Student ID: Name:

Exam IST 441 Spring 2012

- 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.

There are 40 point extra credit questions which **graduate students** have to answer but are **optional** for undergrads.

If you use the back of any paper, please note on the front of that page.

Student ID: Name:

Information Retrieval (15 pts)

- You are asked to design and implement an information retrieval / search engine system for enterprise search for at least 1,000,000 documents that are on an internal internet and is expected to grow.
  - (5pts) Explain to your clients what a search engine is by describing its basic components and what each does.

Student ID: Name:

Information Retrieval - Characteristics & Evaluation

- You are asked to design and implement an information retrieval / search engine system for enterprise search for at least 1,000,000 documents that are on an internal internet and is expected to grow.
  - (5pts) What are the important characteristics you must consider in the design including the size of storage needed?

  - (5pts) 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.

- (5 pts) How does a crawler work? How does the crawler determine from a web site what and how much it can crawl? Give an example.

- (5 pts) Define the different types of search a crawler can perform. Is one better? If so, why?

Student ID: Name:

Document Analysis (35 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:

Stop words : {are, am}

  - D1: You say goodbye, you .
  - D2: Hello goodbye, hello goodbye, I am good.
  - D3: I say hello. How are you?
  - Q1: I hello
  - Q2: Goodbye, hello

- (5 pts) Define for the above documents and queries the token vocabulary. Explain your tokenization.

Student ID: Name:

Document Vectors

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

Student ID: Name:

Inverted index

- (5 pts) Construct the dictionary file and posting file for the 3 documents

Student ID: Name:

Document Similarity and Ranking

- (5 pts) Define the inner product similarity metric between a document and a query.

- (5 pts) Construct the inner product for the term frequency document vectors for all documents with all queries. Make sure you show all the computation.

- (5 pts) Rank the documents for each query.

Student ID: Name:

Document Similarity and Ranking (extra credit – 5 pts)

- (5 pts - extra credit) Define the cosine similarity metric between documents and queries.

Student ID: Name:

Precision-Recall (20 pts)

- (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.

- (3 pts) Which is more important for a search engine, recall or precision? Why?

Student ID: Name:

Precision-Recall Calculation

- (7 pts) Consider the following universe of documents:
  - D1, D2, D3, D4, D5, D6, D7
  - For a particular query, documents D1, D2, D4, D6 are relevant. However our information retrieval system returns D2, D3, D5, D6, D7.
  - Calculate the Recall and Precision for this query. Be sure you show what documents belong in the ratio values for complete credit. Do not just put in a number.

Student ID: Name:

Precision-Recall Calculation (extra credit – 5 pts)

- (5 pts) Your IR system works over a fixed set of documents. The number of returned documents is 100 and number relevant is 50. Recall is greater than precision by 0.2. Calculate precision and recall. Show your work.

Student ID: Name:

