GESIS Fall Seminar in Computational Social Science 2023

Syllabus for week 2:
“Automated Web Data Collection with R”

Lecturers:	Allison Koh	Hauke Licht
Affiliation:	Hertie School of Governance	University of Cologne
Email:	koh@hertie-school.org	hauke.licht@wiso.uni-koeln.de

Date: September 18 – September 22, 2023
Time: 9:30-12:30 and 13:30-16:30

About the Lecturers

Allison Koh is a Ph.D. researcher at the Hertie School’s Centre for International Security. She uses text analysis and computational methods to study state repression, transnational activism, and online disinformation. In her research, she uses APIs and web scraping techniques to collect social media data.

Hauke Licht is a post-doctoral researcher at the Cologne Center for Comparative Politics, University of Cologne, and has received his Ph.D. from the University of Zurich. He develops and applies computational text analysis methods to study political communication, electoral competition, and democratic representation. He also has a strong focus on multilingual analyses. In this research, he frequently relies on collecting textual and audio-visual data at scale by applying different web scraping techniques.

Course Description

The increasing availability of large amounts of online data enables new lines of research in the social sciences. Over the past decades, a variety of information – whether election results, press releases, parliamentary speeches, or social media content – has become available online. Although it has become easier and easier to find such information online, its extraction and reshaping into data formats ready for downstream analyses can be challenging. This makes web data collection and cleaning skills essential for researchers. **The goal of this course is to equip participants to gather online data and process it in R for their own research.**

During the course, participants will learn about the characteristics of web data and their use in social science research. The **main learning objective** is that participants acquire the skills to collect (“scrape”) content from different types of web pages as well as from *application programming interfaces* (APIs) such as those hosted by governments, international organizations, and popular newspapers. However, the course will also demonstrate **programming strategies for sustainable and robust social media data extraction** – a skill that has become all the more important since major social media platforms like Facebook and Twitter have discontinued API access to their data in recent years.

The course is hands-on, with daily lectures followed by exercises. In the exercises, participants will apply and practice these methods in R. While we introduce tools and techniques that help with data collection more generally, the focus will be on three common scenarios:

- scraping data from static and dynamic web pages
- automating the collection of information spread over multiple pages, including by navigating dynamic websites (through simulation of clicking and scrolling behavior)
- interacting with APIs to, for example, collect data from government institutions, news publishing companies, or international organizations.

Keywords

web data, web scraping, APIs, automated data collection, R

Course Prerequisites

- willingness to engage with different web technologies
- basic knowledge of the R programming language (incl. the use of loops and writing custom functions): Participants should make sure before the course that they are familiar with the following R programming concepts and techniques:
    - primary data object classes (vectors, lists, and data frames)
    - data wrangling (manipulating vectors, lists, and data frames; reshaping/pivoting data frames),
    - `for` loops and (ideally) functions in the apply/map families (`map_*` in the `purrr` package)
    - writing simple functions
- knowledge of `tidyverse` R packages (recommended)
- We will briefly recap these topics in the afternoon session of the first day of the course. However, if participants are unfamiliar with these topics, we recommend taking the corresponding free online short tutorials in the SICSS R Bootcamp: https://sicss.io/boot_camp. For those who would like a primer or refresher in R, we recommend taking the online workshop Introduction to R” which takes place from 05-07 September 2023.

Target Group

Participants will find the course useful if:

- they want to collect larger amounts of web data from web pages or APIs
- they want to learn about best practices in automated web data collection
- they want to improve pre-existing web scraping skills by deepening their understanding of common web technologies and learning more about the process of developing robust web scrapers

Participants will be asked to indicate their prior experience with web scraping, their research interests and potential web scraping-related project ideas in a pre-course survey. Based on this survey, the instructors will attempt to include examples in the afternoon tutorial sessions that match participants' research interests and project ideas.

Course and Learning Objectives

By the end of the course participants will:

- know the most important characteristics of web data, including web page content and social media data
- understand of a variety of scraping scenarios: static pages, dynamic pages, APIs, social media data
- be able to parse, clean and process data collected from the web
- be able to write reproducible and robust code for web scraping tasks

Organizational Structure of the Course

The course will be organized as a mixture of lectures and exercise sessions. We will switch between lectures and exercises throughout the morning and afternoon sessions of the course. In the lecture sessions, we will focus on explaining core concepts and methods in web scraping. In the exercise sessions, participants will apply their newly acquired knowledge. Both instructors will be available to answer questions and provide guidance during the entire course.

Software and Hardware Requirements

- Participants should bring their own laptops for use in the course.
- RStudio (or a comparable R IDE)
- the *Google Chrome* web browser

- required R packages (a complete list of packages will be provided before the course)
    - for web scraping: `rvest`, `RSelenium`, `httr`
    - for data processing: `dplyr`, `tidyr`, `purrr`, `stringr`

Recommended Literature to Look at in Advance

A list of required and suggested readings and online resources will be distributed four weeks before the course. All data that will be used throughout the course will be provided by the instructors before course sessions.

While we work with prepared examples and data throughout the course, we encourage participants to develop ideas on using web data in their research. For this purpose, we recommend the following reading as preparation: Salganik, Matthew (2017) *Bit by Bit: Social Research in the Digital Age.* Princeton University Press. (esp. Chapters 1 & 2)

Day-by-day Schedule and Literature

Day 1: Introduction

We will cover what web scraping is and how it can be used in social science and digital humanities research. Participants will be asked to share their expectations of the course and how they plan to use web scraping in their research. We will then introduce the most fundamental concepts, including APIs, the XML and HTML formats, and how websites are commonly organized.

