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

Winter Semester 2022/23
11/16/2022

Exercise sheet 4

Text as Data

Hand-in: 11/20/2022 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 and can be detected using text clustering.

Task 2

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

Task 3

Calculate the tfidf-score for each word in each text.

Task 4

Perform k-means clustering using the tfidf-score with 10 clusters. Compare the results to the true news categories. Do the clusters match the categories?

Recommended packages & functions

R: kmeans
Python: sklearn.cluster.KMeans