Clara Heinrich
John-F.-Kennedy Institute
Freie Universität Berlin
Email: clara.heinrich@fu-berlin.de

32602 –Text Analysis in R
Winter term 2023/24
Room: PC pool, JFKI (Lansstraße 7-9)

Block seminar
Tuesday, 17.10.2023, 10am - noon
Tuesday, 24.10.2023, 10am - noon
Friday, 17.11.2023, 11am - 4.30pm
Friday, 08.12.2023, 12am - 4.30pm
Friday, 02.02.2024, 11am - 4.30pm
Tuesday, 06.02.2024., 10am - noon
Tuesday, 13.02.2024., 10am - noon

Course Description

This block seminar introduces students to text analysis in social sciences research. Starting from basic exploratory analysis, the course will continue to cover dictionary based analysis, sentiment analysis as well as more complex language models.

The course is organized in three „computer lab“- sessions and students are expected to work independently in the weeks between the sessions.

Students are also expected to have basic programming skills in R or, if not, the motivation to learn the basics through self-study. We will also use statistical analysis, and students with no prior knowledge of statistics should be willing to learn the basic concepts on their own (for which case I am happy to provide materials). **TEST**

Course Objectives

The seminar will provide students with a general understanding of:

- Different methods of text analysis and their application in R
  - → Dictionary-based analysis & dictionary construction
  - → Sentiment analysis
  - → Topic modelling
  - → Natural language processing
  - → Discourse network analysis
- Existing textual data sources for social science research
- Approaches to generate textual data from primary data sources

Course Prerequisites

Students are expected to have basic programming skills in R or, if not, the motivation to learn the basics through self-study. Moreover, students with no prior knowledge of statistics should be willing to learn the basic concepts on their own.

Recommended material:

Zoë Field, Jeremy Miles, and Andy Field. *Discovering statistics using R.* London: Sage, 2012

→ for R and statistics

Simon Munzert, Christian Rubba, Peter Meißner, and Dominic Nyhuis. *Automated data collection with R: A practical guide to web scraping and text mining.* West Sussex: John Wiley & Sons, 2014

→ for web-data collection

Politikwissenschaftliche Statistik mit R. By Christoph Garwe, Philipp Meyer, Laura Brune and Christoph Hönnige, open access [in German]

Online series of tutorial introducing students to R and statistics by Prof. Kohl

Introduction to R: Tidyverse Skills for Data Science

R for Data Science: Extensive introduction to R

Specifically on statistics:

Timothy C Urdan. *Statistics in plain English.* New York: Routledge, 2017

Recorded lectures on basic statistics for social sciences by Prof. Kohl

---

Recommended Preliminary Readings

Klaus Krippendorff. *Content analysis: An introduction to its methodology.* Thousand Oaks: Sage publications, 2019

Philip Leifeld. Discourse network analysis. *The Oxford handbook of political networks*, pages 301–326, 2017

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken AC G van der Velden. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18, 2022

Follow RStats Question A Day on Twitter for daily tips and tricks in R.

Office Hours

You are invited to come and talk to me whenever you feel like doing so and I am happy to discuss anything from questions on homework, the course content, reading or your term paper. Regular office hours are held Tuesdays 4-5pm at my office (room 328, JFKI). You need to register for the office hour via email beforehand. To accommodate students with caring responsibilities or other commitments, I try my best to find flexible solutions for those who cannot come to the regular office hours. Please send me an email to discuss possible options.

Course Structure & Modes of Teaching

Communication

This course is meant to be a space in which we all learn with and from each other. My aim is to create an atmosphere in which everyone feels comfortable to speak, ask questions and comment on others in a respectful manner. Therefore, I rely on your feedback if you feel like the material is too difficult/ inaccessible for you, if you lack basic knowledge of certain terms or topics, but also if there is something else that makes you feel uncomfortable. Moreover, I would appreciate it if you could let me know if you are unable to attend a session, as this makes it easier for me to plan the sessions in order to make them as lively and interesting as possible.

Recording the seminar sessions is strictly prohibited.

Course Materials

The basic materials of the course are RScripts, which you can find on BlackBoard (Folder: Scripts), and data files (Folder: Data). In addition to the scripts, there will be data homework, which you should complete by using the respective script (Folder: Homework) and upload it under assignments on BlackBoard. You also find the syllabus and, if applicable, the readings and slides for the individual sessions on BlackBoard. Scripts and data for each session will be uploaded on the day of the session and should be downloaded by students before/at the beginning of class.

Participation and Examination

For participation, students are expected to actively participate in class. While I do not want to force anyone to speak out loud in class, my aim is that everyone participates according their*her*his possibilities.

