



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

SCRAPING WEBSITES WITH R

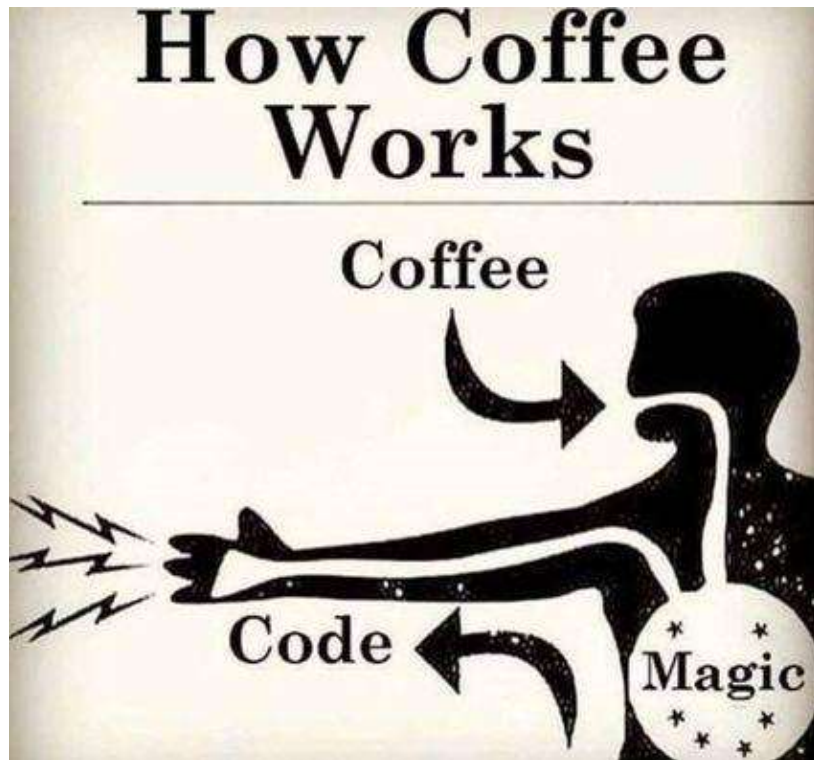
21 October 2024

Dr Jessica Witte



**SUPPORT FOR DATA-LED
AND APPLIED DIGITAL
RESEARCH ACROSS THE
ARTS, HUMANITIES AND
SOCIAL SCIENCES.**





ABOUT ME

- CDCS Digital Research Analyst
- Digital methods: web scraping, text analysis, natural language processing, machine learning, data visualisation, generative AI
- Python, R, HTML, CSS
- 'Self-taught'





THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

SCHEDULE

14:00-15:00

- Introduction & housekeeping
- Review of web scraping, HTML/CSS, and website structure
- Hands-on tutorial: scraping a static webpage

Break

15:10-16:00

- Hands-on tutorial: multi-page web scraping
- Additional resources



www.cdcs.ed.ac.uk

WHY SCRAPE THE WEB?

- To collect data—social media, public records, government data
- To expand, update, or complete datasets
- To examine public discourse about a topic
- To analyse the relationship between online and offline behaviour





THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

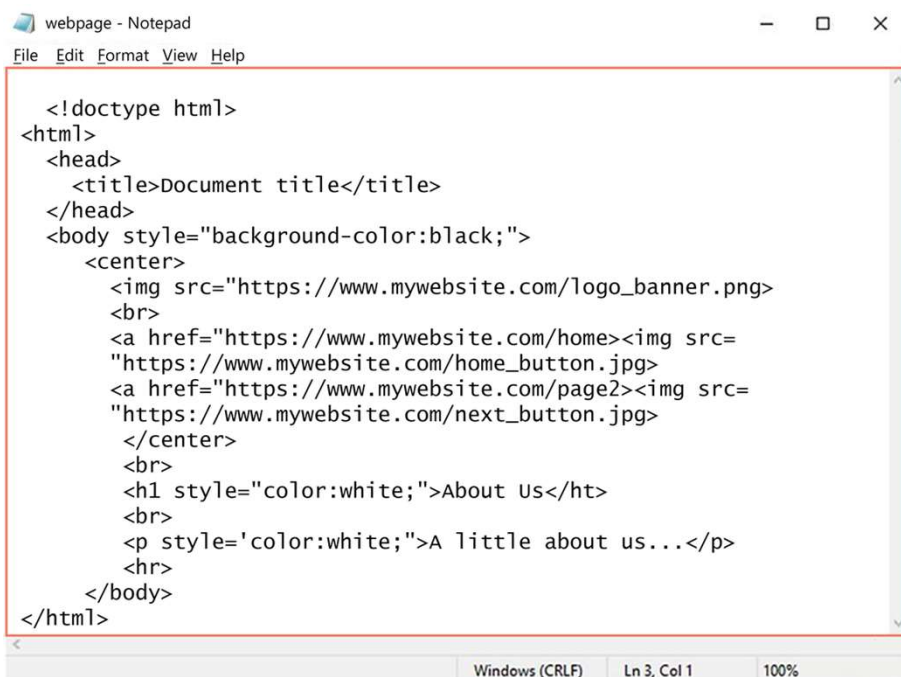
METHODS FOR WEB SCRAPING

- **Scraping and crawling HTML/XML**—generate a list of URLs from which to extract information, selecting relevant sections by HTML, and downloading content
- **APIs**—request data from sites on their own terms. Standard approach for social media data pre-2023, but has become more complicated
- **Browser automation**—using a script to click through a website like a human user and download content page by page



www.cdcs.ed.ac.uk

HTML



```
<!doctype html>
<html>
  <head>
    <title>Document title</title>
  </head>
  <body style="background-color:black;">
    <center>
      
      <br>
      <a href="https://www.mywebsite.com/home"><img src=
"https://www.mywebsite.com/home_button.jpg">
      <a href="https://www.mywebsite.com/page2"><img src=
"https://www.mywebsite.com/next_button.jpg">
    </center>
    <br>
    <h1 style="color:white;">About Us</ht>
    <br>
    <p style='color:white;'>A little about us...</p>
    <hr>
  </body>
</html>
```

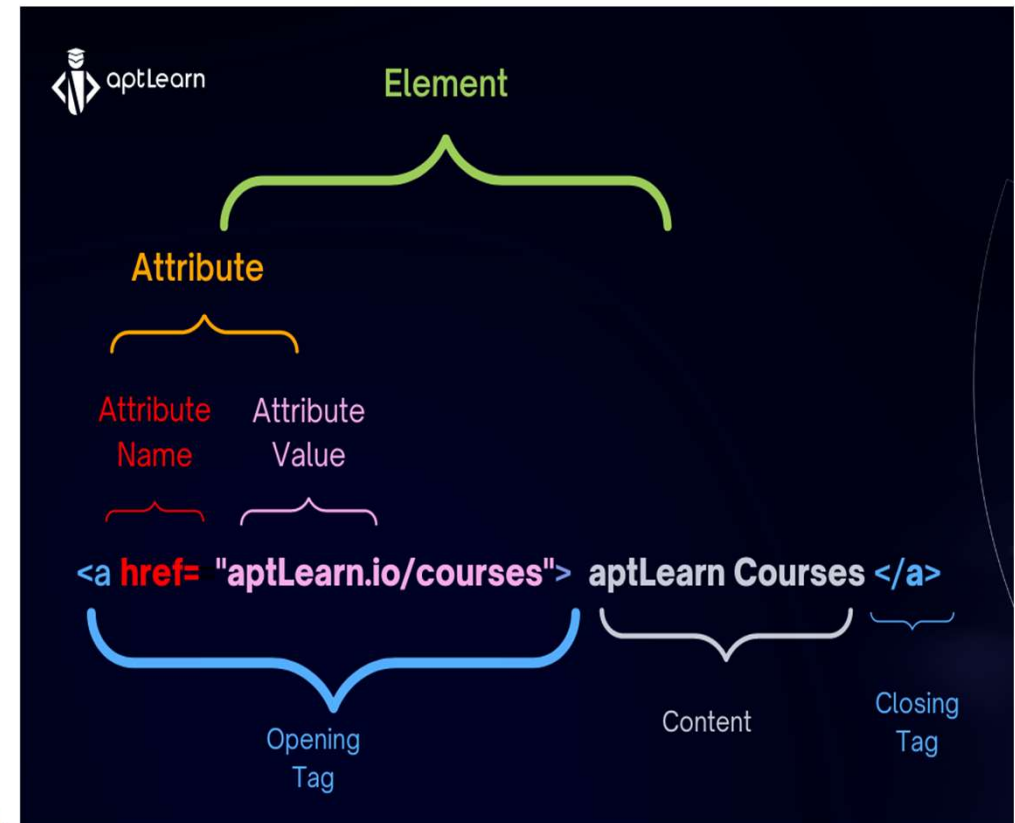
[Image credit](#)

