Text Mining: Homework 1

due on June 12, 2019

This homework asks you to categorize the famous Reuters-21578 corpus which consists of Reuters newswires in 1987. There are 21587 documents stored in 22 files. The original test collection is available from `http://www.daviddlewis.com/resources/testcollections/reuters21578` in the format of sgml. For the convenience, I put the XML versions in e-learning. It would be much easier for coding to preprocess 22 the XML files into 21587 text files.

To perform text categorization, one needs to label the 21587 documents. ClairBee labels the whole document collection into 135 categories with possible multiple categories. She then classify the documents using RWeka. The data and Rcodes are available at `https://github.com/ClairBee/cs909`.

Fabrizio Sebastiani prepares the R8 and R52 which have respectively 8 of the 10 most frequent classes and 52 of the original 90. These are single-labeled dataset available at `http://www.cs.umb.edu/~smimarog/textmining/datasets/`. You might find the code by Ian Kloo useful and Bryan Cole respectively which is available at

`http://data-analytics.net/cep/Schedule_files/Textmining%20%20Clustering,%20Topic%20Modeling,%20and%20Classification.htm`

`https://rpubs.com/bmcole/reuters-text-categorization`.

Do the follows:

1. Pre-process the documents
   - remove punctuation
   - remove digits
   - remove extra white space
   - remove stop words
   - conversion to lower case
2. Produce a weighted version of the TDM by term frequency - inverse document frequency (tf-idf).
3. Use K-Nearest Neighbors (kNN), Support Vector Machine (SVM) and Decision Tree to classify the documents.
4. Summarize your findings.