GESIS Fall Seminar in Computational Social Science 2022

Syllabus for week 1:
"Introduction to Computational Social Science with R"

Lecturers:	Dr Aleksandra Urman	Dr Max Pellert
Affiliation:	University of Zurich	Sony CSL Rome, Complexity Science Hub Vienna, Graz University of Technology
Email:	urman@ifi.uzh.ch	pellert@csh.ac.at

Date: September 05-09, 2022
Time: 09:00-16:00

About the Lecturers

Aleksandra Urman is a postdoctoral researcher at Social Computing Group, University of Zurich. In her PhD dissertation, defended in May 2020 at the University of Bern, she examined comparative aspects of polarization on social media. She also holds a MA in Political Science from Central European University. In her research, Aleksandra employs computational methods to examine various aspects of political communication on social media, with a particular focus on polarization, authoritarian regimes, and far-right groups. In addition, she is interested in algorithmic biases in web search.

Max Pellert has a background in Cognitive Science and Economics (University of Vienna, Austria, and University of Ljubljana, Slovenia). He was a doctoral researcher affiliated to the Complexity Science Hub Vienna and the Medical University of Vienna in the WWTF research group "Emotional Well-Being in the Digital Society" led by David Garcia (Graz University of Technology). He works now as Assistant Researcher at Sony CSL Rome. His research focuses on analyzing the digital traces of individual and collective emotional behavior and affective expression on social media. He is broadly interested in the social sciences and uses traditional and novel computational methods to study emotion dynamics, belief updating, collective emotions, and other interesting phenomena.

Course Description

The course will provide an overview of the methods used in the field of computational social science (CSS) and their real-world applications. It will include both theoretical explanations of different methods and hands-on practical exercises through which the participants will be able to apply the discussed techniques in R. The course is aimed at participants with no or little experience with computational methods. Within the course, topics such as web scraping, foundations of computational text analysis, data visualization, and ethical aspects of CSS will be covered. The course will take place in person and will consist of a combination of lectures and practical exercises. By the end of the course, each participant will have practical experience in R in retrieving web data, applying basic text analysis techniques to it, and visualizing the results. The participants will gain this experience through supervised practical exercises as well as through group projects on which they will work semi-independently, with guidance from the lecturers, throughout the course. To make full use of the course participants should have knowledge of the very basic concepts of programming in R (for example write a loop themselves, read in a CSV file and be familiar with data types such as a data.frame), we link to a self-assessment test below (see Course Prerequisites). To gain that basic knowledge, several pointers to online crash courses on those very basics of R are linked below (see Course Preprequisites). Participants are expected to work through some of those materials before the course should they have never worked with R before at all or only had very limited experience with R.

Keywords

computational social science; R; text analysis; web scraping; visualization

Course Prerequisites

- Basic knowledge of R (if you are unsure if your R knowledge is sufficient, here is a self-assessment test we prepared for you. In case you will see that the test is too difficult for you, we have also included links to several free online R crash courses that you should go through to prepare for our course https://seafile.ifi.uzh.ch/f/63542a5ab4be4d37846d/))
- Working command of English language
- Knowledge of basic statistics (distributions, correlation)
- Basic programming knowledge (variables, loops, conditions) in R (see the self-assessment test above)

Target Group

Participants will find the course useful if:

- They are social scientists with very little or no experience with computational methods who would like to learn more about the methods and potentially use them in their research

Course and Learning Objectives

By the end of the course participants will:

- Be able to define what constitutes the field of computational social science and know which methodologies are commonly utilized in the field as well as which types of research questions can be handled using these methodologies
- Be familiar with the major ethical aspects of conducting computational social science research
- Have hands-on experience gathering digital trace data from online sources through direct web scraping and APIs using R
- Know about the basic computational text analysis methods and have practical experience utilizing some of them using R
- Be able to visualize their data using various techniques in R
- Be equipped to use provided pointers to advanced materials to further improve their skills

Organizational Structure of the Course

The course will consist of a combination of lectures and practical hands-on lab sessions. The lab sessions will consist of two components. The first one is practical scripted exercises related to a specific topic that the participants will be guided through by the lecturers. The second one involves semi-independent group work on the side of the participants and will be constituted by a group project in which the participants will apply the skills gained studying different topics covered in the course. Throughout this project the participants will be supported through individual consultations with the lecturers.

Software and Hardware Requirements

Participants should bring their own laptops and pre-install R and RStudio installed on their laptops, it's highly preferable that R is updated to the latest version. We will let participants know about specific packages necessary to install shortly before the course, and, if necessary, will help them with the specific package installation problems on Day 1 of the course. The lecturers are most familiar with Linux environments (e.g., Ubuntu or Debian) to run R and RStudio, but they can also provide support for Windows and macOS.

Recommended Literature to Look at in Advance

In general, all of the literature listed below (within Day-to-day schedule) should be treated as recommended rather than required; we do not require reading any of the literature for the course but suggest that familiarising yourself with it would provide a better understanding of the CSS field in general, and give participants more advanced knowledge of specific topics. Hence, we would suggest that participants go through some of the recommended literature before/during the course, and then read it in detail to deepen their knowledge after the course.

Day-to-Day Schedule and Literature

Each day includes a mix of lectures and practical exercises, some of which are done in groups. The instructors are always there during the practical and group exercises to help the participants and answer their questions.

