Introduction to Text Mining

University of Kassel
Summer 2021-2022

Instructor: Burcu Ozgun
burcuozgun@uni-kassel.de

Medium of Instruction: English

Time: Tuesday 16:00 – 18:00 Möncheberg 7 Raum 0607

Wednesday 10:00 – 12:00 Untere Königsstraße 86 Raum 2034

Office Hour: Monday 10:45-11:45 (Zoom link) or by appointment via e-mail

Exam: 13 July 2022 Wednesday, 10:00

Overview

This course aims to give an introduction to text mining concepts and applications, and increase student awareness of the power of large amounts of text data and computational methods to find patterns in corpora.

The course is broken into three phases. Phase I gives a grounding in the basics of programming with R. Phase II introduces essential concepts and methods in text mining. Phase III demonstrates some interesting case studies and discusses the state-of-the-art techniques in text mining.

The course is taught as a series of workshops. Methods will be introduced and discussed, applications will be shown, and students will be expected to perform some tasks during the lectures.

There is no official textbook for this course. However, in addition to lecture notes, supplementary materials and relevant readings will be provided for students' reference.

The maximum number of participants (reflecting current capacity constraints) is 25. Places will be made available in the order of registration. To register, please send an e-mail to me and make sure you attend the first (introductory) lecture.

Credits and Grading

6 ECTS can be earned for (see HIS-POS for details) modul 1B (M.Sc. Economic Behaviour and Governance). Grading will be based on the exam.

Course Webpage

All information and materials regarding the course, including announcements, lecture materials etc. will be posted on the moodle. It is students' responsibility to check the course page at regular intervals and be up-to-date with all relevant information.

Outline

Phase I

1. Introduction (0.5 week)
2. Introduction to R (2 weeks)
3. Tidy data (1 week)
4. Visualization (1 week)
5. Good practices in R programming (0.5 week)

Phase II

1. String manipulation and regular expressions (1 week)
2. Text preprocessing (0.5 week)
3. Vector space model (1 week)
4. Similarity (0.5 week)
5. Text classification and text clustering (2 weeks)
6. Sentiment analysis (1 week)
7. Web and social media scraping (1 week)

Phase III

1. Case studies (1 week)
2. A review of state-of-the-art techniques (1 week)