Norges teknisk-naturvitenskapelige universitet
Institutt for datateknikk og informasjonsvitenskap

EKSAMENSOPPGAVE I FAG TDT4117 – INFORMASJONSGJENFINNING

Contact during the exam: Heri Ramampiaro

Phone: 93440

Date: ***1st of December 2010***

Time / length: *3:00 P.M. - 7:00 P.M. / 4 hours*
Allowed aid: *D – No printed or written materials allowed.* ***Only*** *simple, approved calculator is allowed.*

Language: English

Result deadline: 22nd of December 2010

Give **short and concise** answers to all questions. Short sentences are preferred rather than long explanation.

Please read all questions very carefully and consider your answer to each question.

Please explain all your answer to every question.

If you believe some information is missing, please describe any assumptions you make.

Problem I (20%)

a) List the three most important characteristics that distinguish *data retrieval* from *information retrieval.*

b) Why is the choice of "Index terms" important within the context of information retrieval? Justify your answer.

c) Explain what steps or operations you should perform before you can index documents?

d) What do we mean by "Feature extraction" and in what contexts do we use this concept? What three requirements should extraction of features satisfy?

Problem II (10%)

a) Draw a figure of the multimedia data model. What three layers (layers) does this model consist of? Give at least one example of something belonging to each layer.

b) Compare the centralized and distributed web search system. Use or include drawings of the architectures to support your answer.

Problem III (15%)

Table 1 shows a result of the search conducted by Ola.

Rang	Dok ID	Score	Relevance
1	232	1534	
2	111	1240	REL
3	2343	1111	
4	332	1022	REL
5	441	1001	
6	112	0999	REL
7	22	0555	
8	1	0233	REL
9	23	0220	
10	334	0100	

Tabell 1

Assume that we have totally 6 relevant documents in our collection.

a) Explain the concepts ”*precision*” and ”*recall*”?

b) Compute the values of precision and recall using the results in Table 1.

c) What is R-precision? Compute the value of R-precision for the result above.

d) What is MAP (Mean Average Precision)? Explain.

Problem IV (25%)

a) Compare the two models Probabilistic Model and the Language Model for information retrieval. Focus on the principle, advantages and weaknesses.

b) Query expansion.

1. Explain the goal of query expansion.
2. Explain the Standard Rocchios equation.

c) What is the main difference between "Automatic Local Analysis" and "automatic global analysis"?

d) Explain the principles behind the inverted index. Use examples and figures to support your answer.

e) What is the suffix string for the following text? Construct a "suffix Trie" and "suffix tree" based on this text.
"South Korea says it ha Returned four after North Korea fired Dozens of artillery shells that one of its border islands, killing two marines"

Problem V (20%)

a) Explain what we have to do with the indexing and the index terms in order to be able to perform or execute wildcard queries.

b) Explain what the term "*edit distance*" means. What does it do? If we have a text string s_i, where *len*(s_i) is the length of s_i. Show that the "*edit distance*" between s_1 and s_2 can never be greater than *max(len(s_1), len(s_2)).*

c) Explain the term "Jaccard Coefficient." Why doesn't this work very well for ranking of search results?

d) Explain two ways to allow phrase queries. Tip: The focus here is the indexing and posting list.