Webscraping in R

Erin York

Columbia University Methods Workshop

4/21/17

What we will cover:

- Why scrape?
- R packages
  - rvest
  - Rselenium
- Headless browsers
- Being polite. Being unobtrusive

Webscraping: why do it?

- There is so much data on the Internet!
- Sometimes it's not hosted by a particular website (e.g. search results, tweets)
- Sometimes the people that operate the website won't give it to you
- Sometimes they'll give it to you, but in a bad format (pdf instead of csv)

Do you have to use R?

- No.
- But it works quite well for most basic stuff and even some complicated stuff!
- Python's good too.

Two packages - rvest and RSelenium

rvest

- Very simple to use (a Hadley Wickham number)
- Great for grabbing content and returning it in a friendly format
- Requires a reference url
- r-blogger intro

RSelenium

- Powerful but finicky
- For websites that don't like to be scraped (urls not linked to content, online databases, searches, etc.)
- Run a browser from R!

General Webscraping Things

How do you identify the parts of the webpage you care about?

- css selectors (or xpath, name, etc.)
  - Inspect feature in Chrome
  - But you may need to customize selectors for multiple objects or when running an automated script
  - *Fun Game*

How do you get your other work done when your computer is scraping?

- Headless browsers scrape in the background!
- Phantom JS

And now to the code...

Being Polite

Is scraping ethical? Well, it's in the public domain. But you should be nice about it.

- Consider asking for content before scraping.
- Don't overwhelm servers
- Space out your requests (this might be required anyway, especially with an interactive request)