Support vector machines for document classification:

Support Vector Machines (SVM) are popular in the machine learning community as a technique for tackling high-dimensional problems.

Repeat similar steps in HW1 for processing document data:

1. Download the data from http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html (Newsgroup Data).

2. For each data point, please remove the first four lines, i.e. the lines starting with Newsgroup, document_id, From, Subject.

3. Build the word vocabulary for training data, and remove the top 300 most frequent words as stop words from the vocabulary.

SVM implementation:

1) Create feature vector for each document using tf-idf method.

2) Use linear SVM function from sklearn, e.g. sklearn.svm.SVC, for multi-class classification.

3) Implement 5-fold cross validation.

4) Submit Jupyter notebook at ELMS