Syllabus “Toolbox Computational Social Science”

Felix Lennert

Winter 2023/24

Relevant Info

- E-mail: felix.lennert@uni-leipzig.de
- Office hours: Fridays, 10–11, on Zoom
- Course hours: see schedule
- Link to Moodle Course

Description

Recently, a “computational turn” has taken hold of the social sciences. Digital data and novel methods originating from the computer sciences offer important opportunities for sociology. The course starts with teaching practical skills to collect digital trace data online (web scraping, API “harvesting”). As the lion’s share of this material is in textual format, the students will subsequently learn to evaluate large text archives in an automated way through machine learning techniques. The programming of these tools is performed in R, for which basic knowledge (as taught in the methods course of the institute) is required. Students have to write an empirical paper that answers a sociologically relevant research question using at least one of the methods learned and give extensive feedback to others’ projects in class.

What to expect

The course is going to address the following questions (incomplete, unordered list)

- Which tools can I use to wrangle data?
- Why would one ever write a function in R?
- How are online data used to study human behavior?
- How is the web written?
- . . . and how can we use this structure to access and store the data it contains?
- Can I use R to send prompts to ChatGPT?
- How do I teach my computer how to “read”?
- Which words are representative of a text document?
- Can I cluster documents based on their content?
- What sentiments are conveyed in text?
- What the hell is “topic modeling”?
- How do I “measure” things with text?
- What concepts can I measure using text?
- Does the sociology of culture benefit from text analysis?
- How to detect biases in text?
- and many more.

This course is structured so that it provides you – the student – with theoretically heavy lectures and hands-on R sessions. The former aims to introduce you to how you can use techniques to conduct empirical research and will mostly consist of the presentation of the ideas of innovative and/or cutting-edge research that has harnessed such digital data to solve exciting research puzzles. The practical sessions will be delivered as videos. You are expected to watch these thoroughly (there will be a codeword appearing somewhere in the video that you will have to know before the next session; of course, this is by no means a perfect solution) and work on exercises before the next session (deadline for the exercises is right before the beginning of the next session. This session will be held in person and dedicated to going through the exercises and answering questions that may have come up while solving these. Since this will hardly take up the entire 1.5 hours, students are encouraged to seize the remaining time in the classroom to watch the next block of videos and start working on the exercises.

Once the first block of content has been finished (i.e., web scraping), you will need to start developing your research. Here, you are expected to conduct empirical research using digital trace data. This implies that you first acquire your data from an online resource (e.g., a web page or a database). You will need to analyze them using means of computational text analysis, therefore, the data need to be in textual format. Once you have learned about the methods, you will have ample time between the January sessions and the final session in February to develop hypotheses and analyze your data.

At the very end of the course, it is time for "peer review." One of your peers will be assigned your project and provide comments on it. This will work as follows: first, the authors present their work briefly (10 minutes). Then, the assigned opponents will provide their comments (5 minutes). The opponents are supposed to give feedback on the question/motivation, the chosen theoretical angle, the data, and the method. I expect the opponent to summarize each section briefly and point out what they perceive as strengths and weaknesses of the respective component. Moreover, they can suggest how to frame the questions, provide further literature, or point out alternate methodological approaches that may be better suited to test the hypotheses. To facilitate this, you must send presentations presenting the main idea, some theory (including hypotheses), a preliminary testing strategy, and the first results to me by January 29, 2024, so I can distribute them to your peers.

This peer review ensures that you are right on track and that everyone has accomplished the course's learning goals. After that, you are all good to go and can check your analyses and write up your resulting papers so that they resemble a proper empirical research paper (see more in "Expectations" below).

The deadline for the paper is March 15, 2024. Code and data have to be sent to me by email (depending on data size, data can also be shared via Google Drive/Dropbox/Uni Leipzig Wolke). The code should run "out of the box" and contain everything necessary to replicate the results and graphs in the paper. Preferably, this is in chronological order and structured into sections with descriptive titles.

