Tools
- Have tried web scraping

This helps us adjust our pace and explanation depth!

What is HTML?

HTML (HyperText Markup Language) is like the skeleton of a webpage!

```
<!DOCTYPE html>
<html>
  <head>
    <title>My First Page</title>
  </head>
  <body>
    <h1>Hello World!</h1>
    <p>This is a paragraph</p>
  </body>
</html>
```

🤔 **Interactive Question:** What do you think `<h1>` and `<p>` mean?



HTML Elements: Building Blocks

Common HTML elements we'll encounter:

```
<div>A container for other elements</div>  
<span>Inline text container</span>  
<a href="https://wikipedia.org">Link to Wikipedia</a>  
<table>  
  <tr><td>Table cell</td></tr>  
</table>
```

Practice Time:

Open Chrome, right-click, select "Inspect" on any webpage. Can you find these elements?

Understanding CSS

CSS = Cascading Style Sheets (Makes websites pretty!)

```
/* Using class */  
.article-title {  
    color: blue;  
}  
  
/* Using ID */  
#main-content {  
    background: white;  
}
```

 **Discussion:** Why do we need CSS for web scraping?

CSS Selectors: Your Scraping Tools


```
/* Different ways to select elements */  
.class-name      /* Select by class */  
#id-name         /* Select by ID */  
div             /* Select all divs */  
div.special     /* Select divs with class 'special' */  
div > p         /* Select paragraphs directly inside divs */
```

 **Live Demo:** Let's try these selectors on Wikipedia!


Chrome Developer Tools: Your Best Friend

Key Features:

1. Elements Panel (Ctrl+Shift+C)
2. Console (Ctrl+Shift+J)
3. Network Tab
4. Sources Panel

 **Interactive Demo:** Everyone open Dev Tools and follow along!

Finding Elements in Dev Tools

1. Right-click > Inspect
2. Use Element Selector ()
3. Search in Elements (Ctrl+F)

Practice Task:

Find the following on Wikipedia:

- Main article title
- First paragraph
- Table of contents

Live Demo: Wikipedia Article Analysis

Let's visit: [List of Academy Award-winning films](#)

Step-by-Step Together:

1. Open the page
2. Find the main table
3. Inspect table structure
4. Identify useful CSS selectors

Hands-On Exercise

In pairs (5 minutes):

1. Find the table with Oscar winners
2. Identify CSS selectors for:
 - Film titles
 - Year of award
 - Number of awards

Share your findings with the group!

Common Challenges & Solutions

1. Dynamic Content

- Look for "loading" indicators
- Check Network tab for AJAX calls

2. Complex Layouts

- Use multiple selectors
- Try XPath as alternative

 **Discussion:** What challenges did you face in the exercise?

☕ **Break Time! (5 mins)**



Practice Project

Let's extract:

1. Film titles
2. Release years
3. Number of awards

```
import requests
from bs4 import BeautifulSoup

url = "https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films"
# Let's write this code together!
```

Interactive Debugging Session

Common issues we might face:

- Table not found
- Wrong data extracted
- Missing elements

 **Group Activity:** Debug a broken scraper together!



Best Practices

1. Always check robots.txt
2. Use meaningful selector names
3. Handle errors gracefully
4. Document your code
5. Respect website terms of service

Final Challenge

In groups of 2:

1. Go to https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films
2. Identify the `<table>` with `class="wikitable sortable jquery-tablesorter"` to extract
3. Find their selectors
4. Present your approach

Useful Resources

- [MDN Web Docs](#)
- [W3Schools HTML Tutorial](#)
- [CSS Selector Game](#)
- [Chrome DevTools Documentation](#)

 **Thank You!**

Connect with us:

- Ch. Sindhura: <https://www.linkedin.com/in/sindhura-chinoori-710b5165/>
- S Jai Prakash (JP): <https://www.linkedin.com/in/s-jaiprakash/>

Remember: The best way to learn is by doing!