Syllabus: Text as Data

Petro Tolochko
Department of Network Science and Data Science, *CEU*
Email: `petro.tolochko@univie.ac.at`

Semester: Fall 2025

Course Description

This course introduces the theory and practice of using text as quantitative data. Students will learn techniques for preprocessing, representing, modeling, and critically analyzing textual data, with applications in social science. The course emphasizes both technical skill development and critical reflection.

Learning Objectives

By the end of the course, students will be able to:

- Preprocess and structure textual data for analysis
- Represent text using classic and modern embedding techniques
- Apply supervised and unsupervised models to extract patterns
- Validate text corpora
- Apply and critically asses text-as-data methods to socially relevant problems.

Tentative Weekly Schedule

Week 1: Introduction to Text-as-Data, Text Representation

Week 2: Preprocessing (tokenization, stemming, etc.) and Feature Engineering (BoW, TF-IDF, n-grams)

Week 3: Regular Expressions and Dictionary Methods

Week 4: Text Annotation and Supervised Labeling

Week 5: Machine Learning for Text Classification (Naive Bayes, SVM, etc.)

Week 6: Unsupervised Learning: Topic Modeling and Clustering

Week 7: Word Vectors and Semantic Similarity (word2vec, GloVe)

Week 8: Transformers and Large Language Models (BERT, GPT)

Week 9: Prompt Engineering and Zero-shot Learning

Week 10: Model Validation and Human Evaluation

Week 11: Networks approaches to Text-as-Data

Week 12: Ethics, Bias, and Critical Reflection in Text Analysis

Software and Tools

- `Python` or `R` (student choice; class examples provided in `Python`)

Prerequisites

Knowledge of `Python` (or `R`) programming language and basic machine learning knowledge are required to participate in the class.

Assessment

- Homework Assignments: **40%**
- Final Project: **50%**
- Participation: **10%**

Readings

There are no mandatory readings for the class. Recommended texts include:

- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as data: A new framework for machine learning and the social sciences. Princeton University Press.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
- James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning: Python edition.

Academic Integrity and Ethics

Students are expected to adhere to the highest standards of academic honesty. Plagiarism will result in a failed class.

Use of Generative AI Tools

The use of generative AI tools (e.g., ChatGPT, Claude, Copilot, Gemini) is permitted with restrictions in this course. If you use generative AI, you must clearly acknowledge its use (e.g., in a code comment, footnote, or reflection). You are responsible for the integrity, originality, and correctness of your work. Generative tools should support your learning and not replace it.