Web scraping

extract data from web

it is a technique for converting data (in unstructured format) to structured format which can be used.

- web page structure

webpages are usually consist of two types of code;

One focus on the appearance and format of the page called HTML Hyper Text Markup Language

The other one called XML similar to HTML but focus more one on managing data in the web.

The key difference HTML & XML is

HTML displays data and describe the appearance & format of the page, whereas XML manges the data (stores and transfers data)

example of HTML

```
<P> Types of Code: </P>
<ul>
    <li> HTML </li>
    <li> XML </li>
</ul>
```

tags:

 unordered list

 ordered list

 list item

<P> paragraph

heading levels

<h1> - <h6>

Web Scraping → extract data from one or more websites

web Crawling → finding or discovering URLs or link on the web.

usually you need to Combine both or data extraction projects.

TOS Term Of Service

2/

There are 4 ways People download data on the web:

1) click to download cvs/xls/txt file

2) use packages that interact with API

3) use API directly

4) SCRAPE Directly from the web Page

HTML file.

web scraping is a way to get data from web sites without APIs or whose APIs don't Provide to the data one is interested in.

{

35

The {polite} package

- The {polite} package is designed to help scrape without violating Term Of Service or robots.text standards
  - ↳ websites use a standard approach to identify their "desires" or preferences for scraping their sites.

Package has Two main function

- `bow` is used to introduce the client to the host and ask for permission to scrape.
- `scrape` is the main function for retrieving data from the remote server.

Cascading Style Sheet (CSS)

we have to know a little bit about CSS in order to understand how to extract certain elements from a website.

CSS is a formatting language that indicates how HTML files should look.

every website you have been on is formatted with CSS.

Using Selector Gadget with chrome

To install it you may want to so to "Chrome Web store"

Search for "Selector Gadget"

or go to selectorgadget.com

The {rvest} package

steps for scraping data with {rvest}

1) create a local copy of an HTML document from a URL, a file on disk, or a string containing HTML with `read_html()`

2) Select the nodes (elements) of a document you want using CSS selectors with `html_nodes(doc, "table td")`

3) extract components from the selected elements with

- `html_name()` (the name of the tag), or
- `html_text()` (all text inside tag), or
- `html_attr()` (contents of a single attribute) and
- `html_attrs()` (all attributes).
- `html_table()` parses table into data frames

Use tidyverse functions to convert to tibble so you can tidy and clean the data.

You Can also use {rvest} with XML files:

- Parse with `xml()`
- Then extract Component using `xml_node()`,

`xml_attr()`, `xml_attrs()`, `xml_text()` and `xml_name()`

- OTHER SPECIAL USE FUNCTION:
  - use `write_html()` or `write_xml()` to save HTML data to disk
  - Extract, Modify and submit forms with `html_form()`, `set_values()`, and `submit_form()`
  - Detect and repair encoding ~~with~~ problems with `guess_encoding()` and `repair_encoding()`

85

we will use {rvest} to extract element from HTML files

what does read_html() do?

it creates an html document from URL

html_nodes()

Selects parts of a documents using CSS selectors

This will give the HTML source code pull out the actual elements that we want to grab

html_text()

we need to call html_text because there will still be the link tags for the movies for example around the names of movies so HTML_text actually parse the text from within those tags