**COGSCI 131 – Assignment 7**
**DUE: Aug 12th at class start**

We will do some information-theoretic calculations on the English language using a dictionary of words as outcomes and frequencies which we will assume are proportional to P(word). That is, the distribution we consider is P(Word)=frequency(word)/sumOfWordFrequencies. A file containing words and frequencies is posted on bcourses.

1. [10pts] What is the entropy over words in English?

2. [10pts] Using your answer from Q1, decide whether the game "20 questions"[1] is a fair game—can you win more than half the time—assuming (a) if the word being guessed is chosen according to frequency, and (b) if the word is chosen uniformly (e.g. with equal probability for each word).

3. [10pts] Make a plot of the conditional entropy over words in English, conditioning on the first character (e.g. one bar for "a", one for "b", one for "c", etc.). Write this using a loop, not 26 separate chunks of code, please.

4. [15pts] Plot the information about word identity that is conveyed by (1) the first character, (2) the last character, or (3) the first vowel (aeiou). Which of these would you predict then is most important for word recognition?

5. [15pts] For each word length (1, 2, 3, …) plot the average surprisal of words that are that length (i.e. plot the averages, not a point for each word). Generally, what would this plot look like if English were an efficient code in Shannon's sense? Qualitatively describe in 1-2 sentences places where your plot does or does not agree with an efficient code.

6. [10pts Extra Credit] Perhaps the most frequent words are "more optimized" to be like an efficient code. Come up with a measure of how well word length agrees with surprisal (as in Q5), and plot that measure for the most frequent N words, N=10, 20, 30, ... for the entire lexicon (each plotted set here should be cumulative, including words 1 through N). What does your plot indicate about this possibility?

1 If you don't know, "20 questions" is a game where I think of a word and you get twenty yes-or-no questions in order to try to guess it.