Code No: 57057 **R09**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
B. Tech IV Year I Semester Examinations, December - 2014
INFORMATION RETRIEVAL SYSTEMS
(Common to CSE, IT)

Time: 3 Hours **Max. Marks: 75**

Answer any Five Questions
All Questions Carry Equal Marks

- - -

1.a) There are many measures/metrics associated with information systems. For the following scenario, suggest an appropriate measure or metric and justify your choice. A student searching a set of documents to find all of those that pertain to an assignment. All that the student wants is part of an answer or full answer for the assignment, so the student does not mind sifting through all of the documents that were returned.

b) Two retrieval systems, X and Y, are being compared. Both are given the same query, applied to a collection of 1500 documents. System X returns 400 documents, of which 40 are relevant to the query. System Y returns 30 documents, of which 15 are relevant to the query. Within the whole collection there are in fact 50 documents relevant to the query. Tabulate the results for each system, and compute the precision and recall for both X and Y. Show your working.

c) Both precision and recall need to be taken into account when evaluating retrieval systems. Why is it not sufficient to pick one and use only that?

2.a) What is the likely effect of Stemming and Lemmatization on?
i) Vocabulary size: Increase, Decrease, Unpredictable.
ii) Precision: Increase, Decrease, Unpredictable.
iii) Recall: Increase, Decrease, Unpredictable.

b) How do stopping and stemming reduce the size of an inverted index?

c) Use the 2-gram index and 3-gram index for processing the following wildcard queries Tol* and Rea* . Is "Tool" result for the wildcard query Tol*. If the answers yes, solve this problem.

3.a) Given a document with the terms A, B and C with the following frequencies A: 3, B: 2, C: 1.The document belongs to a collection of 10,000 docs. The document frequencies are: A: 50, B:1300, C:250. Compute the normalized tf and the tf-idf and compare them.

b) In the vector space retrieval model, if a query has just one term, which term weighting heuristic will be ineffective. Use no more than two sentences to explain why.

c) What are the three major term weighting heuristics in the vector-space retrieval model? Explain in brief.

4.a) Consider the following D matrix.

D=	1	0	0	1	0	0
	1	1	1	0	1	1
	0	0	0	0	0	1
	0	0	1	0	0	1
	0	0	1	1	0	0

Obtain the corresponding single-link clustering structure (dendrogram). Give the clustering structure approach if the dendrogram is cut at the similarity level 0.45 (note that you will obtain a partitioning structure). For similarity calculations use the Dice coefficient.

b) Explain existing terms clustering with an example.

5.a) Suppose that a user's initial query is "cheap CDs cheap DVDs extremely cheap CDs". The user examines two documents, d1 and d2. She judges d1, with the content CDs cheap software cheap CDs relevant and d2 with content cheap thrills DVDs non relevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

b) Why do commercial web search engines typically not provide relevance feedback functionality? Give at least 3 reasons.

c) What is the main advantage of the cosine similarity measure?

6.a) Explain Boyer-Moore algorithm with a relevant example.

b) Write short notes on "Hardware text search systems".

7. Describe efficient Indexing and Searching in Multimedia Information Retrieval.

8.a) Discuss digital libraries.

b) Explain online IR Systems.

--ooOoo--