University of St.Gallen

Course and Examination Fact Sheet: Autumn Semester 2017

9,338: 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%)
Decentral - examination paper written at home (individual) (20%)

Attached courses

Timetable -- Language -- Lecturer
9,338,1.00 Introduction to Web Mining for Social Scientists -- Englisch -- Matter Ulrich

Course information

Course prerequisites

Econometrics I. Very basic knowledge of R (or another computing environment to work with data such as Python, Stata, etc.). Students are **not expected** to have experience with web technologies or programming.

Course content

Short summary
This course introduces students to the automated collection of data from websites and social media for research purposes. Students get to know basic concepts of web mining for social science research and learn to use tools (mostly in R) that enable them to compile their own data sets from web sources (e.g., for seminar papers or Master Theses).

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.

The collection and preparation of web data for research purposes, however, 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. As this course is at the introductory level, no previous experience with web technologies or programming are expected. 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 (in this context, this course can serve well as a preparation for an empirical Master Thesis or seminar papers). 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.)develop a better understanding of economic and social questions related to web data and the internet in general.

Course 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 (or Python modules) 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 structure

Lectures take place in the computer lab on a weekly basis during the semester. The course is structured as follows:

The Internet as a data source for social science research

1. Introduction to web technologies I: HTTP, HTML, and client/server interaction
2. Introduction to web technologies II: JSON/XML, Web applications, and APIs
3. Web scraping: automated information extraction from websites
    a. R tools for web scraping
    b. Fetching and parsing websites
    c. Searching/filtering HTML
4. Building a scraper
    a. Efficiency, robustness, and good conduct
    b. Crawler strategies and algorithms
5. Collecting data from the programmable web
    a. Social media and web APIs
    b. Parsing/filtering JSON and XML
6. Web mining ethics and legal guidelines
7. Web mining and scientific rigor: data quality, sampling, reproducibility

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 and Text 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.

Matter Ulrich, Stutzer Alois (2015) pvsR: An Open Source Interface to Big Data on the American Political Sphere. PLOS ONE 10(7): e0130501.

Additional course information

Note on the literature: there is no general compulsory reading list for this course. Students are thus not expected to buy the listed books. The listed literature is intended as a reference guide and inspiration for the student's practical work on the exercises and their research paper throughout the course. However, there are also plenty of freely available resources on the web to support student's learning process to which students will be guided throughout the course.

Scripts summarizing the core materials of each lecture will be provided online after each lecture.

Examination information

Examination sub part/s

1. Examination sub part (1/2)

Examination time and form
Decentral - examination paper written at home (individual) (80%)

Remark
Term paper including documented code.

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.
- The documentation of sources (quotations, bibliography) has to be done throughout and consistently in accordance with

the APA or MLA standards. The indications of the sources of information taken over verbatim or in paraphrase (quotations) must be integrated into the text 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. (2015), Lern- und Arbeitsstrategien (11th ed., 4th printing). Aarau: Sauerländer).
- For any work written at the HSG, the indication of the page numbers both according to the MLA and the APA standard is never optional.
- 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.
- For papers in law, the legal standard is recommended (by way of example, cf. FORSTMOSER, P., OGOREK R. et SCHINDLER B. (2014, Juristisches Arbeiten: Eine Anleitung für Studierende (5. Auflage), Zürich: Schulthess, or the recommendations of the Law School).

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%)

Remark
Two compulsory problem sets (each 10%).

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.
- The documentation of sources (quotations, bibliography) has to be done throughout and consistently in accordance with the APA or MLA standards. The indications of the sources of information taken over verbatim or in paraphrase (quotations) must be integrated into the text 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. (2015), Lern- und Arbeitsstrategien (11th ed., 4th printing). Aarau: Sauerländer).
- For any work written at the HSG, the indication of the page numbers both according to the MLA and the APA standard is never optional.
- 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.
- For papers in law, the legal standard is recommended (by way of example, cf. FORSTMOSER, P., OGOREK R. et SCHINDLER B. (2014, Juristisches Arbeiten: Eine Anleitung für Studierende (5. Auflage), Zürich: Schulthess, or the recommendations of the Law School).

Supplementary aids
Students are allowed to consult/use any book, document (online or offline) that might be of help to solve the problem sets. However, they are strictly bound to solve both of the problem sets *on their own*. Both problem sets will be released four days before they are due.

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 sets, students demonstrate their acquired skills by solving different exercises related to the automated collection of data from web sources. The first problem set is aimed at collecting data from websites, the second problem set is focused on topics surrounding the collection of data from web applications related to social media. 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.

Supplementary/voluntary 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 and Text 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.

Matter Ulrich, Stutzer Alois (2015) pvsR: An Open Source Interface to Big Data on the American Political Sphere. PLOS ONE 10(7): e0130501.

University of St.Gallen

Please note

We would like to point out to you that this fact sheet has absolute priority over other information such as StudyNet, faculty members' personal databases, information provided in lectures, etc.

When will the fact sheets become binding?

- Information about courses and examination time (central/decentral and grading form): from the start of the bidding process on 24 August 2017
- Information about decentral examinations (examination-aid rule, examination content, examination relevant literature): after the 4th semester week on 16 October 2017
- Information about central examinations (examination-aid rule, examination content, examination relevant literature): from the start of the enrolment period for the examinations on 06 November 2017

Please look at the fact sheet once more after these deadlines have expired.