Module 1 Programing Assignment

Max Robinson

1 Problem 1

Question: Qualitatively explain the impact of using stemming on each of the following:
(a) vocabulary size;
(b) total number of postings in an inverted file;
(c) average posting list length?
The format of a good answer would besomething like: With stemming XXXX {increases, decreases, doesnt change} by {a lot, a little, at all, roughly zz%} because of YYYY. I want to see a statement about the effect, it's magnitude, and a clear rationale.

2 Problem 2

Question: Express the numbers 8, 14, and 513 three ways: using a 12-bit binary representation, and the gamma and delta codes. You must follow the method for computing gamma/delta described in the text and presented in the lecture materials. I strongly recommend learning to do this by hand, but you may write (and provide) a short computer program if you prefer but do not use a program that you did not write yourself.

3 Problem 3

Question: Below is a bit sequence for a gamma encoded gap list (as described in Chapter 5 of IIR and the lecture materials). Decode the gap list and reconstruct the corresponding list of docids. Spaces are added for ease of reading – the final part only has two bits. Hint: there are four docids. 1111 1100 0100 0111 1111 0010 0000 1111 1010 1011 1111 0100 00

4 Problem 4

Question: True or False – Any bit sequence (i.e., any combination of zeros and ones) can be interpreted as a valid gamma encoded list of integers? Explain why this is true, or give an example showing that it is not.

5 Problem 5

Question: Below is a bit sequence for a set of gaps encoding using Variable Byte encoding as described in Chapter 5 of IIR. Decode the list of gaps and reconstruct the corresponding list of docids. Hint: there are three docids. 1100 0001 0000 0011 1011 0011 0000 0100 0001 1111 1000 0011