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, APRIL 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.
3. Course Outcome : Table

Qn.1	CO1.	Qn.2	CO2	Qn.3	CO3.	Qn.4	CO.4	Qn.5	CO5.

1. a) What is the concept of an inverted file and what is its use in information Retrieval?

   b) i) Describe how skip pointers are used in postings lists. What is the advantage of skip pointers in processing a Boolean query of the form x and y? (4)

   ii) Recommend a query processing order for the following Boolean query: (3)
   (bush OR apricot) AND (pudding OR brown) AND (phones OR ears)
   Assume that the document frequencies of the terms in the above query are:
   ears 213312; phones 87009; pudding 107913; brown 271658;
   bush 46653; apricot 316812
   Give the rationale of 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: Digital Speech Synthesis Systems
   Doc 4: Speech Filtering, Speech Retrieval
   Doc 5: Retrieval Systems
   Query Speech Systems
   Compute the tf-idf representation of these 5 documents. Idf is $\log_{10}(N/dft)$, where N is the number of documents and dft is the document frequency of the term t. tf is the term frequency (un-normalized). Calculate cosine similarity of these documents with respect to the query.

2. a) Suppose we have a large, mostly unknown collection. By inspection of the first one lakh tokens, we find 30,000 distinct terms. About how many terms would we expect to find in the first 10 lakh tokens? Why? Which law allows us to predict this?

   b) i) Please pick the **most appropriate** evaluation metric for the following search tasks. Justify your answer.

1) A businessman searching for New York Time's homepage for his breakfast reading.
2) A lawyer searching for all relevant evidence to one of his cases. The lawyer is evaluated by whether he could win the case and he bills his client by hours. Therefore he does not mind to read through all the documents that are returned by a search engine. (3)

ii) In what way does DCG differ for the evaluation measures precision and recall? Can you rationalize why we go from CG to DCG? (4)

c) Suppose that we have a standard IR evaluation data set containing 1000 documents. Assume that a particular query in this data set is deemed to be relevant to the following 25 documents in the collection:

REL = { d1, d5, d6, d10, d88, d150, d200, d210, d250, d300, d400, d405, d450, d472, d500, d501, d530, d545, d590, d600, d635, d700, d720, d800, d900 }

Two different retrieval systems S1 and S2 are used to retrieve ranked lists of documents from this collection using the above query. The top 10 retrieved documents for these two systems are given below (each list is in decreasing order of relevance).

RET(S1) = d2, d5, d150,d250, d11, d33, d50, d600, d500, d520

RET(S2) = d250, d400, d150, d210, d999, d3, d501, d800, d205, d300

i) First plot an exact recall/precision graph for each system as a function of the number of documents returned (for 1 document returned, 2 documents returned, etc) (recall on the X axis) and then overlay it with a graph where the precision values are interpolated to the standard 11 points.

ii) A single metric that can be used to combine precision and recall is the F Measure. Using the F measure, create graph similar to the above comparing the two systems.

iii) Which system is better? Explain your answer.

3. a) EM uses a mixture of k Gaussian for clustering; what purpose do the k Gaussian serve? What is the task of the E-step of the EM-algorithm?

b) 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" nonrelevant. 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 $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

c) What are the two models for Naïve bayes classification? How does the calculation of posterior probabilities differ for the two models?

Assume we want to categorize science texts into the following categories: Physics, Biology, Chemistry. Consider performing naive Bayes classification with a simple model in which there is a binary feature for each significant word indicating its presence or absence in the document. The following probabilities have been estimated from analyzing a corpus of preclassified web pages:

c	Physics	Biology	Chemistry
P(c)	0.35	0.4	0.25
P(atom|c)	0.2	0.01	0.2
P(carbon|c)	0.01	0.1	0.05
P(proton|c)	0.1	0.001	0.05
P(life|c)	0.001	0.2	0.005
P(earth|c)	0.005	0.008	0.01

Assuming the probability of each evidence word is independent given the category of the text, compute the posterior probability for each of the possible categories for each of the following short texts.

i) The carbon atom is the foundation of life on earth.

ii) The carbon atom contains 12 protons

Assume the categories are disjoint and complete for this application. Note that words are first stemmed to reduce them to their base form, therefore "proton" and "protons" should be considered equivalent. Ignore any words that are not in the table.

4. a) What is named entity recognition? Give an example. Name any two applications of named entity recognition.

b) i) What are the characteristics of web that makes Web IR more challenging than conventional IR? (4)

ii) What are Shingles? Explain with an example. What is its role in duplicate detection? (3)

c) Differentiate user based collaborative filtering and item based collaborative filtering. The table below gives the ratings of six items by five users. The rating scale is 1 (poorest) through 5 (best) and "X" indicates not rated.

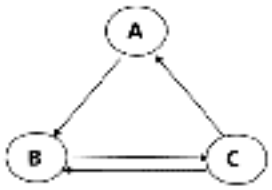
	User 1	User 2	User 3	User 4	User5
Item 1	3	X	2	3	2
Item 2	X	1	2	2	X
Item 3	4	2	X	X	X
Item 4	5	1	X	5	3
Item 5	3	5	5	X	4
Item 6	5	1	X	1	1
Item7	X	X	5	4	3

Predict the ratings given by user3 for item3 based on item-based collaborative techniques.(Consider K=2 and use Cosine similarity for item based calculations).

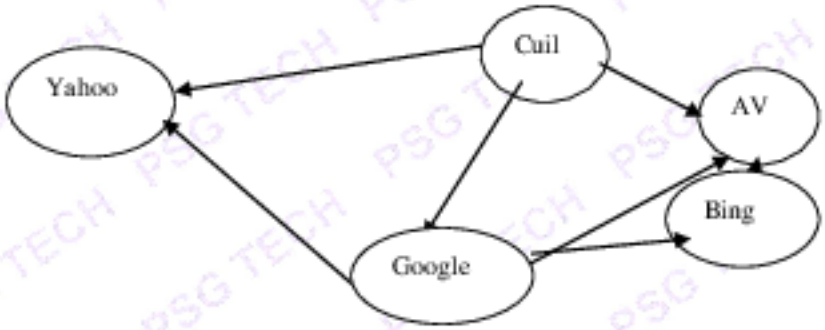
No of Pages : 4 **Course Code : 18XW81**

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

   b) What are hubs and authorities? Starting with hub score 1 at each vertex, make two complete iterations in the calculation of hub and authority scores for the network below.



   c) What are dead ends and Spider traps? How does the google pagerank algorithm deal with them? 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 0.2. Use a 'teleport' (or transition) probability of 0.80. Show the PageRank scores of all pages (write first three iterations).

/END/

FD/JU