Due: 2021/3/5

Homework 1

Student Number:
Name:

Problem 1. (30 points)

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

Consider the documents above,

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. (30 points) Are the following statements true or false?

a. In a Boolean retrieval system, stemming never lowers precision.

b. In a Boolean retrieval system, stemming never lowers recall.

c. Stemming increases the size of the vocabulary.

d. Stemming should be invoked at indexing time but not while processing a query.

Problem 3. (10 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 4. (30 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