National Institute of Technology Hamirpur (H.P.)

Computer Science and Engineering
End Semester Examination

Branch/ Semester: M.Tech (AI, 1st Year) **Semester** 1st
Subject Code: CS-728 **Duration: 120** Minutes

Subject Name: Information Retrieval **Max. Marks:** 50

Date: 15/12/20 **Time**: 3.00 PM to 5.00 PM

Note: All questions are compulsory.

1. (a) What is the soundex code for the following two names, Robert and Reupert? Assume that the alphabets are mapped to numbers as follows
(B, F, P, V→1), (C, J, K, Q, S, X, Z→2), (D, T→3), (L→4), (M, N→5) and (R→6)

(b) Suppose a program for recognizing dogs in scenes from a video identifies 9 dogs in a scene containing 11 dogs and some cats. If 4 of the identifications are correct, but 5 are actually cats, then compute the precision and recall of the program.

[3+3=6]

2. Consider the table below showing how two users rated the relevance of a set of 12 documents to a particular information need (0 = non-relevant and 1= relevant). Assume that you have developed an IR system that for this query returns the set of documents (4,5,6,7,8).

Doc-id	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
User-1	0	0	1	1	1	1	1	1	1	0	0	0
User-2	0	0	1	1	0	0	0	0	1	1	1	1

(a) Calculate the precision, recall and F-measure of your system if a document is considered relevant when both users are agreed.

(b) Calculate the precision, recall and F-measure of your system if a document is considered relevant when either user it is relevant.

[4+4=8]

3. Consider the following documents
Doc1: catholic church in Brisbane
Doc2: garden city church Brisbane
Doc3: brisbane courier garden city
Doc4: where in brisbane catholic

(a) Draw a term-document incidence matrix for this document collection.

(b) Draw the positional inverted index representation for this document.

[5+5=10]

4. Assume the simple term frequency weights are used(with no IDF factor), and the stop words “is”, “am” and “are” are removed. Compute the cosine similarity of the following two documents.[Show term frequency matrix].
Doc1: “Precision is very very high”
Doc2: “High precision is very very important”

[5]

5. Consider a collection made of 800 documents and the number of unique words is estimated to 800. The following things are required for dictionary storage assuming that all the terms are stored as s string: 4 bytes per term per frequency, 4 bytes term pointer to postings, 3 bytes for term pointer and average 8 bytes for term in term string. Estimate the space usage for dictionary without blocking and with block size of K=8. **[8]**

6. (a) What is the posting list that can be decoded from from the following *variable byte-code?*

10001001 00000001 10000010 11111111

(b) What would be the encoding of the same posting list using a gamma-code?

[4+3=7]

7. What is the requirement of crawling? Explain the concept of URL frontier with example.

[6]