Title
Text analysis for Social Science Research: Techniques for Data Collection, Analysis, and Causal Inference

Course Description

This course, designed for advanced undergraduate and postgraduate students in the social sciences, covers the basic concepts and advanced techniques of textual analysis. The course progresses through various aspects of text analysis to equip students with the skills necessary to extract information from text and to obtain empirical findings through hypothesis testing.

The course starts with an overview of natural language processing in the social sciences and teaches the prerequisite methods for data analysis, such as text pre-processing and tokenisation. You will also learn web scraping techniques as a method of data collection, enabling you to collect textual data from various online sources.

In the section on empirical analysis, you will get an overview of machine learning, which is inseparable from natural language processing, and then learn how to apply text analysis techniques. From topic modelling and sentiment analysis as exploratory analysis to statistical modelling of texts, you will learn techniques for deriving meaningful insights from textual data.

At the end of the course, you will learn more advanced topics. You will be introduced to Python for large-scale language models (LLM) and also get an overview of how to conduct causal inference with text as data. This will give you leads on how to apply state-of-the-art technology to your research.

Course Objectives

By taking this course, participants will have a foundation for conducting text analysis in social sciences mainly using R. In particular, they will learn:

- Understand textual analysis fundamentals in social science
- Master basic NLP techniques in preprocessing
- Learn web scraping for data collection
- Acquire the basics of machine learning for text analysis
- Utilize statistical modeling for analysis (e.g. Naïve Bayes, LASSO regression)
- Learn methods for exploratory analysis (e.g. topic modelling, scaling)
- Explore advanced topics: LLMs, causal inference with text

Required Textbooks (and Websites)

- Quanteda Tutorials by Kohei Watanabe and Stefan Muller (https://tutorials.quanteda.io/)

- An Introduction to Statistical Learning with Applications in R (ISL), Second Edition, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (Springer, pdf available at the book website: https://www.statlearning.com/)

Suggested textbook

- R for Data Science (R4DS), by Hadley Wickham (O'Reilly, available at: https://r4ds.had.co.nz/)
- Text as Data: A New Framework for Machine Learning and the Social Sciences by Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart (Princeton UP)
- Natural Language Processing with Transformers by Lewis Tunstall, Leandro Von Werra, and Thomas Wolf (O'Reilly).
- Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining by Simon Munzert, Christian Rubba, Peter Meißner, and Dominic Nyhuis (Weily)

Prerequisites

This course assumes that the participant has some knowledge of data analysis. The knowledge required is:

- Regression analysis (linear regression, logistic regression, etc.)
- Basic knowledge of the statistical language R

For regression analysis, it is recommended that students have taken an undergraduate or graduate course on statistical analysis for social scientists. For R, advanced knowledge is not required, but the ability to read and write files, manipulate data frames, define and use functions, and perform some statistical analysis (e.g. estimate linear regression models or categorical dependent variable models, interpret the results) is desirable. Knowledge of Python is not required but definitely a plus. Please consult with the instructor if you are unsure about any of the prerequisite knowledge.