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

Winter Semester 2022/23
12/07/2022

Exercise sheet 7

Text as Data

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

Task 1

In moodle you will find the files `train.csv` and `text.csv` containing the columns "review" and "sentiment". This is a sentiment analysis-corpus where the sentiment denotes the true baseline sentiment of each document ("1" for positive, "0" for negative).
Load these texts into your console and create a train-test-split so that we have a training set consisting of 75% of the documents and a test data set consisting of 25% of the documents.

Task 2

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

Task 3

Train a Doc2Vec-model with a winow size of 5, a vector-size of 100, 1 epoch on the documents in the train data set.
Then train a classification model of your choice (such as Logistic Regression or Random Forests) with the labels of the train data set as the predictor and the document embeddings created by Doc2Vec as covariates.
Use your already trained Doc2Vec-model to create embeddings for the documents in the test data set. Then, using these embeddings and your classification model, predict the labels for the test data set. How many percent of the documents are correctly classified?

Task 4

Repeat task 3 two additional times with 10 and 20 epochs. Compare the resulting classification rates.

Task 5

In Moodle you will also find the file `WKWSCI.xlsx`. It contains a sentiment lexicon which yields a sentiment value for a large amount of words. Use this lexicon to calculate the sentiment score of each document in the test data set.
Use this sentiment score as a predictor: If the sentiment score is positive, we predict the corresponding document to be positive (label 1) as well and vice versa. Compare the classification rate to the ones resulting from task 4.

Recommended packages & functions

R: `doc2vec, lm()`
Python: `gensim.models.doc2vec, sklearn.linear_model, sklearn.model_selection.train_test_split()`