CS458 - Homework 2
Due: Thursday Sept. 27, At the beginning of class

1. (10 points) Distributed indexing

   Figure 4.5 in the book shows an example MapReduce framework for distributed indexing.

   (a) Given n documents and m machines, describe a good method for splitting up the documents. Justify your answer.

   (b) In the reduce phase, the example suggests partitioning the data by $a - f$, $g - p$, and $q - z$. Is this a good approach? Explain your answer. Describe a better partition of 3 parts (or a better method of how to partition into 3 parts). Why is your partition better?

2. (10 points) Zipf's law

   From our data set from assignment 1, I counted the frequency of each word and sorted them by frequency. Below are a few data points:

Rank	Frequency
1	417,667
10	70,848
100	8,508
1000	842
10,000	37

   (a) Create a plot using these points like Figure 5.2 (on paper is fine, or you can use a program). Do the points seem to follow Zipf's law?

   (b) For all of the data rank 10 or higher from the table above, estimate what the value should be using the point below it (to estimate rank 10, use rank 1, for 100 it would be 10, etc.). Then calculate percentage-wise, how far away the real values are from this estimate. Does Zipf's law seem like an appropriate model for the data?

3. (10 points) Decoding

   For the codes below, show the gaps and the corresponding postings for the encoded strings.

   (a) variable code: 11010110 01011101 10111011 01101010 01110101 01101100 10000011

   (b) gamma code: 1111000111101110011111101011011