IPSDS
INTERNATIONAL PROGRAM IN SURVEY & DATA SCIENCE

SPONSORED BY THE
Federal Ministry of Education and Research

AUFSTIEG DURCH BILDUNG >>
OFFENE HOCHSCHULEN

Web Scraping and APIs
SURV 736
1 credit/2 ECTS
Summer 2020, sec. 1

Instructor

Simon Munzert, PhD, munzert@hertie-school.org

Video lecture by Simon Munzert, PhD

Short Course Description

The short course provides a condensed overview of web technologies and techniques to collect data from the web in an automated way. To this end, students will use the statistical software R. The course introduces fundamental parts of web architecture and data transmission on the web. Furthermore, students will learn how to scrape content from static and dynamic web pages and connect to APIs from popular web services. Finally, practical and ethical issues of web data collection are discussed.

Course and Learning Objectives

By the end of the course, students will…

- have an overview of state-of-the-art research that draws on web-based data collection,
- have a basic knowledge of web technologies,
- be able to assess the feasibility of conducting scraping projects in diverse settings,
- be able to scrape information from static and dynamic websites as well as web APIs using R, and
- be able to tackle current research questions with original data in their own work.

Prerequisites

Students are expected to be familiar with the statistical software R. Besides base R, knowledge about the "tidyverse" packages, in particular, dplyr, plyr, magrittr, and stringr, are of help. If you are familiar with R but have no experience in working with these packages, the best place to learn them is the primary reading "R for Data Science".

Class Structure and Course Concept

This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video-recorded lectures and reading the required literature for each unit prior to participating in mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching videos, participating in online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb you can expect to spend approximately 3h/week on in-class-activities and 9 hours per week on out-of-class activities (preparing for class, readings, assignments, projects, studying for quizzes and exams). Therefore, the workload in all courses will be approximately 12h/week. This is a 1-credit/2 ECTS course that runs for 4 weeks. Please note that the actual workload will depend on your personal knowledge.

Mandatory Weekly Online Meetings

Wednesday, 8:00 AM EDT/2:00 PM CEST, starting June 3, 2020

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on https://www.elms.umd.edu/. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In addition, students are encouraged to post questions about the materials covered in the videos and readings of the week to the "Place to post your questions" forum before the meetings (deadline for posting questions to the forum is Tuesdays, 8:00 AM EDT/2:00 PM CEST before class).

Students have the opportunity to use the Zoom meeting room set up for this course to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Students are encouraged to post the times that they will be using the room to the course website forum to avoid scheduling conflicts. Students are not required to use Zoom and can of course use other online meeting platforms such as Google Hangout or Skype.

Grading

Grading will be based on:

- participation in discussion during the weekly online meetings and submission of questions via the forum (deadline: Tuesday, 8:00 AM EDT/2:00 PM CEST before class) demonstrating understanding of the required readings and video lectures (10% of grade)
- weekly quizzes that check factual knowledge about the course topics (30% of the grade)
- weekly assignments that require students to implement and practice scraping techniques in R (60% of grade)

A+	*100 - 97*
A	*96 - 93*
A-	*92 - 90*
B+	*89 - 87*
B	*86 - 83*
B-	*82 - 80*
Etc.	

The grading scale is a base scale recommended by the IPSDS. Variations for grading on a scale are at the discretion of the instructor.

Dates of when assignment will be due are indicated in the syllabus. Extensions will be granted sparingly and are at the instructor's discretion.

Technical Equipment Needs

The learning experience in this course will mainly rely on the online interaction between students and the instructor during the weekly online meetings. Therefore we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.

Long Course Description

The rapid growth of the World Wide Web over the past two decades tremendously changed the way we share, collect, and publish data. What was once a fundamental problem for the social sciences - the scarcity and inaccessibility of observations - is quickly turning into an abundance of data. In addition to classical forms of data collection (e.g., surveys, lab or field experiments), a variety of new possibilities to collect original data has emerged. The internet offers a wealth of opportunities to learn about public opinion and social behavior. Data from social networks, search engines or web services open avenues for new ways of measuring human behavior and preferences in previously unknown velocity and variety. Fortunately, the open source programming language R provides advanced functionality to gather data from virtually any imaginable data source on the Web - via classical screen scraping approaches, automated browsing, or by tapping APIs. This allows researchers to stay in one programming environment in the processes of data collection, tidying, analysis, and publication.

This short course will provide an overview of web technologies fundamental to gather data from internet resources, such as HTML, CSS, XML, and JSON. Furthermore, students will learn how to scrape content from static and dynamic web pages using state-of-the-art packages of the R software. Also, they will learn how to use R to connect to APIs from popular web services to read out ready-made data. Finally, practical elements of the web scraping workflow as well as ethical issues of web data collection are discussed. The course will have a strong practical component; sessions will feature live R coding and students are expected to practice every step of the process with R using various examples.