- **Hypertext** markup language
- Creates the structure and content of a static webpage
- HTML **documents** are built of **paired HTML tags**
 <html>
 </html>
- **Elements** are objects or features such as images and text
- HTML **attributes** modify the appearance or behaviour of elements



HTML ATTRIBUTES

- **Provide additional information** about HTML elements embedded in tags
- Attributes are specified within the opening tag of an HTML element and consist of a **name-value pair**
- For instance, the "**href**" attribute is used in the "**a**" tag to specify **the target URL** of a **hyperlink**





Ardbeg Uigeadail

Average score from 90 reviews and 338 ratings 91

Reviewed by @Megawatt
♥ 5 💬 5 94/100

Not the first I've tried, mind you, but the first I've owned. Not sure why it took me so long.

On the nose, there is warm, thick, sweet, perfumed, enveloping smokiness. A different quality of smoke that some of the other Islays I've tried. It is not sharp, or ashen, or medicinal. There is a tarry quality to it, and a



Product details

Brand: Ardbeg
Bottler: Distillery Bottling
ABV: 54.2%

Shop for this

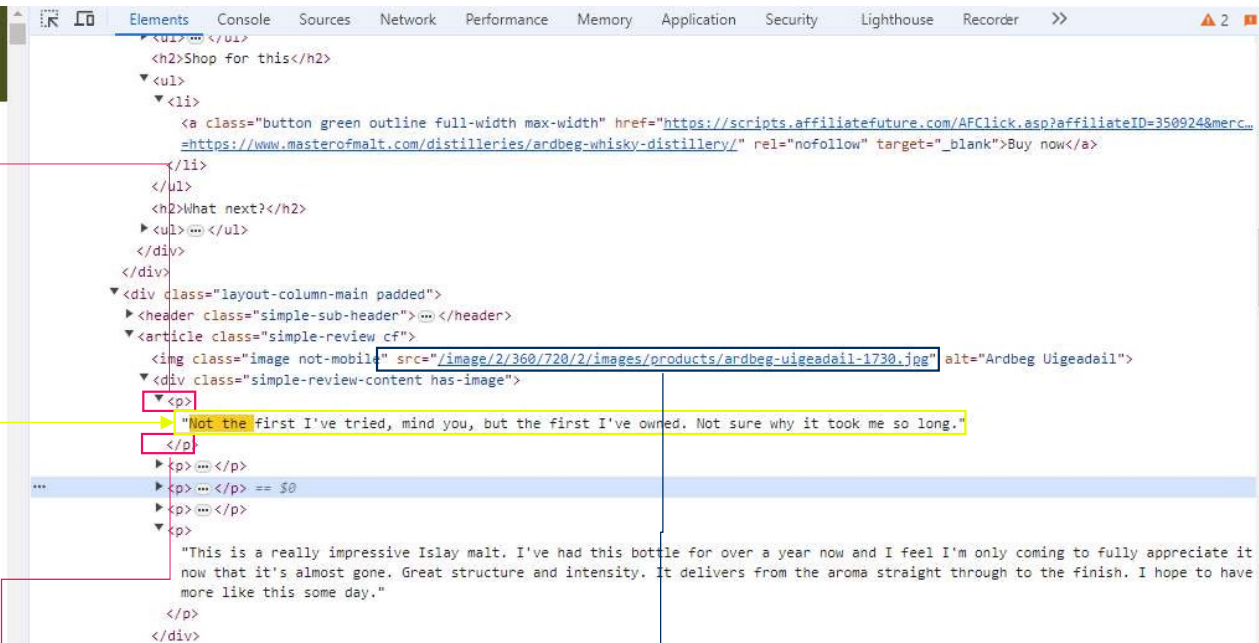
BUY NOW

What next?

