Dr. Simon Munzert
Hertie School of Governance
munzert@hertie-school.org
https://simonmunzert.github.io/

A Primer to Web Scraping with R

Cologne Center for Comparative Politics
University of Cologne
November 12, 2018

Outline

The rapid growth of the World Wide Web over the past two decades tremendously changed the way we share, collect and publish data. Firms, public institutions and private users provide every imaginable type of information and new channels of communication generate vast amounts of data on human behavior. What was once a fundamental problem for the social sciences—the scarcity and inaccessibility of observations—is quickly turning into an abundance of data. But how to efficiently collect data from the Internet with statistical software? In this one-day workshop, we will learn how to scrape content from static and dynamic web pages, connect to APIs from popular web services to read out and process user data, and set up automatically working scraper programs.

Schedule

Time	Topic
09:00 - 09:45	Introduction; a first encounter with the Web using R
09:45 - 12:15	Scraping static webpages
12:15 - 13:15	*Lunch break*
13:15 - 14:00	Scraping dynamic webpages
14:00 - 15:30	Tapping APIs
15:30 - 16:30	Scraping ethics and workflow

Prerequisites and Software

I strongly recommend to bring your own laptop. Furthermore, although no special knowledge of web technologies or programming languages is required, participants are expected to have applied knowledge of R. In particular, I assume working knowledge in

- data structures (lists, data frames, vectors) and basic vocabulary
- data manipulation with dplyr
- iterative programming using for loops and the apply() family

In addition, knowledge on regular expressions, how to write own functions, and experience with the tidyverse would be helpful but is not required.

If you plan to extend or refresh your R knowledge before the course you might want to check out *Swirl*. *Swirl* is an R package that lets you learn R interactively on your machine. It's probably better than any other tutorial on the web and as gentle as it gets. You find more information

of how to get the program running at: http://swirlstats.com/. As a recommendation for this course, you might want to work through the course "R Programming" or "Getting and Cleaning Data".

Before the course starts, you should make several preparations:

1. make sure that the newest version of R (available here) is installed on your computer
2. install the newest stable version of *RStudio* (available here)
3. install the needed packages as outlined on the GitHub repository (see below)

Texts and Materials

The workshop is accompanied by the following book:

Munzert, Simon, *Christian Rubba*, *Peter Meißner*, and *Dominic Nyhuis*, 2015: Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining. Chichester: John Wiley & Sons.

Some things have changed since this book was published. I will make sure to cover packages that are most up-to-date in the R environment. In addition, more materials will be made available online on the following GitHub repository:

https://github.com/simonmunzert/rscraping-cologne-2018

Supplemental Literature

Other useful texts on R and web technologies include:

- *Nolan, Deborah*, and *Duncan Temple Lang*, 2014: XML and Web Technologies for Data Sciences with R. New York: Springer.
- *Murrell, Paul*, 2009: Introduction to Data Technologies. Chapman & Hall/CRC.
- *Gandrud, Christopher*, 2015: Reproducible Research with R and RStudio. Chapman & Hall/CRC, 2nd Ed.
- *Wickham, Hadley*, 2014: Advanced R. Chapman & Hall/CRC.
- *Grolemund, Garrett*, and *Hadley Wickham*, 2016: R for Data Science. O'Reilly.

If you want to dig deeper into web and data technologies, you may want to consider the following books:

- *Beaulieu, Alan*, 2009: Learning SQL. Sebastopol, CA: O'Reilly.
- *Cerami, Ethan*, 2002: Web Services Essentials. Sebastopol, CA: O'Reilly.
- *Holdener III, Anthony T.*, 2008: Ajax: The Definitive Guide. Sebastopol, CA: O'Reilly.
- *Gourley, David*, and *Brian Totty*, 2002: HTTP. The Definitive Guide. Sebastopol, CA: O'Reilly.
- *Crockford, Douglas*, 2008: JavaScript: The Good Parts. Sebastopol, CA: O'Reilly.