Indian Institute of Technology Jammu

Computer Science & Engineering

Mid-Semester Examination

CSC004P1M: Data Organization and Retrieval

Max. Marks: 30 **Time: 45 + 15 = 60 minutes**

Note: It is closed book / notes examination

Answer all questions

Q1. [Marks 5] Given the following queries:

a. Betty AND Cake
b. Betty OR Cake

For which of the queries above, skip lists will be useful - give appropriate reasons.

Q2. [Marks 5] Recommend a query processing order for the following query:

(Tangerine OR Trees) AND (Orange OR Mango) AND (Fruits OR Vegetables)

Given the following postings list sizes:

Term: Postings size - as follows:

Vegetables: 213K, Fruits: 87K, Orange: 107K, Mango: 271K, Tangerine: 46K, Trees: 316K

Also, give the reason for choosing a particular query processing order. Give any (if) modifications you would propose in a basic inverted index with (term_id, doc_id) pairs to support your answer.

Q3. [Marks 7] We have two word query for which the postings list are as given below:

[2,4,5,11,14,17,19,23,28,31,34,67,89,91,145,167,180]

and

[67]

Give the execution snapshot for the algorithm that uses skip lists for computing the intersection of two postings lists and work out how many comparisons would be required. Assume the skip length to be 3. [No algorithm is required]

Q4. [Marks 4] Give map function of map-reduce algorithm that builds a biword index. The input to the algorithm is a set of documents.

Q5. [Marks 9] Given the following document collection:

Doc 1 --- Betty bought a butterscotch-cake

Doc 2 --- The cake was very bitter

Doc 3 --- Betty returned the cake ‘cos the cake was bitter

Doc 4 --- Betty got a new cake and cont’d the party

Draw an inverted index (term_id, doc_id) that considers the following:

a. Normalization
b. Stop words
c. Stemming

Give necessary assumptions about normalization forms, stop list, and stemming rules adopted in support of your answer.