Extensions

Extensions can be granted for some reasons. These involve, among others, internships and sickness. In the case of the former, please give me a quick heads-up so that I can arrange it (preferably with some sort of proof). If you need an extension for another reason than the ones mentioned above, feel free to reach out anytime, and I will do my best to accommodate your needs.

Expectations

- the basics
  - font size 12 pt, 1.5 line spacing
  - no typos, grammatical flaws, etc. (you are living in the age of helpers such as Grammarly, there are no more excuses)
  - length: between 4,000 and 7,000 words
  - cite correctly; my preferred citation style: ASA (it's strongly advised to use Zotero and Quarto/Overleaf; resources can be provided upon request)
- structured as an empirical research paper:
  - the *introduction* contains an empirical social scientific research question that is theoretically and practically motivated (i.e., showing its scientific and real-world relevance)

  - the *theory section* provides a **brief** overview of relevant prior research; clearly testable hypotheses are derived from the literature/goals for exploratory analyses are formulated
  - in *data and methods*, the data (including acquisition strategy), as well as the analysis strategy, are described; in our case, the data consist of text, the analyses are related to the course content; data and methods need to enable valid results
  - *results* need to be visualized through tables and/or (gg)plots and described in the text; tables and visualizations need to be properly labeled so that they can "stand on their own"
  - *discussion* of the results is performed in lieu of the theoretical foundations; potential shortcomings and reach of the paper are outlined
  - the *conclusion* circles back to the introduction and connects it to the results; it needs to clearly answer the research question

Basic rules of behavior

- If anything is unclear, ask me. This probably means that I have failed my job, and your question offers me a second chance to fix this.
- No discrimination. Never. If you witness any, tell me. I will find a way to deal with it.
- THIS IS IMPORTANT: If there are problems, reach out whenever. Do not let them become too big.
- Copy R Code from the internet – but you are responsible for the solution, so please make sure it works and solves your problem.
- ChatGPT is explicitly allowed. In my opinion, it is a tool that is here to stay, and you should use whatever resource you have to get the job(s) done. Plus, writing the right prompt is a skill in itself that you should definitely hone. If you use it for your writing, please make sure to proof-read everything properly, since you will be held accountable for both content and style.
  - Same holds for Grammarly – use it!!!
- Form groups with your peers for working on the material. Everything will be easier and more fun. Except for when you have free riders. Kick them out of your group.
- AGAIN: ask questions if needed. Anytime.

Schedule and Readings

The course will be structured into one Zoom session that will cover all the relevant housekeeping (October 13, 2023) and provide a first introduction, one large block in December/January (December 12, 2023 – January 8, 2024), and one final session in February (February 2, 2024). R-related content will be covered in a script and explained in videos which are embedded in the script. The script is published online in `bookdown` format and can be found here. You are expected to watch the videos – there will be a codeword mentioned at some point to check if you have done so – and work on related exercises. We will go through the exercises at the beginning of the following session, which will provide you with ample opportunities to ask questions.

Literature-wise, we will use a mix of textbooks that contain the basic concepts introduced in the sessions and online resources that over the R implementations; however, everything that is relevant in terms of R content can be found in the script. Moreover, I will add studies that innovatively apply CSS techniques and, hence, serve as inspiration. I do not expect you to read the literature and will do the theoretical part of the sessions in rather "lecture" style. This is because I am convinced that the bulk of the learning happens once you apply the techniques. All relevant readings are either freely available online or will be uploaded to Moodle. Boldened readings are the papers I will use as examples in the session.

The textbooks we use are "Text as Data" by Justin Grimmer, Margaret Roberts, and Brandon Stewart; "Tidy Text Mining with R" by Julia Silge and David Robinson (*available online*); "Supervised Machine Learning with R" by Julia Silge and Emil Hvidtfeld (*available online*); and "Bit By Bit" by Matthew Salganik (*available online.* When it comes to general R-related things (i.e., data wrangling, visualization, etc.), I am happy to refer you to the second edition of "R for Data Science" (henceforth, R4DS) by Hadley Wickham (*available online*).

If you feel particularly inspired – or insufficiently informed, there will also be an extensive reading list that is structured into related readings (i.e., textbook stuff), approachable readings (blog posts and the like), and illustrious literature (i.e., more papers that harness the covered methods in an innovative way).

