PWM-CS-HS3: Computational Social Science III Computational Text Analysis for Social Science Syllabus

Sergei Pashakhin, M.A.

1 Description

Social and political processes are often accompanied by a written text: from bureaucracies, parliament speeches, and print media to job advertisements and medical records. We can consider texts as traces as well as outcomes of such processes. The ever-increasing penetration of digital technologies into daily life dramatically multiplies volumes of available texts and opens new frontiers for social sciences. Advances in computer science (CS) and linguistics (CL) provide a wide range of tools to approach this mass of data and to look at social science questions from new angles. In this seminar, we will learn ways to ask research questions with text data and what tools are available to help us find the evidence. We will read and discuss research papers and book chapters. And we will practice using the most common tools by replicating published analyses.

1.1 Topics

1. Introduction. Text as data: from close reading to content analysis and text mining.

2. Making texts useful for research: preprocessing and models of text as data.

3. Units of analysis: what can we measure and how; and when do we need humans to do it?

4. Relying on text data in research design: classification, prediction, and clustering tasks.

5. Topic modeling and the best practices for applying bleeding-edge tools from CS & CL in the context of social sciences.

6. Extra. Using large pre-trained language models for computational text analysis in `R`.

2 Prerequisites

- Participants *must be* familiar with the basics of quantitative research (data analysis) and the `R` programming environment.
- Participants *must have* access to a computer to write and run code for home assignments.
- If possible, participants are encouraged to bring personal laptops to the seminar.

3 Required software

- The latest version of `R`: `https://cran.r-project.org/`
- The latest free version of `RStudio`: `https://www.rstudio.com/products/rstudio/download/`
- Instead of `RStudio`, it is possible to use other interactive development environments that support `R` and Unicode. Examples include `Jupyter`, `VSCode`, `GNU Emacs`, etc.

4 Recommended literature

- Grimmer, J., & Stewart, B. M. (2013). *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.* Political Analysis, 21(03), 267–297. `https://doi.org/10.1093/pan/mps028`
- Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach.* O'Reilly Media. `https://www.tidytextmining.com/`
- Atteveldt, W. van, Trilling, D., & Arcíla, C. (2021). *Computational analysis of communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* John Wiley & Sons. `https://cssbook.net/`

- Ashish, K., & Avinash, P. (2016). *Master text-taming techniques and build effective text-processing applications with R*. Packt Publishing. https://learning.oreilly.com/library/view/mastering-text-mining/9781783551811/

- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

5 Learning objectives

- Understanding the basics of computational text analysis.
- Knowledge of how to build research design relying on text data.
- Knowledge of best practices for applying the latest methods from computer science and computational linguistics in the context of social science.
- Foundational skills in solving most common problems with text data in the `R` programming environment.

6 Grading

- Participants will get a grade for the course based on their paper (10 pages) developed during the semester. For code review, the grading principles are (1) evidence of independent work and (2) application of the best practices of reproducible research.

7 Language

- English

8 Office hours

- By appointment.

9 Contacts

- sergei.pashakhin@uni-bamberg.de
- Feldkirchenstraße 21, Raum FMA/01.14