Assignment 1: text categorization

Computational Linguistics course

This is an **individual hand-in assignment**. Your report should not be longer than 3 pages

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 followed the tutorial 'working with text data' in sklearn: http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
- 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 on this multi-class classification task, including at least Naïve Bayes.
3. Compare three type of features for your classifiers: counts, tf, and tf-idf.
4. Look up the documentation of the `CountVectorizer` function and experiment with different values for the following parameters:
    a. lowercase
    b. stop_words
    c. analyzer (in combination with ngram_range)
    d. max_features
5. Write one script 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