Text analysis in R
(taught by Dr. Helge Liebert)

Much of human knowledge is stored in unstructured formats. Processing and analyzing unstructured text data is an elementary part of both research in modern social science and data science in industry. This course teaches methods to process and analyze unstructured data, focusing on text and web data. The first part of the lecture reviews tools and concepts for processing text data and introduces the fundamentals of web scraping. The second part focuses on different representation concepts underlying the transformation of unstructured text data into structured formats suited for statistical analysis. The last part introduces statistical models suited for the analysis of text data, focusing on both supervised models for prediction as well as unsupervised models which make it possible to discover structure in unlabeled text data. Throughout the course, I try to emphasize real-world applications of the techniques in research and industry. The methods taught in class are applied to example data sets using the statistical software R. All class material will be provided on a dedicated website.

Course objective

- A thorough understanding of the workflow, tools and models related to the analysis of text data.
- Improve data management workflow related to text.
- Understand the structure of web scrapers and write simple programs independently.
- Understand the advantages and disadvantages of different text data representation concepts.
- Understand the advantages and disadvantages of different models to analyze text data.

Content

- Introduction
- Regular expressions and pattern matching
- Web scraping
- Representing text as data: Count-based approaches
- Representing text as data: Prediction-based approaches
- Analysis of text data: Supervised models
- Analysis of text data: Unsupervised models

Prerequisites

- Basic knowledge of the statistical software “R” and introductory statistics (linear regression), as e.g. provided in the course “Introduction to R”. A basic understanding of predictive modeling concepts (e.g., a class on computational statistics) is helpful,but not required.

Duration

- 3 days on Feb 16 th, 17 th and 18 th (roughly 7*45 minutes each day)

Evaluation

- take home exam: project work to be solved in R

ECTS

- 1.5