3007

No of Pages : 4 Course Code : 18XW81

Roll No:

(To be filled in by the candidate)

PSG COLLEGE OF TECHNOLOGY, COIMBATORE 641 004

SEMESTER EXAMINATIONS, AUGUST 2023

MSc - SOFTWARE SYSTEMS Semester : 8

18XW81 INFORMATION RETRIEVAL

Time : 3 Hours **Maximum Marks :100**

INSTRUCTIONS:

1. Answer **ALL** questions. Each question carries 20 Marks.
2. Subdivision (a) carries 3 marks each, subdivision (b) carries 7 marks each and subdivision (c) carries 10 marks each.

1. a) The tf-idf weighting scheme assigns to term 't' a weight in document 'd'. Is idf of a term same across all the documents? Why or why not?

   b) i) What is meant by 'term independence' in BIM? Give an example. (3)

   ii) Write the query processing order for the following Boolean query. (4)

   (algorithm OR analysis) AND (organize OR structure) AND (file OR data)

   Assume that the document frequencies of the terms in the above query are:

   algorithm 303; analysis 890; data 346; structures 728; organize 245; file 471;

   Give the rationale for the order that you suggest

   c) You have the collection of documents that contain the following index terms:

   Doc 1 : Information Retrieval Systems

   Doc 2 : Information Storage

   Doc 3 : Storage Systems

   Doc 4 : Speech Filtering, Speech Retrieval

   Doc 5 : Retrieval Systems

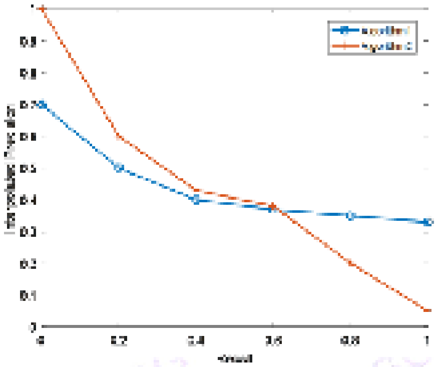
   Query : Speech Systems

   Compute TF*IDF scores (TF is un-normalized term frequency and IDF is inverse document frequency) of the terms. Apply cosine similarity to identify top k(=3) similar documents of the query and rank the documents in the collection for the query.

2. a) State Zipf law and write its significance in IR?

   b) i) Give the expansion of NDCG. Why DCG is preferred to CG? (3)

ii) Consider the following figure which depicts the PR curve of two algorithms Algorithm1 and Algorithm2. Give proper interpretation of the graph. (4)



c) The following list of Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

R R N NNNNN R N R N NN R N NNN R

i. What is the precision of the system on the top 20?

ii. What is the F1 on the top 20?

iii. What is the un-interpolated precision of the system at 25% recall?

iv. What is the interpolated precision at 33% recall?

v. Assume that these 20 documents are the complete result set of the system. What is the MAP for the query?

3. a) Differentiate soft clustering and hard clustering. Give an example,

b) Why does thesaurus-based query expansion typically not work very well?

Corpus C consists of the following three documents:

d1: "new york times"

d2: "new york post"

d3: "los angeles times"

Assume in response to the results of the query "new new times," the user rates the following documents as irrelevant:

"new York times"

"new york post"

Reformulate the query to account for relevance feedback with α = 0:8, β = 0:2, and γ= 0.

c) What is the difference between the two models of Naïve Bayes classification? Assume we want to categorize query into two categories: Systems, Theory. Consider performing NB classification using Bernoulli's method for classification

	docID	Words in document	class
Training Set	1	data search	Systems
	2	disk search heuristic	Systems
	3	theorem algorithm	Theory
	4	algorithm search analysis	Theory
	5	heuristic search	Systems
Query	6	search search algorithm	?

4. a) Differentiate user based Collaborative filtering and item based collaborative filtering

b) i) Consider the document collection.

DocID	Document Text
1	a rose by any name smells sweet
2	a rose is a rose
3	a rose smells sweet

Write 2 Shingles of the collection and calculate the similarity between pair of documents using Jaccard similarity of Shingles. (4)

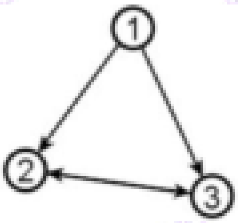
ii) Differentiate between web content mining, web structure mining and web usage mining. (3)

c) Consider the following suggestions collected in the form of ratings from five healthcare professionals (HP1-HP5) for the selection of drugs to *influenza*. The ratings range from 1(high side effects, worst) to 5(less side effects, good). Find out the rating predicted for $P_{3,\ Peramivir}$ based top 3 similar drugs to Peramivir using item-based Collaborative Filtering. Make predictions based on similarities between drugs. Apply dot product similarity for similarity calculations and weighted average for prediction calculation.

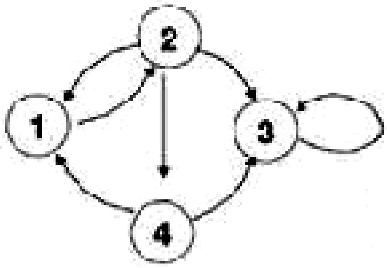
Drug name	HP1	HP2	HP3	HP4	HP5
Arbidol	5	1	?	5	1
Amantadine	2	3	?	3	?
Peramivir	5	?	?	1	4
Zanamivir	2	?	2	3	1
Rimantadine	?	5	3	?	3
Laninamivir	?	4	1	?	?

5. a) When do you say a Markov chain is ergodic? What is its significance in pagerank?

b) What are hub scores and authority scores? For the following graph, findout hub score and authority score according to HITS algorithm. Perform 2 iterations.



c) Discuss random walk modelling in PageRank? Consider the following graph.



Given the directed graph of webpages, perform three iterations of PageRank computations. The arcs indicate outbound links between webpages. Initially give each page a PageRank score of 0.25. Use a 'teleport' (or transition) probability of 0.10. (Put differently, 90% of the time the random surfer follows a link to get to a new page). Show the PageRank scores of all pages.

/END/

FD/RL