CMSC25010 Homework 4

University of Chicago, Winter 2012 due Sunday, February 19, 11:59 pm

The Federalist Papers are a series of 85 articles that were written by Alexander Hamilton, James Madison and John Jay under the pseudonym "Publius" in 1787–1788. For some articles it is clear who their author was, while the authorship of some of the others is still debated.

The file `federalist.tgz` contains 15 articles by Hamilton, 15 by Madison, and 11 articles whose authorship is uncertain. The task in this assignment is to build two separate mutlinomial bag-of-words models for the Hamilton and Madison articles (as discussed in class), and use them to classify the 11 articles of unknown origin.

The multinomial models should be fit either the freqentist way using the maximum likelihood estimator (MLE) with an appropriate pseudocount, or the Bayesian way with a Dirichlet prior and then taking the maximum a posteriori (MAP) estimator. Each further document $\mathcal{D}$ can then be classified as "Hamilton" or "Madison" based on comparing its log-likelihood $\log \ell_H(\mathcal{D})$ under the Hamilton model with its log-likelihood $\log \ell_M(\mathcal{D})$ under the Madison model. The parameters, such as the pseudocounts in the frequentist case or the parameters of the Dirichlet prior in the Bayesian case, should be set by cross-validation. In the present context this is best done by removing a single document from the corpus, training the models on the other 29 documents, and then evaluating performance based on whether the removed document is correctly classified. You can do this for all 30 documents and average the results.

Please submit your work in a single `.tar.gz` including your code and a write-up with the following:

(a) A brief description of the approach taken and the exact method used to set the parameters.

(b) A table showing the likelihoods and predicted labels of the training documents when the models are trained on the other 29 training documents.

(c) A table showing the likelihoods and the predicted labels of the 11 documents of unknown authorship.

Solving this problem requires building a dictionary of all the words featured in the training set. You can either do this as a preprocessing step, or you can do it dynamically, at the same time that you count the number of word occurences (word frequencies). You will also need to make subtle choices about capitalization, hyphenation, the choice of prior, and so on. Some of these choices might have a large effect on performance. Experimenting with these various options is a large part of the assignment, and should be described in the write-up.

For extra credit, you may experiment with truncating the dictionary, removing very frequent words, such as 'a', 'an', 'the', etc., called stopwords, and reducing words to their canonical form by removing common endings such as the 's' for the plural, which is called stemming.