Web scraping part II

Statistical Computing & Programming

Shawn Santo

Supplementary materials

Full video lecture available in Zoom Cloud Recordings

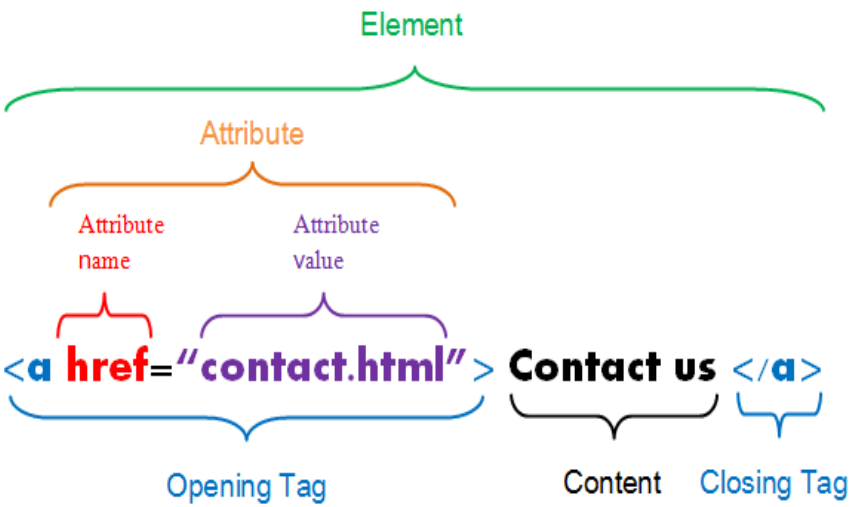
Additional resources

- Web scraping cheat sheet
- `RSelenium` website

Recall

Hypertext Markup Language

- HTML describes the structure of a web page; your browser interprets the structure and contents and displays the results.
- The basic building blocks include elements, tags, and attributes.
  - an element is a component of an HTML document
  - elements contain tags (start and end tag)
  - attributes provide additional information about HTML elements



HTML vs. XML

HTML snippet

```
<tr>
<td class=date>Jul 11, Sat</td>
<td class=where>
   <a class=mapq href="http://www.google.com/
   <span class=more>
   Wellington SP, West Shore Rd, Bristol, NH
   {43.639648,-71.779373}
   <a class="maplink" href="http://bing.com/m
   <a class="maplink" href="http://www.google
   <a class="maplink" href="https://www.mapqu
   <a class="maplink" href="http://www.openst
   </span>
</td>
<td class=name><a href="http://www.swimnewfou
<td class=distance>5 km, 10 km, 10 mi</td>
<td class=more><span class=time>7:00 AM</span
</tr>
```

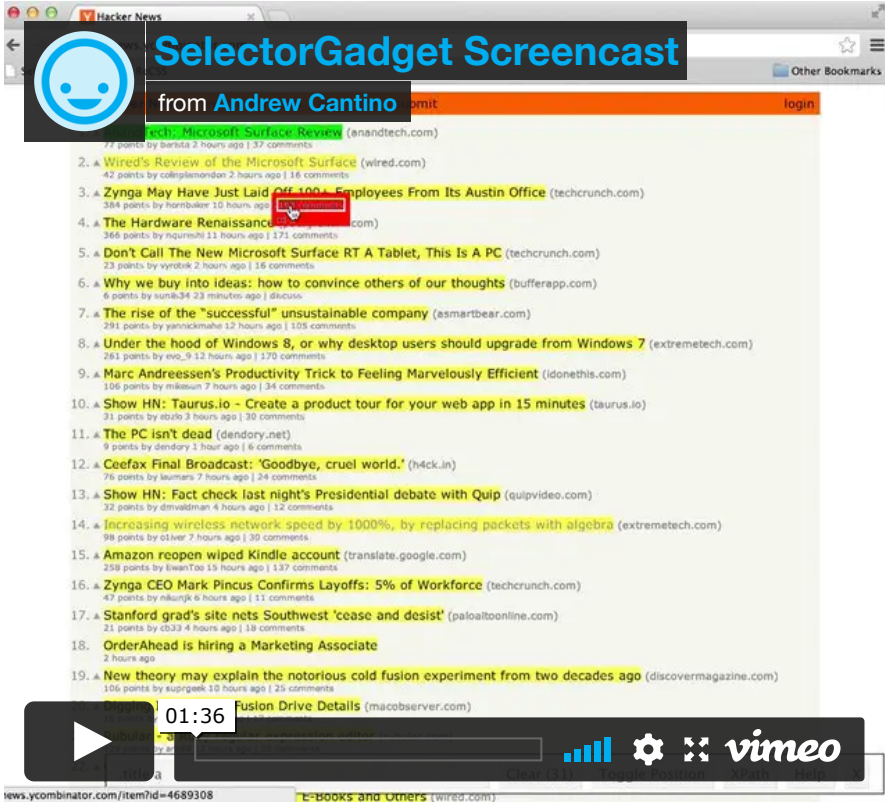
XML snippet

```
<swim>
  Swim with a Mission
  <location>
    Wellington SP, West Shore Rd, Bristol, NH
  </location>
  <link>
    http://www.swimnewfoundlake.com/
  </link>
  <date>
    Jul 11, Sat
  </date>
  <distance>
    5 km, 10 km, 10 mi
  </distance>
</swim>
```

