Université d'Ottawa
Faculté de genie

École d'ingénierie et de technologie
de l'information

uOttawa
L'Université canadienne
Canada's university

University of Ottawa
Faculty of Engineering

School of Information Technology
and Engineering

CSI 4107
Information Retrieval and the Internet

FINAL EXAMINATION

Length of Examination: 3 hours **April 28, 2006, 14:00**
Professor: Diana Inkpen

Family Name: _____

Other Names: _____

Student Number: _____

Signature _____

Important Regulations:
1. **Calculators are allowed.**
2. **A student identification cards (or another photo ID and signature) is required.**
3. **An attendance sheet shall be circulated and should be signed by each student.**
4. **Please answer all questions on this paper, in the indicated spaces.**

Marks:

A	B	C	D	Total
10	10	8	12	40

Part A **[10 marks]**
Short answers and explanations.

1. (2 marks) Explain how could you use text categorization and text clustering together to implement a text classifier? Assume you have a few labeled documents (manually annotated with the class label), and many documents that are not labeled. Assume there are two classes, C1 and C2.

2. (1 mark) What is one way in which Zipf's Law manifests itself on the Web?

3. (2 marks) Compute the longest common subsequence between the following strings. Normalize by the length of the longest string. (A *subsequence* of a string is obtained by deleting zero or more characters.)

"alabaster"
"albastru"

4. (2 marks) Write a regular expression to describe an URL, in **generic** terms. Cover URLs of the following types:

http://www.ibm.com
http://www.uottawa.ca/welcome.html
http://www.site.uottawa.ca/~diana/csi4107/
http://www.site.uottawa.ca/~diana/eacl2006-clki-workshop.html#committee

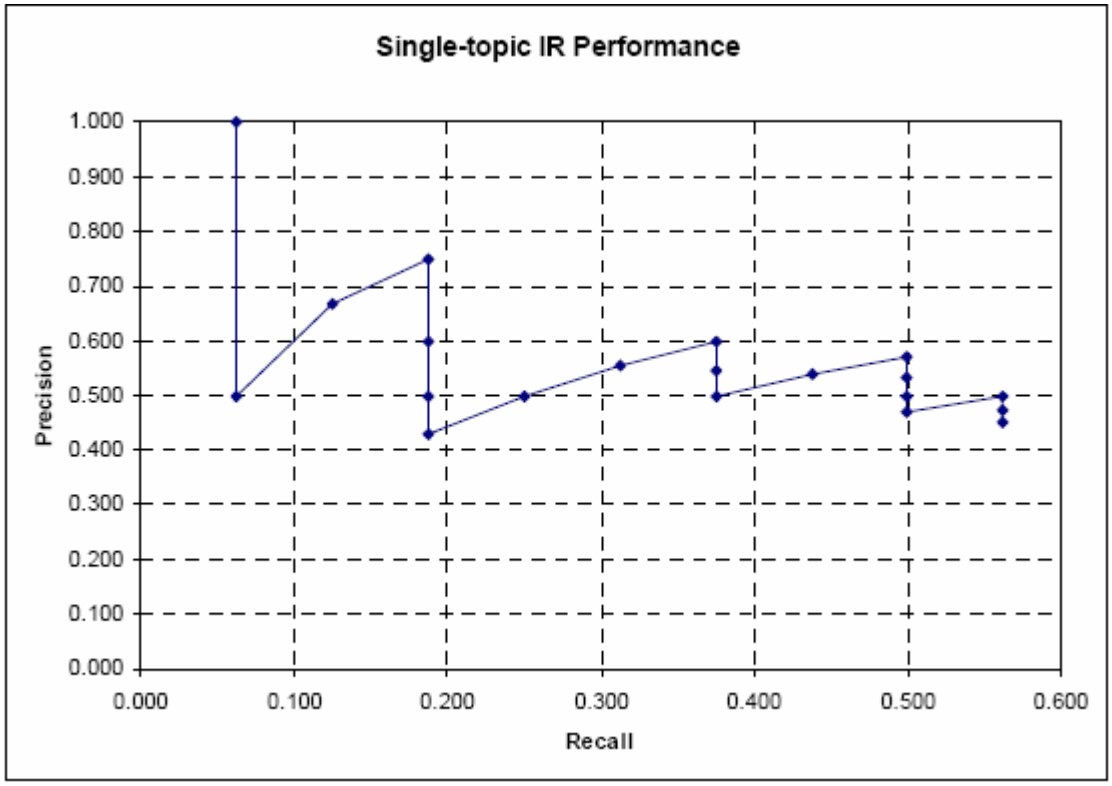
5. (2 marks) Explain how do you compute probabilities of relevance and non-relevance of a document to a query in the probabilistic retrieval model. Explain the idea, not the formulas. How do you move from the probability of a document to counts of terms? How do you deal with zero counts?

