

# Text as Data in Economics

UNIVERSITY OF KASSEL

SUMMER 2024

**Instructor:** Dr. Burcu Ozgun  
burcuozgun@uni-kassel.de

**Medium of Instruction:** English

## Overview

This course aims to familiarize students with economic literature in which text is used as data and provide information on theory, methods and software for working with textual data in economic research. By the end of the course, students are expected to know more about how text can be, and has been, applied to answer economically relevant questions. Students will be able to do some analysis themselves and will be well equipped to take more detailed and advanced courses on the topic.

The course is taught as a series of workshops. Theories and methods will be introduced and discussed, applications will be shown, and students will be expected to perform some tasks during the lectures. This is a very hands-on course, accordingly students are expected to bring their own laptops to class and write and run their codes during the *in-class exercises* part at the end of each session.

## Software Requirements

The software requirements for this course are open-source. We will mainly be using the statistical programming language R. Please install R and RStudio IDE by the start of our first lecture.

## Pre-requisites

The course is aimed at PhD students with a background in economics or related fields. Working knowledge in R is a pre-requisite; the course material assumes that the students are already comfortable with programming with R.

## Evaluation and Grading

Attendance at all sessions is required. Evaluation will be based on performance *in-class exercises*, and two short final assignments, which will be explained during the last lecture and must be completed individually. Successful completion of the course will earn students 6 ECTS.

## Course Materials

There is no official textbook for this course. The lecture notes are pretty detailed and are thus self-contained. In addition to lecture notes, supplementary materials and relevant readings will be provided for students' reference.

All course-related information and materials, including announcements, lecture materials, and datasets, will be accessible through a private shared GitHub repository. Therefore, a GitHub profile is necessary.

## Accessibility

If you have a disability and/or anticipate needing accommodations for this course, please contact me to make arrangements at least one week before the course starts.

## Schedule

17 June 2024 Mon 10:00 – 16:30 in Kurt-Wolters-Straße 5 Room 1012  
18 June 2024 Tue 10:00 – 16:30 in Kurt-Wolters-Straße 5 Room 1012  
19 June 2024 Wed 10:00 – 16:30 in Kurt-Wolters-Straße 3 Room 2207  
20 June 2024 Thu 10:00 – 16:30 in Kurt-Wolters-Straße 3 Room 2207  
21 June 2024 Fri 10:00 – 16:30 in Kurt-Wolters-Straße 3 Room 2207

## Assignment Deadline

12 July 2023 Friday, 18:00.

## Office Hours

Individual slots will be made available for each student, at the beginning of the respective week. Office hours will be held virtually.

1. 26 June 2024 Wednesday 09:30-11:30 (Zoom Link)
2. 2 July 2024 Tuesday 13:30-15:30 (Zoom Link)
3. 8 July 2024 Monday 14:30-16:30 (Zoom Link)

## Outline

- General Introduction
  - Getting to know each other
  - Importance and relevance of text analysis in economics
  - What this course offers
- Good practices in R programming
- String manipulation and regular expressions
- Text representation
  - Text pre-processing
  - Vector space representation
  - The multinomial language model
  - Word embeddings and low dimensional document embeddings
  - Similarity metrics
  - Simple text analysis
- Text clustering
  - K-means clustering
  - Hierarchical clustering
  - Topic modeling
- Text classification
  - Dictionary-based methods
  - Text regression methods / Discriminative models
  - Generative models
- Inference with text data
- Collecting text data
  - Collecting text data from various sources
  - Good practices when working with *found* text data
  - Handling and working with different file formats
- A review and discussion of seminal papers in economics utilizing text analysis & NLP
- Conclusion
  - Overview of the key takeaways and limitations
  - Discussion on final assignments