Exam IST 441

- Last name:_____ First name:_____
- Student ID:

- I acknowledge and accept the University Policies and the Course Policies on Academic Integrity

  _____

This 100 point exam determines 30% of your grade.

Search Engines (10pts)

- (10pts) Draw, connect and explain the 5 basic components of a search engine. Describe what each does.

Information Retrieval (15pts)

- You are asked to implement an information retrieval / search engine system for enterprise search. What are the important characteristics you must consider in the design? Discuss the measures that will help you evaluate its performance.

Size of things (20pts)

(20pts) You have an interview with Google. They know you took a IR/search engine course and want to see what you know about what they do. Give reasons for you answers

1. (5pts) How many web documents have they indexed? How many will they have to index in the future?

2. (10pts) What is the major kind of information growth that they will need to consider? Why?

3. (5pts) What is the size of all the data they have currently indexed? How will it grow?

Document Analysis (30 pts)

- Consider the following documents and queries:
  - D1: run dog run dog
  - D2: run run run cat
  - D3: dog bites cat
  - Q1: run
  - Q2: dog
- (5pts) Define the term vocabulary and its size.

Document Analysis cont.

- (10pts) Construct the document vector for each document and query using term frequency weights.

Document Analysis cont.

- (5pts) Define the inner product similarity metric.

- (10pts) Construct the inner product for the term frequency document vectors for all documents and the queries. Make sure you show all the computation. Which query is most relevant to which document?

Precision-Recall (25pts)

- (10pts) Define relevance; contrast it with importance. Define recall and precision in terms of relevant and irrelevant documents. Use set drawings to explain both.

- (5pts) What is most important for a search engine? Recall or precision? Why?

Precision-Recall cont.

- (10pts) Consider the following universe of documents:
  - D1, D2, D3, D4, D5, D6
  - For a particular query, we know that documents D1, D3, D4, D5 are relevant However, our information retrieval system returns documents D1, D2, and D3.
  - Determine the Recall and Precision for this query. Show your work.

Extra credit (5pts)

- (5pts) What is Moore's law and what will be the impact of Moore's law on the future of ***mobile*** search?