MAT 331 Fall 2017, Project 8
Distinguish English from French

In class we wrote code to test whether decrypted text looked like English or not. This project is to write similar code that will tell if a string of text is more likely English or French. There is a difficulty because French text will have special ASCII codes to represent letters with accent marks and these will involve ASCII codes larger than 127, so you could tell a text was not English by looking for special characters. For this project, when you read in any text, I want you to throw away any ASCII values larger than 127, and work with what is left.

I have put some French texts from Project Gutenberg in the `Crypt` directory of the class webpage (the French original and a English translate of two volumes of "Les Miserables' called `lmf.txt` and `lm.txt` respectively. These should be of about the same length and use similiar words.) You can choose different text (or even a different language) if you prefer.

(1) Using some long texts in English and French, plot the letter frequencies in each sample. Turn each into a probability distribution on $\{1, \ldots, 27\}$ as we did in class (divide each letter count by the total number of letters). Youwill have to throw away non-letters and convert lower case to upper case. How do the distributions look similar or different?

(2) Choose 10 short texts (say around 1000 consecutive letters chosen from one of the texts)) in each language and compute the letter distributions for each (counts for each letters, divided by total number of letters; probabiliteis should sum to 1). Compare each of these to the two distributions and choose the one it is closer to. There are several ways to compare two probability distributions p and q. One way is to compute $|p - q| = \sum_k |p(k) - q(k)|$. Another is to compute the dot product *pḑotq* $= \sum_k p(k)q(k)$ (the closer the vectors are, the larger the dot product will be). How good is each method at choosing the correct language? Report how many correct/incorrect results you got.

(3) Using the same two texts as in (1) compute the matrix of letter pairs (the probability that letter j follows letter k) as we did in class.

(4) Using the same shorter samples as in (2), score each of the samples using the two matrices from (3). Use this to guess the correct language for each sample. Report the results.

(5) Which method seems to work better? If you take shorter samples, say of length 500 or 100 or 50, does one method work better than the other? About how many letters do you need to tell the languages apart?