Algorithms in the Real World (15-853), Fall 04
Assignment #5

Due Dec 8. Please answer all three questions.

Problem 1: Posting Lists (15pt)
We want to analyze the number of bits required to store a posting list (asymptotically). Assume that the posting list has m documents and each document is an integer identifier in the range $0..n$.

1. Show that the lower bound on the number of bits required by any representation of these m documents is $\theta(m \log((n+m)/m))$.
2. Show that the difference coding as described in class matches this lower bound (i.e. is optimal within a constant factor).

It is interesting to note that these bounds are identical to those we analyzed for merging, where n and m were the length of the two input lists.

Problem 2: Set Difference (15pt)
In the class notes we showed pseudocode for union and intersection given split and join operations. Give the pseudocode for set difference, which can be used to process (A andnot B) queries. For two sets of size m and n ($m \leq n$) it must take time $O(m \log((n+m)/m))$. You can assume that $(L, R, f) = \text{split}(A)$ and $\text{join}(L, R)$ take time $O(\log |L|)$.

Problem 3: Treaps (15pt)
In the discussion of merging with Treaps (see lecture 1 slides) we introduced the random variables A_{ij} and C_{ilm} along with their expectations a_{ij} and c_{ilm}. To bound the expected pathlength from the start to the l^{th} element we analyzed the term $\sum_{i=1}^{l} a_{i1}$ but not the term $\sum_{i=1}^{n}(a_{il} - c_{i1l})$. Please analyze this sum by writing an expression of c_{ilm} and taking the sum.