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. Please implement the SVM algorithm by yourself and conduct the document classification.

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. Split data into two groups, and use half data as training data and the other half as testing data. **Note: split the data of each class into two halves.**

4. 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) Implement the linear soft margin SVM. You can use library (https://pypi.org/project/qpsolvers/) to solve the quadratic programming problem. For the rest of implementation, you need write your own code.

3) Document classification is a multi-class classification problem. You can use one-vs-all strategy to conduct multi-class SVM. Set the trade-off parameter C (for slack variables) as 100.

4) Use training data to train the classifiers. Predict each testing data and compare the predicted label to the ground truth label. The average prediction accuracy is calculated.

5) Please implement the polynomial kernel SVM; $K(x,y) = (x^T y+c)^d$ with c=0, d=2

6) Repeat 4) for polynomial kernel SVM

7) Submit Jupyter notebook at ELMS