Assignment 1: text categorization

Text mining course

This is a **hand-in assignment for groups of two students**. Send in via Brightspace **before or on Tuesday October 10**:

- Submit your report as PDF and your python code as separate file. **Don’t upload a zip file containing the PDF** (the Python code might be zipped if it consists of multiple files).
- Your report should **not be longer than 3 pages** (being concise is an important lesson!)
- Do not copy text from sources (other groups, web pages, generative models such as chatGPT). Turnitin is enabled and a large overlap will be reported to the Board of Examiners.

Goals of this assignment

- You can perform a text categorization task with benchmark data in scikit-learn
- You understand the effect of using different types of feature weights
- You can evaluate text classifiers with the suitable evaluation metrics

Preliminaries

- You have completed the tutorial ‘working with text data’ in sklearn: http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html (**exercise week 4**)
- You have all the required Python packages installed

Tasks

1. The tutorial classifies between only four categories of the 20newsgroups data set. Change your script so that it addresses all 20 categories.
2. Compare three classifiers in sklearn on this multi-class classification task, including at least Naïve Bayes.
3. Compare three types of features for your classifiers: counts, tf, and tf-idf. Keep the best combination of a classifier and a feature type for the next task.
4. Look up the documentation of the `CountVectorizer` function and experiment with different values for the following parameters for your best classifier-feature combination. For each of these parameters compare different values and store the results.
    a. Lowercasing (true or false)
    b. stop_words (with or without)
    c. analyzer (in combination with ngram_range), try out a few values
    d. max_features, try out a few values
5. Write one script or notebook for running these experiments and printing the results.

Write a two-page report (3 pages is the hard maximum) in which you:

- describe your methods (classifiers, features);
- show a results table (Precision, Recall, and F1) for the classifiers and features;
- write a brief discussion on which classifier performs the best, with which features

Grading

Maximum 2 points for each of the following criteria:

- General: length correct (2-3 pages) and proper writing + formatting
- Experiments on 20 newsgroups
- Results table for 3 classifiers x 3 feature weights (counts, tf, and tf-idf)
- Results for a. lowercase; b. stop_words; c. analyzer (in combination with ngram_range); d. max_features
- Brief discussion on which classifier performs the best, with which features