Jacob Bohac
ELEC 465
HW3

To test the Huffman Encoder, three datasets where used:

4letters.word	–	the sample data provided	–	10.7 kB
TaleOfTwoCitiesExcerpt.txt	–	a portion of the Tale of Two Cities	–	138.9 kB
DNA.txt	–	randomly generated DNA data	–	10.0 kB

The following table shows the compression rate achieved with each data set. Reference file is the set that was used to generate the code lengths and Encoded file is the file that was compressed.

		Reference file		
		4letters	**TaleOfTwo**	**DNA**
Encoded File	**4letters**	62.6%	72.9%	136.4%
	TaleOfTwo	98.6%	56.7%	147.8%
	DNA	120.0%	132.0%	36.0%

The results when the file to be encoded was used to generate the code lengths was much better than when there was a mismatch. The DNA file had the best compression rate, due to the fact that it was comprised mostly of the 4 letter alphabet {a,c,g,t}. A mismatch between the 4letter.words and a Tale Of Two Cities resulted in a very poor compression rate. This was most likely due to the difference in the number of occurrence in capital letters between the two texts. Mismatches with the DNA file ended up much worse and actually increased the size of the files. This is because a,c,g,t produces codewords of length 2,2,2,3 respectively, leaving the rest of the alphabet to be encoded in at least 11 bits, much worse than the 8 bits they were encoded with originally.

The differences of the code words generated for each file are highlighted in the following tables:

4letters.word	
s	01011
n	01100
t	01011
r	01110
i	01111
l	1000
o	1001
a	1010
e	1011
(new line)	11

TaleOfTwoCitiesExcerpt	
r	01001
s	0101
i	0110
h	0111
n	1000
o	1001
a	1010
t	1011
e	110
(space)	111

DNA	
X	00001111011
W	00001111100
V	00001111101
U	00001111110
R	00001111111
(new line)	0001
T	001
A	01
G	10
C	11