Information Retrieval (CS60092)
Computer Science and Engineering, Indian Institute of Technology Kharagpur

Supplementary Examination

Time: 3 hours
Full Marks: 60

Attempt all questions.
Use of calculator is allowed.

---

1. Consider the following documents.

   D1: Programmer builds software

   D2: Good software has fewer bugs

   D3: Full time software programmer

   D4: Remove bugs software full version

   (a) Find the representations of the above documents in
      i. Boolean model
      ii. Term frequency model

   (b) Consider the query "Software full of bugs". Use tf-idf scores to find the rank order of the above documents for this query. **[(4+4)+8=16]**

2. Let the relevance labels of the first 10 documents for a query be 1, 0, 0, 1, 1, 0, 0, 0, 1, 0. Here, 1 means relevant and 0 means non-relevant. Suppose total number of relevant documents for this query is 5.

   (a) Draw the P-R curve for this resultset.
   (b) Draw the interpolated precision curve.
   (c) What is the average precision for this resultset? **[8+4+4=16]**

3) Consider the following documents:

   D1: English Channel Atlantic

   D2: National Geography Channel English

   D3: Doordarshan national English news

   Using unigram language model, rank the above documents for the query "national news channel English". To compute the model probabilities, combine MLE estimates from documents and the collection giving equal importance to both. **[15]**

4) Suppose we use the notation [a1, a2] to denote the representation of a document in a 2-dimensional vector space of terms. Consider a collection that has 5 documents. For each document, its vector space representation and classification is given in the following table.

Document	Representation	Class
D1	[10, 3]	Technical
D2	[8, 2]	Technical
D3	[4, 0]	Technical
D4	[4, 2]	Non-technical
D5	[2, 1]	Non-technical

Consider another document q whose representation is given as q=[3, 3]. Predict the class (Technical or Non-technical) of q using 1-NN, 2-NN and 3-NN classification. Here k-NN means k Nearest Neighbors approach. Justify your answers. **[8]**

5) Create an index for the collection containing the documents below. Do not index the terms containing less than 4 letters.

D1: Data Definition Language
D2: C programming language
D3: Good programming practice
D4: Programming and Data structures assignments for practice **[5]**

6) Consider the following matrix representing **distance** between five documents:

Document	A	B	C	D	E
A	**0**	**10**	**5**	**11**	**7**
B	**10**	**0**	**12**	**9**	**6**
C	**5**	**12**	**0**	**17**	**13**
D	**11**	**9**	**17**	**0**	**4**
E	**7**	**6**	**13**	**4**	**0**

Compute hierarchical single-linkage clustering of these five documents. Clearly show the matrices at each step of building the dendrogram.
(No marks will be given for showing only the final dendrogram)

[10]