Examination in IN3120/IN4120

Written home exam in IN3120/IN4120
2021 FALL
Duration: 2021-11-29 09:00 to 2021-11-29 13:00

It is important that you read this cover page carefully before you start.

General information:

- Important messages during the exam, if any, are given directly from the course teacher through the course's Mattermost team in the Announcements channel. It is therefore important that you check Mattermost regularly.
- Your answer should reflect your own independent work and should be a result of your own learning and work effort.
- All sources of information are allowed for written home exams. If you reproduce a text from books, online articles, etc., a reference to these sources must be provided to avoid suspicions of plagiarism. This also applies if a text is translated from other languages.
- You are responsible for ensuring that your exam answers are not available to others during the exam period, neither physically nor digitally.
- Remember that your exam answers must be anonymous; do not state either your name or that of fellow students.
- If you want to withdraw from the exam, press the hamburger menu at the top right of Inspera and select "Withdraw".
- You can submit your answers in English or Norwegian. Please use the language you are the most comfortable with.
- When answering, please state any assumptions you make and clearly show how you arrive at your answers. Sound explanations and clear reasoning count positively.
- If you need to ask clarification questions regarding the questions on the exam (“digital trøsterunde”) you can send an SMS to +4747937222 exactly one hour after the start of the exam and request a callback. You will then be called back shortly after.

Collaboration during the exam:

It is not allowed to collaborate or communicate with others during the exam. Cooperation and communication will be considered as attempted cheating. A plagiarism control is performed on all submitted exams where text similarities between answers are checked. If you use notes that have been prepared in collaboration with others before the exam, this might be detected in a plagiarism control. Textual similarities such as these can be considered by graders as a show of low independence or even attempted cheating. Refrain from copying/pasting from notes made in collaboration with others.

Cheating:

Read about what is considered cheating on UiO's website.

Contact information:

User support exam

Background scenario:

All questions in this exam assume the following background scenario:

You have been put in charge of architecting the backend of a complete system for news search. You expect the number of news articles ("documents") to grow to a billion (10^9), and expect that you will need to handle 1000 queries per second. All news articles will be in English, and a typical news article is 2000 words long. An average English word is 5 characters long, and the news articles are all UTF-8 encoded. You expect the significant majority of the queries to be short (usually not more than 3 terms) and typically be the names of people or places mentioned in the news.

1 SIZING [30%]

This question assumes the background scenario listed on the cover page.

Before you start you decide to do some sizing estimates.

(a) [7%] When the system has seen a billion (10^9) documents, how many unique terms do you estimate will have been seen?

(b) [7%] After having collected some statistics across a sample of 10 million documents, you note that the second most frequent term is “of” and that it occurs about 700,000 times in total across the sample. Based on this observation, how many times would you estimate that the most frequent term and the third most frequent term will occur across a billion documents?

(c) [6%] Based on findings from analyzing query logs from a similar system you decide that it makes sense to decorate your posting lists with skip lists, in a non-positional index. How many entries do you estimate that the skip lists for your 5 longest posting lists will have in total when you reach a billion documents? Do not include the original posting lists themselves in your count.

(d) [10%] It makes sense to you to compress both your posting lists and your skip lists, and you decide to use gap coding and then apply Elias gamma-coding to the gaps. How much space do you estimate that your 5 longest posting lists will consume in total when compressed, including their associated skip lists?

Skriv ditt svar her

Format

Words: 0

Maximum marks: 10

2 QUERY SUGGESTIONS [35%]

This question assumes the background scenario listed on the cover page.

After a week with real traffic, you want to add a feature where users of the search system are provided with a suggest-as-you-type query completion feature: For each keypress the system should suggest up to 5 queries that the system has seen before, and that start with whatever characters the user has typed into the search box so far. The suggested queries should be ranked according to how many times the system has seen the suggestion before.

To this end you collect a large set of query logs, where in each log file you have one query string per line. An example excerpt:

…
oslo festival
jonas gahr støre
oslo pizza
jonas gahr støre
bergen ordfører
…

Based on the above excerpt as an example, if the user types *osl* the queries *oslo festival* and *oslo pizza* are two plausible completion suggestions.

To prepare the data material used for suggestions, you want to run a daily job on a MapReduce cluster that from last week's query logs produces a list of (query, frequency) pairs, where frequency is the occurrence count across all your query logs for the query string query.

(a) [5%] Explain what the signatures of the mappers and reducers in your MapReduce job look like. Provide concrete examples of inputs and outputs.

(b) [6%] You do not want to suggest queries to the user that could be perceived as somehow offensive. To this end, you have a big dictionary with about 250,000 offensive words and phrases, and queries that contain at least one of these dictionary entries should not end up in the data material used for suggestions. Explain how you would match the logged queries against your dictionary as efficiently as possible, and where/how you would inject this logic into your MapReduce job.

Given the overall query volume and that suggestions need to be generated per keypress, you aim for an as-efficient-as-possible in-memory solution for serving the suggestions.

(c) [6%] Given the requirement that what the user has typed is an exact prefix of a query in the generated query dictionary, suggest a suitable data structure and describe the associated lookup algorithm.

(d) [6%] Suggest some practical ways to speed up the process of serving suggestions to users, that helps for the cases where the lookup costs are the greatest.

(e) [6%] If the requirement changes from exact prefix matching to allowing up to k edit errors when matching the prefixes, suggest a suitable data structure and describe the associated lookup algorithm. You can assume that the value of k is 0, 1 or 2 and determined as a function of the length of the prefix typed by the user.

(f) [6%] If the requirement changes from matching prefixes to allowing matches to start on any token boundary, suggest a suitable data structure and describe the associated lookup algorithm.

Skriv ditt svar her

Format

Words: 0

Maximum marks: 10

3 RELEVANCE [35%]

This question assumes the background scenario listed on the cover page.

In news search, freshness matters, and news articles are published continuously. Also, in the news domain, the importance of proper names is also more prominent than in many other domains and the authority and trustworthiness of a news site is a significant factor.

(a) [6%] You want news articles to be searchable shortly after they are first published, crawled and pushed to the indexer. Describe a strategy for building and organizing your index so that it gives the user the ability to find documents that have been very recently published.

(b) [6%] Your news search system has an inverted index at its core. Given your knowledge about the news domain, suggest ways of influencing search relevance in beneficial ways. Be as concrete as you can.

(c) [8%] You will launch the news search system using a heuristic ranking function based on cosine similarity with TF-IDF weights and a handful other factors, but aim to eventually replace this with a machine-learnt ranker based on an SVM. Discuss which data you will start collecting to train such a ranker, and explain how you make use of this data to train it.

(d) [7%] Having trained a new model you will want to assess if the trained model performs better than the current model. How would you determine this? Discuss your methodology.

(e) [8%] You want your user to experience low query latency without significantly compromising on relevance. Discuss some tricks and approximations you can do to achieve this.

Skriv ditt svar her

Format

Words: 0

Maximum marks: 10