# INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
## Mid Autumn Semester Examination 2023-24

**Date of Examination:** 22/09/23 **Session:** AN **Duration** 2 hrs, **Marks** = 50

**Sub No:** CS60092 **Sub Name:** Information Retrieval

**Department/Centre/School :** Computer Science and Engineering

**Specific charts, graph paper, log book, etc. required** NO

**Special Instructions:** ANSWER ALL questions. All parts of a single question should be answered together. Answers should be brief and to-the-point. Marks will be deducted for sketchy answers and claims without proper reasoning. In case of reasonable doubt, make assumptions and state them upfront. You can keep probability values in fractional forms.

1. Consider the two postings list A and B in Figure 1. Intersect the two list and answer the following questions.

   (a) How many times is a skip pointer followed during intersection?

   (b) How many comparisons will be made to perform this intersection? List them.

   (c) How many comparisons would have been made if there were no skip pointers in list A?

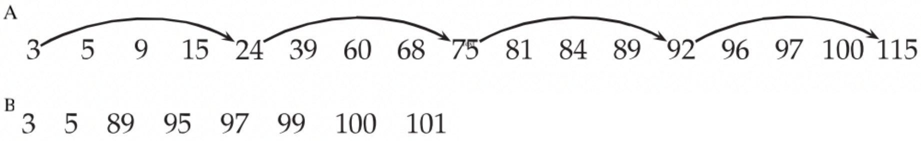
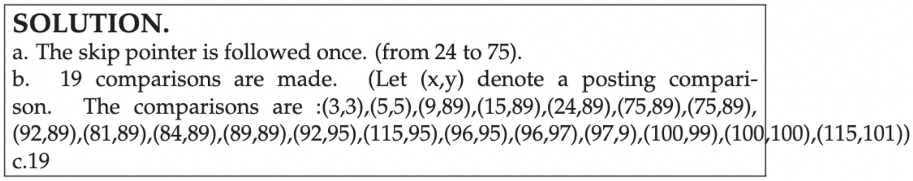


Figure 1: Postings list.

**[1 + 3 + 1 = 5]**

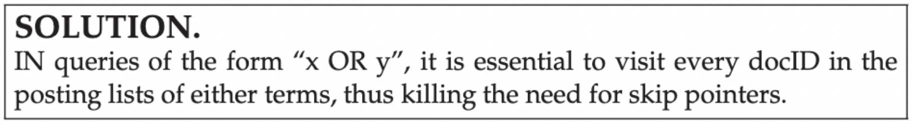


**SOLUTION.**
a. The skip pointer is followed once. (from 24 to 75).
b. 19 comparisons are made. (Let (x,y) denote a posting comparison. The comparisons are :(3,3),(5,5),(9,89),(15,89),(24,89),(75,89),(75,89),(92,89),(81,89),(84,89),(89,89),(92,95),(115,95),(96,95),(96,97),(97,9),(100,99),(100,100),(115,101))
c.19

Figure 2: Answer 1

2. Are skip pointers useful for queries of the form $x$ `OR` $y$? Justify. **[2]**

3. Consider the following fragment of a positional index with the format:
   word: document: <position, position, . . .>; document: <position, . . .>
   . . .



**SOLUTION.**
IN queries of the form "x OR y", it is essential to visit every docID in the posting lists of either terms, thus killing the need for skip pointers.

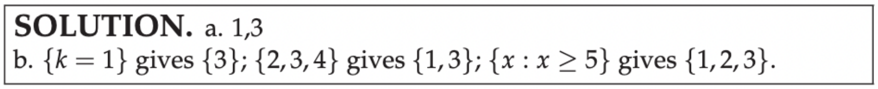
Figure 3: Answer 2

Gates: 1: <3>; 2: <6>; 3: <2, 17>; 4: <1>;
IBM: 4: <3>; 7: <14>;
Microsoft: 1: <1>; 2: <1, 21>; 3: <3>; 5: <16, 22, 51>;
The $/k$ operator, word1 $/k$ word2 finds occurrences of word1 within $k$ words of word2 (on either side), where $k$ is a positive integer argument. Thus $k = 1$ demands that word1 be adjacent to word2.

(a) Describe the set of documents that satisfy the query Gates /2 Microsoft.

(b) Describe each set of values for $k$ for which the query Gates $/k$ Microsoft returns a different set of documents as the answer. Recall that $\{x\}$ and $\{x, y\}$ are different sets.

**[2 + 4 = 6]**



**SOLUTION.** a. 1,3
b. $\{k = 1\}$ gives $\{3\}$; $\{2, 3, 4\}$ gives $\{1, 3\}$; $\{x : x \geq 5\}$ gives $\{1, 2, 3\}$.