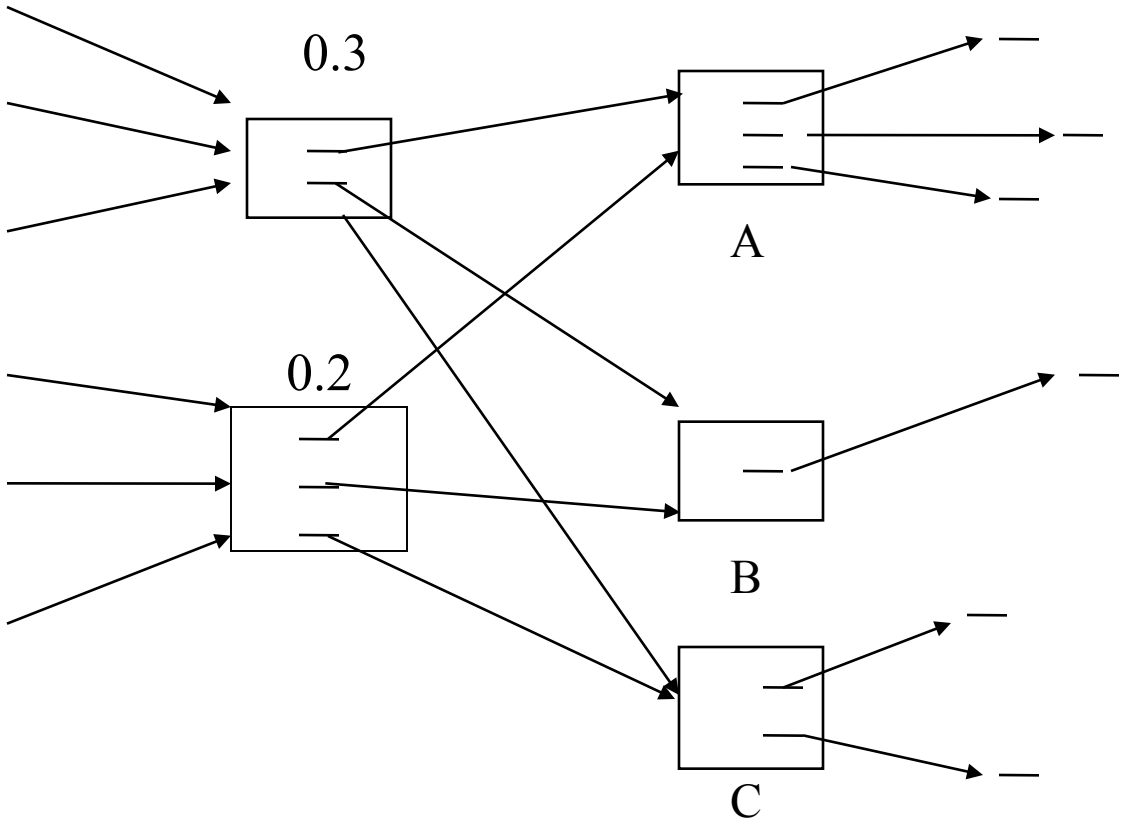
Precision-Recall (extra credit – 5 pts)

- Plot how in general precision and recall will change as the number of documents retrieved is increased. Make sure you give the numerical limits on precision and recall on your plots.

Student ID: Name:

Pagerank Calculation (10 pts)

- (10 pts) 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:

Pagerank vs hubs and authorities (extra credit - 10 pts)

- (5 points extra credit) Give a definition of pagerank and what it means.

- (5 points extra credit) Contrast pagerank with hubs and authorities.

Student ID: Name:

Social network analysis (10 pts)

- Consider the following actors (nodes) and relations (edges)

  Nodes:

  P1: Puck, P2: Greta, P3: Nat, P4: Yang, P5: Kate, P6: Carlo, P7: Nan, P8: John

  Edges:

  P1 <-> P2, P1 <-> P5, P1<-> P8, P2 <-> P3, P2 <-> P5, P3 <-> P6, P3<->P7,P4 <-> P5
  P4 <-> P6, P6 <-> P7, P7<->P8

- (6pts) Consider all edges as undirected. Construct the social network matrix and the graph.

- (2pts) Find the shortest path(s) from P2 to P7. Label the path(s).

- (2pts) Which node(s) has the largest degree and what is its value?

Student ID: Name:

Queries and the Size of things (extra credit 7 pts)

You are trying to determine how many documents a search engine has indexed. You know the search engine is a full text indexer and has a complete Boolean query system.

1. (3 pts -extra credit) What query will give an estimate of the number of documents they have indexed? Why?

2. (4 pts - extra credit) What is the size of Google?

Student ID: Name:

Boolean Queries (extra credit – 8 pts)

- Consider the following documents that have in them the terms or tokens:

{A, B, C, D, E}:

  - D1: (A, B, C)
  - D2: (A, D)
  - D3: (B, C, D)
  - D4: (C, D, E)

- You have the vague query "A OR (NOT B AND C)". What are the 2 interpretations of this query? Then, give all the documents that are retrieved for both interpretations of the query. Explain your reasoning.