Introduction to Web Scraping with R

Overview

Simon Munzert | IPSDS

Web scraping with R

Web scraping. What? Why?

Web scraping

A.k.a. screen scraping, is the business of

- getting (unstructured) data from the web and
- bringing it into shape (e.g., clean, make tabular format)

Web scraping. What? Why?

Web scraping

A.k.a. screen scraping, is the business of

- getting (unstructured) data from the web and
- bringing it into shape (e.g., clean, make tabular format)

A data analyst's view

- data abundance online
- social interaction online
- services track social behavior
- online data meant for display, not download

Web scraping. What? Why?

Web scraping

A.k.a. screen scraping, is the business of

- getting (unstructured) data from the web and
- bringing it into shape (e.g., clean, make tabular format)

A data analyst's view

- data abundance online
- social interaction online
- services track social behavior
- online data meant for display, not download

A pragmatist's view

- financial resources
- time resources
- reproducibility
- updateability

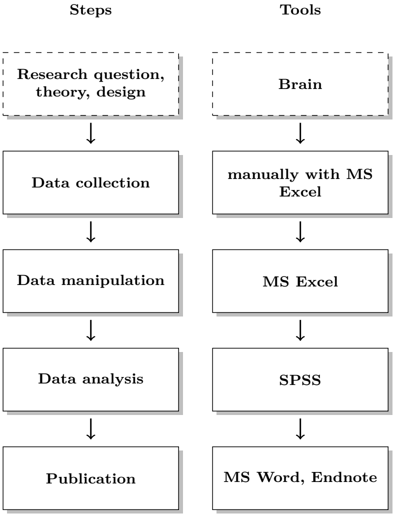
Why R?

Why R?

- free
- open source
- large community
- powerful tools for statistical analysis
- powerful tools for visualization
- flexible in processing all kinds of data/languages
- useful in every step of the workflow

Why R?

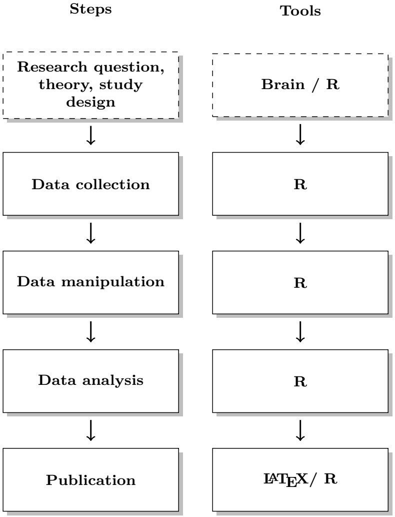
- free
- open source
- large community
- powerful tools for statistical analysis
- powerful tools for visualization
- flexible in processing all kinds of data/languages
- useful in every step of the workflow



Steps	Tools
Research question, theory, design	Brain
Data collection	manually with MS Excel
Data manipulation	MS Excel
Data analysis	SPSS
Publication	MS Word, Endnote

Why R?

- free
- open source
- large community
- powerful tools for statistical analysis
- powerful tools for visualization
- flexible in processing all kinds of data/languages
- useful in every step of the workflow



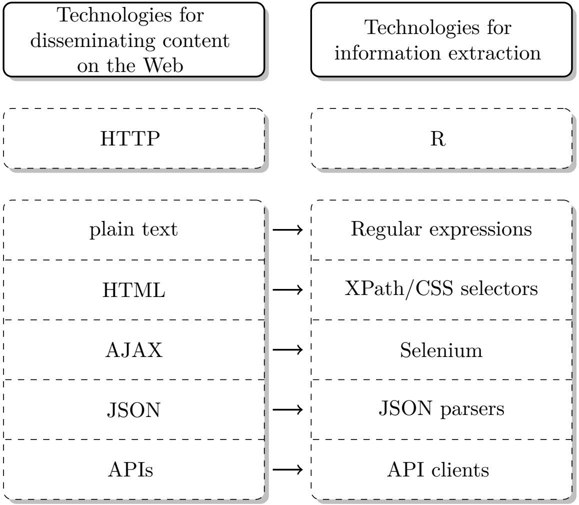
Steps	Tools
Research question, theory, study design	Brain / R
Data collection	R
Data manipulation	R
Data analysis	R
Publication	LaTeX/ R

The philosophy behind web data collection with R

- no point-and-click procedure
- automation of download, parsing, and data extraction procedures
- classical screen scraping
- tapping of web services and APIs
- post-processing of text data
- reproducibility

Technologies of the World Wide Web

Technologies of the World Wide Web



Technologies for disseminating content on the Web

Technologies for information extraction

HTTP

R

plain text ⟶ Regular expressions

HTML ⟶ XPath/CSS selectors

AJAX ⟶ Selenium

JSON ⟶ JSON parsers

APIs ⟶ API clients

Technical Setup

1. make sure that the newest version of R is installed on your computer (available here: https://cran.r-project.org/)
2. install the newest stable version of *RStudio Desktop* (available here: https://www.rstudio.com/products/rstudio/download)
3. follow the instructions on Moodle to install all the necessary packages