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

Winter Semester 2022/23
11/30/2022

Exercise sheet 6

Text as Data

Hand-in: 12/04/2022 until 11:59 p.m. via Moodle

Task 1

In moodle you will find the file `NewsCategorizer.csv`. Load the file into your console. We are interested in the columns ”category,” and ”short_description” and want to see whether the short descriptions match their respective category.

Task 2

Preprocess the texts so that they are fit for an analysis.

Task 3

Train an LDA-model with 10 topics and 15 epochs/iterations/passes on the documents. Do the resulting topics match the categories?

Task 4

Compare the results of task 3 to the results of exercise sheet 4. Did the K-Means clustering or the LDA capture the contents of the texts better?

Task 5

Train 4 additional LDA-models with the same parameters and compare the results of each one.

Recommended packages & functions

R: `tosca::LDAgen, tosca::LDAPrep`
Python: `gensim.corpora.dictionary, gensim.models.ldamodel, pyLDAvis`