University of St.Gallen

Course and Examination Fact Sheet: Spring Semester 2025

8,338: Introduction to Web Mining for Social Scientists

ECTS credits: 4

Overview examination/s

(binding regulations see below)
decentral - Presentation, Analog, Individual work individual grade (20%)
Examination time: Term time
decentral - Written work, Digital, Individual work individual grade (80%)
Examination time: Term time

Attached courses

Timetable -- Language -- Lecturer
8,338,1.00 Introduction to Web Mining for Social Scientists -- English -- Matter Ulrich

Course information

Course prerequisites

Econometrics I (or similar). Basic knowledge of R. Students are not expected to have experience with web technologies or web programming, but should feel comfortable writing data analytics scripts in R.

Learning objectives

- Students will understand the basic concepts of contemporary web technologies relevant for web data mining.

- Students are capable of applying the relevant R packages to effectively and efficiently collect data from different types of web sources.

- Students develop a sense of how to navigate the basic ethical and legal aspects of web data collection for research purposes.

- Students understand different theoretical procedures of web data collection such as snowball sampling and can implement/apply them in a simple context.

- Students have a basic understanding of how Large Language Models (Generative AI) can be used to facilitate key aspects of web mining procedures.

Course content

Short summary

This course introduces students to the automated collection of data from websites and social media. Students get to know basic concepts of web mining for social science research and learn to use tools that enable them to compile their own data sets from web sources.

Description

The diffusion of the Internet has led to a stark increase in the availability of digital data describing all kind of every-day human activities. The dawn of such web based big data offers various opportunities for empirical research in economics and the social

sciences in general. While web (data) mining has for many years rather been a discipline within computer science with a focus on web application development (such as recommender systems and search engines), the recent rise in well-documented open-source tools to automatically collect data from the web as well as AI-driven tools makes this endeavor more accessible for researchers without a background in web technologies. Web mining has recently been the basis for studies in various fields such as labor economics, finance, marketing, political science, as well as sociology.

However, the collection and preparation of web data for research purposes poses new challenges for social scientists. Web data often comes in unusual or unsuitable formats for statistical analysis. Moreover, effective as well as efficient collection of such data demands basic understanding of web technologies. This course introduces students to the necessary basic concepts and practical skills to successfully handle the data collection and data preparation processes for a research project based on web data. While getting familiar with the basics of web technologies, students get in contact with various access points for web based data collection as well as develop ideas for potentially relevant research questions in these contexts. Building on the understanding of where what data is available in the web, students are introduced to basic concepts and practical tools to harvest these data. Practical exercises and problem sets support the learning process at this stage of the course. In the second half of the course students start their own empirical project based on web data in which they empirically tackle a research question of their choice. The term paper is both evaluated with respect to the demonstrated data collection skills as well as the scientific rigor of the empirical approach.

Course Goals

The main goal of the course is to enable students to conduct automated data collection from web sources on their own. Students get familiar with the advantages and disadvantages of extracting information from the Internet for scientific research. Finally, students get an opportunity to think about social science research questions with respect to human behavior that is particularly observable on the web (i.e., in social media, blogs, etc.).

Course structure and indications of the learning and teaching design

The course is structured as a 3-day block seminar during the term break. The following contents will be covered:

1. The Internet as a data source for social science research

2. Introduction to web technologies I: HTTP, HTML, and client/server interaction.

3. Web scraping: automated information extraction from websites

   a. R tools for web scraping

   b. Fetching and parsing websites

   c. Searching/filtering HTML

4. Introduction to web technologies II: JSON/XML, Web applications, and APIs.

5. Collecting data from the programmable web

   a. Social media and web APIs

   b. Parsing/filtering JSON and XML

6. Scrapers, Spiders, Crawlers

   a. Efficiency, robustness, and good conduct

   b. Crawler strategies and algorithms

7. Web mining ethics and legal guidelines

8. Web mining and scientific rigor: some thoughts on data quality, sampling, reproducibility

9. The next frontier: large language models and web mining

Course literature

Textbooks

Matter, Ulrich (2025). An Introduction to Web Mining with Applications in R. (under contract, Springer)

Journal articles

Edelman, Benjamin (2012). Using Internet Data for Economic Research. *Journal of Economic Perspectives*, 26(2): 189-206.

Einav, Liran and Levin, Jonathan (2014). Economics in the Age of Big Data. *Science*, 346 (6210): 1243089-1-1243089-6.

Additional course information

--

Examination information

Examination sub part/s

1. Examination sub part (1/2)

Examination modalities

Examination type	Presentation
Responsible for organisation	decentral
Examination form	Oral examination
Examination mode	Analog
Time of examination	Term time
Examination execution	Asynchronous
Examination location	On Campus
Grading type	Individual work individual grade
Weighting	20%
Duration	--

Examination languages

Question language: English
Answer language: English

Remark

Project Disposition Presentation

Examination-aid rule

Free aids provision

Basically, students are free to choose aids. Any restrictions are defined by the faculty members in charge of the examination under supplementary aids.

Supplementary aids

--

2. Examination sub part (2/2)

Examination modalities

Examination type	Written work
Responsible for organisation	decentral
Examination form	Written work
Examination mode	Digital
Time of examination	Term time
Examination execution	Asynchronous
Examination location	Off Campus
Grading type	Individual work individual grade
Weighting	80%
Duration	--

Examination languages

Question language: English
Answer language: English

Remark

Web Mining Project: Term Paper and Code

Examination-aid rule

Free aids provision

Basically, students are free to choose aids. Any restrictions are defined by the faculty members in charge of the examination under supplementary aids.

Supplementary aids

--

Examination content

In the term paper, students apply web-mining techniques to collect data in order to tackle a social science research question of their choice. Students derive a research question, explain the data collection strategy, describe the collected data, discuss the empirical strategy, execute a short empirical analysis, and discuss the results. The paper should be short and to the point (max. 4000 words). Students also hand in their documented code (including web mining/data collection, preparation, and analysis).

In the presentation, students present and defend (in person, on campus) their project proposal. The presentations take place on the third course day (in March). Students are expected to present a solid plan for their project (duration: 8-15 minutes, depending on number of course participants), and hand in the corresponding slide deck. We will organize the presentations in the style of a research seminar, where each student presents their plan/ideas and gets critical questions and constructive feedback from the audience (the other course participants and myself). The idea is that students can use the feedback gathered during the presentation to substantially improve their project plan and then use the remaining weeks/months of the term to work on their project/paper.

Examination relevant literature

There is no mandatory examination literature.

University of St.Gallen

Please note

Please note that only this fact sheet and the examination schedule published at the time of bidding are binding and takes precedence over other information, such as information on StudyNet (Canvas), on lecturers' websites and information in lectures etc.
Any references and links to third-party content within the fact sheet are only of a supplementary, informative nature and lie outside the area of responsibility of the University of St.Gallen.

Documents and materials are only relevant for central examinations if they are available by the end of the lecture period (CW21) at the latest. In the case of centrally organised mid-term examinations, the documents and materials up to CW 13 (Monday, 25 March 2025) are relevant for testing.

Binding nature of the fact sheets:

- Course information as well as examination date (organised centrally/decentrally) and form of examination: from bidding start in CW 04 (Thursday, 23 January 2025);
- Examination information (supplementary aids, examination contents, examination literature) for decentralised examinations: in CW 12 (Monday, 17 March 2025);
- Examination information (supplementary aids, examination contents, examination literature) for centrally organised mid-term examinations: in CW 14 (Monday, 31 March 2025);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised examinations: two weeks before ending with de-registration period in CW 15 (Monday, 07 April 2025).