Technische Universität Dortmund
Department of Statistics
Dr. J. Rieger
M. Sc. K.-R. Lange

Winter Semester 2023/24
01/15/2024

Exercise sheet 11

Natural Language Processing

Hand-in (voluntarily): 01/19/2024 until 11:59 p.m. via Moodle

---

In moodle you can find the file `arXiv_train.csv` and `arXiv_test.csv`. It contains the titles, abstracts and categories of 107,055 papers uploaded to the preprint-platform arXiv. You are tasked to create a pipeline, which predicts the category (given in the terms column) of a paper. The train corpus contains 100 labeled examples for the 4 classes.

This is an open exercise. You will not be tasked with a specific model to use. Instead, you can use any model we have talked about in the exercises and the lecture so far to solve this task.

Task 1

Argue the usefulness of the following models for the task at hand in 2-3 sentences:

- Dictionary-based analysis
- Latent Dirichlet Allocation
- Word2Vec
- Doc2Vec
- BERT

Are there other analyses that might help to solve this task? Which analyses do you deem to be the best for this task?

Task 2

Preprocess the data so that it is fit for the pipeline you intend to solve the task with. Shortly explain every step of preprocessing and why it is useful for this analysis.

Task 3

Solve the task by creating a pipeline with the model you deem to be optimal for this task.