Web Scraping: Tables, PDFs, OCR

Cleo O'Brien-Udry

Yale University

25 May 2020

Plan

1. short review of html code/basic web-scraping techniques
2. scraping tables from a webpage
3. importing PDFs into R
4. Optical character recognition (pulling text from images into R)

Tools

- RStudio: packages **rvest**, **pdftools**,**tesseract**, **magick**, 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`)

Quick review

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

1. Find web-page
2. Identify location of relevant data on web-page
3. Import into R
4. Clean data
5. Repeat

Research Question: Global Voting Patterns

Research Question: Global Voting Patterns

How have global levels of voting changed over the last 50 years? Which countries show similar patterns of turnout and registration; which show different patterns?

Research Question: Global Voting Patterns

How have global levels of voting changed over the last 50 years? Which countries show similar patterns of turnout and registration; which show different patterns?

Data we need:

- Country voter turnout data
- Covariates (country development indicators, VDEM indicators, etc.)

Research Question: Global Voting Patterns

How have global levels of voting changed over the last 50 years? Which countries show similar patterns of turnout and registration; which show different patterns?

Data we need:

- Country voter turnout data
- Covariates (country development indicators, VDEM indicators, etc.)

Use `https://www.idea.int/data-tools`, which has lots of data.

Plan

1. **short review of html code/basic web-scraping techniques**
2. scraping tables from a webpage
3. importing PDFs into R
4. Optical character recognition (pulling text from images into R)

Plan

1. short review of html code/basic web-scraping techniques
2. **scraping tables from a webpage**
3. importing PDFs into R
4. Optical character recognition (pulling text from images into R)

Plan

1. short review of html code/basic web-scraping techniques
2. scraping tables from a webpage
3. **importing PDFs into R**
4. Optical character recognition (pulling text from images into R)

Plan

1. short review of html code/basic web-scraping techniques
2. scraping tables from a webpage
3. importing PDFs into R
4. **Optical character recognition (pulling text from images into R)**

Other web scraping topics

- Python for web scraping
- clicking links
- remote servers

Thank you!

cleo.obrien-udry@yale.edu