GESIS Fall Seminar in Computational Social Science 2021

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

Lecturers:	Dr. Aleksandra Urman	Max Pellert, MSc
Email:	urman@ifi.uzh.ch	pellert@csh.ac.at

Date: September 13-September 17, 2021
Time: between 10am and 6pm (the allocated time slot includes a 1-hour lunch break between 1pm and 2pm and several 10-15 minute breaks)

About the Lecturers:

Dr. Aleksandra Urman is a postdoctoral researcher at the Institute of Communication and Media Studies of the University of Bern and 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 is a PhD candidate at the Section for Science of Complex Systems at the Medical University of Vienna and a resident scientist at the Complexity Science Hub Vienna affiliated to Graz University of Technology. Max obtained a bachelor’s degree in economics from the University of Vienna and a master’s degree in cognitive science from the University of Vienna and the University of Ljubljana, Slovenia. 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 online and will consist of pre-recorded video lectures combined with live Q&A sessions, exercises and project work. 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 the guidance from the lecturers, throughout the course.

Keywords:

computational social science; R; text analysis; web scraping

Course Prerequisites:

- Basic knowledge of R (in addition, links to online tutorials for those who need to refresh their R skills will be provided to participants ahead of the course)

- Working command of English language
- Knowledge of basic statistics (distributions, correlation)
- Basic programming knowledge (variables, loops, conditions), preferably in R

Target Group:

Participants will find the course useful if:

- They are social scientists with 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

Organisational Structure of the Course:

The course will consist of a combination of pre-recorded lectures, live Q&A sessions and practical hands-on lab sessions that will take place live online. 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 requirements:

All the participants should have 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.
The participants will also need Zoom installed as the online live sessions will be held on Zoom.

Long Course Description:

The course provides an overview of the methods used in the field of computational social science (CSS) and their real-world applications accompanied by hands-on training on the implementation of basic methods used in CSS in R programming language. The main target audience of the course are interdisciplinary researchers who want to learn more about CSS and gain practical experience. The course introduces core CSS methods to enable participants to apply them in their own research and/or make them better understand contemporary developments in CSS with respect to their subfields of interest. The course is suitable for both, participants with no prior experience with CSS and participants that have introductory-level familiarity with one or more methods (i.e., web scraping, text analysis, visualization) but want to gain a more in-depth and structured overview of the techniques they know and learn about new ones. The participants can be at any stage of their academic career, and the only requirements for attending the course are working-level fluency in English and at least basic knowledge of the programming language R.

During the course, we will cover the following topics: general introduction to CSS with examples of CSS research and discussion of ethical considerations when conducting CSS studies; web scraping, including web data extraction directly from websites and data gathering through APIs; foundations of text analysis with a focus on bag-of-words-based methods such as frequency analysis, co-occurrence analysis and LDA topic modeling and an overview of more advanced methods; basics of data visualization in R using ggplot2 for effective visual communication of research results. These topics are selected as they constitute a foundation of a broad range of CSS methods, and are crucial to get introductory-level knowledge and skills in the field. Participants will be able to build on these basic skills if they wish to deepen their expertise in CSS and master more advanced topics in other GESIS courses or through other means.

The course will consist of pre-recorded lectures, live Q&A sessions, and practical exercises as well as group work on a project. As an outcome of the course, participants will gain not only a theoretical understanding of key concepts and methods used in CSS, but practical experience applying such methods. The latter will be achieved not just through programming exercises but also through project group work. By the end of the course, all groups will have completed small projects where they will collect their own textual data from online sources, analyze it using one or more of the covered computational text analysis techniques, and visualize their results. Such project work will enable the participants to gain real-world insight on how CSS research is done, which obstacles one may encounter when doing such research, and which opportunities it offers. The projects will be done under the guidance of the course instructors and, when necessary, with their assistance. The choice of the specific topics of the projects will be up to the group members. To ensure that the groups are balanced and coherent in terms of the participants' backgrounds and interests, we will match the participants based on their interests in the first day of the course (for that, we will ask the participants to fill out a short questionnaire on their interests and background shortly before the course).

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 coursebut suggest that familiarizing oneself 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 to deepen their knowledge after the course.

Day-to-day schedule and literature:

Structure: The course will consist of a combination of instructional sessions and practical sessions. The instructional sessions will be scheduled for the morning and for the most part, will be pre-recorded by the lecturers. The afternoon sessions will consist of practical exercises and group project work online via zoom under the online guidance from the lecturers. The lecturers, however, will be available via Zoom to assist participants with technical problems they might have during mornings too, as well as conduct a daily Q&A session before the scheduled start of the practical sessions.

Day 1: Introduction to computational social science and ethical aspects of doing CSS research

Instructional sessions (morning): During Day 1, we will provide the participants 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.

Practical sessions (afternoon): The participants will be given practical assignments during which they will need to evaluate different CSS study designs from an ethical standpoint and propose ways to mitigate potential harms. The participants will be divided into groups where they will start discussing potential ideas for the project to pursue during the course.

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

Instructional sessions (morning): During Day 2, we will provide participants with a background and practical R skills on how to scrape web data using R (including direct HTML scraping and API-based scraping). We will also discuss comparative pros and cons of different data scraping options, related ethical and legal issues, as well as give participants information about alternative data repositories where they can find CSS-relevant datasets (e.g., those already compiled by other researchers).

Practical sessions (afternoon): Participants will do practical exercises on web scraping. In project groups, they will need to make their project ideas more concrete, decide on the way to scrape the necessary data, and, ideally, start the data scraping process.

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 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 modeling. 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:

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*, *0*(0), 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 visualization with R

Instructional sessions (morning): We will cover the basics of data visualization 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 visualization 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 visualization using R. They will further develop their project ideas and come up with the ways to visualize 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

In the morning, the participants will keep working on the projects they started during the previous practical sessions, and finalize them (under the guidance from the course instructors).

In the afternoon, the participants will present their group projects.