Dr. Simon Munzert
Institute of Social Science
Humboldt University of Berlin
simon.munzert@gmail.com
http://r-datacollection.com

A Primer to Web Scraping with R

EUI Firenze, May 18/19, 2017

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. This turn of events does not come without problems. For example, traditional techniques for collecting and analyzing data may no longer suffice to overcome the tangled masses of data. One consequence of the need to make sense of such data has been the inception of 'data scientists', who sift through data and are greatly sought after by research and business alike.

But how to efficiently collect data from the Internet, retrieve information from social networks, search engines, and dynamic web pages, tap web services, and, finally, process the collected data with statistical software? We will learn how to scrape content from static and dynamic web pages, connect to APIs from popular web services such as Twitter to read out and process user data, and set up automatically working scraper programs. The sessions are hands-on; we will practice every step of the process with R using various examples.

Schedule

Time	Topic
May 18, 10:00 -11:30	Introduction; a first encounter with the Web using R
May 18, 11:45 -13:00	Scraping with regular expressions
May 18, 14:00 -15:30	Scraping static webpages
May 18, 15:45 -17:00	Advanced scraping of static webpages
May 19, 10:00 -11:30	Scraping dynamic webpages
May 19, 11:45 -13:00	Tapping APIs
May 19, 14:00 -15:30	Gathering social media data
May 19, 15:45 -17:00	Workflow, scraping etiquette, and tricks of the trade

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. Ideally, areas you are familiar with include

- data structures and basic vocabulary
- data import and export
- data manipulation with plyr and dplyr
- writing own functions

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 following packages:
   `pkgs <- c('plyr', 'dplyr', 'stringr', 'lubridate', 'jsonlite', 'httr', 'xml2', 'rvest', 'devtools', 'ggmap', 'networkD3', 'RSelenium', 'pageviews', 'aRxiv', 'twitteR', 'streamR')`
4. install the *Chrome* (from here) and *Firefox* (from here) browsers
5. install *Java* (from here)
6. if you want to follow the code on Twitter mining live in the course, please consult the instructions to connect with Twitter as described here (first section, 'Connecting with Twitter')

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 the most useful software available for scraping purposes in the R environment. In addition, more course materials will be made available online prior to, during, and after the course on a GitHub repository.

Credits

It is possible to earn 10 credits by visiting this course. To do so, you will have to solve a take-home assignment after the course. The solutions have to be handed in as an R script by June 2, 2017.

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.