Information Theory – Midterm
Prof. Anderson C A Nascimento
University of Washington

1. Estimating the entropy of English. Download " The Adventures of Tom Sawyer" (http://www.gutenberg.org/files/74/old/sawyr10.txt) and estimate its entropy for the following cases:
    a. Assume that each character in the text is independent of each other, consider solely the 26 letters of the English alphabet. Do not consider punctuation marks, spaces and make no distinction between uppercase and lowercase letters. Compute the frequency of each character and estimate the entropy of the text given the stated assumptions.
    b. Assume that each sequence of two characters (bigrams) is independent from each other. Repeat the previous calculations.
    c. If you estimate the entropy of " The Adventures of Tom Sawyer" by using trigrams (instead of bigrams), will you get a higher or lower value of entropy as a result?
2. Download " The Adventures of Tom Sawyer" (http://www.gutenberg.org/files/74/old/sawyr10.txt) and compress this file by using Huffman codes following these steps (estimate the probability distribution of each ascii symbol present in the file). What is the compression ratio achieved? Compare the value obtained with the results you got in the first question. Are they related in any way? Give the running times for compression and decompression. Repeat these steps now encoding bigrams rather than single characters.
3. Download " The Adventures of Tom Sawyer" (http://www.gutenberg.org/files/74/old/sawyr10.txt) and compress this file by using the tree structured Lempel Ziv compression algorithm as described on page 442 of our textbook. What is the compression ratio achieved? Compare your result with your previous estimation of the entropy of English. Give the running times for compression and decompression.
4. (a) Two teams A and B play a best-of-five series that terminates as soon as one of the teams wins three games. Let X be the random variable representing the outcome of the series, written as a string of who won the individual games (e.g., possible values of X are AAA, BAAA, ABABB, etc.) Let Y be the number of games played before the series ends. Assuming that A and B are equally matched and the outcomes of different games in the series are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, $H(X|Y)$, and $I(X;Y)$. Let pA and qA be the distributions of X and Y , respectively, given that A wins the series. Calculate $D(pA||X)$ and $D(qA||Y)$.
   (b) Suppose X, Y , and Z are each Bernoulli(1/2) and are pairwise independent (i.e., $I(X;Y) = I(Y;Z) = I(X;Z) = 0$). What is the minimum possible value of $H(X,Y,Z)$?

5. The frequency p_n of the nth most frequent word in English is roughly approximated by $p_n=0.1/n$ for n in {1,...,12367} and $P_n=0$ for $n>12367$. If we assume that English is generated by picking words at random according to this distribution, what is the entropy of English (per word)? How do you compare your result with the values you obtained in the first question?

For all the questions that require implementation you should submit the source code you produced and running times.