SCI 199 **Assignment 6**

Due: Monday, March 18

The purpose of this assignment is to gain a little experience with text compression. In particular, you will gain some experience understanding Huffman coding which produces an optimal prefix code.

1. Consider the first one or two pages of the document you submitted when doing the text summarization assignment. Produce a list of English words occurring in those one or two pages and count the number of times each of these words occur. (You can disregard mathematical notation, etc. and just list English words.)

2. Simulate the Huffman algorithm and construct the Huffman tree for the words you have listed.

3. What is the expected length of a code word in your Huffman code. Compare this with $\lceil \log_2 N \rceil$ where N is the number of words in your list.

4. If you have time, try the same exercise on the summary of the document. Intuitively, one would think that the summary should be higher in "information density" and one might guess that the expected code length for the summary is higher than for the document.