UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Exam in INF3800/INF4800 Search Technology
Day of exam: June 8th, 2015
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.

STEMMING (15%)

Are the following statements true or false? Justify your answers.

a) [4 points] In a Boolean retrieval system, stemming never lowers precision.
b) [4 points] In a Boolean retrieval system, stemming never lowers recall.
c) [3 points] Stemming increases the size of the vocabulary.
d) [4 points] Stemming should be invoked at indexing time but not while processing the query.

RELEVANCY EVALUATION (25%)

a) [9 points] You want to quantify how relevant the results returned by a given search engine are, given a set of benchmark queries with relevance judgments. For the three benchmark queries below the search engine's top 10 results are:

   ariana grande
   RNRNNRNNRR

   trine skei grande
   NRNNRNRNNN

   jono el grande
   RNNRNNNNNR

   Here R indicates a document that is relevant to the query and N indicates a non-relevant document. You can assume that all relevant documents are among the top 10 results. What is the search engine's MAP (mean average precision) on this set of benchmark queries?

b) [9 points] Wanting to go beyond binary relevancy judgments, you now produce a benchmark query *grande latte* with relevance judgments on a scale of [0, 16] as follows:

   Document A = 4
   Document B = 2
   Document C = 8
   Document D = 16

   For all other documents the relevancy judgment is assumed 0 for this query. For this query the search engine's top 10 results are:

   grande latte
   ADXCXXXBXX

   Here X indicates a document other than the ones listed above. What is the search engine's DCG (discounted cumulative gain) score at rank 10 for this query? Use a simple logarithmic discounting.

c) [7 points] Explain how you would normalize the DCG score you computed above to arrive at the search engine's NDCG (normalized DCG) score at rank 10.

SIMILARITY FUN (30%)

a) [15 points] Define cosine similarity, and give a geometrical intuition of this distance metric.
b) [15 points] One measure of the similarity of two vectors is the Euclidian distance between them:

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2}$$

Given a query q and documents d_1, d_2, …, we may rank the documents d_i in order of increasing Euclidian distance from q. If q and d_i are all normalized to unit vectors, is the rank ordering produced by Euclidian distance identical to that produced by cosine similarity? If yes, prove that it is. If no, prove that it isn't.

LINK ANALYSIS (30%)

a) [10 points] Give a brief and concise explanation of PageRank and the random surfer model.
b) [10 points] Consider a web graph with three nodes 1, 2 and 3 and links as follows: 1→2, 3→2, 2→1 and 2→3. Write down the transition probability matrix P for the surfer's walk with teleporting, using the teleport probability $\alpha = 0.5$.
c) [10 points] Given P, explain how to use this to compute PageRank values.