For participation credit, the data **homework must be submitted in the week before the upcoming lab session on Wednesday, 11am**. For full credit, students also have to prepare a short student paper pitch in class (10 min maximal) and hand in the **finalized student paper** (MA: 20 pages, 8000 words main text, BA: 12 pages, 5000 words main text) as a PDF via email the latest on **31st March 23:59**. The student paper should develop a research question and answer it using data and techniques learned in class. Discussing your ideas with me during an office hour (prior to the pitch in class) is mandatory.

Class Schedule

Session 1: 17.10.2023, Introductory Session

Topics covered:

- What is this course about?
- Why R and how does the program work?
- A primer on content analysis & text as data

Mandatory readings:

Cornel Ban. Content analysis in international political economy. In *The Oxford Handbook of International Political Economy.* Oxford University Press, 2021

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken AC G van der Velden. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18, 2022

Session 2: 24.10.2023, Why studying texts?

Topics covered:

- Basic concepts of content analysis
- Theoretical reflections on text as data: Power, discourse, social interaction

Mandatory readings:

Susan Strange. *States and Markets: An Introduction to International Political Economy.* London: Pinter. Chapter 2 (Power in the World Economy), 1994

Carsten Schwemmer and Oliver Wieczorek. The methodological divide of sociology: Evidence from two decades of journal publications. *Sociology*, 54(1):3–21, 2020

Cornel Ban. Content analysis in international political economy. In *The Oxford Handbook of International Political Economy.* Oxford University Press, 2021

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken AC G van der Velden. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18, 2022

Further readings:

Paul DiMaggio. Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2):2053951715602908, 2015

Klaus Krippendorff. *Content analysis: An introduction to its methodology.* Thousand Oaks: Sage publications, 2019

→ Especially chapter 1 (Introduction) & chapter 2 (Conceptual foundations)

Submit your data homework I via BlackBoard by Wednesday, 15th November 2023, 11am

---

Session 3: 17.11.2023, Lab session I: Text as data with *quanteda*

Topics covered:

- Data preparation
- Frequency analysis
- Wordclouds
- Key-word in context analysis
- Dictionary-based analysis

Mandatory reading:

Quanteda tutorials, Sections 1-7

Ana Macanovic. Text mining for social science–the state and the future of computational text analysis in sociology. *Social Science Research*, 108:102784, 2022

Qi Deng, Michael J Hine, Shaobo Ji, and Sujit Sur. Inside the black box of dictionary building for text analytics: a design science approach. *Journal of international technology and information management*, 27(3):119–159, 2019

Further readings:

Qihao Ji and Arthur A Raney. Developing and validating the self-transcendent emotion dictionary for text analysis. *PloS one*, 15(9):e0239050, 2020

Gary King, Patrick Lam, and Margaret E Roberts. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 61(4):971–988, 2017

Submit your data homework II via BlackBoard by Wednesday, 6th December 2023, 11am

---

Session 4: 08.12.2023, Lab session II: Sentiment analysis & topic modeling

Topics covered:

- Sentiment analysis
- Topic modeling

Mandatory reading:

Maurizio Naldi. A review of sentiment computation methods with r packages. *arXiv preprint arXiv:1901.08319*, 2019

Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. Keyword-assisted topic models. *American Journal of Political Science*, 2020

Further readings:

Martin Haselmayer and Marcelo Jenny. Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & quantity*, 51:2623–2646, 2017

Rob Churchill and Lisa Singh. The evolution of topic modeling. *ACM Computing Surveys*, 54 (10s):1–35, 2022

Pooja Kherwa and Poonam Bansal. Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), 7 2019. doi: 10.4108/eai.13-7-2018.159623

Submit your data homework III via BlackBoard by Wednesday, 25th January 2024, 11am

---

Session 5: 02.02.2024, Lab session III: More text as data – Networks, speech acts, grammatical parsing

Topics covered:

- Discourse network analysis
- Speech act analysis
- NLP package

Mandatory reading:

Philip Leifeld. Discourse network analysis. *The Oxford handbook of political networks*, pages 301–326, 2017

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019

Philip Leifeld. Policy debates and discourse network analysis: A research agenda. *Politics and Governance*, 8(2):180–183, 2020

Jan Goldenstein and Philipp Poschmann. Analyzing meaning in big data: Performing a map analysis using grammatical parsing and topic modeling. *Sociological Methodology*, 49(1):83–131, 2019

Submit your data homework IV via BlackBoard by Monday, 8th February 2024, 11am

---

Session 6: 06.02.2024, A primer on web-scraping with *rvest*

Topics covered:

- Static vs dynamic web-pages
- Downloading, parsing, extracting from HTML files
- Strategies for trouble shooting

Mandatory reading:

Simon Munzert, Christian Rubba, Peter Meißner, and Dominic Nyhuis. *Automated data collection with R: A practical guide to web scraping and text mining.* West Sussex: John Wiley & Sons, 2014

Session 7: 13.02.2024, Wrap-up

Topics covered:

- Course evaluation
- Term paper pitches
- Concluding discussion
