Examination in IN3120/IN4120

UNIVERSITY OF OSLO
The Faculty of Mathematics and Natural Sciences
Written examination IN3120/IN4120
2022 Autumn
Duration: December 5, 09:00 a.m. - 01.00 p.m. (4 hours)
Permitted aids: None
It is important that you read this front page before you start.

The different questions have different weights, as indicated.

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

1 PAIRWISE PREFERENCES [20p]

You are running a popular search engine and want to quantitatively assess how relevant the results you serve are. Your plan is to use some kind of gold standard relevance judgments for a set of queries, and then compare these with the results you serve. You recall that a pairwise relative preference (d_i, d_j) for a query q tells you that for query q document d_i should be ranked before document d_j, and that Kendall's tau distance for a query q can be used to measure how "close" a ranked document list R_q is to a set P_q of pairwise relative preferences.

a) [7p] To get the gold standard relevance judgments you decide that manually assessing relevance is too laborious and want to produce these from the search engine's clickthrough logs instead of manually assessing relevance: For a query q the clickthrough logs tell you the top 10 results that are served (and their ranks) for q, and also which of these results that a user clicked on. Describe how you could use the clickthrough logs to produce a set P_q of pairwise relative preferences for a query q.

b) [5p] We can define Kendall's tau distance as $K(R_q, P_q) = (X - Y) / (X + Y)$. Define what X and Y are.

c) [5p] For the specific query *foo* you have a set of pairwise preferences P_{foo} = {(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)} and a contender ranking algorithm that produces the ranked document list R_{foo} = [1, 3, 2, 4]. What is the Kendall tau distance between R_{foo} and P_{foo}?

d) [3p] What are the minimum and maximum possible values of the Kendall tau distance? Justify your answer.

Fill in your answer here

Format

Words: 0

Maximum marks: 20

2 SUFFIX ARRAYS [15p]

Consider the questions and statements below about suffix arrays, and select the correct answer. Justify your answers, i.e., explain why you believe that your answer is correct. Answers with no justification give no credit.

a) [5p] What is the suffix array for the string *engineering*?

1. [2, 3, 8, 4, 9, 1, 7, 5, 0, 6, 10]
2. [5, 0, 10, 6, 2, 3, 8, 4, 9, 1, 7]
3. [5, 0, 6, 10, 2, 4, 9, 1, 7, 3, 8]
4. [5, 0, 6, 10, 2, 3, 8, 4, 9, 1, 7]
5. [5, 10, 0, 6, 8, 3, 2, 4, 9, 7, 1]

b) [5p] In IN3120's obligatory assignment on suffix arrays, the naïve comparison-based sorting algorithm used to construct the suffix array for a string *s* runs in $O(n\log n)$ time, where $|s| = n$.

1. True
2. False

c) [5p] Using the suffix array for a string *s* with $|s| = n$, what is the best bound required to locate in *s* the first occurrence of a pattern having length *m*, where $m < n$?

1. $O(nm)$
2. $O(n^2)$
3. $O(mn\log n)$
4. $O(m\log n)$
5. $O(\log n)$

Fill in your answer here

Format

Words: 0

Maximum marks: 15

3 BASIC LINGUISTIC PROCESSING [20p]

a) [7p] What is the goal of both stemming and lemmatization? Explain the difference between stemming and lemmatization and discuss their relative merits.

b) [7p] When building a search engine, a design choice you might face is to select between reduction to base form versus expansion, and if you should process the query or the document or both. Discuss!

c) [6p] Describe the general principle behind how stemming algorithms work. You can use Porter's algorithm as an example.

Fill in your answer here

Format

Words: 0

Maximum marks: 20

4 MIXED GRILL [25p]

a) [5p] Describe what a Bloom filter is and how it works.

b) [4p] Provide an example of where using a Bloom filter can improve a search engine's performance.

c) [4p] In the random surfer model the surfer is allowed to teleport with some probability greater than zero. If you assume that teleportation is not allowed, discuss some of the problems that can arise and their implications.

d) [4p] Explain the difference between collection frequency and document frequency. Give at least one example where each of these concepts are key.

e) [4p] In SVMs the concept of a support vector is key. Describe what a support vector is, and which role support vectors play.

f) [4p] Let R denote a relevant document, and let N denote a non-relevant document. For the query *foo* your search engine returns the ranked list [R, R, N, N, R], and for the query *bar* your search engine returns the ranked list [N, R, N, R, R]. What is the search engine's MAP score for the query set {*foo*, *bar*}?

Fill in your answer here

Format

Words: 0

Maximum marks: 25

5 FIELDS [20p]

A document might have different fields. For example, typical fields for a document include *title*, *body* and *author*. It is sometimes desirable to direct a query only towards specific fields, e.g., “*search for ‘foo’ but only within the ‘title and ‘author’ fields.*” This is sometimes called fielded search.

a) [10p] Using an inverted index, discuss at least two ways to support fielded search. What are their advantages and drawbacks?

b) [10p] Some fields might intuitively be more important than others. For example, if your query matches the content of the *title* field for a document that might be better than a match in the *footnotes* field. Discuss how you could design a relevance function that takes this into account.

Fill in your answer here

Format

Words: 0

Maximum marks: 20