Indian Institute of Technology Jammu

Computer Science & Engineering

End-Semester Examination

CSC004P1M: Data Organization and Retrieval

Max. Marks: 50

Total Marks: 55 **Time: 2 hours**

Note: It is closed book / notes examination

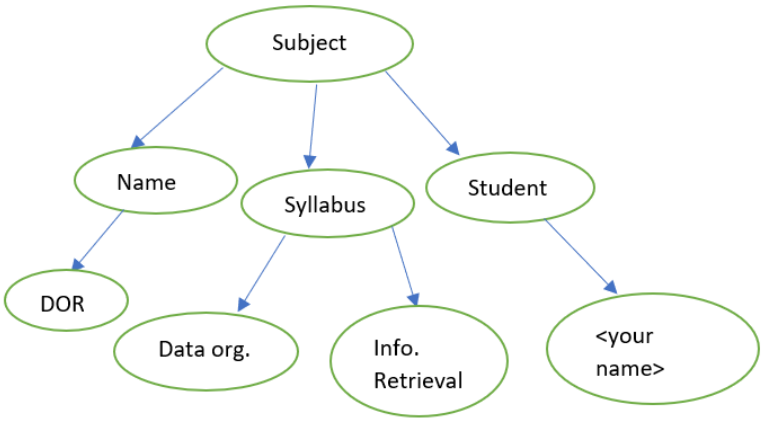
Answer all questions

Q1. [Marks 3] Given a collection of 1 million documents with a vocabulary of 1000 terms. We design a binary term-document incidence matrix with such a scheme. What are the design issues of this scheme? What is the alternative ?

Q2. [Marks 5] Give reason and different examples to use why we can / cannot use RDBMS for structured and unstructured information retrieval.

Q3. [Marks 2] "Stemming is required to be performed during index creation time and not while processing the query". The given statement is true / false. Why?

Q4. [Marks 8] Write all the structural terms appearing in the following XML document.



Q5. [Marks 10] Write pseudocode of map() and reduce() functions in Map-reduce paradigm to generate an index with structure term -> doc_id , term frequency.

Q6. [Marks 10] Given a set of stemmer rules for matching suffix of words and their replacements:

ION -> E

ING -> E
SS -> SS
S ->
MENT ->

a. Based on the above rules, what will be the stem of the following words
    i. Rabbits
    ii. Flying
    iii. Cement
    iv. Ration

Discuss where the given stemmer rules are able / not able to generate meaningful words. Wherever possible, propose a modification to correct the rule as well.

Q7.[Marks 7] Write different possible rotations of the term "Allo*ment" in the permuterm wildcard index. What key(s) would one lookup on? Write different trigrams of the key.

Q8. [Marks 3+7] Construct the minimum loser tree from the given graph.

a. Initialize the tree.
b. Generate the execution sequence for the first two winners (replacing the winning one with 9).