

Homework 1: Vector Search

Naomi Derel 325324994, Gili Cohen 326280815, Renana Shachak 213920010

02.07.2024

1 Faiss

With permission from the course staff, we kept our answers for the original question and not the updated requirements as we finished it prior to the update.

1.1 Running Times

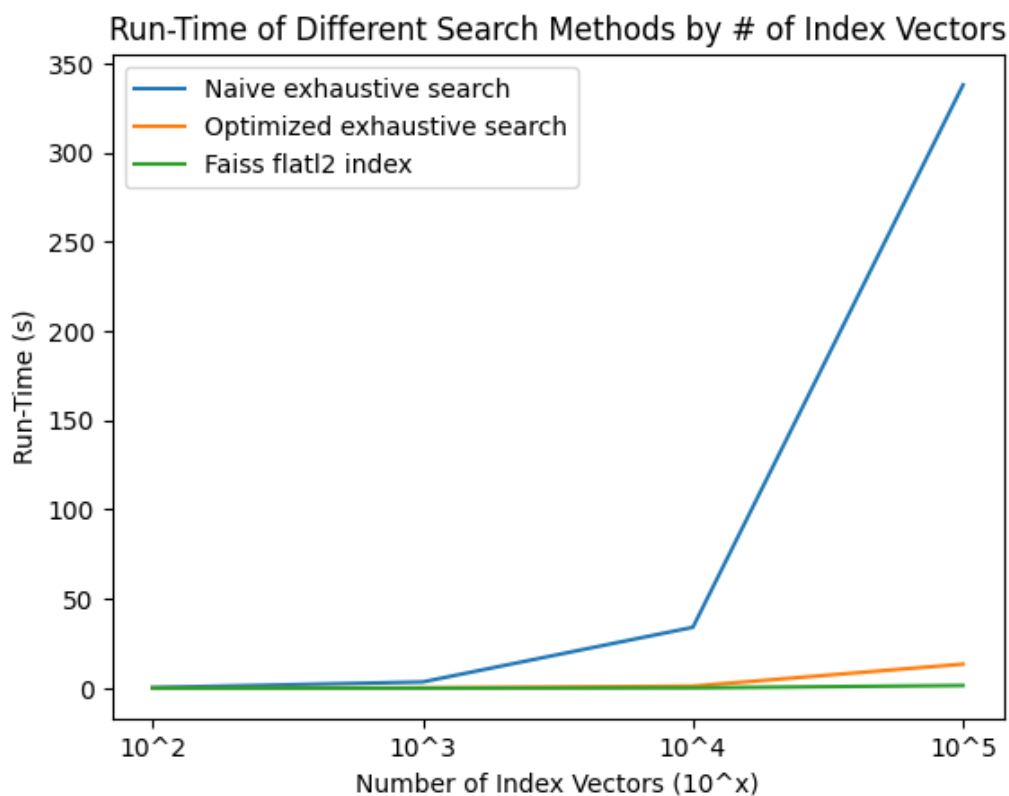


Figure 1.1.1: Running times by Number of Index Vectors

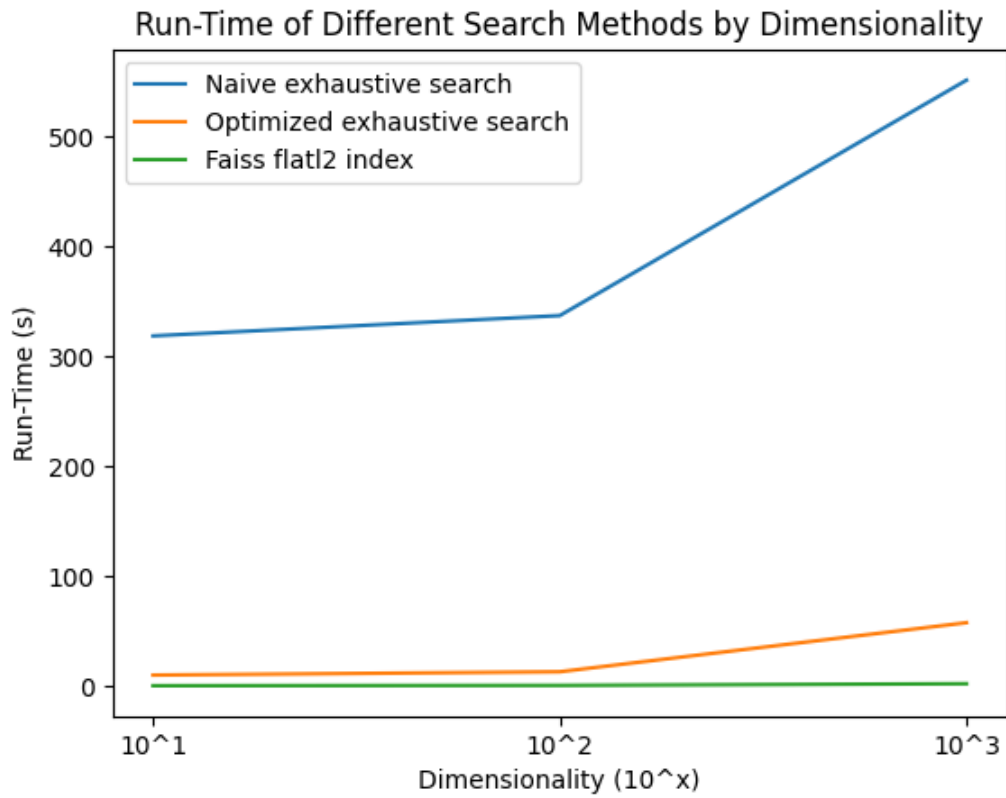


Figure 1.1.2: Running times by Dimentionality

1.2 Faiss LSH

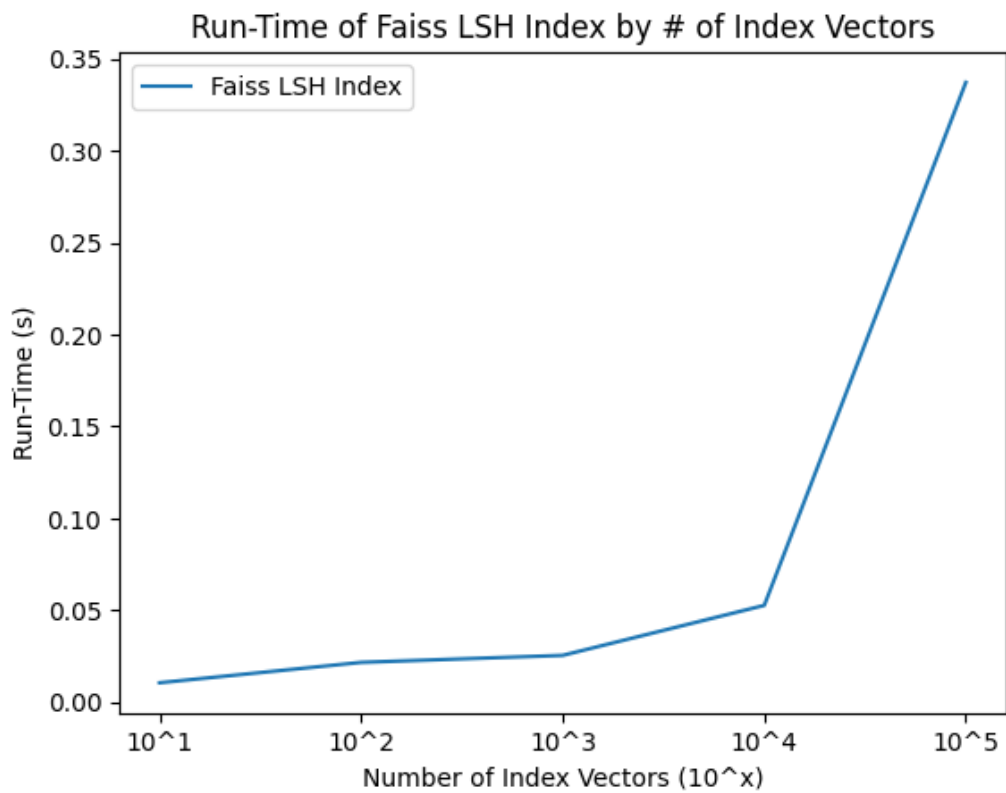


Figure 1.2.1: Running times by Number of Index Vectors



Figure 1.2.2: Running times by Number of Bits

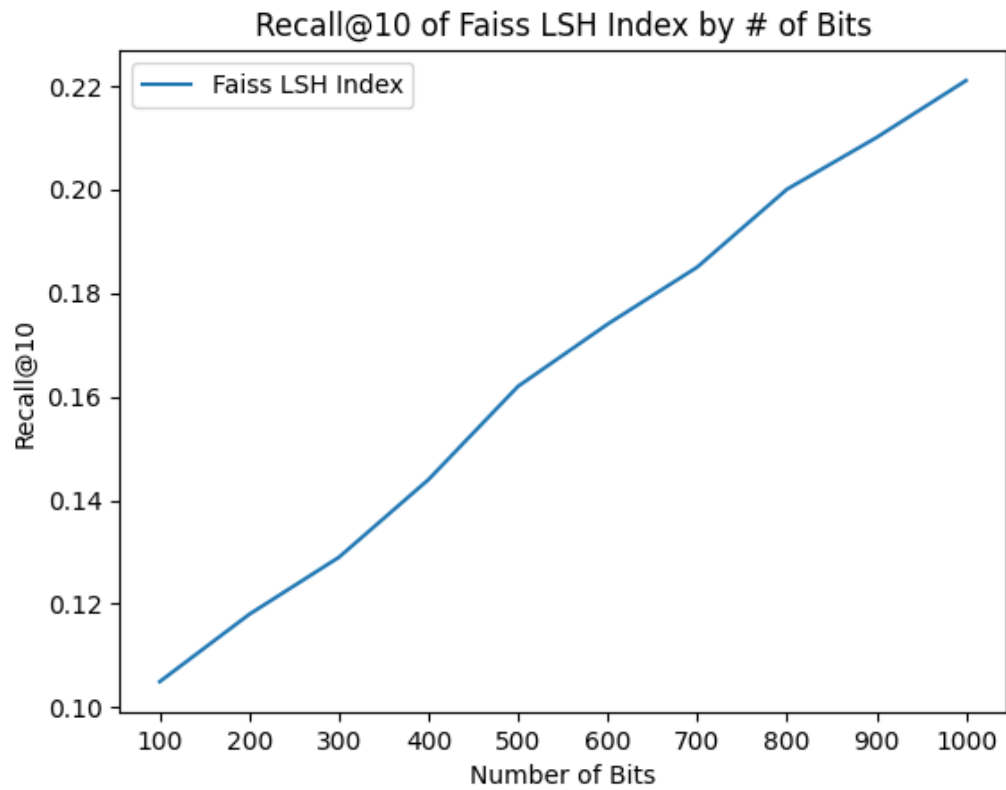


Figure 1.2.3: *Recall@10* by Number of Bits

2 Implementing an Index

2.1 LSH vs. Native Exhaustive Search

1. Running time of the `semi_optimized_exhaustive_search` function was: **4.14 seconds**.
2. Running time of building the LSH index was: **0.62 seconds**.
3. Running time of LSH search over query vectors was: **0.13 seconds**.
4. *Recall@10* for the LSH index was: **0.138**.

2.2 Custom Indexing Algorithm

Implementation Description

