No of Pages : 3 Course Code : 18XW81

Roll No:
(To be filled in by the candidate)

PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

SEMESTER EXAMINATIONS, MAY 2022

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 are the limitations of Boolean model and how is it overcome in vector space model?

   b) i) What is the advantage of taking the Odds ratio rather than calculating the posterior probabilities for the Binary Independence Model?

   ii) We have a two word query. For one term the postings list consists of the following 16 entries. [2, 4, 9, 12, 14, 16, 18, 20, 24, 32, 47, 81, 120, 125, 158, 180] and for the other it is the one entry postings list [81] Work out how many comparisons would be done to intersect the two postings list with the following two strategies.

   - Using standard postings list.
   - Using postings list stored with skip pointers, with the suggested skip length of √P.

   c) Consider the query "oil producing nations", and the three query terms have inverted lists given as follows: list → (dfk, ctfk,(doci , tfik), ...), where dfk – document frequency of the word 'k', ctfk – collection frequency of the term 'k' and tfik - frequency of the term 'k' in document 'i'.
   oil → (5, 18,(1, 4),(4, 3),(6, 1))
   producing → (4, 20,(1, 6),(2, 2),(5, 4))
   nations → (3, 11,(1, 1),(3, 2),(6, 8))
   Further consider a collection of documents with lengths: d1 → 498, d6 → 639, d2 → 627, d3 → 551, d4 → 648, d5 → 621 and the total number of terms in the corpus is 5687.
   What are the scores of the 6 documents using Vector space model with term weighting, where N is the total number of terms in the collection

$$W_{i,k} = \frac{tf_{i,k}}{len_i} \log \frac{N+1}{0.5 + df_k}$$

2. a) With 5, 000 documents and 10, 000 unique vocabulary terms, a bit vector index requires 5×10^7 bits of storage. Suppose documents have 200 terms on average. If

we added 2200 more documents to the collection, roughly how big would the bit vector index become? Use Heaps' law with k = 10 and β = 0.5.

b) i) What are the lower and upper bounds for IDF of a term in a corpus? Why is IDF combined with tf in calculating the weights of terms? (4)

ii) Assuming Zipf's law with a corpus independent constant A = 0.1, what is the fewest number of most common words that together account for more than 18% of word occurrences (i.e. the minimum value of m such that at least 18% of word occurrences are one of the m most common words). (3)

c) Assume that an IR system returns a ranked list of 10 total documents for a given query. Assume that according to a gold-standard labelling there are 5 relevant documents for this query, and that the relevant documents in the ranked list are in the 2nd, 3rd, 4th, and 8th positions in the ranked results. Calculate and clearly show the interpolated precision value for each of the 11 point standard recall levels

3. a) Define Pseudo Relevance feedback. Why might "pseudo-relevance feedback" sometimes transform an imperfect query into one that is incredibly bad?

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: "losangeles times"

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

"new York times"

"newyork post"

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

c) Consider the following document-term matrix with raw frequencies. Assume that documents have been manually assigned to two pre-specified categories as follows:

Cat1 = {Doc1, Doc2, Doc5} Cat2 = {Doc3, Doc4, Doc6, Doc7}

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1
Doc4	0	1	0	2	0	0	1	0
Doc5	0	0	2	0	0	4	0	0
Doc6	1	1	0	2	0	1	1	3
Doc7	2	1	3	4	0	2	0	2

Determine how **Doc8** given below will be classified using Naïve Bayes classification. (Both models)

Doc8	T1	T2	T3	T4	T5	T6	T7	T8
	3	1	0	4	1	0	2	1

4. a) What is a cold start problem in Recommendation systems?

b) Differentiate between web content mining, web structure mining and web usage mining.

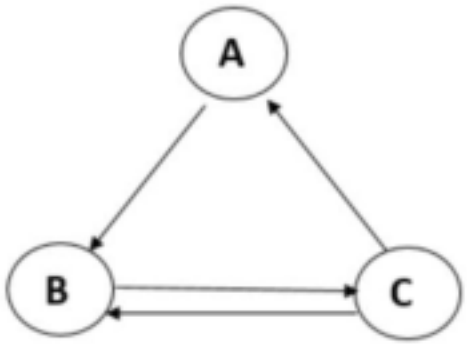
c) Differentiate user based Collaborative filtering and item based collaborative filtering. The table below gives the ratings of six items by three users. The rating scale is 1 (poorest) through 5 (best) and "X" indicates not rated.

	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).

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 web pages and the set of web pages they link to: Page A points to pages B, C, and D.

Page B points to pages A and D.

Page C points to pages B and D.

Page D points to page A. Trace the page rank algorithm for two iterations. Use initial PR values 0.25 for all nodes. Use d=0.85.

FD/JU **/END/**