CMI/Aug-Sep-2019/End-Semester Exam Sep 2019

Chennai Mathematical Institute

Information Retrieval Duration: 120 Minutes. Max Marks: 60.

Roll No.: _____ Date: 27/09/2019

Name: _____

Instructions

- This is a closed book test.
- You are allowed to carry one page (A4 size only) of hand written or printed notes. Write your name and roll number on these notes. Submit the notes along with this booklet.
- Please switch off your mobile phones and any other digital equipment you may have (like laptops and smart watches). Calculators are allowed as long as they are not internet enabled.
- Negative marks apply for first section only.
- Please ensure this booklet has 20 (Ten 2-Mark Questions, Five 3-Mark Questions and Five 5-Mark Questions) questions. All questions are compulsory.
- You are encouraged to make your assumptions if any, explicit.

Section 1: Questions 1 - 10 carry 2 mark each. A negative mark of -1 applies for every wrong answer.

Question 1. Assume that there are N documents in your index. Apply the postings merge algorithm to the query given below. What is the time complexity to merge the postings for the following query: (IIITS OR CMI) AND NOT (DS OR CSC)?

Choose the best answer:

(1) O($logN$)
(2) **O(N)**
(3) O($NlogN$)
(4) O(N^2)

Question 2. Which situation(s) prompt you to prefer simple TF weighting over TF-IDF weighting scheme?

Choose the best answer:

(1) When term specificity is not important
(2) When all the queries have only one term
(3) **Both of the above**
(4) None of the above

Question 3. Consider the following statements:

a. In a boolean retrieval system, stemming cannot improve precision.
b. In a boolean retrieval system, stemming cannot improve recall.

Choose the best answer:

(1) (a) is true; (b) is true.
(2) **(a) is true; (b) is false.**
(3) (a) is false; (b) is true.
(4) (a) is false; (b) is false.

Question 4. An IR system returns 10 relevant documents, and 20 non-relevant documents. There are a total of 20 relevant documents in the collection. What is the F_2 score of this system?

Choose the best answer:

(1) 1/4
(2) 2/5
(3) **5/11**
(4) None of the above.

Question 5. Consider a collection with documents structured into three zones namely title, body and comments. Assume that the corresponding zone weights are distributed in the ratio 1:1:2. If the term “cmi" appears once in title alone, what would be the score of the document according to the zonal weighting scheme discussed in your course?

Choose the best answer:

(1) **0.25**
(2) 0.4
(3) 0.5
(4) None of the above.

Question 6. How many nodes will a suffix tree for a string of length n have, provided that the string has only distinct characters?

Choose the best answer:

(1) $log(n)$
(2) n^2
(3) $n(n+1)/2$
(4) $2n$
(5) None of the above.

Well, note that there are exactly n(n+1)/2 characters in all suffixes. Hence, an efficient way of storing the suffix tree should need only as much nodes. But, if you selected "None of the above" and cited an example, you should get marks too.

Question 7. What is the edit distance between balaji and bajaj if you are allowed to insert, delete and replace a character in one operation?

Choose the best answer:

(1) **2**
(2) 3
(3) 4
(4) 5

Question 8. Our index has 1 Million documents, each containing exactly 20 words. If there are 1000 distinct terms in total, how much space will an uncompressed term-document incidence matrix need if it is stored as a two dimensional bit array?

Choose the best answer:

(1) 10^9 **bits**
(2) $2 * 10^6$ bits
(3) $2 * 10^9$ bits
(4) None of the above.

Question 9. Our index contains only three terms namely chennai, mathematical and institute. In the vector space model, what would be the angle between the vectors representing the sentences “institute" and “chennai"?

Choose the best answer:

(1) 0
(2) 45
(3) 60
(4) **90**

Question 10. For an information need to find cafes in Chennai, a query “chennai cafe" was issued. Three annotators were employed to provide relevance judgment. The result of their work is shown in the following table:

Note that 'R' refers that the corresponding judge found the document to be relevant while 'N' marks non-relevance.

Document	**Judge1**	**Judge2**	**Judge3**
cafe	R	N	R
chennai cafe	R	R	R
No cafe	N	N	N

How many results should a probabilistic retrieval system based on *Bayes Optimal Decision Rule* produce?

Choose the best answer:

(1) 1
(2) **2**
(3) 3
(4) 4

Section 2: Questions 11 - 15 carry 3 marks each. No negative marks.

Question 11. For a query "great CMI", the term "great" has a postings list with the following 16 entries: [4,6,10,12,14,16,18,20,22,32,47, 81,120,122,157,180]. "CMI" is present in only one document whose docID is 47. Using postings list stored with skip pointers of skip length $\sqrt{p}$ where p is the postings list length, how many comparison(s) would be done to intersect the postings list? You are encouraged to draw the postings list with skip pointers to avoid ambiguity.

Answer: Number of comparisons = 6.

Question 12. Consider the query in*ti*te. What boolean query on a bigram index would be generated for this wildcard query?

Answer: The Boolean query is $i AND in AND ti AND te AND e$.

Question 13. Given that the query is "chennai mathematical institute chennai" and the document is "institute of mathematical sciences chennai", compute the cosine similarity between them assuming a bag of words model.

Answer: $4/\sqrt{5}\sqrt{6} = 1.63$

Question 14. What is the front-code for "foresake", "foreshore" and "foresight" if blocks are of size three (k=3)?

