UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Exam in: INF3800/INF4800 Search Technology
Day of exam: Tuesday June 11th, 2013
Exam hours: 14:30-18:30 (4 hours)
This examination paper consists of 3 pages
Appendices: None
Permitted materials: None

Make sure that your copy of this examination paper is complete before answering.
You can answer the questions in English or Norwegian.

QUESTION 1: NAÏVE BAYES

a) What is the equation used for Naïve Bayes classification? Explain the simplifying assumptions that are made. For each such assumption, provide an example from text classification of when the assumption is violated.
b) There are two different ways we can set up a Naïve Bayes classifier. Briefly explain the difference between the multinomial model and the Bernoulli model. Which model would you typically use for text classification, and why?
c) In the table below, show how the parameter estimation for the multinomial model and the Bernoulli model differs for estimating the conditional probability *Pr(bieber|c=canada)*. Use simple add-one smoothing. (Note that you are not asked to compute all probability estimates required for classifying new documents, just a single conditional probability.)

Document ID	Words in document	In *c=canada*?
1	*bieber hockey bieber*	Yes
2	*bieber bieber ontario*	Yes
3	*bieber aboot*	Yes
4	*oslo opera bieber*	No

QUESTION 2: QUERY EVALUATION

a) Explain what skiplists are and how they work. Are skiplists necessarily always beneficial for performance?
b) Explain the difference between term-at-a-time and document-at-a-time evaluation. Which approach would you use for an index where the posting lists are impact-ordered, and why?
c) Postings lists can be ordered according to a static quality score *g(d)*. Explain why this can be beneficial, and provide some examples from web search of what such a static quality score might represent.

QUESTION 3: XML SEARCH

a) What is the structured document retrieval principle?
b) List at least five things that make XML information retrieval more challenging than information retrieval with ”traditional” documents, and explain what those challenges are. Provide one example for each challenge.

QUESTION 4: EVALUATION METRICS

a) How is the balanced *F* measure (also known as F_1) defined?

b) How is the *R*-precision (also known as the break-even point) defined? Relate this to F_1.
c) An information retrieval system returns eight relevant documents and ten non-relevant documents for a search. There are a total of twenty relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

QUESTION 5: SUFFIX ARRAYS

a) Construct and draw a suffix array for a dictionary that contains the two strings *bieber* and *belieber*.
b) Explain how you can use this suffix array to perform an efficient search for the substring *ieb*.

QUESTION 6: PERMUTERM INDEXES

a) Explain how a permuterm index works. Show how the string *bieber* would be represented.
b) Show how the wildcard query *bi*er* would be evaluated.