UNIVERSITY OF OTTAWA
FACULTY OF ENGINEERING
SCHOOL OF IT AND ENGINEERING

CSI 4107

Midterm
March 2, 2006, 4-5:30 pm

Examiner: Diana Inkpen

Name	
Student Number	

Total marks: **47**
Duration: **80** minutes
Total Number of pages**:** **12**

Important Regulations:

1. **Students are allowed to bring in a page of notes (written on one side).**
2. **Calculators are allowed.**
3. **A student identification cards (or another photo ID and signature) is required.**
4. **An attendance sheet shall be circulated and should be signed by each student.**
5. **Please answer all questions on this paper, in the indicated spaces.**
6. **The last two pages contain some extra space for rough work.**

Marks	A	/ 13
	B	/ 10
	C	/ 4
	D	/ 10
	E	/ 10
Total		/ 47

Part A **[18 marks]**
Short answers and explanations.

1. (2 marks) Explain the difference between an information retrieval system and a search engine.

2. (2 marks) Why is *tf·idf* a good weighting scheme? Why are inverse document frequencies (*idf* weights) expected to improve IR performance when added to term frequencies (*tf*)? (Remember that the *idf* value for a term is the number of documents where it appears).

3. (2 marks) Explain what is the difference between relevance feedback and the pseudo-relevance feedback. Which one do you think would achieve better retrieval performance. Why?

4. (2 marks) In IR systems, a possible pre-processing step is stemming the words.
Do you think the performance of the system (the average precision) would be higher with or without stemming? Why?

5. (3 marks) Compute the edit distance between the following strings. Remember that the edit distance is the minimum number of deletions, insertions and substitutions needed to transform the first string into the second.
How would you normalize the score? Why is the normalization needed?

String 1: abracadabra
String 2: nabucodor

6. (2 marks) Below is a sample robot META tag in the HEAD section of an HTML document. Explain what this tag means.

<meta name = “robots” content = “index,nofollow”>

Part B **[10 marks]**

Consider a very small collection C that consists in the following three documents:

d1: “red green rainbow”
d2: “red green blue”
d3: “yellow rainbow”

For all the documents, calculate the *tf* scores for all the terms in C. Assume that the words in the vectors are ordered alphabetically. Ignore *idf* values and normalization by maximum frequency.

Given the following query: “blue green rainbow”, calculate the *tf* vector for the query, and compute the score of each document in C relative to this query, using the cosine similarity measure. (Don’t forget to compute the lengths of the vectors).

What is the final order in which the documents are presented as result to the query?

Part C **[4 marks]**

Assume that you are given a query vector q=(2,0,3,1,0), three documents identified as relevant by a user: d1, d2, d3, and two irrelevant documents: d4, d5.

d1 = (3,1,2,1,0) d4 = (1,3,0,1,2)
d2 = (4,1,3,2,2) d5 = (0,4,0,2,2)
d3 = (1,0,5,0,3)

Compute the modified query, using the Ide regular method. Remember that the Ide regular method is given by the formula:

$$\vec{q}_m = \alpha\vec{q} + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

where *Dr* is the set of the **known** relevant and *Dn* is the set of irrelevant documents. Use equal weight for the original query, the relevant documents, and the irrelevant ones, α=β=γ=1.

q'	

Part D **[10 marks]**

Given a query q, where the relevant documents are d1, d3, d6, d7, d10, d12, d13
an IR system retrieves the following ranking: d2, d6, d5, d8, d3, d12, d11, d14, d7, d13.

1. What are the precision and recall for this ranking at each retrieved document?

	Recall	Precision
d2		
d6		
d5		
d8		
d3		
d12		
d11		
d14		
d7		
d13		

2. Interpolate the precision scores at 11 recall levels.
Remember that the interpolated precision at the *j*-th standard recall level is the maximum known precision at any recall level between the *j*-th and (*j* + 1)-th level: $P(r_j) = \max_{r_j \le r \le r_{j+1}} P(r)$

Recall	Interpolated Precision
0%	
10%	
20%	
30%	
40%	
50%	
60%	
70%	
80%	
90%	
100%	

3. Why is interpolation of precision scores necessary when evaluating an IR system?

4. What is the value of the R-precision? (the precision at first *R* retrieved documents where *R* is the total number of relevant documents)

R-Precision	

5. Assume we have two users that judged the documents before the search. The first user knew before the search that d3, d6, d7, d10, are relevant to the query, and the second user knew that d1, d3, d12 are relevant to the query, what is the coverage ratio and the novelty ratio for these two users? (Remember that the coverage ratio is the proportion of relevant items retrieved out of the total relevant documents known to a user prior to the search. The novelty ratio is the proportion of retrieved items, judged relevant by the user, of which they were previously unaware.)

	Coverage ratio	Novelty ratio
User 1		
User 2		

Part E **[10 marks]**

Consider the following web pages and the set of web pages they link to:

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

E. 1. Run the Hubs and Authorities algorithms on this subgraph of pages. Show the authority and hub scores for each page for two iterations. Present the results in the order A,B,C,D,E. To simplify the calculation, do not normalize the scores.

Remember that the Hubs and Authorities algorithms can be described in pseudo-code as:

Initialize for all $p \in S$: $a_p = h_p = 1$
For i = 1 to No_iterations:
For all $p \in S$: $a_p = \sum_{q:q \to p} h_q$ (*update authority scores*)
For all $p \in S$: $h_p = \sum_{q:p \to q} a_q$ (*update hub scores*)

E.2. For the same graph, run the PageRank algorithm for three iterations.

Remember that one way to describe the algorithm is:
PR(A) = (1-d) + d(PR(T1)/C(T1) + … + PR(Tn)/C(Tn))
where T1 … Tn are the pages that point to a page A (the incoming links), d is damping factor (usually d = 0.85, you can consider it 1 for simplicity), C(A) is number of links going out of a page A and PR(A) is the PageRank of a page A. NOTE: the sum of all pages' PageRank is 1 (but you can ignore the normalization step for simplicity).