Web Scraping

1) Rvest
    a) Download HTML
    b) Select specific parts of a webpage
    c) Extract text, attributes, and tables
    d) Clean and structure the data in R

2) HTML Basics
    a) Webpages consist of **elements**, each with:
        i) A **start tag**: 
        ii) **Attributes**: id='main', class='title'
        iii) Optional **content**: text or more elements
        iv) An **end tag**: 
    b) Special characters use escapes:
        i) < → <
        ii) > → >
        iii) & → &
3) Important HTML Elements
    a) <html> → root of page
    b) <head> → metadata, title
    c) <body> → visible contents
    d) Types:
        i) **Block elements**: <h1>, , <ul>
        ii) **Inline elements**: <b>, <i>, <a>
        iii) **Void elements (no closing tag)**: <img>
    e) Attributes that matter for scraping:
        i) id="" (unique)
        ii) class="" (categorizes elements)
4) Reading HTML in rvest
    a) html <- read_html("https://example.com")
    b) Create manually → minimal_html(" Hello ")
5) CSS Selectors

| Selector | Matches |
|---|---|
| p | all  tags |
| .title | class = "title" |
| p.special |  |
| #main | id = "main" |

- html_element() → one match per input
- html_elements() → all matches

Ex:

```
html |> html_elements("p")      # all <p> tags
html |> html_element("#first")  # id='first'
```

6) Extracting Data
    a) Text → html_text2()
        i) Returns clean, human-readable text
        ii) Collapses whitespace like a browser
            Ex:
            `html |> html_elements("li") |> html_text2()`
    b) Attributes
        i) `html |> html_element("a") |> html_attr("href")`
        ii) `html |> html_element("img") |> html_attr("src")`
        iii) *Attributes always return **strings** (convert as needed).
7) Extracting Tables
    a) HTML tables use <table>, <tr>, <th>, <td>.
        Convert directly to a tibble:

```
html |>
  html_node("table") |>
  html_table()
```

Key functions:
html_elements() → find all rows/units
html_element() → extract each variable consistently
html_text2() → best for readable text
html_attr() → extract attributes (links, images)
html_table() → import tables directly