Student Name: _____

COSC-488 Information Retrieval

Sample Midterm Exam

Note: This is just a sample to give you some ideas about what kind of questions may appear in the exam. It may not be an accurate prediction to how many questions are and how difficult questions will be in the actual exams.

Instruction:

- This is a close book exam.
- Answer all the questions in this exam paper.
- You must write clearly so that your writing can be recognized.
- Your answers should be thorough, complete, and relevant. Points will be deducted for irrelevant details.
- Start from the questions you are more confident with. Then deal with the difficult ones.
- Use the back of the pages if you need more room to write.

Good luck! ☺

Q1. Basic Concepts.

The followings are short answer questions to test basic IR concepts. Please provide the **formula** (if there is any) for each concept, and **2-3 sentence description** of the concept to explain why we use it.

(a) Bag of words.

(b) Document Length Normalization.

Q2. Text Preparation.

(2a) Many types of data studied in physical and social science can be formulated by the Zipf's law. Provide the **formula** and **2-3 sentence description** to Zipf's law.

(2b) Suppose that a web search engine has 100 terabytes of inverted lists. How much percentage in the 100 terabytes are actually the inverted lists for the 3 most frequent words? Justify your answer.

Q3. Vector Space Model.

Write the formula for the vector space retrieval algorithm using lnc.ltc weighting scheme. Provide descriptions to each component and each variable.

Q4. Evaluation.

In class, we talk about many evaluation metrics for search engines. Please pick the **most appropriate** evaluation metric for the following search tasks. Justify your answer.

(6a) A businessman searching for New York Time's homepage for his breakfast reading.

(6b) A lawyer searching for all relevant evidence to one of his cases. The lawyer is evaluated by whether he could win the case and he bills his client by hours. Therefore he does not mind to read through all the documents that are returned by a search engine.

(6c) An American basketball fan searching for information and history for NBA. Some of the returned pages provide a lot of relevant details, for example, team rankings, match scores, the latest news, etc. Some pages are just marginally relevant. Others are less interesting or irrelevant.