Text-as-Data

CEU Political Science Department

Winter Semester 2024

Instructor: Petro Tolochko

petro.tolochko@univie.ac.at

With an increase in the availability of and access to textual information, social scientists have progressively embraced computer-assisted or automated text analysis methods. This paradigm shift has been fueled by the need for efficient tools capable of handling the immense scale of textual information. Consequently, the field of text-as-data has emerged as a vital domain within the broader scope of computational social sciences.

This course delves into the multifaceted realm of text-as-data methods, offering a comprehensive exploration of the techniques and tools employed by social scientists to extract meaningful insights from textual sources. Going beyond traditional manual methods of content analysis, the curriculum emphasizes automated approaches that leverage computational power to enhance efficiency, reproducibility, and validity in drawing inferences from documents.

Throughout the duration of the course, participants will engage with key components of text-as-data analysis, including data collection, data processing, quality control, and the nuanced interpretation of results. The overarching goal is to equip students with a sophisticated skill set that aligns with the evolving landscape of computational methodologies in social sciences.

Prerequisites: Familiarity with statistical modeling. Familiarity with R or Python programming languages.

Materials: Reading material will be assigned on a weekly basis. However, these are the books for general familiarity with text analysis methods:

- **Text as Data: A New Framework for Machine Learning and the Social Sciences.** Grimmer, Roberts, & Stewart, 2022.
- **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** Hastie, Tibshirani, & Friedman, 2009.
- **Foundations of Statistical Natural Language Processing.** Manning & Schütze, 1999.

Assesment: Students will be assessed on course participation and the final paper.

Course Outline

Over the course, we will cover the following topics:

- Text Representation
- Pre-processing
- Regular Expressions and Classification with dictionaries
- Machine Learning and Classification
- Topic modeling/unsupervised clustering
- Multilingual text analysis

- Word Vectors
- Large Language Models and Transformers
- Scraping and Using APIs
- Validation
- Ethics and Data Security
- Critical reflection on the methods