605.744 Information Retrieval: Topics for Final Exam

Like the midterm exam, the final exam will consist primarily of short-answer and problem-solving questions. The exam will be closed-book. You can bring ***two*** sheets of paper (approximately 8.5 x 11 inches) with anything on the front and back (e.g., study guides, equations, sample problems). I recommend that you bring a simple or scientific calculator to aid with arithmetic, but note that no wireless devices or computing devices with substantial storage are allowed; use of smart phones, iPads, e-Readers, etc.. is prohibited. Key topics that you should be familiar with include:

- issues in processing text, such as dealing with punctuation, stemming
- dictionaries, and how they can be represented and compressed
- document frequency, term frequency, IDF, Zipf's law
- Boolean, vector-space (*e.g.,* cosine), binary independence probabilistic model, and statistical language modeling retrieval models
- methods for term weighting (e.g., binary, tf, idf, tf-idf , 1+log(tf))
- indexing process, algorithms for indexing, including inverted file construction when memory is limited or not very limited.
- inverted file data structures and compression techniques (gap-coding; gamma and delta codes)
- how to score and rank documents for a query
- wildcard querying
- evaluation metrics, especially, precision, recall, precision at x docs, interpolated recall-precision graphs, and mean average precision
- test collections, TREC evaluations
- term operations such as stopword removal, stemming, n-gram tokenization; the effect of stopword removal and stemming on lexicon and inverted file sizes
- relevance feedback and Rocchio's method for query modification
- term similarity measures between two terms (*e.g.,* mutual information)

as well as material that we covered since the midterm exam, including:

- text classification using naive bayes, kNN, SVMs; I will not ask questions about the mathematical derivation of the optimization methods used to learn SVM hyperplanes
- how to estimate p(class|document) using the naive bayes, binomial/bernoulli model
- general issues in Web search
- PageRank, and how to compute PageRank scores for pages in a small web graph
- efficient near-duplicate document detection
- multilingual retrieval and cross-language retrieval, including use of character n-grams as indexing terms
- the general area of distributed retrieval
- the general area of multimedia (i.e., non-speech) IR
- issues involved in use of natural language processing techniques for IR.
- general familiarity with assigned readings (IIR: 1-9, 11-15, 19-21; and papers from the course web page).