Nirma University
Institute of Technology
Class Test, February 2021
BTech in Computer Science & Engineering
2CSDE53 Information Retrieval Systems

Roll No. _____ Supervisor's Signature with Date : _____

Time:75 minutes Max. Marks: 35

Instructions:
1. Attempt all questions.
2. Figures to the right indicate full marks.
3. State and make necessary assumptions wherever necessary.
4. Draw neat sketches wherever required.

1. **Answer the following:** **[15]**

   (a) Can the $tf - idf$ value of a term in a document exceed 1? Discuss. [3]

   (b) "The term appearing in a large number of documents in the collection, is probably not significant or discriminative."Give your remarks on this statement with appropriate justification. [3]

   (c) Which of the following pairs of vector is the most similar to each other? Use cosine similarity measure. [4]
   - v_1 (1,0,0,2)
   - v_2 (-1,0,-1,1)
   - v_3 (0,1,1,1)

   (d) What issues of boolean representation are addressed using TF-IDF based document representation? Describe with appropriate example. [5]

2. **Answer the following:** **[8]**

   (a) For the following corpus, do as directed: [8]

   Doc 1: playing Cricket.
   Doc 2: played flute.
   Doc 3: flying Cricket at night.
   Doc 4: watched a Cricket while playing Cricket.

   1. (1 mark) Apply text-preprocessing on this corpus.
   2. (1 mark) Extract and display the list of vocabulary terms.
   3. (4 marks) Represent each document using TF-IDF model and show necessary calculation.
   4. (2 marks) For a given query "playing and watching Cricket, determine the ranking of all documents retrieved from the system.

3. **Answer the following:** **[12]**

   (a) Assume a corpus of size 25. A query is passed through three search engines S1, S2 and S3. Following are the ranked responses by these search systems. Relevant results are marked by 'R' and irrelevant documents are marked by '–'. From ground truth, it is known that there are 8 relevant documents in the corpus corresponding to this query. [12]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
S1	R	R	--	R	--	--	--	R	R	--	--	--	--	--	--
S2	R	--	--	R	R	--	R	--	--	R	--	R	R	--	--
S3	R	R	R	--	--	R	--	--	--	R	--	--	--	R	--

1. (6 marks) Compute precision and recall at each rank position for all the search engines.
2. (3 marks) Calculate average precision for all search engines.
3. (3 marks) Comment on the performance of all the search engines.