Introduction to Web Scraping

Cleo O'Brien-Udry

Yale University

11 May 2020

Plan

1. recognizing basic HTML terms
2. scraping text from a web page
3. scraping sequential web pages
4. storing and cleaning scraped text

Tools

- RStudio: packages **rvest**, tidyverse, plyr, data.table
- Github script, slides, additional resources (`https://github.com/cobrienudry/webscrape`)
- Selector Gadget Chrome Extension (`https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjoemdmbfginb?hl=en`)

Basic intuitions

Web scraping: extract data from websites and store on your computer (or an external server)

- Useful for large amounts of data (otherwise you can just copy-paste)
- Helps create novel datasets/make use of data that has not yet been gathered, organized, and analyzed
- Needs to be adjusted for most websites (one size != all)

Research Question: Who speaks for the State Department?

Research Question: Who speaks for the State Department?

Who releases press statements at the State Department? Does the content of press releases differ by speaker? What topics do individual speakers cover?

Research Question: Who speaks for the State Department?

Who releases press statements at the State Department? Does the content of press releases differ by speaker? What topics do individual speakers cover?

Data we need:

- type of statement
- public statement (title and transcript)
- speaker
- date

Research Question: Who speaks for the State Department?

Who releases press statements at the State Department? Does the content of press releases differ by speaker? What topics do individual speakers cover?

Data we need:

- type of statement
- public statement (title and transcript)
- speaker
- date

Use `state.gov/press-releases/`, which has aggregated statements.

Plan

1. **recognizing basic HTML terms**
2. scraping text from a web page
3. scraping sequential web pages
4. storing and cleaning scraped text

Plan

1. recognizing basic HTML terms
2. **scraping text from a web page**
3. scraping sequential web pages
4. storing and cleaning scraped text

Plan

1. recognizing basic HTML terms
2. scraping text from a web page
3. **scraping sequential web pages**
4. storing and cleaning scraped text

Plan

1. recognizing basic HTML terms
2. scraping text from a web page
3. scraping sequential web pages
4. **storing and cleaning scraped text**

Other web scraping topics

- how to scrape a table
- how to scrape an image
- Python for web scraping
- clicking links
- remote servers

Thank you!

cleo.obrien-udry@yale.edu