Question Paper

Exam Date & Time: 17-Apr-2018 (10:00 AM - 01:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

SCHOOL OF INFORMATION SCIENCES (SOIS)
SECOND SEMESTER ME - (BIG DATA AND DATA ANALYTICS)
DEGREE EXAMINATION- APRIL/MAY 2018
Tuesday, April 17, 2018
Time : 10.00 am to 1.00 pm
Text Retrieval and Search Engines [BDA 616.1]

Marks: 100 **Duration: 180 mins.**

Answer all the questions.

1) a) What is the feast or famine problem? Which Retrieval model faces this problem and why? **(5** Marks) (10)
b) Why don't we use **grep** for information retrieval? **(5 Marks)**

2) a) Define Precision and Recall. **(5** Marks) (10)
b) Compute the Jaccard Similarities of each pair of the following three set: **(5** Marks)
{1,2,3,4} {2,3,5,7} {2,4,6}

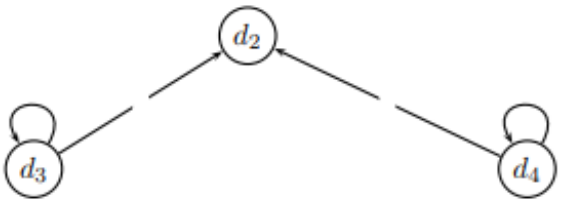
3) What are the basic building blocks of a search engine? (10)

4) Explain Block Sort Based Indexing. (10)

5) Briefly explain the Krovetz Stemmer. (10)

6) Why is PageRank a better measure of quality than a simple count of in-links? (10)

7) (10)



Compute PageRank for the web graph in the figure for each of the three pages. Assume that at each step of the PageRank random walk, we teleport to a random page with probability 0.6 .

8) a) Consider these documents: (10)
Doc 1 breakthrough drug for schizophrenia
Doc 2 new schizophrenia drug

Doc 3 new approach for treatment of schizophrenia
Doc 4 new hopes for schizophrenia patients

Draw the term-document incidence matrix for this document collection. **(5** Marks)

b) Compare the Boolean Retrieval Model vs Ranked Retrieval Model. (**5** Marks**)**

9) You are given the following table that contains the term frequencies for certain novels. (10)

i) Calculate the cosine score between each pair.
ii) What does the cosine score signify?
iii) List the document pairs in order of descending cosine scores.

Take idf score = 1.

Term	SaS	PaP	WH
Affetion	115	58	20
Jealous	10	7	11
Gossip	2	0	6
Wuthering	0	0	38

10) What are the techniques to handle wildcard queries? (10)

-----End-----