

Student ID:

Name:

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.

Student ID:

Name:

Information Retrieval (20 pts)

- You are asked to design and implement an information retrieval / search engine system for enterprise search.
 - Explain to your clients what a search engine is but describing its basic components and what each does.
 - What are the important characteristics you must consider in the design?
 - Define the measures that will help you evaluate its performance.

Student ID:

Name:

Crawling (10 pts)

You are using a web crawler for your project.

1. (5 pts) What does a crawler do? What determines what and how much it can crawl?

2. (5 pts) Define the different types of search a crawler performs. Which is better and why?

Student ID:

Name:

Document Analysis (30 pts)

- (5 pts) What is the document vector model, why is it important and how is it used by information retrieval and modern search engines? Describe what tokens are and how they are represented.

- 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 token vocabulary and its size.

Student ID:

Name:

Document Vectors

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

Student ID:

Name:

Document Similarity and Ranking

- (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.

Student ID:

Name:

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?

Student ID:

Name:

Precision-Recall Calculation

- (10 pts) Consider the following universe of documents:
 - D1, D2, D3, D4, D5, D6
 - 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.

Student ID:

Name:

Google (15 pts)

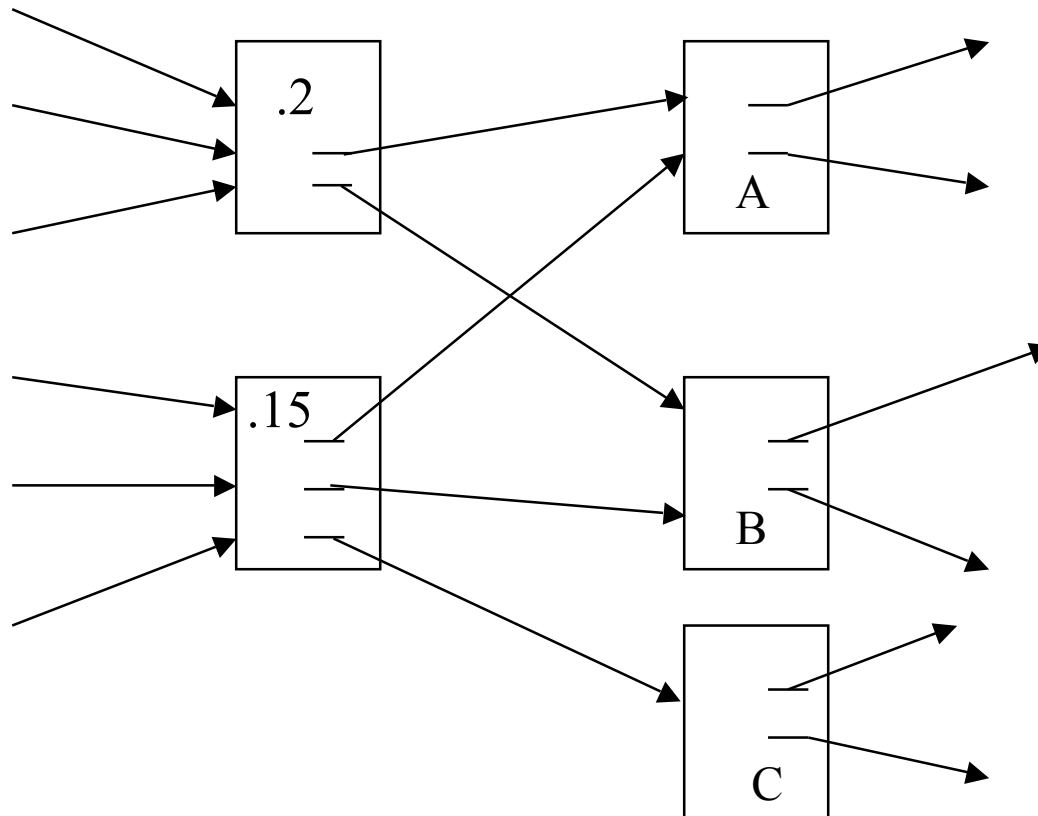
- (7 points) What makes Google different from other search engines, at least when they were first started? Discuss in terms of ranking and coverage.

Student ID:

Name:

Pagerank Calculation

- (8 pts) From from ranked pages, calculate pagerank for the unranked pages, A, B, & C giving values for the links used. Also calculate the values on the outgoing links from pages A, B, & C.



Student ID:

Name:

Extra credit - size of things (5 pts)

- (5 pts) How would you estimate the size of a search engines index? You may or may not assume it has full text indexing.