Day 1: Introduction to computational social science (CSS and the ethical aspects of doing CSS research + Intro to Digital Trace Data

- Morning sessions (9am-12pm):
  - 9am-11pm (incl. a 15-min break) on day 1 we will start with a lecture that presents an overview of the field of computational social science and its development, including examples of CSS research in different subdomains (e.g., CSS research focusing on political processes, economic phenomena, communication science, etc). We will also provide background on the ethical conduct of CSS research and give participants practical guidelines on how to make sure their research adheres to the ethical (and legal) standards.
  - 11.15pm - 12pm: Group task: each group of participants will receive a case study about ethics in CSS. Within this task, the participants will need to evaluate different CSS study designs from an ethical standpoint and propose ways to mitigate potential harms. Before lunch, each group will have ~45minutes to get familiar with the study and start discussing it. After lunch, the groups will first finish their discussions and then shortly present their outcomes (the details and guidelines will be provided during the course).
- 12pm-1pm: Lunch break
- Afternoon sessions (1pm-5pm):
  - 1pm-1.45pm: Group task: participants continue discussing the case studies and prepare short summaries of their discussions to present those to other groups.
  - 2pm-3pm: Group presentations on the group task, joint discussion
  - 3.15pm-4.15pm: Lecture on Digital Trace Data that will provide the participants with background information on what digital trace data is and how it can be leveraged within CSS. This will serve as the basis for the next day's practical lectures and assignments. The lecture is followed by a short introduction to the Group Project work that will take place during the course
  - 4.30-5pm: Participants start discussing ideas for the final group projects

- Optional - Circa 7pm - joint informal dinner (of course, fully optional)

Literature

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. https://www.bitbybitbook.com/en/

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology*, 13(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). SOCIAL SCIENCE: Computational Social Science. *Science*, 323(5915), 721–723. https://doi.org/10.1126/science.1167742

Day 2: Web scraping

- Morning sessions (9am-12pm):
  - 9am-10am: Hands-on tutorial on the use of APIs for data collection using R
  - 10.15am-11.15am: Practical task where participants use R to collect data from an API based on the tutorial
  - 11.30am-12pm: Solutions to the practical task are presented
- Lunch break (12pm-1pm)
- Afternoon sessions (1pm-5pm):
  - 1pm-2pm: Hands-on tutorial on the use of web scraping for data collection using R
  - 2.15pm-3.15pm: Practical task where participants use R to collect data from using web scraping based on the tutorial

- 3.30pm-4pm: Solutions to the practical task are presented
- 4.15pm-5pm: Group work further developing ideas/potentially starting to collect the data for the final project

Literature:

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Kopecký, J., Fremantle, P., & Boakes, R. (2014). A history and future of Web APIs. *It - Information Technology*, *56*(3). https://doi.org/10.1515/itit-2013-1035

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

Wickham, H., & RStudio. (2020). *rvest: Easily Harvest (Scrape) Web Pages* (0.3.6) [Computer software]. https://CRAN.R-project.org/package=rvest

Day 3: Foundations of computational text analysis

- Instructional sessions (morning): During Day 3, we will give the background on contemporary methods commonly employed for computational text analysis. We will discuss and showcase several different methods for sentiment analysis and talk about how to validate results. We will provide participants with practical skills in foundational bag-of-words-approach-based text analysis methods such as frequency analysis, co-occurrence analysis and LDA topic modelling. We will also give participants an overview of existing more advanced methods that they might want to explore if they are interested in the topic.
- Practical sessions (afternoon): Participants will do practical exercises on automated text analysis. In project groups, they will further develop their ideas, and decide on the ways to address question(s) they are interested in using computational text analysis methods they learned in Day 3.

Literature:

Mohammad, S. M. (2021). Sentiment Analysis: Automatically Detecting Valence, Emotions, and Other Affectual States from Text. ArXiv:2005.11882 [Cs]. http://arxiv.org/abs/2005.11882

Atteveldt, W. van, Velden, M. A. C. G. van der, & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, online first, 1–20. https://doi.org/10.1080/19312458.2020.1869198

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774. https://doi.org/10.21105/joss.00774

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

Niekler, A., & Wiedemann, G. (n.d.). *Text mining in R for the social sciences and digital humanities*. Retrieved April 16, 2021, from https://tm4ss.github.io/docs/index.html

Welbers, K., Atteveldt, W. V., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures*, 11(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Day 4: Basics of data visualisation with R

- Instructional sessions (morning): We will cover the basics of data visualisation using R, including different types of plots and diagrams, with a focus on the ggplot2 package. We will also give directions on the more advanced visualisation techniques in R such as interactive graphs (plotly) for participants who are interested in the topic.
- Practical sessions (afternoon): Participants will do practical exercises on data visualisation using R. They will further develop their project ideas and come up with the ways to visualise their group project results.

Literature:

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer Science & Business Media. https://ggplot2-book.org/

Ggplot2-cheatsheet. (n.d.). *RStudio*. https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

Ognyanova, K. *Network visualization in R.* https://kateto.net/network-visualization

Prabhakaran, S. (n.d.). *How to make any plot in ggplot2? | ggplot2 Tutorial*. Retrieved April 16, 2021, from http://r-statistics.co/ggplot2-Tutorial-With-R.html

Day 5: Project work day and Outlook

- In the morning, the participants will keep working on the projects they started during the previous practical sessions, and finalise them (under the guidance from the course instructors). In the afternoon, the participants will present their group projects.
- Optional: Participants can choose to take part in a short session giving an overview of more advanced packages and methods for data analysis in R (like “data.table”) and a quick introduction to versioning control with git.