

GESIS Workshop Series

“Introduction to Computational Text Analysis with R”

Lecturers:	Jan Schwalbach	Lea Kaftan	Jens Wäckerle
Affiliation:	GESIS – Leibniz-Institut für Sozialwissenschaften	GESIS – Leibniz-Institut für Sozialwissenschaften	Universität zu Köln
Email:	jan.schwalbach@gesis.org	lea.kaftan@gesis.org	jens.waeckerle@uni-koeln.de

Date: 01.04.-03.04.2025

Time: 09:00-12:00 and 13:00-16:00 (including a 15-minute break per session)

Location: online

About the lecturers

Lea Kaftan is a postdoctoral researcher at GESIS – Leibniz Institute for the Social Sciences in Cologne. She obtained her doctorate in Political Science at the University of Cologne. Her research focuses on the preferences of voters and the strategies of parties. For her research, she has applied a variety of computational text analysis methods.

Jan Schwalbach is a postdoctoral researcher in the team Archiving at the GESIS department Data Services for the Social Sciences. His research focus revolves around digital behavioral data and legislative politics. This also includes a strong interest in methodological approaches such as computational text analysis or survey experiments, as well as the provision and harmonization of large data sets.

Jens Wäckerle is a post-doctoral researcher at the Cologne Center for Comparative Politics and currently the professorial substitute at the Chair of European Politics. His research focuses on women in politics, legislative politics, quantitative text analysis and the European Union. He holds a Master degree from the University of Essex and a PhD from the University of Cologne.

Course description

Computational text analysis is a fast-growing and widely used methods field, allowing researchers to structure and study large corpora of texts in-depth from a multitude of theoretical backgrounds. The workshop introduces key concepts, standard methods, and the research logic underlying computational text analysis. It is designed for researchers with no or limited experience with computational text analysis who want to analyze text data for their own research.

Participants will learn to pre-process their text data and to apply and validate basic supervised and unsupervised methods such as dictionaries (e.g. sentiment), classification methods (e.g. support vector machines or random forest), and clustering methods (e.g. topic models) to their own text data using the programming language R. Participants will additionally learn the benefits and pitfalls of research designs based on the quantitative analysis of large text corpora, including ethical issues, potential biases, and implications of choices regarding document selection and pre-processing. After the course, participants will be able to choose appropriate designs for their own research questions and apply standard methods of text analysis in R.

Keywords

computational text analysis, R, topic model, sentiment, machine learning, text mining

Categories (please tick all that apply)

Topics

- ☒ Computational Social Science
- ☐ Data Analysis
- ☐ Research Data Management
- ☐ Survey Methodology

Software

- ☐ MPlus
- ☐ Python
- ☒ R
- ☐ Stata

Level

- ☒ Beginner
- ☐ Advanced

Course prerequisites

- Basic knowledge of R is a prerequisite. The course cannot provide a general introduction to R. Participants should know how to import data, such as csv files, install packages, and work with objects in R.

Target group

Participants will find the course useful if:

- they want to work with text data in their research for the first time.
- they want to refresh prior basic knowledge in text analysis.
- they want to learn how to preprocess text data for quantitative analyses.
- they want an overview of standard computational approaches to analyzing text data.

Course and learning objectives

By the end of the course participants will:

- be able to design and execute their own research based on large text corpora.
- be able to preprocess, analyze, and visualize text data using R.
- understand the principles of supervised and unsupervised methods for computational text analysis.
- be able to independently familiarize themselves with more advanced methods.

Organizational structure of the course

The workshop contains both lectures and practical exercises. The lectures aim at providing a general understanding of the research logic and methods. The practical exercises deepen the participants' understanding of the methods and teach them how to apply the methods to their own data. During the exercises, the lecturers will be available for support, troubleshooting, and questions. Participants can bring their own data or work with data provided by the lecturers.

Software requirements

R (at least version 4.0.0) and RStudio. Prior to the workshop, participants will receive an R script to install all required packages.

Day-to-day schedule and literature

Day 1:

Morning session

- Welcome and introduction
- Theoretical foundations of text analysis (text selection, data generating process)

Afternoon session

- Introduction to R and its packages for text analysis
- Overview and application of text pre-processing techniques (e.g., data cleaning and tokenization)
- Hands-on activities to practice pre-processing on a sample text dataset

Day 2:

Morning session

- Regular expressions for text analysis
- Introduction to bag of words models
- Dictionary approaches and constructions (e.g., sentiment dictionaries)

Afternoon session

- Supervised classification of documents
- Hands-on application and recap

Day 3:

Morning session

- Unsupervised learning approaches
- Introduction to topic models
- Hands-on application

Afternoon session

- Scaling algorithms
- Beyond bag of words approaches
- Pros/cons of different text analysis methods
- Outlook and recap

Additional recommended literature

Hvitfeldt, Emil, and Julia Silge. (2021). *Supervised machine learning for text analysis in R*. CRC Press

Silge, Julia, and David Robinson. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Inc.

Welbers, Kasper, Wouter Van Atteveldt, and Kenneth Benoit. (2017) "Text Analysis in R". *Communication Methods and Measures*, 11 (4), 245-265