

Natural Language Processing

Version: 24 February 2025

Instructor: Dr. Jens Wäckerle

Time: Friday, 9. May 2025, 14:00 - 18:00, Saturday, 10. May 2025, 9:00 - 18:00, Friday, 16. May 2025, 14:00 - 18:00, Saturday, 17. May 2025, 9:00 - 18:00

Email

jens.waeckerle@uni-koeln.de

Office hours

Available on request, please send an email.

Course description

This module provides an introduction to Natural Language Processing (NLP). It covers classical approaches to NLP and gives an overview of current methods. Its main emphasis will be on the discussion and analysis of diverse NLP tasks, algorithms for solving them successfully, and techniques for assessing their performance. The module shows how text data sources can be used for social data science.

Data and empirical applications that we will use in class primarily focus on political science, but don't require any previous knowledge of political science. Most of the class will use the ParlLawSpeech dataset (<https://parllawspeech.org/data/>), as a running example. Throughout the course, students will become familiar with analysing text data using R.

Learning Objectives/General Skills:

- Students develop an understanding of natural language processing tools.
- Students develop a general literacy in natural language processing, which allows them to critically evaluate existing literature in the field.
- Students learn to use natural language processing to empirically test theoretical relationships in political science.
- Students are prepared to complete their own research projects autonomously with the help of the software R.

Prerequisites

Students should have a good understanding of quantitative methods as well as R.

Software

The open-source statistical programming language R is particularly suited for carrying out state-of-the-art computer-based simulations and data exercises. It also generates really nice publication-quality graphics. The software runs under a wide array of operating systems. R can be downloaded for free at <http://www.r-project.org/>. Learning R might seem a bit challenging at first, but you will realise that it is incredibly powerful. The lab sessions will be devoted to learning data management and analysis techniques using R.

In addition, we will use RStudio. This is a more user-friendly interface for R with integration to other packages we will use throughout the course. RStudio is available for free at <https://www.rstudio.com/>. In order to avoid spending too much time on installing the programmes in the beginning, students must have installed R and RStudio before the first session. If you are running into errors when installing, contact the instructor before the first session.

If you need a refresher on R, you can check out this website (<https://github.com/jenswaeckerle/IntroRUzK>), where I have provided self-learning modules on the basics of R as well as visualization of data.

Readings

The course is structured as an applied introduction to quantitative analysis. As readings, we will focus on textbook sections as well as scientific articles employing the various approaches. All readings will be provided through Moodle.

Course Requirements

Grading will be based on the following requirements:

- **Homework assignment (25%):** After the first block of classes (9./10. May), there will be a take-home homework assignment with a deadline at the start of the following class (16. May). This homework assignment recaps the application we discussed throughout the first weekend and can be completed in R. The assignments will be uploaded on Moodle. I encourage you to work in small groups on the assignments. Usually 2-3 people per group seems to work best. If you have worked with another student, please indicate with whom you did so on your homework. Moreover, you are strongly encouraged to seek advice during office hours or by posting questions in the forum.
- **Final Paper (75%):** The final research paper of 4,000 words (+/- 10%) is due XXX weeks after the last session of the course (Deadline: XXX). You are free to choose any research question from social science and use any of the methods presented in the course. During the last session, we will have time to talk about your research questions and methods. In the first session, I will give more details on the final paper.

Schedule

Day 1, Session 1 (09.05.2025, 14:00-16:00): Introduction to Natural Language Processing

- What are quantitative text analysis and natural language processing?
- What are the expectations for this course?
- How is textual data different and similar to other data?

Literature:

- K. Benoit (2020). ["Text as Data: An Overview"](#). *Handbook of Research Methods in Political Science and International Relations*. Ed. by L. Curini and R. Franzese. Thousand Oaks: Sage: 461–497.

Day 1, Session 2 (09.05.2025, 16:00-18:00): Dealing with text in R

- How can we get texts into R and into a format we can analyze?
- What are tokens, types and features?
- What preprocessing steps do we need to do?

Literature:

- J. Grimmer, M. E. Roberts, and B. M. Stewart (2022). [Text as Data: A New Framework for Machine Learning and the Social Sciences](#). Princeton: Princeton University Press: chapter 5.
- K. Watanabe and S. Müller (2023). Quanteda Tutorials. url: <https://tutorials.quanteda.io>: chapter 2–3.

Day 2, Session 1 (10.05.2025, 09:00-11:00): Dictionaries and Sentiment

- What can we learn by counting words in texts?
- What is sentiment, how is it measured?

Literature:

- S.-O. Proksch, W. Lowe, J. Wäckerle, and S. N. Soroka (2019). “[Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches](#)”. Legislative Studies Quarterly 44 (1): 97–131.

Day 2, Session 2 (10.05.2025, 11:00-13:00): Complexity

- How complex are texts?
- Who speaks in a more complex way?

Literature:

- K. Benoit, K. Munger, and A. Spirling (2019). “[Measuring and Explaining Political Sophistication Through Textual Complexity](#)”. American Journal of Political Science 63 (2): 491– 508.

Day 2, Session 3 (10.05.2025, 14:00-16:00): Similarity and Scaling

- What can we learn by knowing which texts are more similar to each other?
- What are word embeddings and how can we use them?

Literature:

- P. L. Rodriguez and A. Spirling (2022). “[Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research](#)”. The Journal of Politics 84 (1): 101–115.

Day 2, Session 4 (10.05.2025, 16:00-18:00): Development of research questions for natural language processing

- What questions that interest you can (only) be answered using NLP?
- What is the data generating process we have to model?

Day 3, Session 1 (16.05.2025, 14:00-16:00): Document Classification and Human Coding

- How can we use machine learning to code documents?
- What are measures to evaluate classification quality?

Literature:

- S. Lacy, B. R. Watson, D. Riffe, and J. Lovejoy (2015). "[Issues and Best Practices in Content Analysis](#)". Journalism & Mass Communication Quarterly 92 (4): 791–811.

Day 3, Session 2 (16.05.2025, 16:00-18:00): Introducing LLMs

- What are Large Language Models, how do they work, and how can and should we use them in the social sciences?

Day 4, Session 1 (17.05.2025, 09:00-11:00): Using LLMs for document classification

Literature:

- F. Gilardi, M. Alizadeh, and M. Kubli (2023). "[ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks](#)". Proceedings of the National Academy of Sciences of the United States of America 120 (3): e2305016120.

Day 4, Session 2 (17.05.2025, 11:00-13:00): Topic Models

- What can we learn about the content of text from topic models?

Literature:

- T. Gessler (2022). "Topic Models". Elgar Encyclopedia of Technology and Politics. Ed. by A. Ceron. Cheltenham: Edward Elgar Publishing: 108–111.

Day 4, Session 3 (17.05.2025, 14:00-18:00): Developing your research question and Wrap-up

- What else are you interested in? What are the next steps?
- Talk about your final paper: What are your ideas?