COMP6714 ASSIGNMENT 1

DUE ON 14:59 15 NOV, 2021 (MON)

Q1. (*30 marks*)

(1) Suppose the inverted index only stores the document IDs (e.g., no gap). And suppose a word appears at least once in all N documents in a collection. What is the compression ratio that could be achieved by Elias-γ encoding and Elias-δ encoding?

(2) Suppose the inverted index stores the document IDs (using gap) and the term frequencies. And suppose a word appears in every 5th document, and it appears 10 times in each of those documents (i.e. 10 times in document 1, 10 times in document 6, ...). What compression ratio would be achieved by Elias-γ encoding and Elias-δ encoding?

Q2. (*30 marks*)

Consider the scenario of dynamic inverted index construction. Assume that t sub-indexes (each of M pages) will be created if one chooses the no-merge strategy.

(1) Show that if the logarithmic merge strategy is used, it will result in at most $\lceil \log_2 t \rceil$ sub-indexes.

(2) Prove that the total I/O cost of the logarithmic merge is $O(t \cdot M \cdot \log_2 t)$.

Q3. (*40 marks*)

(1) Prove that stemming will not hurt recall.

(2) Prove or disprove that stemming always helps or hurts the F1 score.

You need to give a formal proof for each of the above two statements.

Submission Instructions

You need to write your solutions to the questions in a pdf file named `ass1.pdf`. You **must**

- include your **name** and **student ID** in the file, and
- the file can be opened correctly on CSE machines.

You need to show the key steps to get the full mark.

Note: Collaboration is allowed. However, each person must independently write up his/her own solution.

You can then submit the file by `give cs6714 ass1 ass1.pdf`. The file size is limited to 5MB.