Data Scraping with R

Ryan Metcalf

I didn't know what I was doing. I just knew the outcome I wanted.

Presentation Goals:

1. Share my experience so that you may benefit.
2. Provide you with resources to get you scraping sooner.

Why scrape the web?

1. Acquire structured and text data for analysis in R.
2. Control how you structure and store your data.

What can you scrape?

- City Police Calls For Service: http://www.slmpd.org/cfs.shtml
- Stock Financials: Apple's financials
- Trump's Tweets: https://twitter.com/realDonaldTrump
- Bank Lending: https://fred.stlouisfed.org/release/tables?rid=22&eid=5168
- PDF Table: http://www.usda.gov/oce/commodity/wasde/Secretary_Briefing.pdf

rvest Resources:

- Online Tutorials -
  - R-Bloggers - IMDB site and Wikipedia table
  - Rpubs - website table
  - rvest on GitHub
  - Google it
- Documentation -
  - CRAN and RDocumentation
- Stack Overflow
  - rvest questions
  - You might even find mine here and here.
- StLRUG Slack Channel

How does rvest work?

It traverses the tags of a website and extracts the associated content. So I expect it works on pages that are HTML, XML, CSS, JSON. Basically anything that has tags and nodes.

What does this stuff look like?

If it's webpage (Apple's financials), right-click on content and select page source and find what you're looking for.

If it's an embedded table (http://www.slmpd.org/cfs.shtml), right-click on content and select view frame source.

If you're saying, "This is completely foreign to me. I need help...", please go to the next slide.

CSS Diner

Interactive website that teaches you CSS. It's fantastic. I stopped coding and spent a few hours with it, and I was able to progress much faster.

Here's a few more good resources:

- https://css-tricks.com/how-css-selectors-work/
- http://www.w3schools.com/css/
- http://www.htmlhelp.org/reference/css/

Once you've got reading CSS down it becomes easier to read HTML or XML or JSON next. Also, remember when you're going through these tutorials, you're not going to be a programmer so you don't have to master the content. You just need to know CSS concepts to use rvest.

There's an app for that...

Google Chrome extension SelectorGadget interactive tool that helps you find the elements on a webpage.

Let's parctice with the IMDB tutorial and the Calls For Service page.

I find it best to toggle between the page with SelectorGadget active and the page source. You have to click around with SelectorGadget to get down to exactly the piece(s) of information you are interested in. You can read the source code and put tags/nodes/elements in the SelectorGadget box.

Getting to the content you're interested in can include selecting a larger section of the page and then unselecting elements in that section to leave what you really want.

Pipes vs Functions

You will find many tutorials using pipes, %>%, instead of functions. Using pipes requires the magrittr package.

str1 = "A scratch? Your arm's off."

str2 = "I've had worse."

str1 %>% paste(str2) %>% toupper()

is the same thing as

toupper(paste(str1,str2))

Extracting Tables from PDFs

The tabulizer package has an algorithm which interprets the structure of a table in a PDF. You need to follow the installation instructions exactly. I found the following to get the package installed (requires ghit package).

ghit::install_github(c("leeper/tabulizerjars", "leeper/tabulizer"), INSTALL_opts = "--no-multiarch", dependencies = c("Depends", "Imports"))

Use and documentation of tabulizer is much more sparse than rvest. This tutorial is useful for getting tabulizer you going. As is my included R script.

Tabulizer will import all tables in a PDF and put them in a list.

rvest Odds and Ends

You might need to close your port/connection. I think you can max out the open connections you have and you'll get an error. I haven't bothered to figure this out (I just trapped the error and moved on). If you figure it out, can you post it in the StLRUG channel?

Comprehensive scraping will most likely require you to follow links / clicking through tables. So figuring out how to capture the href link, its CSS element, or setting values and submitting a form are well worth mastering. Depending on the site you are scraping you may need to know one or all of these things to track down data.