Exercise 2:

This exercise may be done in groups of 2 – 3 people. If so, please explicitly state all group member names on your hand-in!

For this exercise you should first find a suitable text classification dataset from any source, e.g. from

- https://www.kaggle.com/datasets
- https://archive.ics.uci.edu/ml/index.php
- https://www.tensorflow.org/datasets
- https://datasetsearch.research.google.com/
- https://github.com/niderhoff/nlp-datasets

Do the following tasks:

- Shortly describe your dataset:
  - Short description in words (source, use case, trivia, etc.).
  - What are the classes? What are the class distributions? What is the baseline for this dataset (what accuracy would a classifier achieve if it only predicted the most frequent class)?
- Split your dataset into a training and test dataset – if that is not already done for you.
  - If you want to optimize hyperparameters later on, you should also make a validation split.
  - If your dataset is very big or if any of the tasks take too long, take a smaller sample of your original dataset.
- Clean up your texts if necessary e.g. removing special characters.
- Preprocess your texts and choose and apply any feature representation (count matrix, TF-IDF, doc2vec, or any other that we did not discuss).
- Cluster the texts in your training and test dataset and show a silhouette plot (for both training and test) of the best clustering you could find.
- Classify the texts in your training and test dataset and show a confusion matrix of the predictions (for both training and test) of the best classifier you could find. **If your group consists of 3 people you are expected to try at least two different kinds of classifiers (e.g. naiveBayes and kNN).**
- Summarize your results. What worked and what didn't work? (2 – 3 sentences is enough).
  - The performance of your classifier is not influencing the grading of this exercise ;-)

You are expected to hand in a jupyter notebook with at least minimal comments in .html format!

Hand-in is due 31st of January 2024, 23:59;

Points: __/12