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

Winter Semester 2023/24
01/29/2024

Exercise sheet 13

Natural Language Processing

No Hand-in

In moodle you will find the file `movies.txt`. In it, you will find summaries of different movies. We do however not know the genres of these books. Your task is to perform an unsupervised analysis to make an educated guess, which genres might be part of the data set.
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, except for Transformer models. This exercise sheet can be seen as a preperation for the practical part of the exam: You should be able to solve this exercise within 60 minutes on your own, without the help of any AI-assistant (you are allowed to use your previous solutions and Google though).

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

Which method 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.