Code No. 5309 / MP

FACULTY OF INFORMATICS

B.E. 4/4 (I.T.) II-Semester **(Make-up)** Examination, August 2012

Subject : **Information Retrieval Systems**
(Elective – IV)

Time : 3 Hours Max. Marks: 75

Note: Answer **all** questions of Part – A. Answer any **five** questions from Part-B.

PART – A (25 Marks)

1. What is stemming? Give the name of a particular stemmer.

2. Write down the formula for cosine similarity between query q and document d.

3. An IR system returns 3 relevant documents, and 2 irrelevant documents. There are a total of 8 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

4. Define Levenshtein edit distance.

5. What is the difference between adhoc retrieval and relevance feedback?

6. Consider a collection of 5000 documents and 2000 unique vocabulary terms. Suppose documents have 200 terms on average. If we added 2200 more documents to the collection, roughly how many unique vocabulary terms would be added? Use Heaps' law with k = 10 and β=0.5.

7. Why do some of the search engines in the web prefer to index all the terms in the documents?

8. What is a suffix tree?

9. Which of the parallel architectures (MIMD/SIMD) are more suitable for parallel implementation of signature files?

10. Does the portioned parallel processing on a MIMD machine improve query response time or throughput? Justify your answer.

PART – B (50 Marks)

11. Consider the following collection of 4 documents:
Doc1: Information Retrieval Systems
Doc2: Information Storage
Doc3: Digital Speech Synthesis Systems
Doc4: Speech Filtering, Speech Retrieval
and consider the vector model using tf-idf weights, and the cosine measure of similarity.
(a) Compute all non-zero components of the vector representing Doc1.
(b) Compute the ranking of all four documents with respect to the query "Speech Systems".
(c) Compute the similarity between (i) Doc1 and Doc2; (ii) Doc3 and Doc4

12.(a) Consider the Boolean query "A OR B AND C". What are the two different interpretations of this query?
(b) What are different hierarchical models? Briefly describe any one of these models.

13.(a) Explain the working of complete link document clustering algorithm.
(b) How do you use user feedback information to estimate probabilities in probabilistic model?

14.(a) Explain how context queries are processed using inverted files.
(b) Describe the criteria for index selection.

15.(a) What are the four main steps of query processing in distributed IR?
(b) Use the Boyer-Moore algorithm to search for the term FANCY in text string FANCIFUL FANNY FRUIT FILLED MY FANCY.

16.(a) Explain how automatic thesaurus helps in improving effectiveness of information retrieval system.
(b) According to Zipf's law, which of the following strategies is more effective for reducing the size of an inverted index?
(i) reduce k common words;
(ii) remove k rare words.

17. Write short notes on the following:
(a) Models for browsing
(b) Markup languages
(c) Document clustering

*****