06-002-107-3: Forschungsseminar B – Experimentelle Soziologie und Computational Social Science

Felix Lennert

Winter 2025/26

- E-mail: felix.lennert@uni-leipzig.de
- Course hours: Tuesdays and Thursdays, 15:15 – 16:45, NSG, SR 325; for exact dates, see schedule
- Readings: see schedule
- Course materials: see website
- Student hours: are to be set up individually via email; my office is H3 1.07; there is a multitude of valid reasons why you should come see me – some are listed here:
  - things are unclear and you need help with the material
  - you want to discuss a research idea
  - you have come across a cool new paper that I might deem interesting
  - you have recommendations for me in terms of general course resources/references/structure/behavior
  - you want some career advice from someone roughly your age
  - you forgot your mensa card at home and want some coffee/tea
  - in case you need some free period supplies, no email is required, you can just get them at Leonie Steinbrinker's office (I also have a key if she's not there), H3 1.06

Course description

The Forschungsseminar in Computational Social Science (CSS) equips you with the tools to analyze human behavior, predict social trends, and tackle complex societal issues using cutting-edge data science techniques. From web scraping to AI-powered text analysis, you'll learn to use your computer in new ways to gain insights into social phenomena.

The curriculum covers a range of topics including data management, web scraping, speech-to-text, and computational text analysis. Students will hone their R and develop skills in Python, applying these languages to real-world social science problems. The course progresses from fundamental concepts to advanced techniques, including the use of state-of-the-art AI models for text analysis.

The course structure consists of one lecture and one lab session per week, providing a balance of theoretical knowledge and practical application. Throughout the semester, students will benefit from hands-on coding exercises, one-on-one mentoring, and collaborative projects. The course culminates in a research paper, allowing students to apply their new skills to a topic of their choice.

This course is ideal for social scientists looking to enhance their computational skills. It is geared toward 2nd year Master's students who are enrolled in the reformed Sociology Master's program. Interested Bachelor's students are also very welcome to attend, they shall send an email to the instructor no later than October 10, 2025. However, Bachelor's students will not be able to earn credits with their attendance.

What to expect

This course is structured so that it provides you – the student – with theoretical lectures (on Tuesdays) and hands-on coding labs (Thursdays). The lectures aim to introduce you to how you can use techniques

to conduct empirical research and will mostly consist of the presentation of innovative and/or cutting-edge research that has harnessed digital data to solve exciting research puzzles. They also might contain some practical demonstrations already. The practical sessions then showcase how to perform similar things yourself and give you ample opportunities to perform the analyses on your own data. Exercises will be provided, but you are not expected to hand them in – I could not care less and not doing them is simply your loss. However, I am of course always available to help you with the exercises if you get stuck.

The main objective of this course is that students **perform their own research as the course unfolds**. To this end, we will talk about your interests. Study groups of matching research interests might be formed. Students can of course also work individually. Students are expected to continue working on their projects as the course unfolds.

A tried-and-tested strategy is to work on a topic you could later use for your Master's thesis. However, this is not a requirement. In this case, this course would give you more time for the tedious data acquisition and cleaning phase, which is often underestimated.

At the very end of the course, it is time for a "peer reviewed presentations" of your project. One of your peers will be assigned your project and provide comments on it. 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 actual "peer-review" will work as follows: first, you (the author of the project) present your work briefly (10 minutes). Then, the assigned reviewer group will provide their comments (5 minutes). 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 30, 2026**, so I can distribute them to your peers.

As a peer reviewer, you have 5 minutes to provide feedback on four components:

(1) Research Question & Motivation: Is it clear and compelling? Are there ways to sharpen the focus?
(2) Theory & Hypotheses – Is relevant literature cited? Are there gaps or alternative frameworks to consider?
(3) Data & Methods – Is the approach appropriate? Are there concerns about data quality or better-suited methods?
(4) Preliminary Results – Are visualizations clear? Do results address the hypotheses? What additional analyses might help?

Please be constructive rather than critical, frame suggestions as questions ("Have you considered. . . ?"), be specific with your feedback, and acknowledge that everyone is working within constraints. Aim to identify 2-3 main strengths and 2-3 areas for improvement, and suggest specific papers, methods, or framings where helpful.

