Indian Institute of Technology Jammu

Computer Science & Engineering

Class Test 2

CSC004P1M: Data Organization and Retrieval

Max. Marks: 15 **Time: 45 + 15 = 60 minutes**

Note: It is closed book / notes examination

Rough sheets also need to be submitted.

Answer all questions

Q1. [Marks 2] Compute vector space similarity between query "smart tortoise" with document 1 "tortoise with smart brains" and document 2 "lazy tortoise". Given idf values of "smart" and "tortoise" as 1.65 and 2 respectively. The total scores of the query for documents 1 and 2 are:

a. 1.65 and 2 respectively
b. 3.65 and 2 respectively
c. 1.65 and 3 respectively
d. 2 and 2 respectively

Q2. [Marks 2] Consider the following documents:

D1: Jack and Jill went to hill

D2: Jack and Jill stumbled on a rock and Jill hurt herself

Consider the vocabulary vector (with stop words removed) as:

V=[herself, hill, hurt, Jack, Jill, rock, stumbled, went]

Assume a binary text representation for the query and documents where a vector's value for a particular dimension equals 1 if the term appears at least once and 0 otherwise.

Given a query "Jill is hurt". Using inner product as the similarity measure, the similarity score of query with documents D1 and D2 are:

a. 1 and 1 respectively
b. 1 and 2 respectively
c. 2 and 2 respectively
d. 2 and 1 respectively

Q3. [Marks 2] Let g(d) be the query independent score for a document d. While referring to postings in an index, we go in which order

a. Ascending order of g(d)
a. Descending order of g(d)
b. Order of g(d) does not matter
c. None of the above

Q4. [Marks 1] Model of information retrieval where we do not use relevance feedback is called

a. Boolean retrieval
b. Ad Hoc retrieval
c. Ranked retrieval
d. Proximity retrieval

Q5. [Marks 2] Match the following

I. Champion list
II. Impact ordering
III. Cluster pruning
IV. Index elimination

1. Consider documents with multiple query terms
2. Compute similarity with leaders
3. Identify top documents with high term frequency
4. Order documents in posting list of a term by decreasing order of term frequency

a. I-1, II-3, III-4, IV-2
b. I-2, II-3, III-4, IV-1
c. I-3, II-1, III-4, IV-2
d. I-2, II-4, III-2, IV-1

Q6. [Marks 3] Consider two IR systems whose top five retrieved documents are given as:

IR1:

IR2:

In the figures, the darker regions indicate that the document is relevant. Using average precision, determine which retrieval system is better at k=5.

a. IR1
b. IR2
c. Cannot determine
d. Both are equal

Q7. [Marks 1] In which of the following situations the relevance feedback based model does not give satisfactory results?

a. User does not give negative feedback
b. Term weights are negative
c. Mismatch between user's need and query
d. All of the above

Q8. [Marks 2] A thesaurus aids in

a. Objective of removing affixes and allowing the retrieval of document containing syntactical variations of query term
b. Objective of filtering out words with very low descrimination value for retrieval purpose
c. Objective of treating digits, hyphens, punctuation marks, and the case of letters
d. Method of allowing the expansion of original query with related term