Web Data Collection with R

Wednesday 11:45 – 13:15, Online

University of Konstanz
Department of Politics and Public Administration
BA/MA Seminar
Winter Semester 2020/21

Sascha Göbel
sascha.goebel@uni-konstanz.de
Office hours by appointment

Course Description

Finding and collecting data is a central yet often challenging and time consuming part of the research process. An increasingly popular source of information for social scientists is offered by the World Wide Web. Every day, governments, companies, and individual persons share and create large quantities of information on the internet, such as administrative records, search engine queries, website traffic, interactions on social networks, etc. The scale at which data is available on the web precludes manual data collection. Fortunately, the ways in which data are stored online often allow for automating the collection process. In this course students will learn how to identify and efficiently collect information from the web. We will cover basic knowledge about web architectures, legal and ethical concerns, the tools required to extract data from static and dynamic web pages, how to tap APIs, and how to process and explore the collected data. Most of the sessions are hands-on using R.

Prerequisites

Though not a strict requirement, prior knowledge of R will be helpful. The course starts with a primer that covers the basic concepts and functionalities of R. Students without a background in R are encouraged to work through Wickham and Grolemund (2016).

Assignments

In preparation of each class, seminar participants are expect to have read the weekly readings prior to the class for which they are assigned.

Starting from session 3, seminar participants are expected to bring one new idea for a research project based on web data to class each week. This should briefly consist of a research question and the potential source of web data required for addressing the question. At the end of each session, some of you will be asked to briefly present their idea, followed by a discussion in class. This serves as preparation for your term paper projects.

To obtain a grade and credits, you have to (1) present your planned research project in class toward the end of the course (max. 10 minutes) and (2) hand in a term paper (and corresponding R code) that formulates a political science-related research question, describes the implementation of the corresponding web data collection, and presents a first exploration of the data with reference to the research question (approx. 10 pages).

Registration

Registration for this seminar is conditional on the commitment to obtain a grade and credits. To register for this seminar, please join the first meeting on November 4th. Please confirm your participation by sending an email to sascha.goebel@uni-konstanz.de including your program of study subsequent to the first meeting and latest by November 9th. I will add you to thre course on ILIAS, where you will find the Zoom link and password for the rest of the sessions. If the number of participants exceeds 20, a lottery will take place.

Software and Course Material

The seminar takes place online and in real time via Zoom. The link for the introductory session is https://zoom.us/j/96517662229. Registered participants will receive the link and password for the other sessions subsequent to the introductory session on November 4th.

The R software environment required for this class can be downloaded for free at https://cran.r-project.org/. RStudio Desktop is recommended as integrated development environment and can be downloaded for free at https://rstudio.com/products/rstudio/download/. R Code and weekly readings are supplied via ILIAS.

Course Schedule

4 Nov. 2020 – Introduction

- Freelon (2018): Sections "Introduction" and "Web Scraping"
- Monroe (2013)
- Nyhuis (2017): Section "Introduction" and "Conceptual Challenges"

11 Nov. 2020 – R Primer

- r-primer.html

18 Nov. 2020 – Basic Web Technologies

- Munzert et al. (2015): Chapter 2 "HTML" and chapter 3 "XML and JSON"

25 Nov. 2020 – Ethical and Legal Considerations

- Salganik (2018): Chapter 6 "Ethics"
- Freelon (2018): Section "Terms of Service"
- Steinert-Threkeld (2018): Chapter 6.2 "Ethics"
- Munzert et al. (2015): Chapter 9.3 "Web Scraping: Good Practice"

2 Dec. 2020 – Collecting Data from Static Websites: XPath

- Munzert et al. (2015): Chapter 4 "XPath" and Chapter 9.2.2 "XPath"
- Nyhuis (2017): Section "Web Scraping"

9 Dec. 2020 – Collecting Data from Static Websites: XPath practice session

16 Dec. 2020 – Regular Expressions and Data Cleaning

- Munzert et al. (2015): Chapter 8 "Regular expressions and essential string functions" and Chapter 9.2.1 "Regular Expressions"

23 Dec. 2020 – Collecting Data from APIs I

- Nyhuis (2017): Section "Application Programming Interfaces (APIs)"
- Munzert et al. (2015): Chapter 5 "HTTP", Chapter 9.1.10 "Retrieving Data from APIs", Chapter 9.1.11 "Authentication with OAuth", and Chapter 9.2.3 "Application Programming Interfaces"

13 Jan. 2021 – Collecting Data from APIs II/Collecting Data from Dynamic Websites

20 Jan. 2021 – Collecting Data from Dynamic Websites

- Munzert et al. (2015): Chapter 6 "AJAX" and Chapter 9.1.9 "Scraping Data from AJAX-Enriched Webpages with Selenium/Rwebdriver"

27 Jan. 2021 – Presentations in class

3 Feb. 2021 – Presentations in class

10 Feb. 2021 – Presentations in class

References

Freelon, Deen. 2018. “Computational Research in the Post-API Age.” *Political Communication* 35(4):665–668.

Monroe, Burt L. 2013. “The Five Vs of Big Data Political Science. Introduction to the Virtual Issue on Big Data in Political Science.” *Political Analysis* 21.

Munzert, Simon, Christian Rubba, Peter Meißner and Dominic Nyhuis. 2015. *Automated Data Collection with R.* Chichester: Wiley.

Nyhuis, Dominic. 2017. Web Data Collection. Potentials and Challenges. In *The SAGE Handbook of Research Methods in Political Science and International Relations*, ed. Luigi Curini and Robert Franzese. London: Sage pp. 387–403.

Salganik, Matthew J. 2018. *Bit By Bit. Social Research in the Digital Age.* Princeton: Princeton University Press.

Steinert-Threkeld, Zachary C. 2018. *Twitter as data.* Cambridge: Cambridge University Press.

Wickham, Hadley and Garrett Grolemund. 2016. *R for Data Science.* Sebastopol, CA: O’Reilly.