ISTA 456/556: Asignment #1 (60 pts)

Due by 11:59 P.M., September 10
(upload to D2L or, if on paper, turn it in in class)

Problem 1 (15 points)

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

1. Draw the term-document incidence matrix for this document collection.
2. Draw the inverted index representation for this collection, as in Figure 1.3 in IIR.
3. What are the returned results for these queries:
   (a) schizophrenia `AND` drug
   (b) for `AND NOT`(drug `OR` approach)

Problem 2 (20 points)

1. Write out a postings merge algorithm, in the style of Figure 1.6 in IIR, for an x `OR` y query.
2. Write out a postings merge algorithm, in the style of Figure 1.6 in IIR, for an x `NOT` y query.

Problem 3 (10 points)

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

Problem 4 (15 points)

How should the Boolean query x `OR NOT` y be handled? Why is naive evaluation of this query normally very expensive? Write out a postings merge algorithm that evaluates this query efficiently.