Text as Data

Lecture & Tutorial (4+2)

Jonas Rieger & Kai-Robin Lange

WiSe 2023/24

Organization

Formalities

- English, in person, Python (or R, but we suggest Python)
- Modules: BS 14, MS 6/7, MD E1 (Methods), ME 7
- Lecture + Tutorial (4+2 hours, 9 CP)
- Written exam, 120 minutes (probably 15.02.24 and 18.03.24)

Lecture

- Tue 16:00 - 17:30 (CT ZE 15) — first lecture: 17.10.2023
- Wed 16:00 - 17:30 (EF 50 HS 2) — last lecture: 31.01.2024

Tutorial/Exercise

- Monday 10-12, Tuesday 12-14, Wednesday 10-12

Contents

- Text data handling (e.g., encoding) and visualizations
- **Preprocessing**: tokenization, stopwords, stemming, lemmatization, n-grams, Regex, tf-idf, Zipfs law, filtering
- Part-of-speech (POS) tagging and named entity recognition (NER)
- Sentiment analysis
- (Static) embeddings (**word2vec**, fastText, GLoVE, ...)
- (Probabilistic) topic models (pLSA, **LDA**, CTM, STM, ...)
- Neural and transformer-based topic models (e.g., **BERTopic**)
- Transformer-based (pretrained) language models (e.g., BERT, **(Chat)GPT**, ...)
  - Fine-tuning, Parameter-Efficient Fine-Tuning (PEFT) & Low-Rank Adaptation (LoRA), few-shot learning, near-domain training, transfer learning, ...

Links

- Moodle-room with password "AttentionIsAllYouNeed": `https://moodle.tu-dortmund.de/course/view.php?id=41822`
- Moodle Link can also be found in the LSF

Literature and material

- Machine Learning for Text, DOI:10.1007/978-3-319-73531-3
- Text Mining with R, `https://www.tidytextmining.com/`
- R packages: see `https://www.tidytextmining.com/preface.html`
- Python libraries: NLTK, Gensim, spaCy, CoreNLP, TextBlob, Scikit-learn, torch, transformers, ...
- Online class (StanfordNLP): `https://web.stanford.edu/class/cs224n/`
- Illustrations and explanations of transformers: `https://jalammar.github.io/`

Questions

rieger@statistik.tu-dortmund.de
or
kalange@statistik.tu-dortmund.de