

# 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.

## Information Retrieval (20 pts)

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



# Document Analysis (30 pts)

- (5 pts) What is the document vector model, why is it important and how is it used?
  
- Consider the following documents D and queries Q for the following questions:
  - D1: goodbye hello goodbye hello
  - D2: you say goodbye
  - D3: I say hello
  - Q1: I hello
  - Q2: hello goodbye
  
- (5 pts) Define for the above documents and queries the term vocabulary and its size.

## Document Analysis cont.

- (5 pts) Using the 3 documents and 2 queries, construct the document vector for each document and query using term frequency weights.

## Document Analysis cont.

- (5 pts) Define the inner product similarity metric between documents and queries.
- (10 pts) Construct the inner product for the term frequency document vectors for all documents with all queries. Make sure you show all the computation. Rank the documents by each query.

## Precision-Recall (25pts)

- (5 pts) Define relevance; contrast it with importance.
- (5 pts) Define recall and precision in terms of relevant and irrelevant documents. Use set drawings to explain both.
- (5 pts) Which is more important for a search engine, recall or precision? Why?

## Precision-Recall cont.

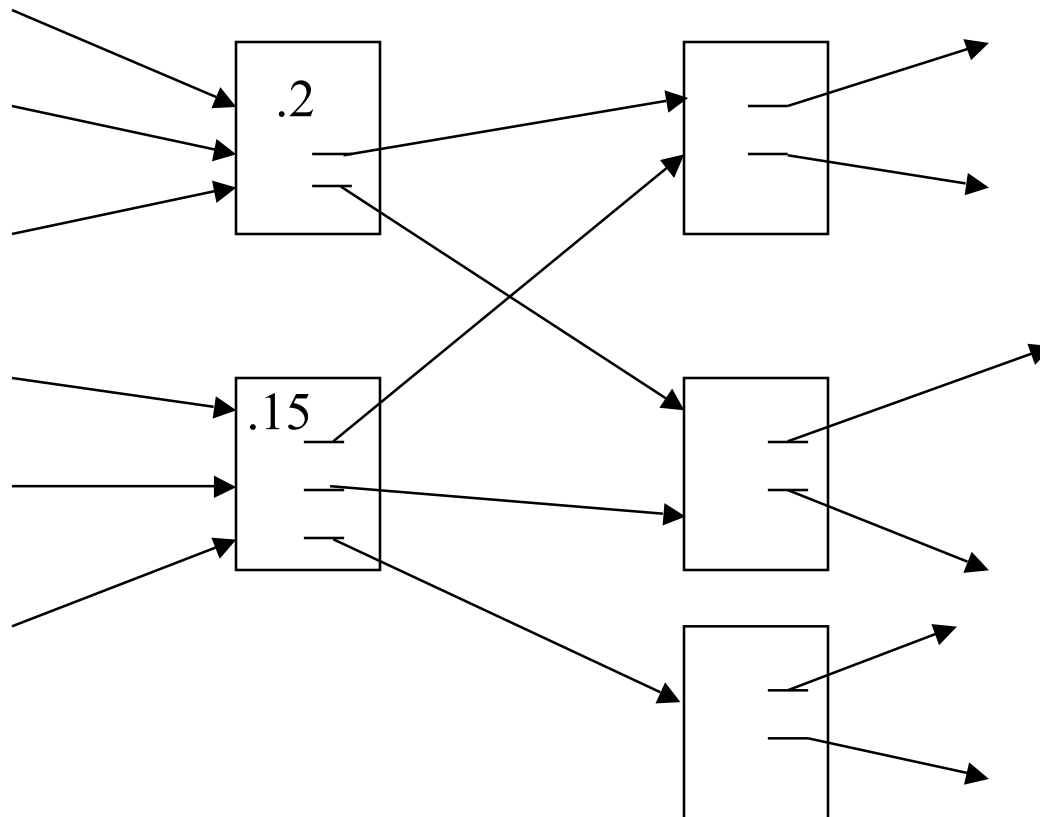
- (10 pts) Consider the following universe of documents:
  - D1, D2, D3, D4, D5
  - For a particular query, documents D1, D2, D4 are relevant. However, our information retrieval system returns D1, D2, D3, D5.
  - Calculate the Recall and Precision for this query. Be sure you show your work for complete credit.

## Pagerank (15 pts)

- (7 points) Give a definition of pagerank and what it means.

## Pagerank continued

- (8 pts) Calculate pagerank for the unranked pages; give values for the links used.



## Extra credit - robots.txt (5 pts)

- (5 pts) What is robots.txt and what does it do?