Answer: 8fores*ake4hore4ight

Question 15. Compute the variable byte code for the postings list (100, 228, 484, 996).

Answer:

Data	100	228	484	996
Gaps	100	128	256	512
Sequences	<100>	<1,0>	<2,0>	<4,0>

The encoded VB code is: 11100100 00000001 10000000 00000010 10000000 00000100 10000000

Questions 16 - 20 carry 5 marks each. No negative marks.

Question 16. An IR system returns the following ranked results where R represents a relevant document and N represents a non-relevant document. There are a total of 8 relevant documents in the collection.

```
Ranked Results (Ranks 1 .. 20)
R R R N N N N R R N N N N N R N N N N R
```

What is the MAP score of this system?

Answer: MAP $= \frac{1+1+1+1/2+5/9+2/5+7/20+0}{8} = 0.6$.

Question 17. Suppose that a user's initial query is cheap CDs cheap DVDs cheap CDs. The user examines two documents, d1 and d2. He judges d1 with the content *cheap software cheap CDs* relevant and d2 with content *cheap thrills DVDs* non-relevant. Assume that we are using the raw term frequency (with no scaling and no document frequency). Do not length-normalize the vectors. Use Rocchio relevance feedback as discussed in the class. What would the revised query vector be after applying relevance feedback? Assume $\alpha = 1$, $\beta = 0.7$, $\gamma = 0.3$.

	cheap	CDs	DVDs	software	thrills
q	3	2	1	0	0
d_1	2	1	0	1	0
d_2	1	0	1	0	1

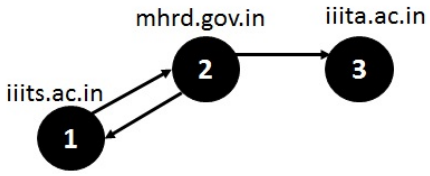
Modified query vector $= \alpha q + \beta d_1 - \gamma d_2 =$ (4.1, 2.7, 0.7,0.7,-0.3)

Since negative values do not exist on our vector space, we make the last dimension zero.

So, we have, (4.1, 2.7, 0.7, 0.7, 0) as the modified query vector after applying Rocchio feedback.

Question 18. Draw the transition probability matrix for a document collection containing HTML pages represented by the graph below. Here, each node represents a page and each edge represents a hyperlink in the source page pointing to the target page. Assume a random surfer model with a teleportation probability $\alpha = 0.4$. Thus, determine the page rank for these pages.

Answer: The Adjacency Matrix, A $= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$



The Transition Probability Matrix, P =

$$\begin{pmatrix} \alpha/3 & \alpha/3+(1-\alpha) & \alpha/3 \\ \alpha/3+(1-\alpha)/2 & \alpha/3 & \alpha/3+(1-\alpha)/2 \\ \alpha/3 & \alpha/3 & \alpha/3 \end{pmatrix}$$

$$= \begin{pmatrix} 2/15 & 11/15 & 2/15 \\ 13/30 & 2/15 & 13/30 \\ 2/15 & 2/15 & 2/15 \end{pmatrix}$$

Question 19. Given the following postings list sizes, what is the best query processing order for (cmi or iit) and (cmu or harvard) and (kellog or sloan)?

Term	**Postings Size**
sloan	220
kellog	870
cmu	107
harvard	700
cmi	466
iit	316

Answer:
cmi or iit adds up to 466 + 316 = 782 postings.
cmu or harvard adds up to 107 + 700 = 807 postins.
kello or sloan adds up to 870 + 220 = 1090 postings.

We process the query terms in the increasing order of their frequency. Therefore, the best order is **cmi or iit** and **cmu or harvard** and **kellog or sloan**.

Interestingly, **kellog or sloan** and **cmi or iit** and **cmu or harvard** would also lead to same number of comparisons and is also a good answer.

Question 20. In a graded retrieval setting, a grade of 0 implies non-relevant, 1 implies poor relevance and so on up to 3 which implies excellent relevance. We design two retrieval systems whose results for the same query are as shown in the table below.

Compute the NDCG for these systems and based on that determine which system is better.

Note: Many of you have taken log with base 10. Hence, this answer is based on log_{10}.

The DCG computation is shown below.

Rank	System 1	System 2
1	3	3
2	0	2
3	1	0
4	0	2
5	1	1
6	3	3

r	log(r +1)	rel	rel/(log(r+1))
1	0.3	3	10
2	0.48	0	0
3	0.60	1	1.66
4	0.69	0	0
5	0.78	1	1.28
6	0.84	3	3.57

r	log(r +1)	rel	rel/(log(r+1))
1	0.3	3	10
2	0.48	2	4.17
3	0.60	0	1.67
4	0.69	2	2.9
5	0.78	1	1.28
6	0.84	3	3.57

For System 1: $\sum_{r=1}^{6} \frac{rel}{log(r+1)} = 16.54$.

For System 2: $\sum_{r=1}^{6} \frac{rel}{log(r+1)} = 23.59$.

Ideal DCG can be computed for a system that outputs (3,3,2,2,1,1) since we don't see more than two documents with relevance better than 3, and so on. But, if only two documents exist, simply taking the $max(DCG_1, DCG_2)$ is also acceptable.

Normalizing the results with max DCG gives, NDCG for system 2 as 1 and NDCG for system 1 as $16.54/23.59 = 0.7$. Clearly, system 2 is better as per NDCG.