Humanwissenschaftliche Fakultät

Workshop: Web scraping

Felix Speckmann and Tobias Wingen
Faculty of Human Sciences, Social Cognition Center Cologne

7th December, 2018, 2 pm to 4 pm

In 2016, people generated 16 zettabytes (16*10^16 megabytes) of digital data - an amount which is growing each year. By 2025, this number is estimated to be 10 times higher.
These huge amounts of data are of potential interest to researchers. But even though a large portion of this data is publicly available, it is typically not presented in an easy-to-use format for research. For example, while researchers could easily check the selling price of a specific eBay item manually over many different auctions, checking all prices over the course of a year by hand is virtually impossible. Fortunately, free and easy software solutions can help us with such tasks, through a process called web scraping.

Web scraping is the process of extracting data from websites, often in automated ways. Using automated web scraping scripts, a user can download large amounts of data with relatively little effort. This is especially interesting for researchers, as it allows for the analysis of datasets that are too large to manually prepare. Types of data that can be publicly accessed and extracted are manifold, such as Amazon reviews, online articles from newspapers, ratings from a movie database, or pictures and personal information from blogs. In our web scraping workshop, we will explain how to use this method to access data that might have previously been too unwieldy to work with, effectively supplying researchers with additional approaches within their field of research.

Our workshop will focus on the use of the package “rvest” (https://cran.r-project.org/web/packages/rvest/rvest.pdf) in conjunction with the popular programming language “R”.

The theoretical introduction to web scraping will be accompanied by practical exercises. As part of those exercises, participants will write their own basic scripts to extract data from the web.
By the end of the workshop, participants will have all the relevant knowledge to conduct web scraping projects in their field of research.

Previous knowledge of the software “R” or a similar programming language is helpful but not necessarily required for participation.
However, participants are asked to bring a laptop with them on which the software programs “R”, “R-Studio” and a web browser (Firefox or Chrome) are installed.

Date:
7th December, 2018, 2 pm to 4 pm

Venue:
Seminar room of the graduate school
“City-Passage Lindenthal”, 2nd floor, Dürener Strasse 89, D-50931 Cologne

Registration:
Please register until 30.11.2018 via e-mail to: Graduiertenschule-HF@uni-koeln.de