605.744 Information Retrieval: Topics for Exam

The final exam will consist primarily of problem-solving and short-answer questions. I recommend that you have available a scientific calculator (or similar program) to aid with calculations. I am not claiming that this is a completely exhaustive list, but the 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), cover density ranking, 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; variable byte codes, gamma and delta codes)
- how to score and rank documents for a query
- wildcard querying, tolerant retrieval
- 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 effects of stopword removal and stemming on lexicon and inverted file sizes
- query expansion, relevance feedback, and Rocchio's method for query modification
- term similarity measures between two terms (*e.g.,* mutual information)
- text classification using Naïve 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 Naïve Bayes binomial/bernoulli model
- general issues in Web search
- PageRank, how to compute PageRank scores for a page in a web graph
- efficient near-duplicate document detection using shingling
- multilingual retrieval and cross-language retrieval, including use of character n-grams as indexing terms
- general issues involved in use of natural language processing techniques for IR.
- familiarity with assigned readings: IIR: 1-9, 11-15, 19-21; papers (i.e., required readings) assigned during the course.