ADD TO CABINET

Open Tag

Close Tag

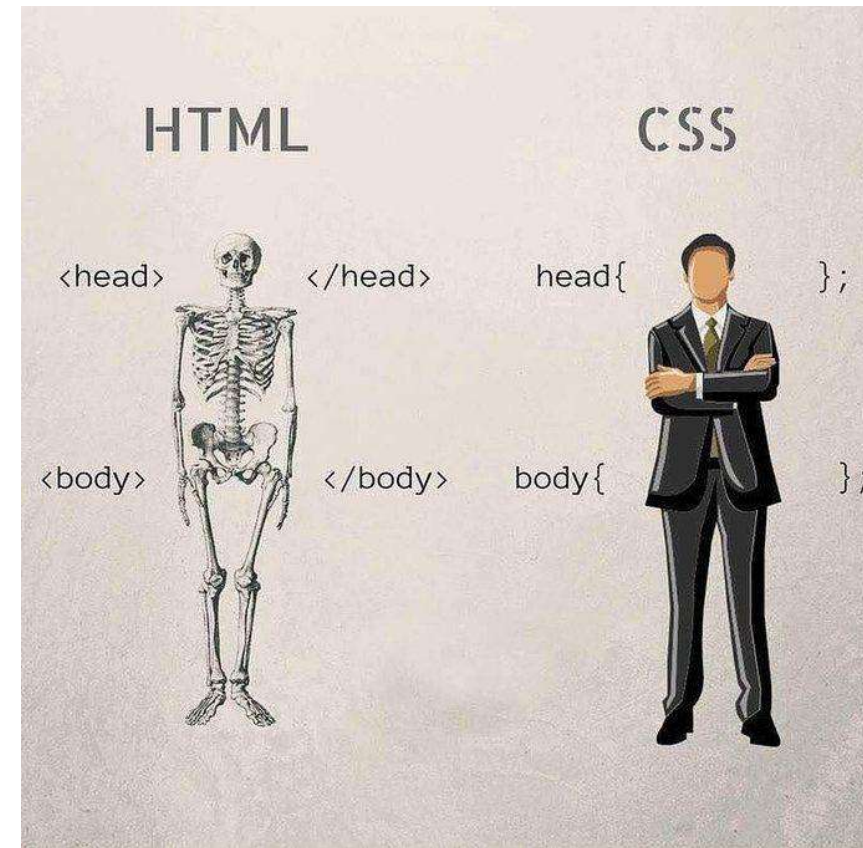


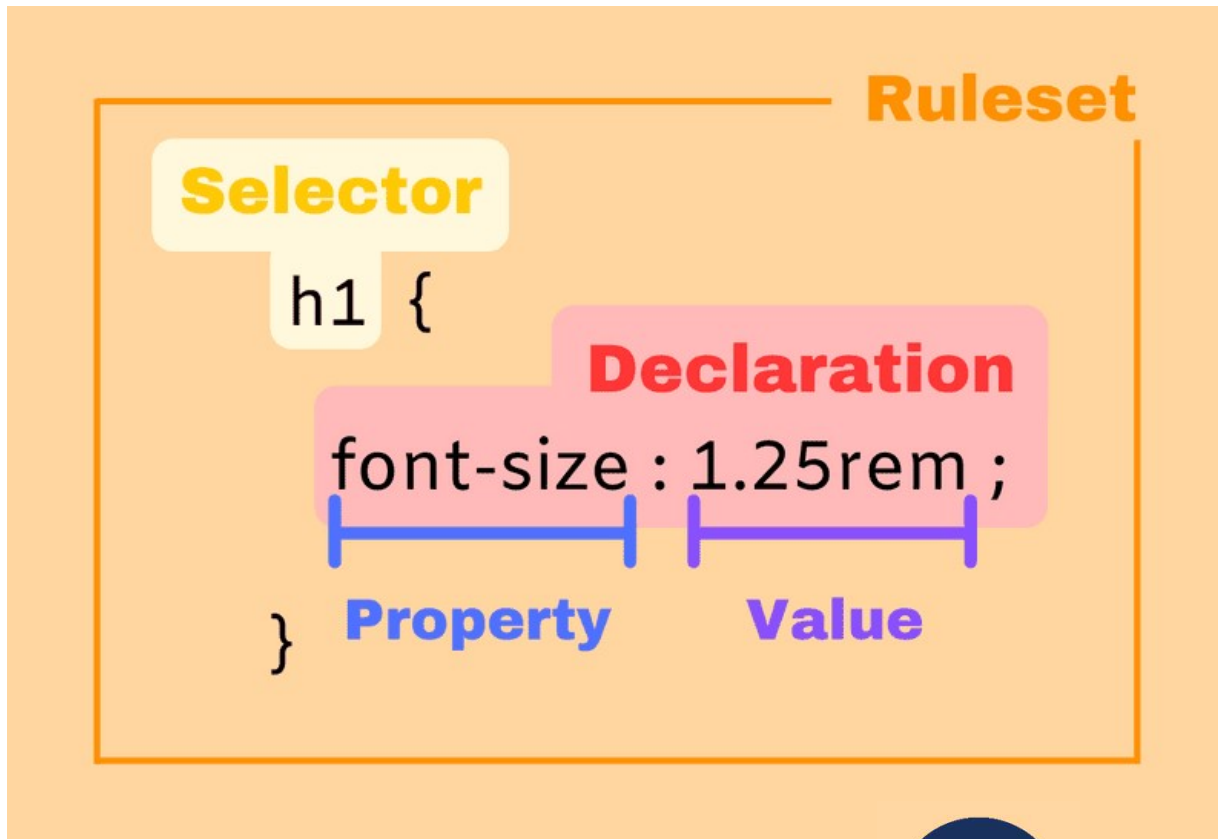
Attribute: SRC



CSS

- Cascading style sheets
- CSS defines **the style** and **appearance** of HTML elements on a webpage
- Linked to the HTML file



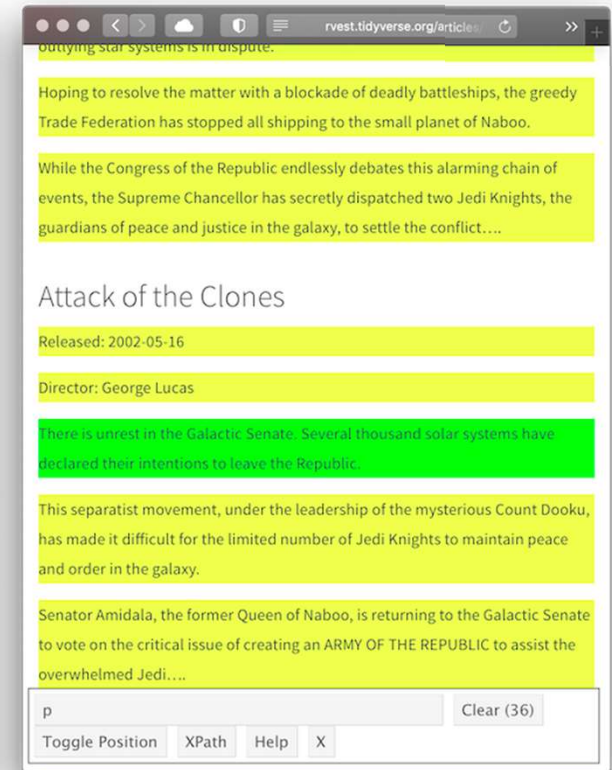