In the afternoon tutorial session, we will first ensure that all participants have a working setup. We will then have a series of coding exercises designed to ensure that all participants are comfortable with basic R programming concepts and techniques (see *Prerequisites* section above).

Recommended readings (only for inspiration, not required):

Golder, S. A., & Macy, M. W. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*, *40*(1), 129–152. https://doi.org/10.1146/annurev-soc-071913-043145

Strohmaier, M., & Wagner, C. (2014). Computational Social Science for the World Wide Web. *IEEE Intelligent Systems*, *29*(5), 84–88. https://doi.org/10.1109/MIS.2014.80

Nagler, J., & Tucker, J. A. (2015). Drawing Inferences and Testing Theories with Big Data. *PS: Political Science & Politics*, *48*(1), 84–88. https://doi.org/10.1017/S1049096514001796

Jungherr, A., & Theocharis, Y. (2017). The empiricist's challenge: Asking meaningful questions in political science in the age of big data. *Journal of Information Technology & Politics*, *14*(2), 97–109. https://doi.org/10.1080/19331681.2017.1312187

Margetts, H. (2017). The Data Science of Politics. *Political Studies Review*, *15*(2), 201–209. https://doi.org/10.1177/1478929917693643

Wallach, H. (2018). Computational social science ≠ computer science + social data. *Communications of the ACM*, *61*(3), 42–44. https://doi.org/10.1145/3132698

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170

Day 2: Scraping static websites

On day 2, we will introduce how to web scrape *static* websites. Building on our general discussion of HTML (Day 1), we will cover how to systematically extract web data by introducing the *CSS selector* and *Xpath* methods.

In practical applications, we will use the `rvest` R package to show how to *(i)* extract data (text, hyperlinks, tables, images, and other media, as well as metadata) from web pages and *(ii)* how to automatically navigate between and scrape multiple pages of a website.

In the afternoon tutorial session, participants will learn how to apply this knowledge to different web pages.

Recommended readings:
CSS Selectors Reference. https://www.w3schools.com/CSSref/css_selectors.asp

Day 3: Scraping of dynamic websites
On the third day of the course, we will go one step further and discuss how to scrape *dynamic* websites. We will first explain what makes a page "dynamic" and show how to recognize dynamic web elements in the wild.

We will then introduce the `RSelenium` package and show how it enables systematic interaction with dynamic web elements. This will include how to set up a web driver in R (Google Chrome), how to click on web elements (e.g., to unfold/collapse drop-down elements) in an automated way, how to navigate dynamic elements (e.g., accordion elements), how to switch between windows (e.g., a main page and a pop-up), and how to automatically download files. In the afternoon, participants will have the opportunity to practice these skills.

Recommended readings:
Street, A., Murray, T. A., Blitzer, J., & Patel, R. S. (2015). Estimating Voter Registration Deadline Effects with Web Search Data. *Political Analysis*, *23*(2), 225–241. https://doi.org/10.1093/pan/mpv002
Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support. *Social Science Computer Review*, *35*(3), 336–356. https://doi.org/10.1177/0894439316631043
Tjaden, J. D., Schwemmer, C., & Khadjavi, M. (2018). Ride with Me—Ethnic Discrimination, Social Markets, and the Sharing Economy. *European Sociological Review*, *34*(4), 418–432. https://doi.org/10.1093/esr/jcy024
Pradel, F. (2021). Biased Representation of Politicians in Google and Wikipedia Search? The Joint Effect of Party Identity, Gender Identity and Elections. *Political Communication*, *38*(4), 447–478. https://doi.org/10.1080/10584609.2020.1793846

Day 4: APIs & collecting social media data
Building on the content discussed during the previous days, we will deepen participants' understanding of APIs, discussing common APIs for data sharing. Using the Mastodon API as an example, we will then show how to use the `rtoot` package to query social media data. This part of the session will also include a primer on authentication, pagination, API rate limits, and ethics.

To enable participants to potentially also interact with APIs for which no R package exists (yet), we will show how to send requests to APIs using the `httr` R package using the example of the *Dad Jokes* API (https://dadjokes.io). In the context of this example, we will also explain the JSON format – the data format commonly returned by APIs.

In the afternoon tutorial session, participants will learn how to apply this knowledge with a small project using the *News API* (https://newsapi.org).

Recommended readings:
Burgess, J., & Bruns, A. (2015). Easy data, hard data: The politics and pragmatics of Twitter research after the computational turn. *Compromised Data: From Social Media to Big Data*, *95*.
González-Bailón, S., & Wang, N. (2016). Networked discontent: The anatomy of protest campaigns in social media. *Social Networks*, *44*, 95–104. https://doi.org/10.1016/j.socnet.2015.07.003
King, G., Schneer, B., & White, A. (2017). How the news media activate public expression and influence national agendas. *Science*, *358*(6364), 776–780. https://doi.org/10.1126/science.aao1100
Munger, K. (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5
Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506
Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political Communication*, *35*(1), 50–74. https://doi.org/10.1080/10584609.2017.1334728

Bruns, A. (2019). After the ‘APIcalypse’: Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447
Puschmann, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information, Communication & Society*, *22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

Day 5: Advanced topics
On the last day, we will begin with a recap of what we have learned during the previous four days. Specifically, we will provide a condensed, systematic overview of the common programming techniques applied to automate web data collection from static websites, dynamic websites, and APIs, respectively.

We will then walk through some advanced topics in web scraping, including web sessions, user agents, proxies, login, and other topics participants might be interested in. We will also discuss tools for the advanced parsing of webpage content, including regular expressions.

Recommended readings:
Sepulveda, M. V., & Beasley, W. (2023). *CRAN Task View: Web Technologies and Services*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/view=WebTechnologies