Birla Institute of Technology & Science, Pilani
Second Semester 2022-2023
Information Retrieval (CS F469)
Mid-Semester Test

Date	: 13th March 2023	Duration: 90 minutes
Nature of Exam	: Closed Book	Weightage: 30%
Total Marks	: 45 marks	

Instructions:

1. **There are a total of 7 questions. All questions are compulsory.**
2. **Write important intermediate steps in numerical. Directly writing the final correct answer is not sufficient to obtain full marks.**
3. **All questions must be attempted**

Q 1. Suppose the inverted index of an IR system is created using the biword indexes only. How will this system retrieve relevant documents for the following phrasal queries:

i) mercedes benz
ii) Stanford university united states of America

Will it be guaranteed that the retrieved documents for the above queries will have the exact same phrase mentioned? Why or why not? **[2 + 2 = 4 marks]**

Q 2. Consider you have been asked to create an index of all of the Wikipedia pages for your IR assignment using BSBI method on a system with very low RAM. Is it actually possible to do this task? If so, how? If not, why? **[2 marks]**

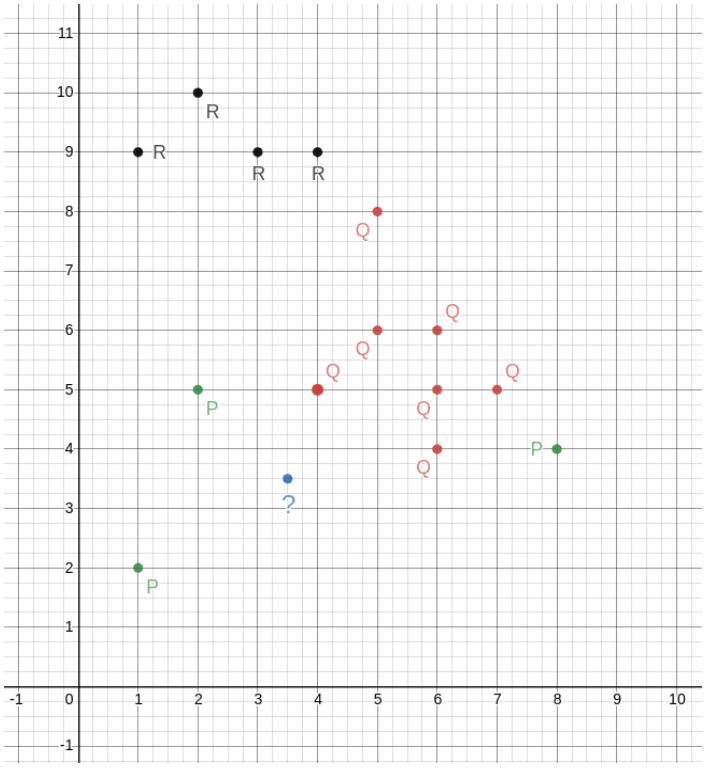
Q 3. Let's assume a search engine is designed such that it can retrieve the top 50 documents from the data collection based on scores computed using a ranking function R(Query, Document). At the user end your user interface is allowed to show only ten results, but can pick any of the top 50 documents retrieved by the search engine. Why would you choose to show the user something other than the top 10 documents from the retrieved document set? State two reasons.
[2 marks]

Q 4. Consider the following documents that have in them the terms or tokens: {M, N, O, P, Q}.

Doc1: (M, N)
Doc 2: (M, Q)
Doc 3: (N, Q, P)
Doc 4: (O, P, Q)

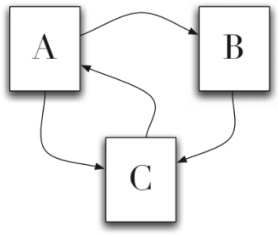
You have the vague query "M OR N AND Q". What are the two interpretations of this query? Give all the documents that are retrieved for both interpretations of the query. Explain your reasoning.
[2 + 3 = 5 marks]

Q5. Suppose documents are represented as two-dimensional vectors. The vector representation and the class of the documents are shown in the figure below:



a) What is the class assigned to the point marked as ? by the Rocchio classification algorithm? **[3 marks]**
b) What is the class assigned by the point marked as ? by the 3-NN classification algorithm? **[3 marks]**

Q 6. Consider a directed graph G given below, where the node represents the web pages, and the edges represent the hyperlinks connecting the web pages. Set up the equations to compute PageRank for G. Assume that the teleportation value is 0.2



a) Compute the link matrix, transition probability matrix(before teleport) and transition probability matrix(after teleport operation). **[1+1.5+1.5=4 Marks]**
b) Assume that the initial PageRank values for all pages are uniformly distributed. Compute the page rank of the graph G using power method (only upto two iterations) before and after teleportation. **[1+4=5 Marks]**

Q 7. a) Construct a positional index with the specified format for the mentioned documents. A positional index captures the position of the terms occurring in the document in the posting lists. For this question, assume that all the pre-processing steps have been applied to the documents. Thus the mentioned documents consist of a list of terms separated by spaces. **[4 marks]**

Positional index format:

word: document: (position, position, …); document: (position, …);
word: document: (position, position, …); document: (position, …);
:

The documents are:

Doc 1: The furious storm pounded the seaside town.
Doc 2: Her voice was a gentle whisper in the dark. The sweet voice was bright.
Doc 3: The bright sun shone on the blooming flowers in a bright way .
Doc 4: The raging fire consumed everything in its path in the town.

b.) Which of the documents will match the phrasal query: "**bright voice in the fire**", and what will be the position of the matches based on positional index? **[2 mark]**

c) What is the term frequency (natural term frequency without logarithm) of the term "**the**" in Doc 1 and the term "**bright**" in Doc 3? **[2 mark]**

d) What is the inverse document frequency (idf) (with logarithm base 10) for the query terms "**bright**" and "**voice**"?

[2 mark]

e) Convert the documents into vector space representation, the terms act as a axis, and the magnitude of that axis is the tf-idf score (product of natural term frequency and idf). **[3 mark]**

f) Compute the similarity of the query "**bright voice in the fire**" with Doc 2 and Doc 3.

[4 mark]

For this similarity computation assume the followings:

i) Use the **nnc.ntc** scoring criteria. Here it means that for the document use, natural term frequency (i.e. no logarithm), no document frequency and cosine normalization. And, for the query use natural term frequency (i.e. no logarithm), idf weighting and cosine normalization.

ii) A document and query will be represented as a vector, where the axis are unique terms of the vocabulary.

iii) There are only these three documents in the corpus.

iv) Use the log to the base 10, if a logarithm is used in the calculations.

v) If floating-point numbers are involved, round the number up to two decimal places.

**********END**********