On MathLAN, the directory
/home/stone/courses/computational-linguistics/corpora
contains files named mysterious-key.txt and eight-cousins.txt, each containing the full text of a short work by Louisa May Alcott. In this exercise, we’ll consider each text as a sequence of experiments in which the sample space comprises the twenty-six letters of the Roman alphabet. We’ll posit a random variable that produces each of these outcomes, and take the relative frequency of each letter as an estimate of the probability that the random variable produces it, and estimate the per-letter entropy of each text accordingly. We’ll also compute the Kullback-Leibler divergence between the probability mass functions for the two posited random variables (one for each text).

Since we’ll be using the files containing the texts but won’t need to modify them, in this exercise, you may either make copies of these files in your home directory or read their contents directly from mine.

0. Define and test a Scheme procedure letter-spectrum that takes as its argument an input port, already opened for reading, and returns a vector of length 26 in which the element in position 0 is the number of occurrences of the letter A (both upper- and lower-case) in the text to which the port provides access, the element in position 1 is the number of occurrences of B, and so on. All non-alphabetic characters in that text should be read but not tallied.

1. Define and test a Scheme procedure normalize-letter-spectrum that takes a vector of the kind that letter-spectrum produces and constructs and returns a similar vector of the same length containing the relative frequency of each letter (the relative frequency of A in position 0, the relative frequency of B in position 1, and so on). All non-alphabetic characters in that text should be read but not tallied.

To obtain the relative frequencies from the tallies that letter-spectrum compiles, divide each tally by the total number of letters in the entire text. The normalize-letter-spectrum procedure requires (as a precondition) that this total be positive, and normalize-letter-spectrum should invoke the error procedure to terminate the program if this precondition is not met.

2. Show informally that the vector of relative frequencies that normalize-letter-spectrum returns is, by construction, a probability function under the definition on pages 40–41 of our textbook.

3. Define and test a Scheme procedure lg that computes and returns the base-2 logarithm of any given positive real number.

4. Define and test a Scheme procedure entropy that takes as argument a vector that is a probability function and computes the entropy of a random variable with that probability function.

It is a precondition of this procedure that none of the elements of the vector is 0, and entropy should invoke the error procedure to terminate the program if this precondition is not met.

5. Define and test a Scheme procedure letter-entropy-of-file that takes as argument a string giving the name of a readable file and computes the per-letter entropy of the alphabetic characters contained in that file.

6. Define and test a Scheme procedure kl-divergence that takes two vectors of equal length, each a probability function, and computes the Kullback-Leibler divergence of the second vector (the probability function called q in presentation on page 72 of our textbook) from the first (p).
Exercise #0 — page 2

It is a precondition of this procedure that, for every position $i$ in the second vector at which a zero element occurs, the element at position $i$ in the first vector is also zero. If this precondition is not met, $\text{kl-divergence}$ should invoke the $\text{error}$ procedure to terminate the program.

7. Using the procedures specified above, write a Scheme program that computes the per-letter entropy of the letters in each of the Alcott texts and the Kullback–Leibler divergence between the (estimated) probability functions for the (posited) random variables for those texts.

This exercise will be due at the beginning of class on Friday, September 16.