Exercise 1.2

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

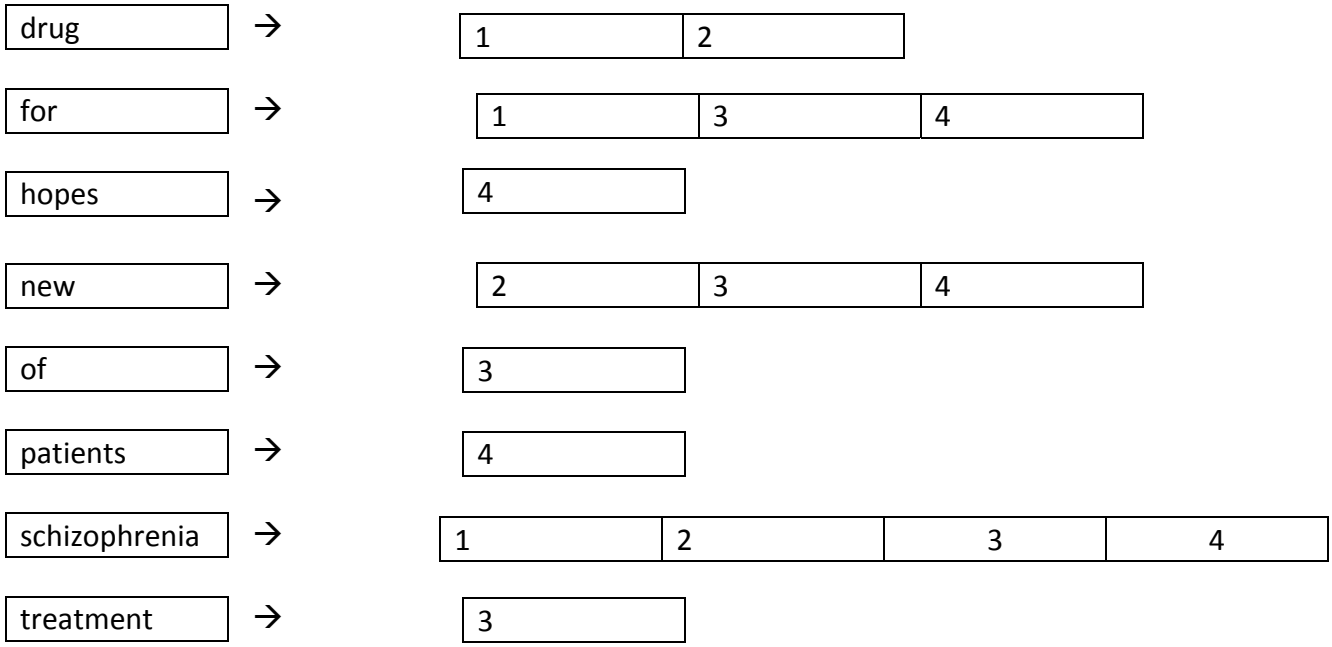
a. Draw the term-document incidence matrix for this document collection.
b. Drawthe inverted index representation for this collection, as in Figure 1.3 (page 6).

a. Term-document incidence matrix

	Doc1	Doc2	Doc3	Doc4
approach	0	0	1	0
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
hopes	0	0	0	1
new	0	1	1	1
of	0	0	1	0
patients	0	0	0	1
schizophrenia	1	1	1	1
treatment	0	0	1	0

b. inverted index representation for this collection (change the order between "hopes" and "for")

approach	→	3
breakthrough	→	1



Exercise 1.3

For the document collection shown in Exercise 1.2, what are the returned results for these queries:

a. schizophrenia AND drug

b. for AND NOT(drug OR approach)

a. Doc1,Doc 2

b. Doc 4

Exercise 1.8

If the query is:

e. friends AND romans AND (NOT countrymen)

how couldwe use the frequency of countrymen in evaluating the best query evaluation order? In particular, propose a way of handling negation in determining the order of query processing.

We always use the frequency of countrymen to evaluate the best query evaluation order.

Exercise 1.10

Write out a postingsmerge algorithm, in the style of Figure 1.6 (page 11), for an x OR yquery.

```
UNION(p1, p2)
  answer<- <>
while p1 != NIL or p2 != NIL
do
      ifp1 = NIL
            ADD(answer, docID(p2))
            p2 <- next(p2)
      else if p2 = NIL
            ADD(answer, docID(p1))
            p1 <- next(p1)
      else
            ifdocID(p1) = docID(p2)
                  ADD(answer, docID(p1))
                  p1 ← next(p1)
                  p2 ← next(p2)
            elseif docID(p1) <docID(p2)
                  ADD(answer, docID(p1))
                  p1 ← next(p1)
            else
                  ADD(answer, docID(p2))
                  p2 ← next(p2)
return answer
```

Exercise 2.1

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.

a. False
b. True
c. False
d. False

Exercise 2.3
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

c. marketing/market should not be conflated
d. university/universeshouldnot be conflated