Assignment -1

1. Consider these documents:

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

a. Draw the term-document incidence matrix for this document collection.

b. Draw the inverted index representation for this collection

2. Recommend a query processing order for

(tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes)

given the following postings list sizes:

Term	Postings Size
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

3. For a conjunctive query, is processing postings lists in order of size guaranteed to be optimal? Explain why / why not.

4. Extend the postings merge algorithm to arbitrary Boolean query formulas. What is its time complexity? For instance, consider:

(Brutus OR Caesar) AND NOT (Antony OR Cleopatra)

Can we always merge in linear time? Linear in what? Can we do better than this?

5. For the Porter stemmer rule group:

a. What is the purpose of including an identity rule such as SS →SS?

b. Applying just this rule group, what will the following words be stemmed to?

circus canaries boss

c. What rule should be added to correctly stem pony?

d. The stemming for ponies and pony might seem strange. Does it have a deleterious effect on retrieval? Why or why not?

6. Why are skip pointers not useful for queries of the form x OR y?

7. 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

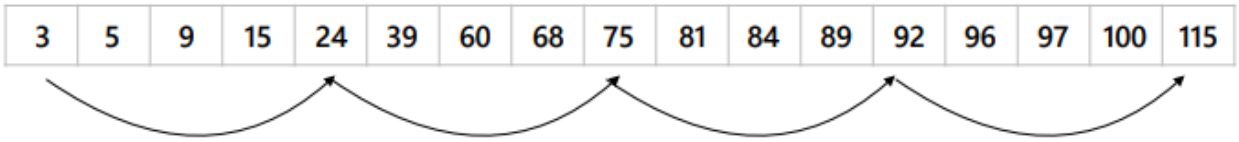
[81]

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.

8. Consider a postings intersection between this postings list, with skip pointers:



And the following intermediate result postings list (which has has no skip pointers):

3	5	89	95	97	99	100	101

Trace through the posting's intersection algorithm.

A. How often is a skip pointer followed (i.e., p1 is advanced to skip(p1))?

B. How many postings comparisons will be made by this algorithm while intersecting the two lists?

C. How many postings comparisons would be made if the postings lists are intersected without the use of skip pointers?

9. How is the inverted index used for the document retrieval and how this inverted index updated with new documents?

10. How are positional indexes different from traditional inverted indexes and what are the benefits of using positional indexes?