IR Assignment #2 (Due 6th Sept)

Construct an inverted index structure in Java (or python). You have to read the given “Wikipedia-sentences.csv” and construct inverted index using the words as terms. The URLs in the first column (separated by tabs) are not to be indexed. Each line can be regarded as a record/document and you can number the documents sequentially as you read the input file. You can use Java Collection Classes. The format of each posting is of the form:

Term1: { docid1, docid2, docid3 }

Term2: { docid1, docid3, docid4, docid7 } ………

See examples in pages in text book.

Implement the following search queries on your index structure.

1. Given a query with two keywords, compute the ids of all documents (or records) that contain BOTH the keywords.

2. Given a query with two keywords, display “k” documents to the user. Display the URL (first column) and the document contents (second column). For eg. if k=10, display the first 10 documents. You should make k as a parameter to your program so that you can vary k.