Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

472867-HS2024-0: NLP and Text Mining

Project Task 1: Data Exploration and Processing

juergen.vogel@bfh.ch, murat.sariyar@bfh.ch

Project Task 1: Data Exploration and Processing

- Explore your dataset by
  - calculating basic statistics
    - number of samples and number of samples per class: is your dataset balanced?
    - min / avg / max length of text
  - determining the national language(s) used
  - reading through 100+ samples: noteworthy style, vocabulary, spelling, …
- Establish a structured and flexible (configurable) processing pipeline with steps for
  - reading the dataset
  - tokenizing
  - normalizing (lowercasing, lemmatizing/stemming, …)
  - token filtering (stop words, …)
- Calculate TFDs for different variants of your processing pipeline
  - Are there differences per class?
- Submit your Jupyter notebook with code and findings via Moodle