Assignment 4 - Text Classification

The goal of this assignment is to optimize the basic pipeline provided in lab05 (text classification using 20newsgroup dataset) to improve the classification performance on the test set. In the code provided, the dataset is already divided into train and test. You should keep the test set for the final evaluation.

To be able to evaluate and select various preprocessing and feature extraction techniques as well as different model parameters, you must split the training set into a new training set and a validation set (e.g., 70/30).

Any machine learning should start by conducting some data exploration on training.

To optimize the overall pipeline you need to select the "best" preprocessing techniques (and order) on validation:

- Find the preprocessing technique that has the most impact on validation performance
- Apply and assess the impact of other preprocessing techniques (e.g., stemming instead of lemmatization)
    - You may compare different libraries (optional)
- Try to find the best sequence of preprocessing techniques that can improve model performance on the validation set

Next, after selecting the preprocessing techniques (and their order) you must select the best feature extraction techniques. You could try to:

- Use the n-gram frequency instead of term (unigram) frequency in TF and TF-IDF
- To reduce the feature size of unigram or n-gram TF and TF-IDF
- Apply a different feature extraction technique using one pre-trained word embeddings method of your choice

Now since you have selected the "best" preprocessing and features techniques, you need to optimize model parameters (on validation). For instance, you can try:

- Experiment with a range of C values in linear SVM or using another kernel and optimize its parameters
- Use KNN and find the best N
- Use ComplementNB instead of MultinomialNB (use different alpha values)

Finally, after selecting the best model parameters for each model, you can retrain the three models on the combined data (train + valid) and evaluate their performances on the **test** set.

Deliverable:

- **One zip file** including Notebook + supporting python files.
    - **userid_full_name_assgin4.zip** (wk7_wael_khreich_assign4.zip)
- At the bottom of the notebook (after reporting the results of the models on the test), describe briefly (**500 words**) your experimental **methodology**, **findings**, and **recommendations**.