Readings

Primary readings will be from the following volume:

Simon Munzert, Christian Rubba, Peter Meißner, and Dominic Nyhuis, 2015: *Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining*. Chichester: John Wiley & Sons.

Moreover, there are online resources listed for each unit that can be used to further explore or practice the treated topics.

Academic Conduct

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at

https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf (University of Maryland) and

https://www.uni-mannheim.de/en/research/good-research-practice/ (University of Mannheim).

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Accommodations for Students with Disabilities

In order to receive services, students at the University of Maryland must contact the Accessibility & Disability Service (ADS) office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office at 301.314.7682; https://www.counseling.umd.edu/ads/.

Students at the University of Mannheim should contact the Commissioner and Counsellor for Disabled Students and Students with Chronic Illnesses at http://www.uni-mannheim.de/studienbueros/english/counselling/disabled_persons_and_persons_with_chronic_illnesses/.

Course Evaluation

In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.

Class Schedule

Please note that assignments and dates are subject to change. Information (e.g., articles and assignments) posted to the course website supersedes the information noted here.

Unit 1: Introduction – Web Technologies

Video lectures: available Wednesday, May 27, 2020

01.01	Introduction
01.02	Overview
01.03	Case study
01.04	HTML
01.05	Regular expressions: basics
01.06	Regular expressions in R
01.07	String manipulation
01.08	Summary

Online meeting: Wednesday, June 3, 8:00 AM EDT/2:00 PM CEST

Assignment 1 (4 problems to be solved with R): due Tuesday, June 9, 7:00 AM EDT/1:00 PM CEST

Quiz 1: due Tuesday, June 9, 7:00 AM EDT/1:00 PM CEST

Readings:
From textbook:
Munzert et al. (2015): Chapters Preface, 1, 2, 8

Further online resources:

- A regular expressions cheat sheet: https://www.rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf
- A brief tutorial to regular expressions in R: https://stat545.com/block022_regular-expression.html
- Interactive ways to learn regular expressions: http://play.inginf.units.it/#/, https://regexcrossword.com/, https://alf.nu/RegexGolf/

Unit 2: Scraping static webpages

Video lecture: available Wednesday, June 3, 2020

02.01 Inspecting the HTML tree
02.02 XPath I
02.03 XPath II
02.04 Scraping HTML tables
02.05 Using SelectorGadget
02.06 The scraping workflow
02.07 Scraping multiple pages
02.08 Summary

Online meeting: Wednesday, June 10, 8:00 AM EDT/2:00 PM CEST

Assignment 2 (4 problems to be solved with R): due Tuesday, June 16, 7:00 AM EDT/1:00 PM CEST

Quiz 2: due Tuesday, June 16, 7:00 AM EDT/1:00 PM CEST

Readings:
From textbook:
Munzert et al. (2015): Chapters 3 (3.1-3.4), 4, 9 (9.1.1-9.1.5; 9.2.1-9.2.2)

Unit 3: Scraping dynamic webpages and good practice

Video lectures: available Wednesday, June 10, 2020

03.01 Dynamic webpages
03.02 AJAX technologies
03.03 The Selenium software: basics
03.04 Scraping case study
03.05 Legal issues
03.06 Good practice of web scraping
03.07 Summary

Online meeting: Wednesday, June 17, 8:00 AM EDT/2:00 PM CEST

Assignment 3 (4 problems to be solved with R): due Tuesday, June 23, 7:00 AM EDT/1:00 PM CEST

Quiz 3: due Tuesday, June 23, 7:00 AM EDT/1:00 PM CEST

Readings:
From textbook:
Munzert et al. (2015): Chapters 6, 9 (9.1.9, 9.3)

Further online resources:

- On the ethics of web scraping: http://robertorocha.info/on-the-ethics-of-web-scraping/
- The state of the law on data scraping: http://blog.galkinlaw.com/weblaw-scout-blog/legality-of-data-scraping
- Web scraping and crawling are perfectly legal, right? https://benbernardblog.com/web-scraping-and-crawling-are-perfectly-legal-right/

Unit 4: Tapping APIs

Video lectures: available Wednesday, June 17, 2020

04.01 APIs
04.02 API clients
04.03 Basic JSON
04.04 Accessing APIs from scratch
04.05 API authentication
04.06 Summary

Online meeting: Wednesday, June 24 8:00 AM EDT/2:00 PM CEST

Assignment 4 (4 problems to be solved with R): due Tuesday, June 23, 7:00 AM EDT/1:00 PM CEST

Quiz 4: due Tuesday, June 23, 7:00 AM EDT/1:00 PM CEST

Readings:
From textbook:
Munzert et al. (2015): Chapters 5 (5.1), 9 (9.1.10, 9.1.11, 9.2.3)

Further online resources:

- The ROpenSci project: http://ropensci.org https://github.com/ropensci/opendata
- The CRAN Task View of Web Technologies: https://cran.r-project.org/web/views/WebTechnologies.html