6. (1 mark) What is the main advantage of Latent Semantic Indexing?

Part B **[10 marks]**

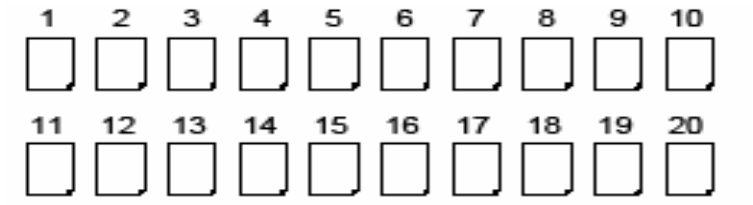
Evaluation of IR systems

The following is the uninterpolated Precision-Recall Graph of an IR system, for one topic. You know that 20 hits were retrieved, and that there are 16 relevant documents for this topic (not all of which are retrieved).



A. (2 points) What does the interpolated graph look like? Draw neatly on the graph above.

B. (4 points) In the diagram below, each box represents a hit. Based on the above Precision-Recall graph, which hits are relevant? Write an "R" on the relevant hits. Leave the non-relevant hits alone.



C. (2 points) What is the Uninterpolated Average Precision?
(defined as the average over the precision values for the points where relevant documents are found)

D. (1 points) What is the R-precision?

E. (1 points) What is Precision at 10?

Part C **[8 marks]**

Consider the following web pages and the links between them:

Page A points to pages B, and C.
Page B points to page E.
Page C points to pages B and D.
Page D points to page E.
Page E points to pages A and C.

Show the order in which the pages are indexed when starting at page A and using a spider (with duplicate page detection). Assume links on a page are examined in the order given above.

1. What is the traversal order for a breadth-first spider?

Order	

2. What is the traversal order for a depth-first spider?

Order	

3. Do you recommend breadth-first or depth-first strategy? Why?

Recommended Why?	

4. What other traversal strategy might be better?

Better traversal strategy	

Part D **[11 marks]**

1. Assume the following documents are in the training set, classified into two classes:

Dog: “cute puppy”
Dog: “big puppy”
Cat: “cute white kitten”
Cat: “white kitten”

a. (4 marks) Apply the Rocchio algorithm to classify a new document: “white puppy”
Use tf without idf and normalization, with cosine similarity. Assume the words in the vectors are ordered alphabetically. Show the prototype vectors for the two classes, and their similarities to the new document.

b. (3 marks) Apply kNN with k=3 to classify the new document: “white puppy”
Use tf without idf and normalization, with cosine similarity.
Would the result be the same if k=1? Why?

2. (5 marks) Cluster to following documents using K-means with K=2 and cosine similarity.

Doc1: “fly eagle”
Doc2: “go eagle fly”
Doc3: “eagle fly eagle”
Doc4: “go fly”
Doc5: “go eagle”

Assume Doc1 and Doc4 are chosen as initial seeds. Use tf (without idf and normalization) and cosine similarity. Assume the words in the vectors are ordered alphabetically.
Show the clusters and their centroids for each iteration. How many iterations are needed for the algorithm to converge?

Space for rough work: