Assignment 2: text categorization

Text mining course

This is a hand-in assignment. Send in via Blackboard **before or on Tuesday October 16**:

- (go to course documents, hand-in assignments, assignment 2)
- Please use the following filename for your uploaded file: assignment_2_YourName.pdf

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 uses only four categories of the 20newsgroups data set. Change your script so that it uses all 20 categories.
2. Compare three classifiers on this multi-class classification task, including at least Naïve Bayes.
3. Compare three feature weights for your classifiers: counts, tf, and tf-idf.
4. Write one script for running these experiments and printing the results.

Write a two-page report in which you:

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

Additions for more points:

- Discuss the differences between the categories. Which categories are the most difficult to classify? Which categories are the most often confused?
- Find out what the most important features are for your best classifier (for a few example category).