INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

End-Spring Semester Examination 2022-23

Duration: 3 hrs. Full Marks: 70

Subject No.: CS60092. Subject: Information Retrieval

Department/Center/School: Computer Science & Engineering

Specific charts, graph paper, log book etc., required N/A

Special Instructions (if any):

1. This question paper contains four (04) printed pages.
2. Attempt all questions. *All parts of the same question must be answered together.*
3. Be short, precise. *Essays will not be graded.* Explain with examples, mathematical derivations or numeric calculations.
4. No clarifications can be provided during the examination. Make reasonable (logical) assumptions if necessary, and state any assumptions made clearly.
5. All workings must be shown. You can use calculators.

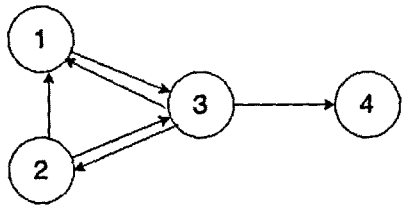
Question 1 **[3+6+5 = 14 Marks]**

a) Discuss the merits and demerits of using duplicate detection technique to implement queries such as "find pages like this one". Propose an improved solution.

b) Assume you have a social network of researchers. Researchers post (in text) about their research, their interest in movies and books, along with their social lives. In a post, one can share someone else's post (only one and assume no hashtag being used). Each researcher may have some friends (not from similar fields) and colleagues (working in similar areas).

Imagine you are designing an IR system for such a researcher, where query topic can vary from individual research interest, movies and books (not research related). How would use the PageRank algorithm (or its variants) to design such a system? Clearly state what will be the documents, how to get links between the documents, and how to encode the documents (to maximally encode required information). Give (nontrivial) examples of queries which your system will *not* be able to handle.

c) The web graph below contains 4 web-pages. Write the row-normalized adjacency matrix, and convert it into the transition matrix as shown in class (as done for PageRank). Show each step.



All pages are initially equally important. The probability of the teleportation is 0.2. Assume that the PageRank is computed by multiplying the transition matrix with the PageRank probability vector iteratively (the power method). Compute the PageRank scores after the third iteration $\pi^{(3)}$. ($\pi^{(0)} =$ [0.25 0.25 0.25 0.25])

Question 2 **[2x3+6 = 12 Marks]**

Consider a document collection having

Number of terms: 2,50,000

Number of tokens: 15,000,000

Bytes per token: 12

Bytes per term id / doc id: 4

Consider a system with the typical system parameters shown below.

► **Table 4.1** Typical system parameters in 2007. The seek time is the time needed to position the disk head in a new position. The transfer time per byte is the rate of transfer from disk to memory when the head is in the right position.

Symbol	Statistic	Value
s	average seek time	$5\ \text{ms} = 5 \times 10^{-3}\ \text{s}$
b	transfer time per byte	$0.02\ \mu\text{s} = 2 \times 10^{-8}\ \text{s}$
	processor's clock rate	$10^9\ \text{s}^{-1}$
p	lowlevel operation (e.g., compare & swap a word)	$0.01\ \mu\text{s} = 10^{-8}\ \text{s}$
	size of main memory	several GB
	size of disk space	1 TB or more

We want to construct an index for the document collection using external storage. We apply BSBI indexing for this purpose. Compute the total time required for BSBI indexing for each of the following steps. Assume we have 10 blocks to read data from or write data to.

a) Reading collection of data from disk

b) Total sorting time of blocks

c) Total time for writing sorted blocks to disk

d) Suppose there is a secondary storage whose seek time is 2.5ms, transfer time 0.01microseconds, size is somewhere between main memory and disk space. Can you modify the BSBI algorithm *optimally* to make use of the secondary storage? Write the algorithm, discuss the time and space complexity.

