# Subject Information Retrieval C3 Date: 26/11/2022 Time: 6.45 pm to 8.15 pm MM: 25

Write down answer of all question in one or maximum two sentences in google form

1. Calculate Retrieval Status Value (RSV) for term\_2 (based on given term-document matrix) using Binary Independence Model for a query Q .(2.5)

Suppose for each term corresponding document value (1 or 0) showed that given term is present or not. Relevant row (second row) show that Documents are relevant or not for query Q

Total number of Docs = 8

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| term/Doc | Doc\_1 | Doc\_2 | Doc\_3 | Doc\_4 | Doc\_5 | Doc\_6 | Doc\_7 | Doc\_8 |
| Relevant | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| term\_1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| term\_2 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| term\_3 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| term\_4 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| term\_5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| term\_6 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |

2. What is main idea behind BIM (Binary Independent Model). (1 marks)

3. For a given matrix A=[[2, 0, 1], [0, 1, 0], [0,0, 0] ] , identify the values of U, ∑ and V'. (1+0.5+1 marks)

4. For a query, number of total relevant Documents are 17, number of total retrieved relevant document are 10 and Total number of retrieved documents are 14, Calculate Matthews Correlation Coefficient(mcc). (1.5 marks)

5. Calculate P(A/B) if P(B/A) = 0.2 P(A) = 0.3 and P(B) = 0.3 using Bayes theorem. (1 marks)

6. For given `Terms-Document matrix` (Each column is corresponding to a Document and each rows are correspond to frequency of terms present in corresponding Documents) find out the pair of two different terms which has largest co-occurrence. (1 marks)

|  |  |  |  |
| --- | --- | --- | --- |
| Term/Document | Docum\_1 | Docum\_2 | Docum\_3 |
| term\_1 | 0 | 0 | 1 |
| term\_2 | 2 | 0 | 1 |
| term\_3 | 0 | 1 | 1 |

7. Apply edit distance (Levenshtein distance) b/n given two words. Write down edit distance matrix of final state. (2 marks)

a. BEEGEGD b. FEABDFD.

8. What is the interpretation of diagonal of co-occurrence matrix. (0.5 marks)

9. Compute tf-idf vector for term\_1 using steps 1: find term\_frequency for documents step 2 normalize terms\_frequency for document to unit length. step 3 find idf using formula log(n/n\_i) step 4: multiply tf and idf. Term-Document graph is given below. (2.5 marks)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term/Document | Docum\_1 | Docum\_2 | Docum\_3 | Docum\_4 | Docum\_5 |
| term\_1 | 3 | 14 | 5 | 1 | 11 |
| term\_2 | 6 | 7 | 2 | 16 | 17 |
| term\_3 | 10 | 17 | 14 | 15 | 3 |
| term\_4 | 1 | 11 | 0 | 9 | 9 |

10. What is use of n-gram overlapping in edit distance algorithm. (1 marks)

11. A fair coin is tossed, What is the a priori probability of landing a head? (0.5 marks)

12. What is different between contextual and global word embeddings. (1 marks)

13. Find Jaccard coefficient for given two sets.

set1 = { F, J, P, W, O, X, C, E, Y, Q, } set2 = { F, J, K, N, W, H, M, V, L, Q, }. (1 marks)

14. Calculate modified query vector using Rocchio Algorithm, based on given vectors. (2 marks)

original query vector =

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 0 | 1 | 0 | 0 |

Average of sum of known relevant Document-vectors =

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 1 | 1 | 0 | 1 |

Average of sum of known irrelevant Document-vectors=

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2 | 1 | 0 | 2 | 2 |

Let alpha = 0.75, bita = 0.70, gama = 0.25

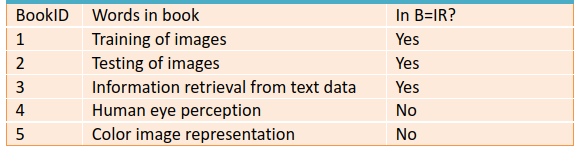
15. Write down all trigram index of word LXLELI . (0.5 marks)

16. Write down one challenge in computing Recall. (0.5 marks)

17. Let there are 6 sets A, B, C, D, E, F length of set A len(A) = 100, len(B)=110, len(C)=120, len(D)=130, len(E)=140, len(F)=150 What will be order of execution for query Q= (A or F) and (B or E) and (C or B). (1 marks)

18. Suppose, you have some set of books in the IIIT Allahabad library, let’s say 5 books are there and among them 3 books are of IR and 2 book are of Computer Vision. Each book has book id (BookID) and some set of words where they belongs to either IR or Computer Vision, please see the below mentioned table for more detail.

Now, from this information compute the prior of P(B) and P(B'). (1.5 marks)



19. What is Thesaurus-based Query Expansion. Write down one example. (1 marks).