SRN □□□□□□□□□□□□

PES University, Bengaluru-85
(Established under Karnataka Act No. 16 of 2013)

UE17CS412

October 2020: B. TECH, VII SEMESTER

ISA-I

UE17CS412 – ALGORITHMS FOR INFORMATION RETRIEVAL

Time: 2 Hrs. | Answer All Questions | Max Marks: 60

Provide full calculation for the numerical problems. No partial marking will be done.

1 a) **The question below has one correct answer.**
In your answer script, write your <u>chosen correct answer with reason in 2-3 sentences</u>:

1) Stemming increases retrieval precision
2) Stemming decreases retrieval precision

2

b) Wildcard **shi*pi*** is given. List the steps to find the **document ids** that contain the wildcard pattern using 3-gram index. You can use the additional alphabet $.

3

c) **Consider these documents:**
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

Using the term-document incidence matrix, find the result of the Boolean query
"for AND NOT(drug OR approach)"

5

2 a) **The question below has one correct answer.**
In your answer script, write your <u>chosen correct answer with reason in 2-3 sentences</u>:
For addressing word error correction in query. An Information Retrieval System should implement

1) Isolated word correction and contextual error correction in both query and corpus
2) Isolated word correction and contextual error correction only in query but not in corpus
3) Isolated word correction in query and contextual error correction in corpus

2

b) Compare the two methods of wild card query support for Boolean Retrieval in the three aspects as indicated below:

Method	Dictionary size	Posting List Size	Post Filtering Need
K gram			
Permuterm			

3

SRN

	c)	Shown below is a portion of a positional index: angels: 2: <36,174,252,651>; 4: <12,22,102,432>; 7: <17>; fools: 2: <1,17,74,222>; 4: <8,78,108,458>; 7: <3,13,23,193>; fear: 2: <87,704,722,901>; 4: <13,43,113,433>; 7: <18,328,528>; in: 2: <3,37,76,444,851>; 4: <10,20,110,470,500>; 7: <5,15,25,195>; rush: 2: <2,66,194,321,702>; 4: <9,69,149,429,569>; 7: <4,14,404>; to: 2: <47,86,234,999>; 4: <14,24,774,944>; 7: <199,319,599,709>; tread: 2: <57,94,333>; 4: <15,35,155>; 7: <20,320>; where: 2: <67,124,393,1001>; 4: <11,41,101,421,431>; 7: <16,36,736>; Which document(s) (if any) match both these queries at which positions? **"fools rush in" AND "angels fear to tread"**	5
3	a)	**The question below has one correct answer. In your answer script, write your <u>chosen correct answer with reason in 2-3 sentences</u>:** When T is the total number of postings and n is the size of auxiliary index, in case of logarithmic merge implementation of dynamic index as compared to naïve method, (a) there is a benefit in both index construction time and query processing time (b) there is a benefit in index construction time but a loss in query processing time	2
	b)	Show that the size of the vocabulary is finite according to Zipf's law and infinite according to Heaps' law.	3
	c)	Assume that the total number of documents in a **corpus** is **1024** and that the following words occur in the following documents in : "Computer" occurs in **32** documents "software" occurs in **8** documents "intelligent" occurs in **16** documents "robust" occurs in **1024** documents Document **D** : ***"Computer intelligent software robust computer software"*** Query **Q** : **"Intelligent Software"** **Assume a simplified TF-IDF weight formula as "tf * $\log_2$ (N/df)" where N is the document frequency.** (a) Calculate the **TF-IDF weighted term vector** for the **document D** without any normalization. (b) Assuming that query vector is computed <u>just in terms of TF weights (no IDF weights)</u>, and similarity is measured by the cosine metric, what is the similarity between Q and D ?	5
4	a)	**The question below has one correct answer. In your answer script, write your <u>chosen correct answer with reason in 2-3 sentences</u>:** 1. Both BSBI and SPIMI have same time complexity 2. BSBI and SPIMI have different time complexity	2

SRN

b) In the following table, 1st column shows the "efficient scoring/ranking" method employed and 2nd to 4th column lists the possible heuristics category. Put tick mark in the appropriate boxes with 1-2 sentences explanation for each row in the table. No part marking will be done for each row. 3

Method	Champion List	Index Elimination	No Heuristics
Only consider documents containing 75% of the query terms			
Consider a net score consisting of static quality score and TF IDF score			
We consider candidate documents containing at least one query term			

c) From the following sequence of γ-coded gaps, reconstruct first the gap sequence and then the postings sequence: 111000111010101 5

5 a) **The question below has one correct answer.**

In your answer script, write your <u>chosen correct answer with reason in 2-3 sentences</u>: 2

1) Rocchio algorithm for text categorization never shows any anomalous behaviour.
2) Rocchio algorithm for text categorization can show anomalous behaviour.

b) For XML based Information Retrieval, for a query q and a document d : 3

(a) What are the **dictionary elements** in the inverted index

(b) Define **context resemblance**

c) For the problem described in **question 3(c)** above, answer the following: 5

Suppose the user is shown D in response to the query Q, and the user says that D is relevant to his query. If we now use relevance feedback to modify Q using SMART algorithm , what will the query vector become? Assume that alpha=1, beta=0.5 and gamma=0.5

6 a) **The question below has one correct answer.**

In your answer script, write your <u>chosen correct answer with reason in 2-3 sentences</u>: 2

1) ROC curve and Precision Recall curve for a web search system are exactly same

2) Precision Recall curve represents only a small part of ROC curve for an IR system

b) Mention briefly one commonality and two differences between Language Model and Vector Space Model approach in Information Retrieval. 3

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. 5

List : **R R N NNNNN R N R N NN R N NNN R**

(a) What is the F1 on the top 20?

(b) Assume that these 20 documents are the complete result set of the system. What is the **MAP for the query?**