Kick Off (Fri, 13 October 2023)

11:15–12:45 (Zoom): theoretical introduction and getting to know each other // Recap and Regular Expressions

- Lazer et al. (2009)
- Edelmann et al. (2020)
- Salganik (2018), p. 13–84. (–> find it *online*)

13:30–15:00 (videos): Recap (RStudio Projects, RMarkdown, `dplyr`, `tidyr`, `ggplot2`, `purrr` functional programming), and `stringr` and Regular Expressions

This may sound like *a lot*. However, for the recap, there will be no mandatory exercises – you will need these packages down the road anyways.

The corresponding chapters in the *R4DS book*:

- RStudio Projects – chapter 7
- `RMarkdown/Quarto` – chapters 28 & 29
- `dplyr` – chapter 4
- `tidyr` – chapter 6
- `ggplot2` – chapters 2 & 10 & 11 & 12
- `purrr` & loops in different flavors – chapter 27
- functional programming – chapter 26
- `string` & Regular Expressions – chapters 15 & 16

Digital Trace Data and basic web scraping (Fri, 08 December 2023):

11:15–12:45 (NSG, SR 124): Digital trace data – promises and pitfalls

- Golder and Macy (2014)
- Salganik (2018), p. 13–84. (–> find it online)

13:30–15:00 (videos): Basics of web scraping with `rvest`

- Webscraping 101 by the developers of`rvest`
- Chapter 24 in the R4DS book

Intro to Web Scraping (Mon, 11 December 2023):

This session will be pretty hands-on.

13:15–15:30 (NSG, SR 205; videos): Web scraping in practice // APIs

- Chapter 24 in the R4DS book
- Webscraping 101 by the developers of `rvest`
- Munzert et al. (2014)
- Overview of R packages for accessing APIs
- The `httr` documentation

Intro to Text Mining // Text Preprocessing (Fri, 15 December 2023):

11:15–12:45 (NSG, SR 124): Theoretical Introduction to Text as Data // How to Measure Things with Text?

Theoretical background:

- Cointet and Parasie (2018)
- DiMaggio (2015)
- Evans and Aceves (2016)
- Grimmer and Stewart (2013)
- Grimmer, Roberts, and Stewart (2022), chapter 2, 15
- **Michel et al. (2011)**
- Underwood (2012) (online)

13:30–15:00 (videos): Practical Introduction to Text as Data

- Silge and Robinson (2017), chapters *1*, & *4*, and *5*

Sentiment Analysis and Feature Extraction (Mon, 18 December 2023)

13:15–14:00 (NSG, SR 205): Dictionary-based Analysis

- Grimmer et al. (2022), chapters 16
- **Garcia and Rimé (2019)**

14:15–15:00 (NSG, SR 205): Feature Extraction

- Grimmer et al. (2022), chapters 11
- Monroe, Colaresi, and Quinn (2008)
- **Bail (2016)**

15:00–16:30 (videos): Sentiment Analysis and Feature Extraction in R

- Bail (n.d.a) (*online*)
- Silge and Robinson (2017), chapter 2 (*online*) & 3 (*online*)

Supervised and Unsupervised Text Classification (Fri, 05 January 2024)

11:15–12:00 (NSG, SR 104): Supervised Classification

- **Bonikowski, Luo, and Stuhler (2022)**
- Grimmer et al. (2022), chapters 17–20

12:15-13:00 (NSG, SR 104): Unsupervised Classification

- Blei (2012)
- DiMaggio, Nag, and Blei (2013)
- Grimmer et al. (2022), chapters 10, 12–3
- **Heiberger, Munoz-Najar Galvez, and McFarland (2021)**

13:30-15:00 (videos): Supervised and Unsupervised Text Classification in R

- Hvitfeldt and Silge (2022), chapter 6 (*online*) & 7 (*online*)
- Silge and Robinson (2017), chapter 6 (*online*)
- Silge and Hvitfeldt (2019) (*online*)

New Directions in Text Analysis (Word Embeddings, Large Language Models) – and how Social Scientists can use them (Mon, 08 January 2024)

13:15–14:45 (NSG, SR 205)

- Do, Ollion, and Shen (2022)
- Törnberg (2023)
- Ollion et al. (2023)

15:00–16:30 (videos)

- Bail (n.d.b)

Final session: project presentation (Fri, 02 February 2024)

11:15–15:00 (NSG, SR 124)

no readings.