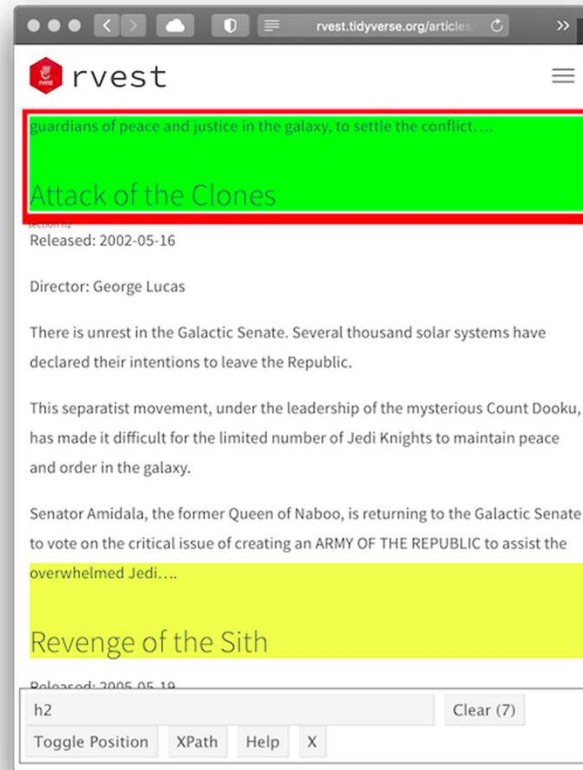


CSS RULESETS

- The **selector** points to a particular element of the webpage
- A **property** is an aspect of the selected element that will be defined in the **value**
- Together, property and value pairs compose a **declaration**

