Text as Data

Lecture & Tutorial (2+1)

Jonas Rieger & Kai-Robin Lange

WiSe 2022/23

Formalities

- English
- modules: BS 14, BD W2, MS 6/7, MD E1, ME 7
- lecture + tutorial (2+1 hours, 4.5 CP)
- module exam as an oral examination (approx. 20-30 minutes)
  - in German or English as you like

Organization

- in person
- moodle (TBA)
- 4+2 format for 7.5 weeks (21.10. until 09.12.)
- mid of November: assignment as admission to the oral exam

- dates: Tue 10:15 - 11:45 (M/E 21) + Fr 8:30 - 10:00 (M/E 21)
- tutorial: Wed 10:15 - 11:45 (CDI 121) in Python or R as you like
- max. 32 participants
- registration via LimeSurvey (until 30.09.):
  `https://umfragen.tu-dortmund.de/index.php/769792?lang=en`

- recommended prerequisite for seminar "Advanced Text Mining Methods",
  cf. `https://lwus.statistik.tu-dortmund.de/en/teaching/courses/`

Contents

- text data handling (e.g., encoding)
- visualizations
- preprocessing: tokenization, stopwords, stemming, lemmatization, n-grams, Regex, tf-idf, Zipfs law, filtering
- part-of-speech (POS) tagging
- named entity recognition (NER)
- sentiment analysis
- embeddings (word2vec, fastText, GLoVE, ...)
- (probabilistic) topic models (pLSA, LDA, CTM, STM, ...)
- transformer based (pretrained) language models (e.g., BERT)

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/`

Questions

rieger@statistik.tu-dortmund.de