

R Coding Session: A Journey into Data Manipulation and Visualisation

Day 3: Web Scraping using `rvest`

Indira Puteri Kinasih

June 7, 2023



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

- 1 Web Scraping
- 2 Scraping Procedures
- 3 Important Notes
- 4 Scraping Demonstration: House Advertisement Listing Website
- 5 References

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 Crawling Parsing	<code>rvest::read_html()</code> <code>purrr::map()</code> <code>rvest::html_nodes()</code> , <code>rvest::html_text()</code> , <code>rvest::html_attr()</code>

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

- 1 Web Scraping
- 2 Scraping Procedures
- 3 Important Notes
- 4 Scraping Demonstration: House Advertisement Listing Website
- 5 References

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

- 1 Web Scraping
- 2 Scraping Procedures
- 3 Important Notes**
- 4 Scraping Demonstration: House Advertisement Listing Website
- 5 References

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

- 1 Web Scraping
- 2 Scraping Procedures
- 3 Important Notes
- 4 Scraping Demonstration: House Advertisement Listing Website
- 5 References

Demonstration

for the demonstration, I will use Rmarkdown file to run some code chunks in Housing Data Repository

Contents

- 1 Web Scraping
- 2 Scraping Procedures
- 3 Important Notes
- 4 Scraping Demonstration: House Advertisement Listing Website
- 5 References

References