Question 3 **[3+7+2x2 = 14 Marks]**

Assume that the among N documents, following relevance judgements (0-2) are collected from three users. Answer the following questions.

q1: "machine learning", q2: "information retrieval", q3: "learning and retrieval"

d1: machine learning provides a set of techniques.

d2: information retrieval often uses learning techniques.

d3: learning to rank algorithms are used to rank and retrieve.

Quey	Doc	User 1	User 2	User 3
q1	d1	1	1	2
q1	d2	0	1	0
q1	d3	2	0	0
q2	d1	0	0	1
q2	d2	2	0	1
q2	d3	1	0	0
q3	d3	1	0	0
q3	d2	1	1	1

(1) What is the Retrieval Status Value of d3 with respect to q3? Show steps. Assume lemmatization can be applied. Apply smoothing *if necessary* ($\epsilon = 0.01$).

(2) Use tf-idf vectors for ranking for query 1 and 2. Then use one round of Rochhio's feedback algorithm (with alpha = beta = gamma = 1) to re-rank. Show the revised queries and re-ranked documents for Query 1 and Query 2. Use majority voting to utilize multiple user judgements.

(3) In the equation $q_{opt} = argmax_q\ (sim(q, \mu(D_r))- sim(q, \mu(D_{nr}))$, is D_{nr} necessary? If yes, why? Assume that you only have feedback on which documents are relevant. You cannot also automatically assume that other documents are not relevant (for example for q3, you cannot assume d1 is not relevant). For any query, provide an automatic way to approximate non-relevant documents.

Question 4 **[2.5x2+2.5x2 = 10 Marks]**

a) Compute (i) variable byte codes and (ii) gamma codes for the postings list 115, 421, 1445, 17829. Use gaps instead of docIDs for all except the first entry. Give the solution for variable byte codes as a sequence of 8-bit blocks. Give the solution for the gamma codes as a sequence of pairs of bit strings, where the first bit string of each pair corresponds to a length and the second to an offset.

b) During index construction, index compression maybe useful to fit as many postings in memory. Consider the γ codes of the two runs of postings for two terms (note we encode gaps other than the first).

γ encoded gap sequence of run 1: 1101011111001100111111101000111111110010
γ encoded gap sequence of run 2: 1101011110100011110010011111000100111111100100011

What is the merged sequence? Show each step (the docIDs for each encoded sequence, then the final merged sequence).

Question 5 **[3+(3+1)+(2+1) = 10 Marks]**

a) Consider the task of predicting the next word from a sequence of words using an RNN. Let the vocab size is 200 and each word is represented using one-hot encoding. The dimension of the hidden layer is 25. Calculate the total number of parameters (ignoring the bias).

b) Assume a language model $P(S) = \Pi_i P(w_i|w_{i-1} \ldots, w_1)$. Clarify all the assumptions. Be brief. Show examples (sentences/phrases) where the assumptions do not hold. Does the Transformers-based LM have some assumptions as well? Explain (in brief).

c) For the given Query vector Q, Key vector K and Value vector V, Attention is calculated as shown below: $softmax(Q * K^T / \sqrt{d})V$, where $d_K = d_Q = d_V = d$. Assume that $E(q_i) = 0 = E(k_i)$ and $Var(q_i) = 1 = Var(k_i)$ for all $i \in \{1, \ldots, d\}$. Calculate the mean and variance of $Q * K^T$. Then explain why it is divided by $\sqrt{d}$.

Question 6 **[3 + 4 + 3 = 10 Marks]**

Answer the following questions. The answers should be brief, to the point.

a) Use of idf in vector space model helps in assigning a higher weight to the rare terms in comparison to the common terms. How is a similar effect achieved in language model? Justify your answer.

b) Relevance feedback produces a more distributed query representation. Can you provide a way to make the representation sparse? Explain the disadvantage of such a query representation (compared to sparse query representation such as tf-idf) in terms of precision-recall.

c) "RNN is similar to a n-gram language model." Is the statement correct? Say "Yes" or "No". Then, explain mathematically. (Hint: show next word probability equations)