FAU – School of Business, Economics and Society
Chair of Statistics and Econometrics

Seminar: Web Scraping

Summer term 2024 – Syllabus

Overview: Web scraping, the automated extraction of data from websites, has emerged as a powerful tool for gathering a variety of structured and unstructured data types from the internet. In today's digital age, vast amounts of valuable information are embedded within web pages, ranging, for instance, from real estate listings and customer reviews to job postings or social media data.

Web scraping enables individuals and businesses to harness this wealth of data, allowing them to use the information for a wide range of research applications and business cases. Nowadays, the use of web scrapers for tasks such as real-time data monitoring or market research plays an important role and is therefore an integral part of the day-to-day business of many companies and the foundation of many research projects in economics.

This hands-on seminar offers an opportunity to learn how to scrape data from websites in R. During the first part of the semester, students will teach each other the necessary tools and skills by means of seminar presentations of roughly 35 minutes. During the second part of the semester, students implement a web scraper for a project of their choice for scraping data that allows them to analyze web data in the context of a chosen research question (possible examples: product price comparisons, analysis of text from speeches of monetary policy makers, features of job vacancies).

Instructors:

Prof. Dr. Jonas Dovern
Office: LG 4.169
E-mail: jonas.dovern@fau.de
Office hours: by appointment

Johannes Frank
Office: LG 4.174
E-mail: johannes.jf.frank@fau.de
Office hours: by appointment

Seminar meetings: Tuesday, 11:30h – 13:00h (23.04. to 25.06.), LG 0.224

Registration: Please send an e-mail to wiso-oekonometrie@fau.de if you want to register for this seminar. **Deadline for registration is 15. March 2024!**

- Please state clearly i) your name and surname, ii) your student ID, iii) your study program, and iv) your first, second, and third choice for a topic that you want to work on in the seminar (see list of topics below).
- **Please register as early as possible – we distribute topics according to a first-come-first-served principle** (potentially you can present a topic with groups of two) and the number of students is restricted to a maximum of 16. We'll make sure that basic topics which are the most important to enable you to implement a web-scraping project are distributed in any case.

- **You will have to register officially for the seminar on campo at the beginning of the semester during a special registration period (I'll send a reminder!). After that you will not be able to unsubscribe from the seminar anymore.**

Grading: Your grade for the module depends on the seminar paper that documents your web scraping application (60 %) and the presentation of your topic (40%). We will distribute information about the formal requirements regarding length and layout of the seminar papers during the first seminar meeting.

Deadline for seminar papers: Your seminar paper is due on **31. August 2024 at 18:00h**.

List of topics:

1. Introduction to Web Scraping
    a. Basics of web scraping
    b. Importance in economics/marketing research
2. Introduction to html syntax
3. Introduction to CSS selectors
4. Navigating html with XPath expressions
5. Ethical and legal aspects of web scraping
    a. Responsible scraping practices
    b. Legal implications and ethical considerations
6. Case studies of ethical and unethical web scraping
7. The rvest package
8. Scraping single static websites
9. Dealing with pagination and multiple pages
10. Understanding and handling dynamic websites
11. The RSelenium package
12. Introduction to APIs
13. Working with APIs in R
14. Post-processing web-scraping results using the dplyr package
15. Extracting data using regular expressions (incl. use of ChatGPT)
16. Working with RMarkdown to produce reports

Course requirements: Course participants are required to …

- **Attend.** Students can only pass the course if they attend the seminar because discussions of the presentations are an essential part of the seminar. Students will also need the skills conveyed during the presentations to implement their own web-scraping applications.
- **Be interested in data handling.** The seminar includes the practical application of web-scraping methods to a project of your choice.
- **Register on StudOn.** We will make course material available through the course website on StudOn (available from mid-April). We will also make all announcements using this platform.