Course and Examination Fact Sheet: Spring Semester 2021

8,338: Introduction to Web Mining for Social Scientists

ECTS credits: 4

Overview examination/s

(binding regulations see below)
Decentral - examination paper written at home (individual) (80%)
Examination time: term time
Decentral - examination paper written at home (individual) (20%)
Examination time: term time

Attached courses

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

Course information

Course prerequisites

Econometrics I (or similar). Very basic knowledge of R. Students are not expected to have experience with web technologies or programming.

Learning objectives

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

- Students will know how to apply the relevant R packages to effectively and efficiently collect data from different types of web sources.

- Students understand the basic ethic and legal aspects of web data collection for research purposes.

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

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 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 simple 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

The course is structured as a 3-day block seminar during the term break. Given the current Covid-relate situation, the course format is hybrid (seminar in persona and live stream). 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: data quality, sampling, reproducibility

NOTE ON LECTURES: Given the current COVID-19 situation, lectures will take place in person but the number of students attending the lectures in person might be restricted (division into alternating groups). All students accepted to the course will be informed about the specifics of the attendance in due time. In any case, the lectures will also be broadcasted via StudyNet/Zoom.

GENERAL NOTE ON HYBRID-FORMAT OF THIS COURSE: Detailed lecture notes for each lecture as well as code and data examples will be made available throughout the course. The aim is to facilitate a well-guided learning experience with online learning material closely fitting the structure of this course.

Course literature

Textbooks

Liu, Bing (2011). *Web Data Mining*. New York, NY: Springer. Mitchell, Ryan (2015). *Web Scraping with Python*. Sebastopol, CA: O'Reilly.

Munzert, S. and Rubba, C. and Meißner, P. and Nyhuis, D. (2014). *Automated Data Collection with R: A Practical Guide to Web Scraping andText Mining*. Chichester, UK: Wiley.

Russell, Mathew A. (2014). *Mining the Social Web*. Sebastopol, CA: O'Reilly.

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

In the case of the President's Board having to implement new directives due to the SARS-CoV-2 pandemic in SpS2021, the course information listed above will be changed as follows:

- The course is conducted online via the platform Zoom.
- The recordings of the course are available for 30 days.
- The lecturer informs via StudyNet on the changed implementation modalities of the course.

The examination information listed above would be changed as follows:

- There are no changes necessary to the examination information.

Examination information

Examination sub part/s

1. Examination sub part (1/2)

Examination time and form

Decentral - examination paper written at home (individual) (80%)
Examination time: term time

Remark

--

Examination-aid rule

Term papers

Term papers must be written without anyone else's help and in accordance with the known quotation standards, and they must contain a declaration of authorship which is a published template in StudentWeb.

The documentation of sources (quotations, bibliography) has to be done throughout and consistently in accordance with the chosen citation standard such as APA or MLA.

For papers in law, the legal standard is recommended (by way of example, cf. FORSTMOSER, P., OGOREK R. et SCHINDLER B., Juristisches Arbeiten: Eine Anleitung für Studierende, newest edition respectively, or according to the recommendations of the Law School).

The indications of the sources of information taken over verbatim or in paraphrase (quotations) must be integrated into texts in accordance with the precepts of the applicable quotation standard, while informative and bibliographical notes must be added as footnotes (recommendations and standards can be found, for example, in METZGER, C., Lern- und Arbeitsstrategien, newest edition respectively.

For any work written at the HSG, the indication of the page numbers is mandatory independent of the chosen citation standard. Where there are no page numbers in sources, precise references must be provided in a different way: titles of chapters or sections, section numbers, acts, scenes, verses, etc.

Supplementary aids

-

Examination languages

Question language: English
Answer language: English

---

2. Examination sub part (2/2)

Examination time and form

Decentral - examination paper written at home (individual) (20%)
Examination time: term time

Remark

One compulsory problem set (20%)

Examination-aid rule

Term papers

Term papers must be written without anyone else's help and in accordance with the known quotation standards, and they must contain a declaration of authorship which is a published template in StudentWeb.

The documentation of sources (quotations, bibliography) has to be done throughout and consistently in accordance with the chosen citation standard such as APA or MLA.

For papers in law, the legal standard is recommended (by way of example, cf. FORSTMOSER, P., OGOREK R. et SCHINDLER B., Juristisches Arbeiten: Eine Anleitung für Studierende, newest edition respectively, or according to the recommendations of the Law School).

The indications of the sources of information taken over verbatim or in paraphrase (quotations) must be integrated into texts in accordance with the precepts of the applicable quotation standard, while informative and bibliographical notes must be added as footnotes (recommendations and standards can be found, for example, in METZGER, C., Lern- und Arbeitsstrategien, newest edition respectively.

For any work written at the HSG, the indication of the page numbers is mandatory independent of the chosen citation standard. Where there are no page numbers in sources, precise references must be provided in a different way: titles of chapters or sections, section numbers, acts, scenes, verses, etc.

Supplementary aids

--

Examination languages

Question language: English
Answer language: English

Examination content

- In the term paper, students apply web-mining techniques to collect data in order to tackle a simple 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 web-mining code.

- In the problem set, students demonstrate their acquired skills by solving different exercises related to the automated collection of data from web sources. The problem set is about collecting data from websites as well as topics surrounding the collection of data from web applications related to social media. The problem set includes the task of preparing a short disposition of the planned research paper. The problem sets are generally aimed at deepening the material covered in class as well as improve the students' practical web mining skills.

Examination relevant literature

There is no mandatory examination literature.

Please note

Please note that only this fact sheet and the examination schedule published at the time of bidding are is 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 12 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, 28 January 2021);
- Examination information (regulations on aids, examination contents, examination literature) for decentralised examinations: in CW 12 (Monday, 22 March 2021);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised mid-term examinations: in CW 12 (Monday, 22 March 2021);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised examinations: two weeks before the end of the registration period in CW 14 (Thursday, 8 April 2021).