SelectorGadget

In CSS, selectors are patterns used to select the element(s) you want to style.

SelectorGadget makes identifying the CSS selector you need as easy as clicking on items on a webpage.



Web scraping workflow

1. Understand the website's hierarchy and what information you need.
2. Use SelectorGadget to identify relevant CSS selectors.
3. Read html by passing a url and subset the resulting html document using CSS selectors.

```
read_html(url) %>%
  html_nodes(css = "specified_css_selector")
```

4. Further extract attributes, text, or tags by adding another layer with

```
read_html(url) %>%
  html_nodes(css = "specified_css_selector") %>%
  html_*()
```

where `*` is `text`, `attr`, `attrs`, `name`, or `table`.

Example with `html_table()`

http://www2.stat.duke.edu/courses/Spring21/sta323.001/schedule.html

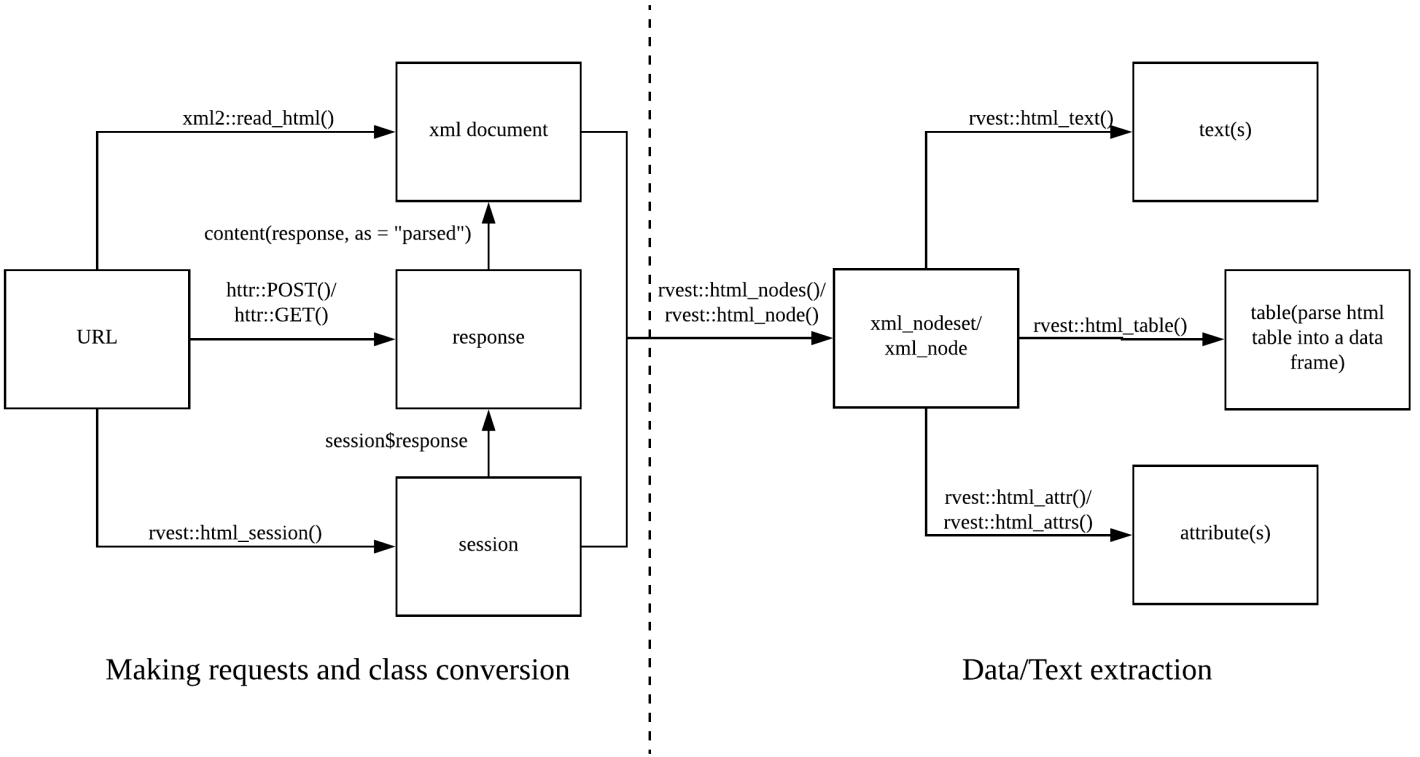
```
library(rvest)
library(tidyverse)

url <- paste0("http://www2.stat.duke.edu/courses/",
              "Spring21/sta323.001/schedule.html")

read_html(url) %>%
  html_nodes("table") %>%
  html_table(header = TRUE) %>%
  .[[1]] %>%
  janitor::clean_names() %>%
  as_tibble()
```

```
#> # A tibble: 51 x 8
#>    week   date     topic         html_slides pdf_slides exercises lab   homework
#>    <chr>  <chr>    <chr>         <lgl>       <lgl>      <lgl>     <lgl> <lgl>
#>  1 "Week… "Wed, J… Introduction… NA          NA         NA        NA    NA
#>  2 ""     "Fri, J… Vectors and … NA          NA         NA        NA    NA
#>  3 "Week… "Mon, J… Lab 01        NA          NA         NA        NA    NA
#>  4 ""     "Wed, J… Shell and ve… NA          NA         NA        NA    NA
