Due: 2023/2/24 12:00

Homework 1

Student Number:
Name:

Problem 1. (30 points)
Consider these documents:

Doc1 breakthrough drug for schizophrenia

Doc2 new schizophrenia drug

Doc3 new approach for treatment of schizophrenia

Doc4 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.

c. For the document collection, what are the returned results for these queries:

i schizophrenia AND drug
ii for AND NOT (drug OR approach)

Problem 2. (20 points) For a conjunctive query, is processing postings lists in order of size guaranteed to be optimal? Explain why it is, or give an example where it isn't.

Problem 3. (20 points) The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue shouldn't be conflated. Give your reasoning.

a. abandon/abandonment

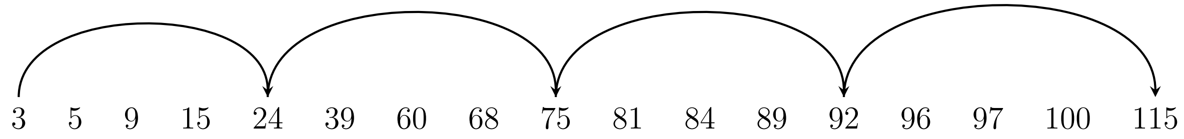
b. absorbency/absorbent

c. marketing/markets

d. university/universe

e. volume/volumes

Problem 4. (30 points) Consider a postings intersection between this postings list, with skip pointers:



and the following intermediate result postings list (which hence has no skip pointers):
3 5 89 95 97 99 100 101
Trace through the postings intersection algorithm(pdf of lecture 1, under section Skip Pointers)

a. How often is a skip pointer followed?

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?