References

Bail, Christopher A. 2016. “Combining Natural Language Processing and Network Analysis to Examine How Advocacy Organizations Stimulate Conversation on Social Media.” *PNAS* 113(42):11823–28.

Bail, Christopher A. n.d.a. “Dictionary-Based Text Analysis in R.”

Bail, Christopher A. n.d.b. “Word Embeddings.”

Blei, David M. 2012. “Probabilistic Topic Models.” *Communications of the ACM* 55(4):77–84.

Bonikowski, Bart, Yuchen Luo, and Oscar Stuhler. 2022. “Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models.” *Sociological Methods & Research* 51(4):1721–87.

Cointet, Jean-Philippe and Sylvain Parasie. 2018. “Ce que le big data fait à l’analyse sociologique des textes: Un panorama critique des recherches contemporaines.” *Revue française de sociologie* 59(3):533.

DiMaggio, Paul. 2015. “Adapting Computational Text Analysis to Social Science (and Vice Versa).” *Big Data & Society* 2(2):205395171560290.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. “Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding.” *Poetics* 41(6):570–606.

Do, Salomé, Étienne Ollion, and Rubing Shen. 2022. “The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy.” *Sociological Methods & Research* 004912412211345.

Edelmann, Achim, Tom Wolff, Danielle Montagne, and Christopher A. Bail. 2020. “Computational Social Science and Sociology.” *Annual Review of Sociology* 46(1):61–81.

Evans, James A. and Pedro Aceves. 2016. “Machine Translation: Mining Text for Social Theory.” *Annual Review of Sociology* 42(1):21–50.

Garcia, David and Bernard Rimé. 2019. “Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack.” *Psychological Science* 30(4):617–28.

Golder, Scott A. and Michael W. Macy. 2014. “Digital Footprints: Opportunities and Challenges for Online Social Research.” *Annual Review of Sociology* 40(1):129–52.

Grimmer, Justin, Margaret Roberts, and Brandon Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton: Princeton University Press.

Grimmer, Justin and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21(3):267–97.

Heiberger, Raphael H., Sebastian Munoz-Najar Galvez, and Daniel A. McFarland. 2021. “Facets of Specialization and Its Relation to Career Success: An Analysis of U.S. Sociology, 1980 to 2015.” *American Sociological Review* 86(6):1164–92.

Hvitfeldt, Emil and Julia Silge. 2022. *Supervised Machine Learning for Text Analysis in R.* First edition. Boca Raton London New York: CRC Press, Taylor & Francis Group.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. 2009. “Life in the Network: The Coming Age of Computational Social Science.” *Science* 323(5915):721–23.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. “Quantitative Analysis of Culture Using Millions of Digitized Books.” *Science* 331(6014):176–82.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict.” *Political Analysis* 16(4):372–403.

Munzert, Simon, Christian Rubba, Peter Meißner, and Dominik Nyhuis. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining.* Chichester, West Sussex, United Kingdom: Wiley.

Ollion, Etienne, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. *ChatGPT for Text Annotation? Mind the Hype! Preprint.* SocArXiv.

Salganik, Matthew J. 2018. *Bit By Bit: Social Research in the Digital Age.* Princeton: Princeton University Press.

Silge, Julia and Emil Hvitfeldt. 2019. “Predictive Modeling with Text Using Tidy Data Principles.” in *useR2020.*

Silge, Julia and David Robinson. 2017. *Text Mining with R: A Tidy Approach.* First edition. Beijing ; Boston: O’Reilly.

Törnberg, Petter. 2023. “How to Use LLMs for Text Analysis.”

Underwood, Ted. 2012. “Where to Start with Text Mining.”