#>  5 ""     ""       Homework 1 a… NA          NA         NA        NA    NA
#>  6 ""     "Fri, J… Data structu… NA          NA         NA        NA    NA
#>  7 "Week… "Mon, F… Lab 02        NA          NA         NA        NA    NA
#>  8 ""     "Wed, F… Object-orien… NA          NA         NA        NA    NA
#>  9 ""     ""       Homework 2 a… NA          NA         NA        NA    NA
#> 10 ""     "Fri, F… Data manipul… NA          NA         NA        NA    NA
#> # … with 41 more rows
```

Overview



Source: https://github.com/yusuzech/r-web-scraping-cheat-sheet/blob/master/README.md

Recall previous exercise

Scrape the first page of books from each genre in the side bar on the website http://books.toscrape.com/. Scrape the same information as before and include the genre.

```
# A tibble: 517 x 5
   title                                price  rating available genre
   <chr>                                <chr>  <chr>  <chr>     <chr>
 1 It's Only the Himalayas              £45.17 Two    In stock  Trav…
 2 Full Moon over Noah’s Ark: An Odyss… £49.43 Four   In stock  Trav…
 3 See America: A Celebration of Our N… £48.87 Three  In stock  Trav…
 4 Vagabonding: An Uncommon Guide to t… £36.94 Two    In stock  Trav…
 5 Under the Tuscan Sun                 £37.33 Three  In stock  Trav…
 6 A Summer In Europe                   £44.34 Two    In stock  Trav…
 7 The Great Railway Bazaar             £30.54 One    In stock  Trav…
 8 A Year in Provence (Provence #1)     £56.88 Four   In stock  Trav…
 9 The Road to Little Dribbling: Adven… £23.21 One    In stock  Trav…
10 Neither Here nor There: Travels in … £38.95 Three  In stock  Trav…
# … with 507 more rows
```