The (tentative) deadline for the paper ("Forschungsbericht") is **March 19, 2026**. Please send it – in PDF format – to Simone Müller (muellers@uni-leipzig.de) and CC me. Code and data have to be sent to me via 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. I have no preferences regarding the programming language (R, Python), therefore your project can be in either language or a mix of both (e.g., using `reticulate` in R or `quarto` notebooks – the latter being my suggestion).

Extensions

Extensions can be granted for particular 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 a different 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
  - written in English
  - file format: PDF
  - 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 8,000 words
  - cite correctly and in a uniform manner. My preferred citation style is ASA. It is strongly advised to use Zotero and Quarto/Overleaf; resources can be provided upon request.
- Structured resembles 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 the light 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 code from the internet – but you are responsible for the solution, so please make sure it works and solves your problem.
- Generative AI (i.e., ChatGPT et al.) 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 – and only you – will be held accountable for its content.
- 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

As stated above, Tuesdays are lectures and Thursdays are lab sessions. Please bring a laptop to all of our meetings (if you don't have one, feel free to reach out and we will try our best to lend you one). It's also a good idea to charge it before, since outlets appear a bit scant.

Most of you will use RStudio. I personally use Positron, which is basically the same but more accomodating to Python. You can download it for free from posit.co. However, RStudio is perfectly fine, too.

Literature-wise, we will use a mix of textbooks and review papers to introduce theoretical concepts and related studies to illustrate. In terms of programming, we will mostly rely on online resources. However, everything that is relevant in terms of R content (and more!) can be found in the R script.

Every reading will be either provided online or linked to in this syllabus (just click on *"online"*) – the link is hidden.

I do not expect you to read the literature and will do the theoretical sessions in a "lecture" style. This is because this is an applied course and not every piece of content has the same relevance for everyone. Having been a student myself, I think that students should not be overwhelmed by having to read everything while working on their projects. I will upload additional readings on top of the ones this syllabus mentions in case you want to read more and need some inspiration. I also recommend looking at the references of the papers we read to find more relevant literature or just ask me anytime.

Week 1: Kick Off

Welcome & Housekeeping (Tue, 14 October 2025; 15:15 – 16:45; NSG, SR 325)

No readings.

Setting up your workstation (Thu, 16 October 2025; 15:15 – 16:45; NSG, SR 325)

- Acquire access to *sc.uni-leipzig.de*
- R recap (the corresponding chapters can be found in the R4DS book – _online)
  - `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
  - set up `reticulate` in RStudio – *online*

Week 2: Brief Intro to Python & Regexes

Python & Regexes(Tue, 21 October 2025; 15:15 – 16:45; NSG, SR 325)

- Brief intro to Python
  - *reticulate* – how to run Python in R studio
  - data types
  - loops
  - functions
  - `pandas`

Regular Expressions (Thu, 23 October 2025; 15:15 – 16:45; NSG, SR 325)

- `stringr` & Regular Expressions – R4DS book, *online*, chapters 15 & 16

Week 3: Data Acquisition I

How the Web is Written and Ethics (Tue, 28 October 2025; 15:15 – 16:45; NSG, SR 325)

- Stoltz and Taylor (2024) – chapter 5
- Blog post on CSS selectors – *online*
- Blog posts on API calls – *online*

`rvest`

- `rvest` Web scraping 101 – *online*

Week 4: Data Acquisition II

Dynamic Pages and Forms (Tue, 04 November 2025; 15:15 – 16:45; NSG, SR 325)

- `selenium` documentation – *online*

APIs (Thu, 13 November 2025; 15:15 – 16:45; NSG, SR 325)

- `httr2` documentation – *online*

Week 5: Data Acquisition III

Intro to OCR and Transcription (Thu, 13 November 2025; 15:15 – 16:45; NSG, SR 325)

- Stoltz and Taylor (2024) – chapter 5

Week 6: Data Acquisition IV

Optical Character Recognition and Transcription (Tue, 18 November 2025; 15:15 – 16:45; NSG, SR 325)

- `Tesseract` documentation – *online*
- OpenAI Whisper Python package documentation – *online*

Data Acquisition Recap and Project Discussion (Thu, 20 November 2025; 15:15 – 16:45; NSG, SR 325)

Week 7: Student Project Week

Students are expected to show up to class and work on their projects.

