Automated Data Collection of Web and Social Data

Dominic Nyhuis
Goethe University Frankfurt

Course description

An increasingly vast wealth of data is freely available on the web -- from election results and legislative speeches to social media posts, newspaper articles, and press releases, among many other examples. Although this data is easily accessible, in most cases it is available in an unstructured format, which makes its analysis challenging. The goal of this course is to gain the skills necessary to automate the process of downloading, cleaning, and reshaping web and social data using the R programming language for statistical computing. We will cover all the most common scenarios: scraping data available in multiple pages or behind web forms, interacting with APIs and RSS feeds such as those provided by most media outlets, collecting data from Facebook and Twitter, extracting text and table data from PDF files, and manipulating datasets into a format ready for analysis. The course will follow a "learning-by-doing" approach, with short theoretical sessions followed by "coding challenges" where participants will need to apply new methods.

The first part will offer an introduction to the course and then dive into the basics of webscraping; that is, how to automatically collect data from the web. This session will demonstrate the different scenarios for webscraping: when data is in table format (e.g. Wikipedia tables or election results), when it is in an unstructured format (e.g. across multiple parts of a website), and when it is behind a web form (e.g. querying online databases). The tools available in R to achieve these goals – the *rvest* and *RSelenium* packages – will be introduced in the context of applied examples in the social sciences

NGOs, public institutions, and social media companies increasingly rely on Application Programming Interfaces (API) to give researchers and web developers access to their data. The second day will focus on how we can develop our own set of structured http requests to query an API. We will start by discussing the components of an API request and how to build our own authenticated queries using the *httr* package in R, taking the New York Times API (to query newspaper articles) as our running example. Then, we will learn how to use the most popular R packages to query social media APIs: *rtweet*, *streamR*, and *Rfacebook*. These packages allow researchers to collect tweets filtered by keywords, location, and language in real time, and to scrape public Facebook pages, including likes and comments. The process of collecting and storing the data will be illustrated with examples from published research on social media.

After the course, students will have an advanced understanding of the web and social data available for social science research, and will be equipped with the technical skills necessary to collect and clean such datasets on their own.

Software

The course will use the open-source software R, which is freely available for download at https://www.r-project.org/ . We will interact with R through RStudio Server. Each participant will be given access to the server used for the course, where all packages will be already installed, and which can be accessed through any web browser. Therefore, participants are not required to install any software before the course.

Prerequisites

The course will assume familiarity with the R statistical programming language at the level of the introductory course offered by the summer school. Participants should be able to know how to read datasets in R, work with vectors and data frames, write functions and loops, and run basic statistical analyses, such as linear regression.

Schedule

July 5th, 2018

Time	Topic
14.00-14.30	Introductions and course overview
14.30-15.00	Scraping the web: motivation, general rules, scenarios and strategies
15.00-15.30	Application: scraping web data in table format
15.30-16.00	Challenge 1: scraping tables from Wikipedia
16.00-16.30	Break
16.15-17.00	Application: scraping web data in unstructured format using *rvest* and *RSelenium*
17.15-18.00	Challenge 2: scraping a newspaper website

July 6th, 2018

Time	Topic
14.00-14.45	Application Programming Interfaces (APIs): definition, logic, how to build an API request
14.45-15.15	Application: scraping the New York Times API
15.15-15.45	Challenge 3: interacting with the Clarifai API
15.45-16.15	Break
16.15-16.45	Social media data: Twitter and Facebook
16.45-17.30	Application: Downloading Twitter and Facebook data from their APIs
17.30-18.00	Challenge 4: building a dataset of social media posts by politicians

References

Course largely based on:

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063-1064.

Barberá, P., & Rivero, G. (2014). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 0894439314558836.

Wickham, H., & Grolemund, G. (2016). *R for Data Science.* O’Reilly

Ravindran, S. K., & Garg, V. (2015). *Mastering social media mining with R*. Packt Publishing Ltd.

Short bio

Dominic Nyhuis (PhD in Political Science, University of Mannheim) is a Researcher at the Department of Social Sciences, Goethe University Frankfurt. Prior to this, he was affiliated with the Universities of Vienna, Mannheim, and Mainz. His research focuses on comparative legislative studies with a particular emphasis on questions of representation, small-area policy preferences, and municipal politics. He is also interested in techniques for automated data collection and quantitative methods for the social sciences. His work has been published in high-impact journals such as Electoral Studies, Political Science Research and Methods, and Party Politics. More information is available at http://www.fb03.uni-frankfurt.de/59181596/.