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

Winter Semester 2023/24
11/27/2023

Exercise sheet 6

Text as Data

Hand-in (voluntarily): 12/01/2023 until 11:59 p.m. via Moodle

Task 1

In moodle you will find the file `NewsCategorizer.xlsx`. 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 and can be detected using text clustering.

Task 2

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

Task 3

Train an LDA model on this data with $K = 10$ and 200 iterations (if this takes too long on your hardware, you can also use 50 iterations).

Task 4

Calculate the tfidf-score for each word in each text and perform k-means clustering using the tfidf-score with 10 clusters.

Task 5

Compare the clusters of the k-means clustering with the true news category labels. Do the clusters represent the categories well? How about the LDA soft-clusters – does the content of the topics match the categories?

Recommended packages & functions

R: `readxl::read_xlsx, kmeans, tosca::LDAgen, tosca::LDAPrep`
Python: `pandas.read_excel, sklearn.cluster.KMeans, gensim.corpora.dictionary, gensim.models.ldamodel`