UNIVERSITÄT LUZERN

Web-Scraping with R

Tutor	Resul Umit, University of Oslo
Organization	Digital Skills, University of Lucerne
Language	English
ECTS-Points	1
Contact	nadia.buehler@unilu.ch
Dates and time	March 4/5 from 9.15AM-4.45PM/9.15AM-1.00PM
Content	**Content** This workshop will guide you in automating the process of extracting data from websites for analysis, in the software environment R. Some data are available in documents, ready to be downloaded from websites. Others are embedded into one or more webpages. This workshop will focus on the latter case, which poses specific challenges to researchers — in terms of extracting, cleaning, and tidying data. You will learn about how to deal with these challenges. To do so, we will work with a number of R packages — most importantly *rvest*, *RSelenium*, and some of the packages in *tidyverse*. **Prerequisites** The workshop assumes basic knowledge of R programming language. If you have no experience with R, but willing to familiarise yourself with it before the workshop, please contact Resul in advance to discuss your options.

Reading list

Steve Pittard, S. (2020). *Web Scraping with R*. Available at
https://steviep42.github.io/webscraping/book/

Schedule

- Part 1. Getting the Tools Ready
  - e.g., installing software
- Part 2. Preliminary Considerations
  - e.g., ethics of web scraping
- Part 3. HTML Basics
  - e.g., elements and attributes
- Part 4. CSS Selectors
  - e.g., selecting an element
- Part 5. Scraping Static Pages
  - e.g., getting text from an element
- Part 6. Scraping Dynamic Pages
  - e.g., clicking to create an element