UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Exam in INF3800/INF4800 Search Technology
Day of exam: May 30th, 2016
Exam hours: 14:30-18:30 (4 hours)
This examination paper consists of 3 page(s)
Appendices: None
Permitted materials: None

Make sure that your copy of this examination paper is complete before answering.

You can answer in Norwegian or English. Please use the language that you are most comfortable with.

SQUEEZE (20%)

Provide clear and concise explanations of how the following compression algorithms work. Give examples, if possible. You can assume that sequences of integers are already gap-encoded, if appropriate.

a) [5 points] Variable-byte (VB) encoding
b) [5 points] Elias γ-encoding
c) [5 points] Simple9 (S9) encoding
d) [5 points] Patched frame-of-reference (PFOR) encoding

LUCY IN THE SKY WITH DIAMONDS (15%)

a) [8 points] Recommend an efficient query processing order for the Boolean query below, given information about the size of the posting lists. You can assume that the expression is evaluated as written, i.e., no rewriting of the Boolean expression takes place. Justify your answer.
b) [7 points] Would skip pointers be useful for making the processing of this query more efficient? Justify your answer.

(*tangerine* OR *trees*) AND (*marmalade* OR *skies*) AND (*kaleidoscope* OR *eyes*)

kaleidoscope	87K entries
eyes	213K entries
marmalade	107K entries
skies	271K entries
tangerine	46K entries
trees	316K entries

NEW YORK DOLLS (20%)

a) [7 points] Heaps' law is given as $M = kT^b$. Explain what M, k, T and b are.
b) [13 points] Indexing New York Times newswire from 1991–1995 reveals that it contains about 400 million word tokens, and a lexicon of size about 1 million. What would be a good prediction of how many word tokens and what lexicon size one would get in indexing New York Times newswire from 1991–2000? You can assume that all decisions related to text processing (term normalization, lowercasing, treatment of numbers, and so on) are fixed, that the New York Times' rate of publishing new content is constant during the entire period, that $b = 0.5$ in Heaps' law, that $\sqrt{2} \approx 1.41$, and that the student sitting next to you has a shoe size of 44.

BAYES CITY ROLLERS (25 %)

a) [5 points] What is the mathematical expression for the naïve Bayes classification rule? Outline which assumptions the naïve Bayes model makes when used for text classification.
b) [4 points] Explain what smoothing is and its purpose.

c) [8 points] Consider the documents $\{d_1, d_2, d_3, d_4\}$ below. Using these as the training set for a multinomial naïve Bayes classifier, compute the probabilities used in the trained model. Use simple add-one or Laplace smoothing.

d) [8 points] Consider the naïve Bayes model developed above and the document d_5 below. Show the computations involved in classifying d_5.

d_1: longmuir nobby longmuir	*c = byebyebaby*
d_2: longmuir longmuir faulkner	*c = byebyebaby*
d_3: longmuir mckeown	*c = byebyebaby*
d_4: hendrix lennon longmuir	*$c \neq$ byebyebaby*
d_5: longmuir longmuir longmuir hendrix lennon	*c = ?*

BEACH BOYS (20%)

a) [5 points] Give a brief and concise explanation of PageRank and the random surfer model.

b) [8 points] Consider a graph with three nodes *brian*, *dennis* and *carl* and links as follows: *brian*→*dennis*, *carl*→*dennis*, *dennis*→*brian* and *dennis*→*carl*. Write down the transition probability matrix P for the surfer's walk with teleporting, using the teleport probability $\alpha = 0.5$.

c) [7 points] Given P, explain how to use this to compute PageRank values.