NATIONAL UNIVERSITY OF SINGAPORE

SCHOOL OF COMPUTING

SEMESTER I (2010-11)
FINAL EXAM FOR

CS2309: CS Research Methodology

November 2010 Time Allowed: 120 Minutes

INSTRUCTIONS TO CANDIDATES

1. This examination paper consists of **SIX (6)** questions and **FOUR (4)** printed pages including this page. Answer all questions.
2. Do not spend too much time on any problem. Read them all through first and attack them in the order that allows you to make the most progress.
3. Show your work, as partial credit will be given. You will be graded not only on the correctness and efficiency of your answers, but also on your clarity. Be neat.
4. This is an **OPEN BOOK** examination

For Examiner's Use Only		
	Max Marks	Earned Marks
Problem 1	10	
Problem 2	20	
Problem 3	20	
Problem 4	20	
Problem 5	20	
Problem 6	10	
TOTAL:	100	

Next Generation Search Engine. Inspired by Google's success, we wish to build the next generation search engine. To work out new ideas for the search engine, we will formulated a research project, proposed potential solutions, and run some experiments.

Problem 1. Natural language queries [10 points]

To do better than Google, we propose to let the users use natural language to form their queries. So, if the user wants to know who the Prime Minister of Singapore is, the user will type in "Who is the Prime Minister of Singapore?" instead of "prime minister singapore" as is currently done. Instead of displaying some sentence fragments containing the query words together with each returned link, the user will be shown a complete sentence containing the answer together with each returned link.

Give one advantage and one disadvantage (with supporting arguments) of the new system compared to current search engines. Discuss the appropriateness of the new system as a general search engine.

Problem 2. Scoring sentences [20 points]

For our system, we will retrieve documents in the same way as current search engines. We will then further process the top k documents in order to rank each sentence in the documents. We will rank the sentences by their *edit distance* from the query sentence. The edit distance of a sentence s from the query q is the cost of the least expensive sequence of operations required to sequentially transform the query into the sentence. The operations are started at the first position of both the query and sentence. Let i be the current position in the query and j be the current position in the sentence. The operations and effects are:

- Delete: delete the current word in the query - increment i but leave j the same, at a cost of 1.
- Insert: use the word from the current position in the sentence - increment j but leave i the same at a cost of 1.
- Replace: replace the current word in the query with the word in the sentence - increment i and j at a cost of 1.
- Copy: copy the current word in the query to the current word in the sentence - increment both i and j at the cost of 0.

(a) Give a dynamic programming algorithm (pseudo-code) for computing the edit distance from a query to a sentence.

(b) Write down the loop invariants for the loops in your code.

Problem 3. Improve speed [20 points] Assume that your algorithm for computing the edit distance has a running time of $\Theta(mn)$ where m is the length of the query and n is the length of the sentence. We would like to improve the running time of the system as much as possible.

We will first derive a lower bound for the edit distance. Let $w_1, \ldots, w_k$ be the words in the vocabulary. Let $c_q(w_i)$ be the number of occurences of w_i in the query and $c_s(w_i)$ be the number of occurences of w_i in the sentence. Let $d(q, s)$ be the edit distance between query q and sentence s.

(a) Argue that $\sum_{i=1}^{k} |c_q(w_i) - c_s(w_i)| \leq 2d(q, s)$.

(b) Describe how the inequality $\sum_{i=1}^{k} |c_q(w_i) - c_s(w_i)| \leq 2d(q, s)$ can be used to possibly improve the overall running time of the system assuming that you wish to return the top 10 sentences with the smallest edit distances.

Problem 4. Alternative formulation [20 points] Your colleague pointed out that many sentences may have large edit distance from each other despite having the same meaning. For example, "Lee Hsien Loong is the Prime Minister of Singapore" and "The Prime Minister of Singapore is Lee Hsien Loong". He is worried that the method would not work well because of that.

(a) Give a plausible reason for why the method is likely to work well for some types of queries despite his concern.

(b) Your colleague would like to give a different formulation. He reasons that words that are close together in the query should generally also be close together in an answer sentence but large groups of them may be shifted by a large distance from the corresponding position in the query.

He propose the following measure of goodness for a sentence. Let $s_1 s_2 \ldots s_n$ be the sentence, where s_i denotes word at position i in the sentence. Let $\pi(i)$ be a permutation function such that $s_{\pi(1)} s_{\pi(2)} \ldots s_{\pi(n)}$ is a permutation of the original sentence. The cost of a permutation π is $\sum_{i=1}^{n-1} c_q(s_{\pi(i)}, s_{\pi(i+1)})$ where $c_q(w_i, w_j)$ is the smallest difference in the positions of words w_i and w_j within the query q if both w_i and w_j exists in q, or $c_q(w_i, w_j) = C$ otherwise (for some constant C). The goodness of a sentence (smaller is better) is the cost of the best permutation: $\min_\pi \sum_{i=1}^{n-1} c_q(s_{\pi(i)}, s_{\pi(i+1)})$.

Reduce the problem of computing the goodness of a sentence to a travelling salesman problem (what are the vertices, edges and edge weights?).

(c) Your colleague said, "Oh no, travelling salesman is NP-complete. Since my problem can be represented as a travelling salesman problem, it is intractable". Comment on that statement.

Problem 5. Does natural language queries work? [20 points] To evaluate the new system, you hired some NUS undergraduate students to help you do an experiment. Each student is asked to search for the answer to an information need scenario by

- First use a normal search engine and measure the amount of time required to find the answer.
- Then use the new system and measure the amount of time required to find the answer for the same information need.

(a) Find two weaknesses in the experiment setup described above.

(b) Redesign the experiment to remove the weaknesses.

(c) Describe how hypothesis testing can be done to check whether the measured effect is statistically significant. Describe the null and alternative hypotheses and the type of test to be used.

Problem 6. Does edit distance work well? [10 points] Your colleague is not convinced that edit distance works well for this problem. Design an experiment to evaluate whether it works well. Specify the statistical hypothesis testing procedure that you would use as well.