Student ID: Name:

Exam IST 441 Spring 2016

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

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

  _____

  (Signature)

This 100 point exam determines 30% of your grade.

There are 50 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 10,000,000 documents that are on an internal internet and is expected to grow.
  - (5pts) Explain to your clients what an enterprise search engine is by describing its basic components and what each does. What differences are there between enterprise search and a web search engine such as Google?

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 10,000,000 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? What is the storage for a mixture of documents and images?

  - (5pts) Define the measures that will help you evaluate its performance, including how you can design for scaling.

Student ID: Name:

Crawling (5 pts)

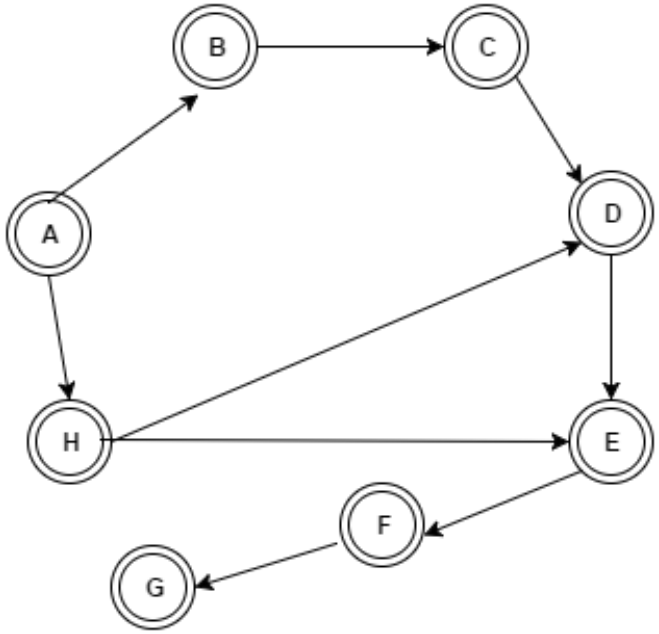
You are using a web crawler for your project.

- (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.

Student ID: Name:

Crawling (10 pts)

For the following graph give the breath first and depth first paths starting from node A. For full credit, show work.



Student ID: Name:

Crawling (extra credit: 5 pts)

- Which is better, a BFS crawler or a DFS crawler. Explain 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 query Q for the following:

Stop words : {do, to, me}

  - D1: You say goodbye to me.
  - D2: A sad goodbye to you.
  - D3: Do I say hello?
  - Q1: I say goodbye.
  - Q2: I say hello.

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

Student ID: Name:

Document Vectors

- (5 pts) Using the 3 documents and the 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. Show work.

Student ID: Name:

Document Similarity and Ranking

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

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

- (3 pts) Rank the documents for the query.

Student ID: Name:

Document Similarity and Ranking (extra credit – 10 pts)

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

- (6 pts – extra credit) Calculate the cosine similarity between each document and query. Show your work.

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, D8, D9
  - For a particular query, documents D1, D3, D4, D5 are relevant. However our information retrieval system returns D1, D2, D3, D5, 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 the number of **total** relevant document is 50 (*these are not JUST the documents that are both relevant and retrieved*). 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 and explain why.

Student ID: Name:

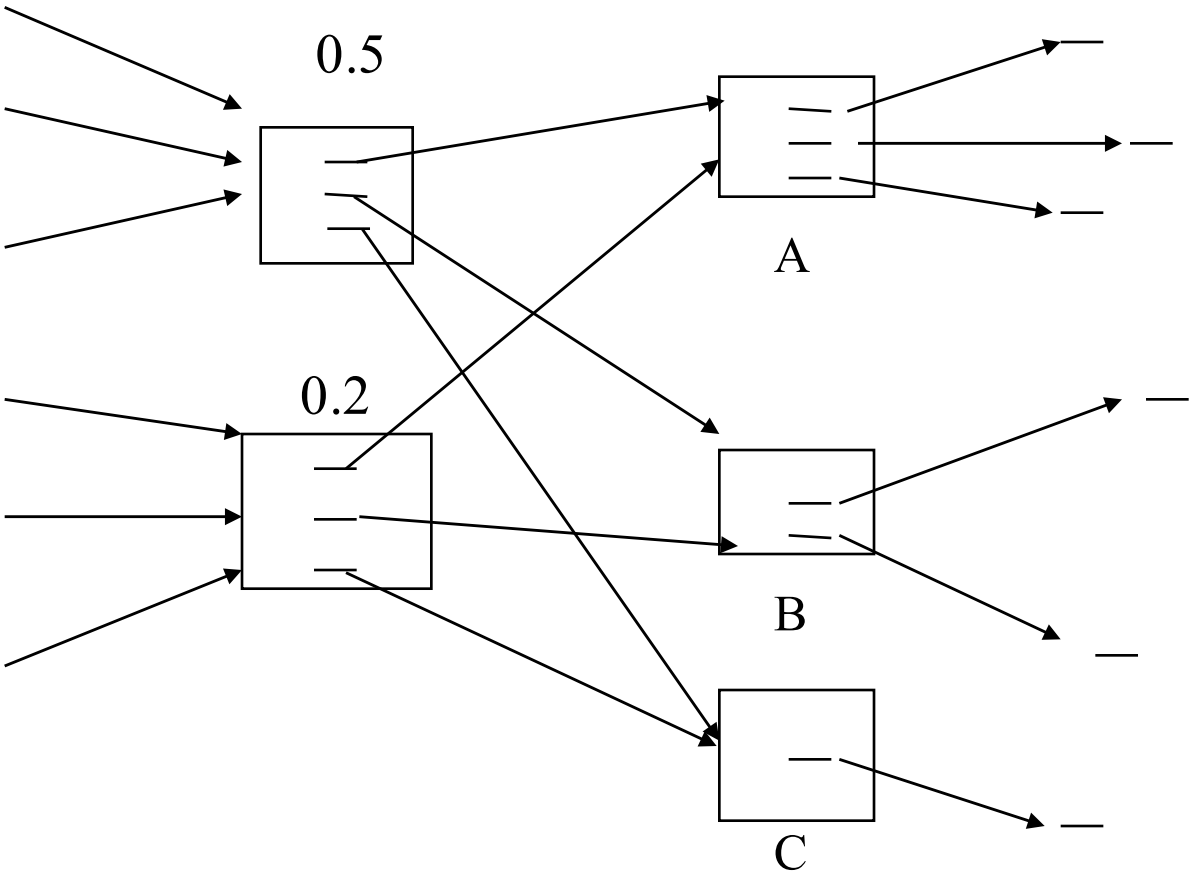
Text processing (5 pts)

- What is Zipf's law and what does it mean?

Student ID: Name:

Pagerank Calculation (10 pts & 5 pts extra credit)

- (10 pts) From ranked pages, calculate simple 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. Assume no normalization.
- (5 pts extra credit) Normalize the outgoing values (just show your work).



Student ID: Name:

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

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

- (5 points extra credit) Define mathematically hubs and authorities; contrast with pagerank.

Student ID: Name:

Complexity (extra credit 10 pts)

Give the Big O complexity for each term. Give the order of complexity of n and Label each as reasonable/unreasonable for scaling. The most complex requires the most work. All unspecified terms are undetermined positive constants

		O(n)	Order of O(n)	Scaling
a.	$100\ n^3 + 10^6\ n$			
b.	0.06			
c.	$10\ a^n + n^3$			
d.	$100\ k^n$			
e.	$.06\ n^2 + n\log n$			
f.	$20 + k \log n$			