Information Retrieval (CS60092)

Mid-Semester Examination

Maximum Marks 50

This question paper has 2 pages and 9 questions

Note: There are no clarifications. In case of doubt, you can take a valid assumption, state that properly and continue.

Question 1

a) Explain in one sentence why it is important for a crawler to detect whether two pages that it has downloaded are "near duplicates".

b) State two reasons why one would be interested to identify such near duplicates.

c) How would you take into account the presence of near duplicate pages while computing the PageRank for preferential crawling? **[1+2+2=5]**

Question 2

Suppose that P, Q, and R are different web pages.

a) Explain how it can happen that adding a link from P to Q can raise the PageRank of R.

b) Explain how it can happen that adding a link from P to Q can lower the PageRank of R.

In both cases, you should show a specific graph where this happens, though you need not work out the actual numerical values. **[6]**

Question 3

If all the hub and authority scores are initialized to 1, what is the hub/authority score of a node after one iteration? Also assume α and $\beta$to be 1 each. **[3]**

Question 4

Consider a web-graph with three nodes, A, B and C, and transition probabilities as follows. From node A, the next node is B with probability 1. From B, the next node is either A with probability p_A, or node C with probability $1 - p_A$. From C, the next state is A with probability 1. For what values of $p_A \in [0, 1]$ every node is reachable from every other node (may not be in a direct hop)? **[3]**

Question 5

Explain with an example the idea of vertex copying in web-graphs. **[3]**

Question 6

We discussed the BSBI approach for index construction. For the questions below, please present arguments using approximate numbers / bounds.

a) Take a corpus for your reference -- the corpus has 8,000,000 documents, 1,000,000 unique terms and 5,000,000,000 non-positional postings. Your server has a main memory of 15 GB. Can the indexing be done either fully using memory or fully using the disk?

b) Suppose now that you use BSBI and divide the data into *n* blocks (Provide some approximate range of *n*). You then sort each of these blocks, and proceed for the multiway merge. At this point, you can take some "*decent-sized*" chunks from each of the sorted blocks. Why can't you take i) all the blocks together? or ii) only one word at a time from each of the n blocks?

[4+4=8]

Question 7

Answer the following questions *(Answer to the point. Essays will not be graded.)*

a) In the permuterm index, we use a boundary marker. Why is it required? How would you search a word like mag*cent using permuterm index? You need to show both the indexing as well as the query sides.

b) In the context-sensitive spelling correction, suppose the query contains words w_1, w_2, w_3 and w_4. Each of these words are in the vocabulary. With an edit distance threshold of 1, you find 5, 7, 9 and 3 candidates for each of these words. How many candidates you will have to work with for the spelling correction? If you take a simplifying assumption (state the assumption), how many candidates will be remaining?

c) In the inverted index data structure, why do we need document frequency? Also, in the postings, why are the documents put in sorted order?

[3+3+3=9]

Question 8

Match the list of surnames below with their corresponding SOUNDEX codes (given in arbitrary order), and also restore the missing characters.
Surnames: Allaway, Anderson, Ashcombe, Buckingham, Chapman, Colquhoun, Evans, Fairwright, Kingscott, Lewis, Littlejohns, Stanmore, Stubbs, Tocher, Tonks, Whytehead
Soundex Codes: S312, T␣6␣, ␣5␣3, C42␣, T520, L␣42, A536, C155, ␣623, S356, ␣252, ␣152, ␣330, A251, A400, L2␣0

You are also provided the required letter to digit mappings (and it is expected that you remember the rest of the algorithm, if not, you can take this question as a puzzle and recreate it.)

$$B, F, P, V \rightarrow 1$$
$$C, G, J, K, Q, S, X, Z \rightarrow 2$$
$$D, T \rightarrow 3$$
$$L \rightarrow 4$$
$$M, N \rightarrow 5$$
$$R \rightarrow 6$$

[8]

Question 9

Consider words w_1 and w_2 in your vocabulary, and assume that they have posting lists of size x and y, respectively. We discussed in the class that for the query, (w_1 AND w_2), we can merge the two postings in O(x+y). What about the query (w_1 AND (not w_2))? Suppose N denotes the number of documents in the corpus. Give a small explanation to justify your answer.

[5]