Code No. 5309 / M

FACULTY OF INFORMATICS

B.E. 4/4 (I.T.) II-Semester **(Main)** Examination, May / June 2012

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

Time : 3 Hours Max. Marks: 75

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

PART – A (25 Marks)

1. What is the tf-idf weight of terms t in document d?
2. What is the difference between the concepts of query and information need?
3. List the three major parts of TREC collections.
4. What is an easy way of maximizing the precision of an IR engine?
5. Suppose we have a collection of N documents and V unique terms. For simplicity, assume that each term appears at most once per document. Assuming Zipf's law holds, What proportion of terms appear D times (i.e. in every document)?
6. Why is relevance feedback not used by most search engines?
7. What is the advantage of word based Huffman coding over Character based Huffman code?
8. What are the four different stemming strategies? Which of these strategies is simple to implement?
9. What is false drop problem in signature files?
10. In distributed IR. What is the main drawback of treating each distributed document collection as single large document for the purpose of source selection?

PART – B (5x10=50 Marks)

11.(a) What are the drawbacks of Boolean model?

(b) The table bellow shows a frequency table for three documents (D1, D2 and D3) in an information retrieval system. A user, who is interested in sport, in particular football and rugby, enters the query 'sports, football, rugby' into this system.

	D1	D2	D3
Sports	4	1	3
Politics	1	5	0
Leisure	0	3	0
Football	3	1	0
Rugby	0	1	5
Economics	1	4	1

What results would the system return if it used the Vector Space model of information retrieval with cosine distance as a similarity metric?

12.(a) Suppose the relevance status of the top-8 ranked results from a system is [+, +, +, -,-,-,-,+]. Here, + indicates non-relevant document. Suppose there are in total 10 relevant documents in the collection. Compute the following evaluation measures for this result: (i) Precision (ii) Recall (iii) Precision at 5 documents (iv) Average Precision.

(b) Discuss the query formulation based on the concept of pattern.

13.(a) Compute the edit distance between following terms :
raicurent, recurrent

(b) Briefly explain query expansion through local clustering.

14.(a) What are the five distinct text preprocessing operations ? Explain whether these operations improve retrieval performance or not.

(b) Consider the collection made of the 3 following documents :
D1 : out of the clear blue sky
D2 : the blue car next to the entrance
D3 : sky news : information retrieval in nice
give the inverted index of this collection

15.(a) Describe the parallel implementation of Inverted file on SIMD architecture.

(b) Discuss the criteria for collection portioning in distribution IR.

16.(a) Us the Knuth-Pratt-Morris algorithm to search for the term FANCY in text string FANCIFUL FANNY FRUIT FILLED MY FANCY.

(b) Describe generic software architecture of IR system.

17. Write short notes on the following :

(a) Suffix Automation

(b) Query protocols

(c) Inverted file compression