CMSC 476/676 Information Retrieval
Midterm Exam – Spring 2014

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.

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

1. **T F** The vector space model of IR assumes that the order in which terms occur in a document is not important for retrieval.

2. **T F** Stoplist processing is sometimes used to reduce run-time storage requirements.

3. **T F** The probabilistic model of retrieval cannot be used unless the probabilities of terms occurring in the documents is known in advance.

4. **T F** In a retrieval system using the vector space model, stemming tends to decrease recall.

5. **T F** In a retrieval system that uses character n-grams rather than words, stemming is unnecessary due to the larger term space.

6. **T F** There is a principle of duality in IR, in the sense that queries can be regarded as just another type of document, and documents can be used as queries.

7. **T F** Compression of entries in the term-document matrix can be used to reduce run-time storage and execution-time requirements.

8. **T F** N-grams are useful in processing documents which may have been subject to typos and OCR errors.

9. **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.

10. **T F** The entropy of a string is related to the extent to which that string can be compressed.

11. **T F** Latent Semantic Analysis makes it possible to retrieve relevant documents even if those documents have no terms in common with the query.

12. **T F** The BM25 ranking formula is based on relaxing some assumptions in the vector space model.

13. **T F** Document length normalization was intended to avoid the problem of short documents being ranked too highly.

14. **T F** The concept of a "scored corpus" refers to a set of documents, a set of queries, and a set of terms and their corresponding weights.

15. **T F** In general, as documents in a result set are listed in descending order of estimated relevance, precision goes down as recall goes up.

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 when or why this would be a good idea.

B. (10 points) In processing a user query, which would then be used to calculate similarity scores between that query vector and a set of document vectors, we might assume the terms in the query are of equal weight. We might instead use the same term weighting formula in the query as was used in the documents. Describe one advantage and one disadvantage to this approach.

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 might the vector space representation be adapted to handle this situation? Describe any obvious advantages or disadvantages to this adaptation.

D. (10 points) Suppose we have a corpus in which Zipf's Law holds. If the most frequent term occurs one million times, and the next most frequent term occurs 250,000 times, how often should we expect to see the third most frequent term? And the fourth most frequent term?

E. (10 points) Calculate the Levenshtein distance between the words "rose" and "rice". To do so, fill in the appropriate (24) blanks in this table:

		R	O	S	E
	0				
R					
I					
C					
E					

F. (10 points) In the evaluation of a question answering system, why might the traditional recall measure be unsatisfactory? Describe a measure that might be used instead, and why it may be more satisfactory.

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