Week 8: Text as Data I

Bag of Words(Tue, 02 December 2025; 15:15 – 16:45; NSG, SR 325)

- Evans and Aceves (2016)
- Grimmer, Roberts, and Stewart (2022), chapters 3–5, 11, & 15
- Stoltz and Taylor (2024), chapters 4–9

Sentiment Analysis, TF-IDF, and NER/POS (Thu, 04 December 2025; 15:15 – 16:45; NSG, SR 325)

- Grimmer et al. (2022), chapter 11
- Jurafsky and Martin (n.d.), chapter 21 – *online*
- Silge and Robinson (2017) – *online*, chapters 2 & 3

Week 9: Text as Data II

Supervised Machine Learning in Theory (Tue, 09 December 2025; 15:15 – 16:45; NSG, SR 325)

Supervised ML

- Barberá et al. (2021)
- Grimmer et al. (2022), chapters 17–20
- Stoltz and Taylor (2024), chapters 9 & 12

Supervised Machine Learning in Practice (Thu, 11 December 2025; 15:15 – 16:45; NSG, SR 325)

- Hvitfeldt and Silge (2022) – *online*, chapters 6 & 7
- Silge and Hvitfeldt (2019) – *online*

Week 10: Text as Data III

Unsupervised ML in Theory and Practice (Tue, 16 December 2025; 15:15 – 16:45; NSG, SR 325)

- Blei (2012)
- DiMaggio, Nag, and Blei (2013)
- Grimmer et al. (2022), chapters 10, 12–3
- Stoltz and Taylor (2024), chapters 10 & 11
- Silge and Robinson (2017) – *online*, chapter 6

Remote Counseling pre-Christmas Break (Thu, 18 December 2025; 15:15 – 16:45; NSG, SR 325)

1-on-1 project counseling available on Zoom.

—- CHRISTMAS BREAK —-

Week 11: Text as Data IV

Measuring Similarity and the Distributional Hypothesis (Tue, 06 January 2026; 15:15 – 16:45; NSG, SR 325)

- Jurafsky and Martin (n.d.), chapter 6 – *online*
- Stoltz and Taylor (2021)

Word Embeddings (Thu, 08 January 2026; 15:15 – 16:45; NSG, SR 325)

- Hvitfeldt and Silge (2022) – *online*, chapter 5
- Stoltz and Taylor (2024), chapter 11
- `text2map`: R Tools for Text Matrices – *online*

Week 12: Text as Data V

Supervised Learning on Steroids: BERT (Tue, 13 January 2026; 15:15 – 16:45; NSG, SR 325)

- Do, Ollion, and Shen (2022)
- Laurer et al. (2024)
- Törnberg (2023)
- Wankmüller (2022)

Active Learning with BERT (Thu, 15 January 2026; 15:15 – 16:45; NSG, SR 325)

- set up environments
- Augmented Social Scientist tutorial – *online*
- BERTopic – *online*

Week 13: Text as Data VI

LLMs for information extraction (Tue, 20 January 2026; 15:15 – 16:45; NSG, SR 325)

- Stuhler, Ton, and Ollion (2025)

Local LLMs – a primer (Thu, 22 January 2026; 15:15 – 16:45; NSG, SR 325)

- `ellmer` documentation – *online*
- Tutorial from IC2S2 by Etienne Ollion, Emilien Schultz, Julien Boelaert – *online*

Week 14: Presentation Preparation Week

No classes. Deadline for sending presentations: January 30, 6PM.

Week 15: Presentation & Wrap Up Week

Presentations (Tue, 03 February 2026; 15:15 – 16:45; NSG, SR 325)

No readings.

Presentations & Wrap-up (Thu, 05 February 2026; 15:15 – 16:45; NSG, SR 325)

No readings.

Deadline Forschungsbericht

March 19, 2026 (tentative).

References

Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29(1):19–42.

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

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.

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

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.

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.

Jurafsky, Dan and James Martin. n.d. "Speech and Language Processing."

Laurer, Moritz, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. "Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI." *Political Analysis* 32(1):84–100.

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.

Stoltz, Dustin S. and Marshall A. Taylor. 2021. "Cultural Cartography with Word Embeddings." *Poetics* 88:101567.

Stoltz, Dustin S. and Marshall A. Taylor. 2024. *Mapping Texts: Computational Text Analysis for the Social Sciences*. New York, NY: Oxford University Press.

Stuhler, Oscar, Cat Dang Ton, and Etienne Ollion. 2025. "From Codebooks to Promptbooks: Extracting Information from Text with Generative Large Language Models." *Sociological Methods & Research* 54(3):794–848.

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

Wankmüller, Sandra. 2022. "Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis." *Sociological Methods & Research* 004912412211345.