Compression and IR

Questions

Question 1

If we use the variable-byte encoding scheme, what is the largest gap that can be encoded using 1 byte?

Question 2

From the following sequence of γ-coded gaps of docids, reconstruct the gap sequences and then the posting sequences:

000000001000000000000100010001100

Question 3

In the course we have seen how to construct a Huffman tree in time $O(n \log n)$, where $n = |S|$ is the size of the alphabet. Now suppose the symbols in S have already been sorted by their respective probability, when they arrive at the encoder. Design an algorithm that builds a Huffman tree in time $O(n)$.

Question 4

The Simple-9 compression method groups sequences of Δ-values into 32-bit machine words, reserving 4 bits for a selector value. Consider a similar method, Simple-14, that uses 64-bit words instead. Assuming 4 bits per 64-bit word are reserved for the selector value, list all possible splits of the remaining 60 bits. Describe the type of docid lists on which Simple-9 will lead to better/worse results than Simple-14.