CMSC 476/676 Information Retrieval
Midterm Exam – Spring 2009

Name:_____

You may consult your notes and/or your textbook. This is a 75 minute, in class exam. If there is information missing in any of the question formulations, state your assumptions and proceed with your answer.

If you decide to quote from the textbook, give the page number of your quotation. Answers written in your own words may carry more weight than quotations from the textbook.

The exam questions total 100 points.

1. **(20 points) The following True/False questions are worth 2 points each. Please circle either T or F.**

T F The vector space model of IR assumes that the dimensions are independent.

T F Stopword removal is sometimes used to increase recall in IR systems.

T F Boolean search engines are easy for end users to understand and use.

T F In a retrieval system using the vector space model, stemming tends to increase precision

T F In a retrieval system using the vector space model, stemming tends to increase recall

T F The MapReduce programming system may be used on single- as well as multi-processor systems

T F Compression of the dictionary file is rarely used in practice because of the overhead involved in keeping the pointers up to date.

T F N-grams are useful in processing documents written in multiple languages.

T F Once the documents in a collection have been indexed, it sometimes makes sense to compress them until they're needed in response to a user's query.

T F According to Zipf's Law, the most common token in a corpus should occur about twice as often as the next most common token, and so forth.

2. **Short answer questions**

A. (10 points) In a term weight calculation such as tf.idf, we sometimes use the logarithm of some quantity instead of the quantity itself. Give two examples of this usage of logarithms, and explain the rationale behind it.

B. (10 points) In turning a user query into a vector, which would then be used to calculate similarity scores between that query vector and a set of document vectors, we might assume the query terms are of equal weight. Describe two other principled approaches to weighting of query terms.

C. (10 points) Consider the case of a query term *t* that is not in the set of indexed terms for a corpus. That is, although the term *t* may occur in the collection, *t* is not represented in the vector space created by that collection. Aside from ignoring the term *t*, how would the vector space representation be adapted to handle this situation? Describe any obvious advantages or disadvantages to this adaptation.

D. (10 points) Describe the technique of n-gram overlap as applied to the problem of detecting and/or correcting spelling errors. In handling spelling errors, should one use Levenshtein distance, or n-gram overlap? Explain.

E. (10 points) Describe how skip pointers are used in postings lists. What is the advantage of skip pointers in processing a Boolean query of the form x and y?

F. (10 points) Calculate the Levenshtein distance between the words "shoe" and "sock". To do so, fill in the appropriate blanks in the skeleton table given below:

		S	H	O	E
	0				
S					
O					
C					
K					

G. (10 points) Consider a large corpus to which new documents are being added all the time. Aside from rebuilding the index from scratch with each new document, how can the index be kept current?

H. (10 points) Suppose we have a large, mostly unknown collection. By careful inspection of the first million tokens, we find 30,000 terms. About how many terms would we expect to find in the first 100 million tokens? Why?