R Coding Session: A Journey into Data Manipulation and Visualisation

Day 3: Web Scraping using `rvest`

Indira Puteri Kinasih

June 7, 2023

Universiti Brunei Darussalam

Aims

The main aims of this session are:

- Introduce web scraping basic technique using `rvest`
- Introduce `library(rvest)` to scraping websites/HTML
- Conduct a short scraping practice/demonstration to a real estate website to collect house prices data

Contents

Web Scraping

Scraping = Crawling + Parsing

- Scraping: **collect** or acquire webpage data for the purpose of analysis;
- crawling: systematically **navigating** through a website, extracting data from multiple pages;
- parsing: **Extracting** the necessary data from HTML by isolating and retaining the relevant syntactic elements (**title**).

Web Scraping

Process	Libs & Function
Scraping	`rvest::read_html()`
Crawling	`purrr::map()`
Parsing	`rvest::html_nodes()`,
	`rvest::html_text()`,
	`rvest::html_attr()`

Table 1: Library and Function for Web Scraping

HTML & CSS

- HTML: Hyper Text Markup Language;
- it uses a set of tags to define the structure and layout of content on a web page, such as headings, paragraphs, links, images, tables, and more

```
<html>
      <body>

            <h1>My first Heading</h1>
            <p>My first paragraph</h1>
            <a href = "https://www.myweb.com">
            link </a> to myweb.com
            </a>
```

HTML & CSS

- CSS: Cascading Style Sheets;
- used for describing the visual appearance and formatting of HTML documents.

Contents

Scraping Procedures

Basically, scraping process will follow this procedures: Development and Production (**johnlittle2020**). The development part consist of (Little John, 2020)

- Import raw HTML of a single target page (page detail, or ”leaf”);
- Parse the HTML of the test page to gather the data you want;
- In a web browser, manually browse and understand the site navigation of the scrape target (site navigation, or ”branches”);

Scraping Procedures

- Parse the site navigation and develop an iteration plan;
- Iterate: write code that implements iteration, i.e. automated page crawling (or for my case: use `for` loop)
- Perform a dry run with a limited subset of the target web site;
- check robots.txt, terms of use, and construct time pauses (to avoid DNS attacks)

Scraping Procedures

Whereas the production part employ an iteration process:

- Crawls the site navigation (branches);
- Parse HTML for each detail page (leaves)

Contents

Important Notes

- in scraping process, we will need to know the exact css selector for the web element that we targeted;
- for that purpose, we can use several technique:
  - View through page source;
  - Inspect the source
  - Use Selector Gadget (**Wickham2020**) to identify the specified element selector (see here)

Contents

Demonstration

for the demonstration, I will use Rmarkdown file to run some code chunks in Housing Data Repository. Below is the basic script to harvest an element (i.e Location) using `rvest`

```
location <- pages %>%
        html_nodes("css selector") %>%
        html_text() %>%
        str_trim()
```

Contents

References