Automated Data Collection Using R

Max Mantei

Short Bio

I did my B.A. in Political Science and Economics at the University of Bremen (Germany) and Aarhus University (Denmark). After that I received my M.Sc. in Politics, Economics, and Philosophy from the University of Hamburg (Germany) and began my Ph.D. at the Graduate School for the Economic Analysis of the Internationalization of the Law at the Institute of Law and Economics at Hamburg University.

Throughout my studies I had several research assistant roles and worked as a freelancer doing data analysis for local businesses. I have been using R since 2011, but also enjoy programming in Python. My main interests are Bayesian Statistics and Machine Learning/Deep Learning.

Course Description

This course will teach you the basics of R and programming in R. It will teach how we can use R to work with data using the powerful tidyverse ecosystem. You will learn how to plot your data with the ggplot2 library. Finally, we will cover the basics of text mining and working with text data in R using the tidytext package. This course is concise and we will not have the time to cover any advanced Machine Learning or Natural Language Processing. However, after this course you should be able to start your own test-data driven projects.

Course Outline

Part 0: Introduction and Set-Up

- Introduction and Outline
- Installing R and RStudio

Part 1: Objects and Data Structures I

- Numbers and Arithmetics
- Strings and Variables (Assignment)
- Booleans and Relational Operators
- Vectors and Matrices

Part 2: Data I/O

- Read user input and text files

- Read and write csv files
- Packages in R
- Other

Part 3 Objects and Data Structures II

- Data Frames
- Lists

Part 4: Programming Basics

- Conditional Logic
- Loops
- Functions

Part 5: Data transformation in the tidyverse

- What’s tidy data?
- What’s in the tidyverse?
- The pipe %>%
- Data transformation with dplyr
  - mutate
  - filter
  - group_by
  - summarize
  - arrange
- Tidy data with tidyr
  - Cleaning with drop_na
  - Using pivot
  - Columns: separate and unite
- More: Factors with forcats and strings with stringr

Part 6: Plotting data with ggplot2

- Basics of ggplot2
- Histograms and count plots
- Scatterplots and line plots
- Plot more with facets

Part 7: Basic Statistics/Econometrics (optional)

- R’s formula syntax
- Setting up model matrices with model.matrix
- Linear regression
- Generalized Linear Models (GLM) in R

- Overview: Packages for statistical modeling
    - IV and FE panel models in R (AER, plm)
    - Linear Mixed Models and Generalized Linear Mixed Models in R (lme4)
    - Generalized Additive Models (GAM) in R (mgcv)
    - Bayesian models (brms)

Part 8: Working with Strings and Texts

- Strings and the stringr package
- Text Mining with tidytext