Introduction to Web Scraping with R
SURV 736
1 credit/2 ECTS
Spring 2018, sec. 1

Instructor(s)
Simon Munzert, munzert@hertie-school.org

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 and then "attending" mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor. Just like in an on-site course, homework—often in form of small R coding tasks—will be assigned and graded and there will be a final exam at the end of the course.

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, for each credit offered by a course, students can expect to spend one hour per week on in-class activities and three hours per week on out-of-class activities over the span of a full 12-week term. This is a 1-credit course that runs for 4 weeks. Hence, the total average workload is about 12 hours per week.

Mandatory Weekly Online Meetings

Tuesday, 11 AM (EST)/5 PM (CET), starting March 6
Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on http://jpsmonline.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 Mondays, 11 AM (EST)/5 PM (CET) 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: Monday, 11 AM (EST)/5 PM (CET) 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)

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.

Additional required and recommended readings will be made available on the course website:
http://jpsmonlinedev.umd.edu/

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

http://www.graduate.umaryland.edu/policies/misconduct.html (University of Maryland) and

https://www.uni-mannheim.de/1/english/research/Good%20Research%20Practice/141119-Satzung%20wiss%20FV%20Senat_en.pdf (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 Disability Support Services (DSS) office to register in person for services.

Please call the office to set up an appointment to register with a DSS counselor. Contact the DSS office at 301.314.7682; http://www.counseling.umd.edu/DSS/.

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

Please keep in mind that daylight saving time begins in the US on March 11, 2018 and clocks are turned forward 1 hour. Daylight saving time begins in the EU on March 25, 2018. Therefore, look carefully at the times of meetings and deadlines!

Online meeting (Simon Munzert): Tuesday, March 6, 11 AM (EST)/5 PM (CET)

Assignment 1 (4 problems to be solved with R): due Monday, March 12, 12 AM (EST)/5 PM (CET)

Quiz 1: due Monday, March 12, 12 AM (EST)/5 PM (CET)

Video lectures (Simon Munzert): available online Tuesday, February 27

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

Readings:

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

Unit 2: Scraping static webpages

Online meeting (Simon Munzert): Tuesday, March 13, 12 AM (EST)/5 PM (CET)

Assignment 2 (4 problems to be solved with R): due Monday, March 19, 12 AM (EST)/5 PM (CET)

Quiz 2: due Monday, March 19, 12 AM (EST)/5 PM (CET)

Video lecture (Simon Munzert): available online Tuesday, March 6

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

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

Please keep in mind that daylight saving time begins in the US on March 11, 2018. Daylight saving time begins in the EU on March 25, 2018 and clocks are turned forward 1 hour. Therefore, look carefully at the times of meetings and deadlines!

Online meeting (Simon Munzert): Tuesday, March 20, 12 AM (EST)/5 PM (CET)

Assignment 3 (4 problems to be solved with R): due Monday, March 26, 11 AM (EST)/5 PM (CET)

Quiz 3: due Monday, March 26, 11 AM (EST)/5 PM (CET)

Video lectures (Simon Munzert): available online Tuesday, March 13

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

Readings:

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

Unit 4: Tapping APIs

Online meeting (Simon Munzert): Tuesday, March 27, 11 AM (EST)/5 PM (CET)

Assignment 4 (4 problems to be solved with R): due Monday, April 9, 11 AM (EST)/5 PM (CET)

Quiz 4: due Monday, April 2nd, 11 AM (EST)/5 PM (CET)

Video lectures (Simon Munzert): available online Tuesday, March 20

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

Readings:

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

Note: Student access to the course website will be revoked two weeks after the final exam.

Please keep in mind that daylight saving time begins in the US on March 11, 2018. Daylight saving time begins in the EU on March 25, 2018. Clocks are turned forward 1 hour. Therefore, look carefully at the times of meetings and deadlines!

	Unit 1	**Unit 2**	**Unit 3**	**Unit 4**
Video available	Tuesday, February 27, 2018	Tuesday, March 6, 2018	Tuesday, March 13, 2018	Tuesday, March 20, 2018
Online meeting	Tuesday, March 6, 2018, 11 AM (EST)/5 PM (CET)	Tuesday, March 13, 2018, 12 AM (EST)/5 PM (CET)	Tuesday, March 20, 2018, 12 AM (EST)/5 PM (CET)	Tuesday, March 27, 2018, 11 AM (EST)/5 PM (CET)
Online quiz due	Monday, March 12, 2018, 12 AM (EST)/5 PM (CET)	Monday, March 19, 2018, 12 AM (EST)/5 PM (CET)	Monday, March 26, 2018, 11 AM (EST)/5 PM (CET)	Monday, April 2nd, 2018, 11 AM (EST)/5 PM (CET)
Assignment due	Monday, March 12, 2018, 12 AM (EST)/5 PM (CET)	Monday, March 19, 2018, 12 AM (EST)/5 PM (CET)	Monday, March 26, 2018, 11 AM (EST)/5 PM (CET)	Monday, April 2nd, 2018, 11 AM (EST)/5 PM (CET)
Final exam due				---