Web scraping considerations

Best practices

- Abide by a site's terms and conditions.
- Respect robots.txt.
    - https://www.facebook.com/robots.txt
    - https://www.wegmans.com/robots.txt
    - https://www.google.com/robots.txt

```
robotstxt::paths_allowed("https://www.facebook.com/")
robotstxt::paths_allowed("https://www.wegmans.com/")
robotstxt::paths_allowed("https://www.google.com/")
robotstxt::paths_allowed("https://www.google.com/search")
```

- Cache your `read_html()` chunks. Isolate these chunks.
- Avoid using `read_html()` in code that is iterated.
- Do not overload the server at peak hours.
    - Implement delayed crawls: `Sys.sleep(rexp(1) + 4)`
- If available, use a site's API.
- Do not violate any copyright laws.

A failure to abide

Dear Professor Santo,

We were recently forwarded several slides of yours describing how to scrape data from websites.

Although Slide 8, "Best Practices", states "Abide by a site's terms and conditions", Slide 17 specifically shows how to scrape data from PredictIt.org .

From "Web Scraping Part II, Statistical Programming (Shawn Santo, 10—03-19)".

You may not be recall or be aware that PredictIt's terms of service prohibit expressly data scraping, which has the effect of slowing down the site.

We believe that using PredictIt in your scraping instructions is causing and will continue to cause "Best Practices" not to be used on our site. This will result in degradation of the site's performance.

We ask that you abide by the terms of service in this regard, remove the reference to PredictIt in your instructions, and replace it with a different target.

If you have any questions about this, please reply to this message.

Respectfully,

Other considerations

- Disguise your IP address.
  - `httr::use_proxy()`

- Avoid scraping behind pages protected by log-in, unless it is permitted by the site.
  - `html_session()`

- Watch out for honey pot traps - invisible links to normal visitors, but present in HTML code and found by web scrapers.

Beyond `rvest` and static sites

Limitations of using `rvest` functions

- It is difficult to make your code reproducible long term. When a website or the HTML changes, your code may no longer work.
    - CSS selectors change
    - Contents are moved
    - Switch from HTML to JavaScript

- Websites that rely heavily on JavaScript

What is JavaScript?

- Scripting language for building interactive web pages
- Basis for web, mobile, and network applications
- Every browser has a JavaScript engine that can execute JavaScript code.
    - Chrome and Edge: V8
    - Safari: JavaScriptCore
    - Firefox: SpiderMonkey

If `read_html()` is meant for HTML, what can we do?

Possible solutions

1. Use your browser's developer tools (Chrome is easiest)
2. Execute JavaScript in R
3. Use package `Rselenium` or other web drivers
    - http://ropensci.github.io/RSelenium/

