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

Date	: 6th May 2023	Duration: 3 hours
Nature of Exam	: Open Book	Weightage: 35%
Total Marks	: 65	

Instructions:
1. Write important intermediate steps in numerical. Directly writing the final correct answer is not sufficient to obtain full marks.
2. Attempt the questions in the answersheet in the given order only.

Q1: **a) [3 marks]** Suppose, you are designing a recommendation system for an online bookstore that has been launched recently. The bookstore has over 1 million book titles, but its rating database has only 20,000 ratings. Which of the following would be a better recommendation system? a) User-user collaborative filtering b) Item-item collaborative filtering c) Content-based recommendation.
In One sentence, justify your answer.

(b) [3 marks] Suppose the bookstore is using the recommendation system you suggested above. A customer has only rated two books: ”Convex Optimization” and ”Compiler design” and both ratings are 5 out of 5 stars. Which of the following books is highly likely to be recommended?Explain why? **a) ”Computer programming” b) ”A Tale of Two Cities” c) ”Two states” d) It depends on other users’ ratings.**

(c) [3 marks] After some years, the bookstore has attracted enough ratings after it started using a more advanced recommendation system. Suppose the mean rating of books is 3.4 stars. Alice, a faithful customer, has rated 350 books and her average rating is 0.4 stars higher than average users’ ratings. Animals Farm, is a book title in the bookstore with 250,000 ratings whose average rating is 0.7 higher than global average. What would be a baseline estimate of Alice’s rating for Animals Farms?

Q2. [6 marks] In a collection of 5,000 documents, 1,000 are relevant to a specific query. A probabilistic information retrieval system retrieves 500 documents, of which 300 are relevant. Calculate the system's precision, recall, F1-score, and accuracy.

Q3. [8+4 = 12 marks] Use the similarity matrix given below and perform a single and complete link hierarchical clustering algorithm to build the hierarchical clustering dendrogram.

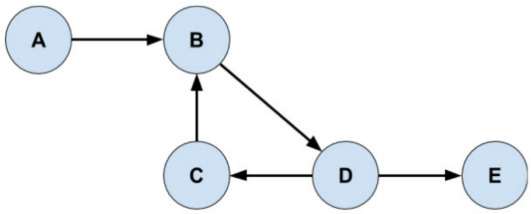
- **a)** After merging the cluster clearly shows the updated proximity matrix corresponding to each iteration of the algorithm.
- **b)** Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Q4. [3 marks] In a collection of 10,000 documents, 3,000 contain the term "apple," 4,000 contain the term "banana," and 1,500 contain both terms. If you perform a Boolean search for documents containing either "apple" or "banana," how many documents would you retrieve?

Q5. [10 Marks] Consider the graph given below, where the node represents the web pages, and the edges represent the hyperlinks connecting the web pages.

a) Compute link and transition matrices (before and after teleport) **[6 marks].**
b) Compute the page rank for all the pages in the graph using the power method (only do three iterations). The initial PageRank should be initialized with a uniform probability distribution. Assume that the PageRank teleport probability is 0.5, and all edges shown in the graph are equally important. Show every step in detail. **[4 marks]**

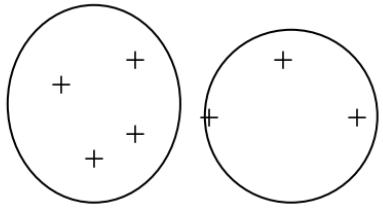


Q6. [3 marks] In a dictionary-based query translation for cross-language information retrieval, you have a query with 4 terms, and the bilingual dictionary contains translations for 3 of those terms. If each term has 2 possible translations, how many translated queries will be generated using all possible combinations?

Q 7. [4 marks] Consider the one dimensional dataset shown below: Classify the data point **x=5.0** according to its 1, 3, 5, and 9 nearest neighbors approach. Clearly justify your classification decisions.

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

Q8. [4 marks] Are the two clusters shown below well separated? Yes or No? Justify your answer in one or two sentences.



Q9. [4 marks] Generally association rules with high confidence are considered as more interesting. However, often we will not be interested in association rules that have a confidence of 100%. Why? Then specifically explain why association rules with 99% confidence may be interesting (i.e., what might they indicate)?

Q 10. [10 marks] In a text corpus, you have a collection of 15,000 documents in three different languages: ***English, Spanish***, and ***French***. There are 7,000 documents in English, 5,000 in Spanish, and 3,000 in French. You are developing a cross-language information retrieval (CLIR) system with a multimedia component that supports searching based on both textual and visual features.

For the textual component, you use the tf-idf weighting scheme. In the English documents, the term "computer" appears in 600 documents, the term "ordenador" (Spanish translation of "computer") appears in 450 Spanish documents, and the term "ordinateur" (French translation of "computer") appears in 300 French documents.

For the visual component, you use color histograms as features to compare images. You have an image with the following color histogram: [0.3, 0.1, 0.2, 0.4], and another image with the color histogram: [0.2, 0.3, 0.1, 0.4].

Compute the following:

a) **[3 marks]** Calculate the inverse document frequency (IDF) values for the term "computer" and its translations in the Spanish and French documents. Use the base 10 logarithm.
b) **[2 marks]** Suppose you have a query containing the term "computer" and an image with the color histogram: [0.25, 0.25, 0.25, 0.25]. Calculate the Euclidean distance between the query image's color histogram and the two given images' color histograms.
c) **[3 marks]** For a specific document, the term frequencies for "computer" and its translations are 4, 3, and 2, respectively. Calculate the tf-idf weighting for the term "computer" and its translations in that document, considering the calculated IDF values.
d) **[2 marks]** Based on the calculated Euclidean distances and tf-idf weightings, suggest a way to combine these two scores to rank documents for a multimedia query containing both the term "computer" and the image with the color histogram: [0.25, 0.25, 0.25, 0.25].