National University of Computer and Emerging Sciences, Lahore Campus

Course:	Information Retrieval and Text Mining	Course Code:	CS567
Program:	MS(Computer Science)	Semester:	Fall 2018
Duration:	85 Minutes	Total Marks:	25
Paper Date:	1-Nov-18	Weight	20%
Section:	CS	Page(s):	2
Exam:	Midterm	Roll No:	

Q3) Suppose we have the following documents:

Document	Words
D1	a b b a b b c
D2	a a b a b a
D3	b b b b b b c c

How will a basic inverted index for this corpus look like?

i. a => D1 -> D2 -> D3;
b => D1 -> D2 -> D3;
c => D1 -> D2 -> D3

ii. a => D1 -> D1 -> D2 -> D2 -> D2 -> D2;
b => D1 -> D1 -> D1 -> D1 -> D2 -> D2 -> D3 -> D3 -> D3 -> D3 -> D3 -> D3;
c => D1 -> D3 -> D3

iii. a => D1 -> D2;
b => D1 -> D2 -> D3;
c => D1 -> D3

iv. a => D1:1,4 -> D2:1,2,4,6;
b => D1:2,3,5,6 -> D2:3,5 -> D3:1,2,3,4,5,6;
c => D1:7 -> D3:7,8.

Solution: Option 4 is correct

Q3) If I use simple TFxIDF for ranking, which will be the order generated for the regular query (a b)?

i. D1, D2, D3
ii. D2, D1, D3
iii. D1, D3, D2
iv. D2, D3, D1.

Solution

Idf of a = lg 3/2(0.58) , b = 0, c = lg 3/2(0.58)

TF*IDF of D1 = 1.3*0.58 + (log 4)
TF*IDF of D2 = 1.6*0.58 + log (2)
TF*IDF of D3 = 0
Option 2 is correct

Q3) If my main concern is producing the fastest possibly dictionary search, which will be the best possible data structure to use?
a) An array of words and word IDs over a contiguous chunk of memory, because this would allow very fast sequential passes through all dictionary words.
b) A sorted linked list with skip pointers, because this would allow both very fast sequential passes and very fast look-ups.
c) A hash map, because this would allow very fast look-ups.

Solution
Option c

Q3) Apply the SPIMI algorithm to the following collection:
d1: bsbi use term id
d2: sort term id doc id
d3: spimi use term
d4: no term id sort

Assume that main memory can only hold two documents at a time, i.e., the SPIMI algorithm will write to disk each time after two documents, a block, have been processed.
Write out the content of each block just before merging and the result after merging in the following format:
Block 1:
bsbi → 1
...
term → 1, 2

Solution

Block 1:
bsbi → 1
doc → 2
id → 1, 2
sort → 2
term → 1, 2
use → 1

Block 2:
id → 4
no → 4
spimi → 3

sort→ 4
term → 3, 4
use → 3

After Merge
bsbi → 1
doc → 2

id → 1, 2,4
no → 4
spimi → 3
sort → 2, 4
term → 1, 2, 3 , 4
use → 1, 3

Q3) What is the largest number that can be stored in 4 bytes using unary encoding?

Solution
4 bytes = 32 bits
Largest number is 31

Q3) Assume that postings lists are delta gap encoded using Elias Gamma codes. Using this encoding, suppose that the postings list for the term information is the bit sequence:
1111 1111 1011 1100 1101 0011 1110 0000 0

and the postings list for the term retrieval is the bit sequence:
1111 1111 1100 0000 0011 1110 0111 1101 111

What docids match the following boolean query:
information AND NOT retrieval

Solution
postings list for the term information

1111 1111 10 1111100110 10 10 111110 100000
1990 2 64
1990 1992 2056

postings list for the term retrieval

1111 1111 110 10000000111 110 101 11110 11111
2051 5 31
2051 2056 2087

Answer = 1990 and 1992

Q3) Is smoothing more important for long or short queries? Justify your answer.

Solution

It is more important for long queries because long queries will have larger number of words and there are are chances that a relevant document might not have all query words. There are more chances of missing query words in documents.

Q3) Give 2 advantages of using average precision as compared to precision@5. Justify your answer with some example of rank lists.

Solution:

AP is system centric, P@5 is user centric

List 1
R
R
N
N
N

List 2
N
N
N
R
R

List 1 has better quality s compared to list 2.
AP is higher for list 1, while P@5 is same for both lists

Q3) Computer average precision of following list of documents. Leftmost document is at top. Total relevant documents for this query is 15.

R N N R N R N N

Solution:

(1 + 2/4 + 3/6)/15
= 0.13

Q3) Consider the following hypothetical information retrieval scenario. Suppose it has been found at a hospital that due to equipment malfunction, the results of blood tests taken on 2016-12-01 are unreliable for diabetic patients. The hospital would like to contact all diabetic patients who had any kind of blood test on that day, to repeat the test. The hospital uses an information retrieval system to identify these patients. Suppose the collection of patients' medical records contains 10000 documents, 150 of which are relevant to the above query. The system returns 250 documents, 125 of which are relevant to the query.

For the given scenario, which measure do you think is more important, precision or recall? Why?

Solution:
Recall is more important because if a patient's diabetic test is unreliable and he has diabetes but the system did not retrieve his record then the person may not know that he is suffering from diabetes.