Figure 4: Answer 3

4. Consider two IR systems S1 and S2 that produced the outputs show in Figure 5 for the 4 reference queries q1, q2, q3, q4. Here dXX refer to the documents that do not appear in the referential (i.e., relevant documents for a given query).

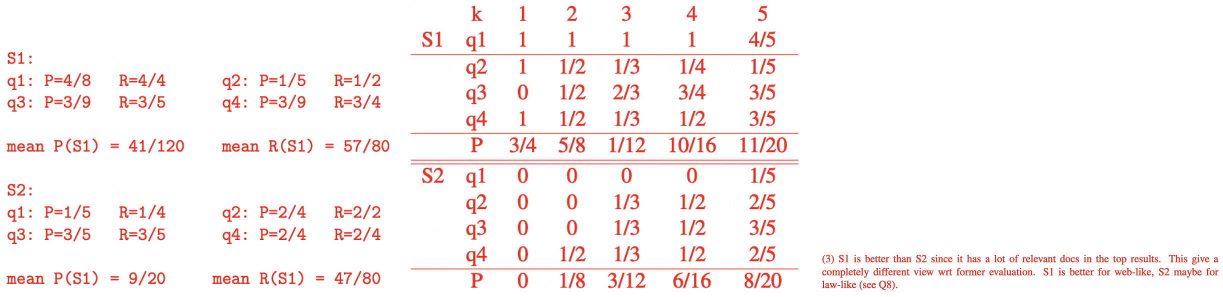
   (a) For each of the two systems, compute the precision and recall for each query (provide the results as fractions). Finally average the precision and recall for all the queries and compare S1 and S2 based on precision. Explain all the steps of your computation.

   (b) Compute the average P@k values for k between 1 and 5 for the IR systems S1 and S2 above. Based on these results, what is your relative evaluation of the two systems? How does it compare to your previous observation based on binary precision. **[3 + 1 + 4 + 1 = 9]**

```
S1:                                       | referential:
 q1: d01 d02 d03 d04 dXX dXX dXX dXX      |  q1: d01 d02 d03 d04
 q2: d06 dXX dXX dXX dXX                  |  q2: d05 d06
 q3: dXX d07 d09 d11 dXX dXX dXX dXX dXX  |  q3: d07 d08 d09 d10 d11
 q4: d12 dXX dXX d14 d15 dXX dXX dXX dXX  |  q4: d12 d13 d14 d15

S2::                                      | referential:
 q1: dXX dXX dXX dXX d04                  |  q1: d01 d02 d03 d04
 q2: dXX dXX d05 d06                      |  q2: d05 d06
 q3: dXX dXX d07 d08 d09                  |  q3: d07 d08 d09 d10 d11
 q4: dXX d13 dXX d15                      |  q4: d12 d13 d14 d15
```

Figure 5: Results from system S1 and S2.



```
S1:
q1: P=4/8    R=4/4       q2: P=1/5    R=1/2
q3: P=3/9    R=3/5       q4: P=3/9    R=3/4

mean P(S1) = 41/120      mean R(S1) = 57/80

S2:
q1: P=1/5    R=1/4       q2: P=2/4    R=2/2
q3: P=3/5    R=3/5       q4: P=2/4    R=2/4

mean P(S1) = 9/20       mean R(S1) = 47/80
```

|  | k | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| S1 | q1 | 1 | 1 | 1 | 1 | 4/5 |
|  | q2 | 1 | 1/2 | 1/3 | 1/4 | 1/5 |
|  | q3 | 0 | 1/2 | 2/3 | 3/4 | 3/5 |
|  | q4 | 1 | 1/2 | 1/3 | 1/2 | 3/5 |
|  | P | 3/4 | 5/8 | 1/12 | 10/16 | 11/20 |
| S2 | q1 | 0 | 0 | 0 | 0 | 1/5 |
|  | q2 | 0 | 0 | 1/3 | 1/2 | 2/5 |
|  | q3 | 0 | 0 | 1/3 | 1/2 | 3/5 |
|  | q4 | 0 | 1/2 | 1/3 | 1/2 | 2/5 |
|  | P | 0 | 1/8 | 3/12 | 6/16 | 8/20 |

(3) S1 is better than S2 since it has a lot of relevant docs in the top results. This give a completely different view wrt former evaluation. S1 is better for web-like, S2 maybe for law-like (see Q8).

Figure 6: Answer 4

5. List three situations where an IR system's MAP and MRR performances will be equal. **[3]**

- When there is no relevant document
- Where there is only one relevant document
- When all the documents are relevant and they have been returned.

Figure 7: Answer 5