Study Questions on Brin & Page 1998

Referring to “The Anatomy of a Large-Scale Hypertextual Web Search Engine”…

A) in section 2.1.2 paragraph 1, re “add the damping factor $d$ to … a group of pages”, how can this make it hard for people to “deliberately mislead the system”?

B) in 3.2 paragraph 2, re “metadata efforts have largely failed” … what is meant by “metadata” here?

C) in 4.1, paragraph 2, the “anchors file” … what operations does this support?

D) how do the data structures described in sections 4.2.5, 4.2.6, and 4.2.7 relate to Figure 1?

E) in 4.2.6 paragraph 1, “each barrel holds a range of wordIDs” … what does this mean? Is it necessary for each “range of wordIDs” to be contiguous?

F) in 4.2.3 the Document Index is described … is this used at query time?

G) in 4.2.5, why is it important that the lexicon fit in main memory?

H) in 4.2.5 paragraph 2, the “capitalization bit” … what is this used for?

I) in 4.2.5 paragraph 2, give pseudocode for the “limited phrase searching”

J) in 4.2.7, paragraph 2, what would be the advantages and disadvantages of keeping each set of docIDs sorted by PageRank

K) in 4.2.7, paragraph 2, last sentence, explain a scenario where this strategy (only checking larger barrels if there are not enough matches in the first set) gives an undesirable outcome.

L) in 4.3, paragraph 2, “the crawlers are implemented in Python” … Python is not a high-performance language; explain why this language may still have been a good choice.

M) in 4.5, do a rough complexity analysis for each step of the algorithm in Figure 4

N) in Figure 4, which steps are likely to generate disk seeks?

O) in 4.5.2, last sentence, why is this method “far from perfect”?

P) when is the term frequency (tf) computed and where is it stored?

Q) when is the inverse document frequency (idf) computed and where is it stored?