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

Winter Semester 2023/24
01/08/2024

Exercise sheet 10

Natural Language Processing

Hand-in (voluntarily): 01/12/2024 until 11:59 p.m. via Moodle

---

Task 1

In moodle you will find the files `fake_train.csv` and `fake_test.csv`. They each contain a set of news articles that either contain factual news or fake news. We will try to differentiate fake news from real news by comparing their document embeddings. For this, we will train a document embedding model on the whole corpus. Then, we will use the embeddings of our train-corpus to train a logistic regression model that tries to predict the labels of the test-corpus given their embeddings.

Task 2

Preprocess the texts so that they are fit for an analysis. Argue the use the preprocessing steps you take for the given analysis.

Task 3

Train a Doc2Vec model on all documents from both the training and test corpus with a window size of four and a vector dimension of 300.

Task 4

Create a data frame of all document embeddings of the documents within the training corpus and the label of the respective document. Use this data frame to train a logistic regression that uses the embeddings to predict the label of the document.
Use it to predict the labels of all documents in the test-corpus using their embeddings. Compare the resulting labels to the true labels and return the classification rate. How well does the model perform?

Task 5

Repeat tasks 3 and 4 with one adjustment: Train your initial Doc2Vec model only on the train-corpus. This way, the test-corpus is entirely unobsorved for our model. Compare the resulting classification rate on your test-corpus with the classification rate you got from task 4.

Recommended packages & functions

Python: `gensim.models.doc2vec.Doc2Vec, gensim.models.doc2vec.Doc2Vec.infer_vector(), gensim.models.doc2vec.TaggedDocument`