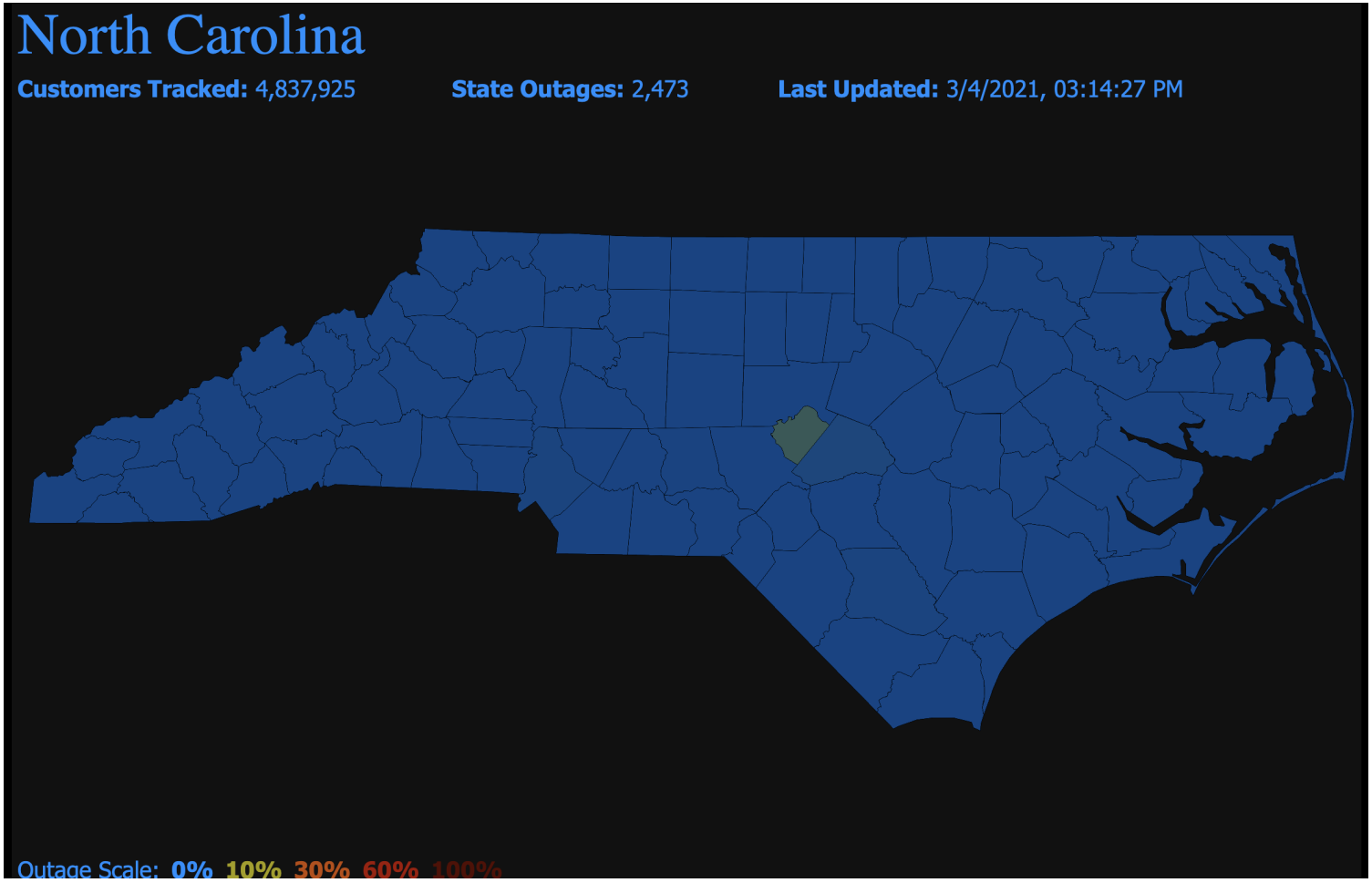
We'll focus on the first option...

> In order for the information to get from their server and show up on a page in your browser, that information had to have been returned in an HTTP response somewhere.
>
> It usually means that you won't be making an HTTP request to the page's URL that you see at the top of your browser window, but instead you'll need to find the URL of the AJAX request that's going on in the background to fetch the data from the server and load it into the page.
>
> Hartley Brody

Live demo

US power outages

https://poweroutage.us/area/state/north%20carolina



Exercise

https://toscrape.com has a website on quotes for scraping. This site-scraping-sandbox provides various endpoints that present different scraping challenges. Try and scrape the first 50 quotes and authors at the following endpoints.

- http://quotes.toscrape.com/scroll
- http://quotes.toscrape.com/js/

First, try using the typical approach with `rvest` to understand what is going on.

References

1. Scraping Sandbox. (2021). http://toscrape.com/.

2. SelectorGadget: point and click CSS selectors. (2021). Selectorgadget.com. https://selectorgadget.com/.

3. yusuzech/r-web-scraping-cheat-sheet. (2021). https://github.com/yusuzech/r-web-scraping-cheat-sheet.