Information Retrieval Tutorial - Set 1
09.09.2022

1. A. Draw the term-document incidence matrix and the inverted index representation for the following document collection:
   - Doc 1 : breakthrough drug for schizophrenia
   - Doc 2 : new schizophrenia drug
   - Doc 3 : new approach for treatment of schizophrenia
   - Doc 4 : new hopes for schizophrenia patients

   B. What are the returned results for these queries-
   - schizophrenia AND drug
   - for AND NOT(drug OR approach)

2. What is its time complexity of the postings merge algorithm to arbitrary Boolean query formulas? For instance, consider:
   (Brutus OR Caesar) AND NOT (Antony OR Cleopatra)

3. Write out a postings merge algorithm that evaluates this query efficiently-
   x AND NOT y

4. We have a two word query. For one term the postings list consist of the following 16 entries.

   [2, 4, 9, 12, 14, 16, 18, 20, 24, 32, 47, 81, 120, 125, 158, 180]

   and for the other list it is the one entry postings list
   [81]

   Work out how many comparisons would be done to intersect the two postings list with the following two strategies.

i. Using standard postings list.
ii. Using postings list stored with skip pointers, with the suggested skip length of √P (P=length of the list).

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

   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.

   Describe the set of documents that satisfy the query -> Gates /k Microsoft for k=1, 2

6. If |S| denotes the length of string S, show that the edit distance between s1 and s2 is never more than max{|s1|, |s2|}.

7. Compute the edit distance between paris and alice. Write down the 5 × 5 array of distances between all prefixes.