Problem 1 [6 Points]. Exercise 2.2

Problem 2 [6 Points]. Exercise 2.3

Problem 3 [8 Points]. Create the inverted index for a collection of short documents given below. (See slide #10 in the lecture notes of Chapter 2 for an example.) You must show all the intermediate steps in creating the inverted index.

Doc 1 new home sales top forecasts
Doc 2 home sales rise in July
Doc 3 increase in home sales in July
Doc 4 July new home sales rise
Doc 5 single family and multiple family home sales rise

Problem 4. Consider the following short documents:

Doc 1 breakthrough drug for headache
Doc 2 new headache drug
Doc 3 new approach for treatment of headache
Doc 4 new hopes for headache patients

(a) [5 Points]. Create the term-document incidence matrix for the document collection listed above. (See slide #11 in the lecture notes of Chapter 2 for an example.)

(b) [5 Points]. Create the term-term correlation matrix for the term-document incidence matrix created in Part (a).

Problem 5 [5 Points]. Indices speed web query processing; however, it is usually a bad idea to create indices on every keyword and every combinations of keywords. Explain why.

Problem 6. Suppose we have a repository of 1000 documents, and we wish to build an inverted index with 10,000 keywords. A disk block can hold ten keyword-pointer pairs or 50 pointers to either a document or a position within a document. Further assume that the number of occurrences of the $i^{th}$ most frequently occurred keyword is $\frac{100,000}{\sqrt{i}}$, for $i = 1, 2, \ldots, 10,000$.

(a) [5 Points]. What is the average number of keywords per document?
(b) [5 Points]. Suppose our inverted index only records for each keyword that is in all the documents. What is the \textit{maximum} number of blocks we could need to hold the inverted index?

(c) [5 Points]. Suppose our inverted index holds pointers to each \textit{occurrence} of each keyword. How many blocks do we need to hold the inverted index?