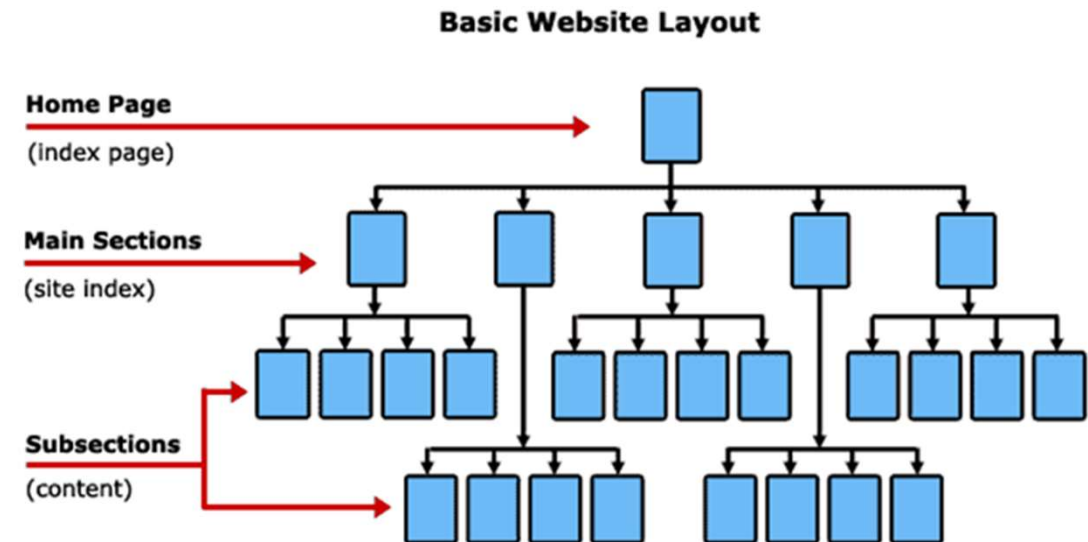
SELECTOR GADGET

- Chrome browser extension plugin
- Makes it easier to select the underlying HTML/CSS behind a webpage



WEBSITE STRUCTURE

- HTML websites are structured like trees
- A single URL can include several sub-URLS
- Sites composed of multiple sections/pages with static content often follow this structure—news sites, online forums, blogs

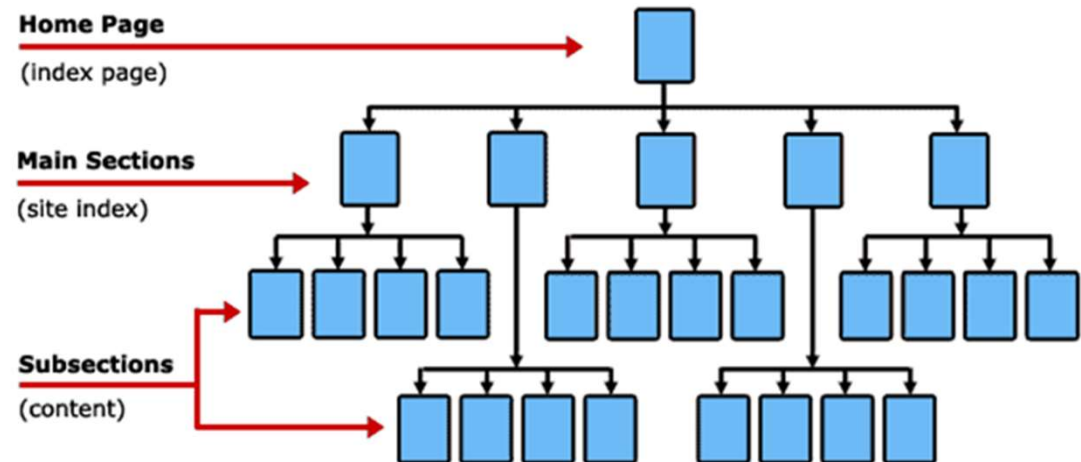


SCRAPING HTML TREES

Process:

- Start with base URL
- Scrape URLs for sub-forums or news topic
- Generate list of URLs for pages
- Scrape URLs from each page
- Scrape content from each page
- Create data frame with each article/post as a row, and each element from the article/post in its own column

Basic Website Layout





THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



TIME FOR R