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

A Primer to Web Scraping and Text Mining with R

Geschwister-Scholl-Institut
LMU Munich
July 16/17, 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. 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 and process the collected (text) data with statistical software? On the first day, 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. The second day is focused on quantitative text analysis. In particular, we learn how to work with text data in R, supervised text classification, and unsupervised topic modeling.

Schedule

Date	Time	Topic
July 16	09:00 - 09:45	Introduction; a first encounter with the Web using R
	09:45 - 10:45	Regular expressions and string manipulation
	10:45 - 11:00	*Coffee break*
	11:00 - 12:15	Scraping static webpages I
	12:15 - 13:30	*Lunch break*
	13:30 - 14:30	Scraping static webpages II
	14:30 - 15:30	Scraping dynamic webpages
	15:30 - 15:45	*Coffee break*
	15:45 - 17:00	Tapping APIs and gathering social media data
July 17	09:00 - 10:15	Scraping ethics and workflow
	10:15 - 10:30	*Coffee break*
	10:30 - 12:15	Working with text data in R
	12:15 - 13:30	*Lunch break*
	13:30 - 14:15	Sentiment analysis and (dis)similarity measures
	14:15 - 15:30	Supervised classification
	15:30 - 15:45	*Coffee break*
	15:45 - 17:00	Unsupervised topic modeling

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. Ideally, areas you are familiar with include

- data structures and basic vocabulary
- data import and export with readr and haven or rio
- data manipulation with 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 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-munich-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.