For our solution, we decided to implement an index based on the HNSW algorithm we saw in the lecture. Our intuition for this choice comes from the assumption that the HNSW is a very fast algorithm based on the skip-list data structure, and yet it is more accurate than the LSH algorithm.

We supported this intuition by testing the implementation of the HNSW algorithm in the Faiss library, and found that an ideal implementation can reach the 0.8 recall@10 mark in a very short time.

For our implementation, we...

Performance Measures

- **creating custom indexing algorithm:** x seconds, half the time of section 2.1.2
- **searching with custom indexing algorithm:** x seconds, a third of the time of section 2.1.3
- **recall@10 for custom indexing algorithm:** x, at least 0.8

3 Pinecone VectorDB and RAG

3.1 Dataset Selection

We opted to use the Stanford Question Answering Dataset (SQuAD) as the dataset for this task. It consists of questions posed by crowdworkers on a set of Wikipedia articles, with corresponding reading passages. In total, SQuAD 1.1 contains 100,000+ question-answer pairs on 500+ articles.

The segments of contexts are repeated across questions, and after eliminating duplicates, we are left with 20,958 unique contexts.

The answers for the questions are usually short, precise answers that can be found in the context, which makes them potentially easier to compare with the generated answer. However, there are also some questions that do not have an answer in the context.

3.2 Standard Model Evaluation

The standard QA model from cohere may fall short in answering factual questions in specific domains it might have not been exposed to enough. We present a few anecdotal examples of such cases:

1. **How many student news papers are found at Notre Dame?** The correct answer is 3, but the model answered: "There are two student newspapers at the University of Notre Dame: 1. The Observer: This is the main student newspaper on campus and is published daily during the academic year. It covers a wide range of topics, including news, sports, opinion, and entertainment. 2. Irish Informer: This is a newer, independent student-run newspaper that was founded in 2013. It publishes weekly and focuses on providing an alternative perspective on campus news and events. Both newspapers are valuable sources of information and showcase the diverse talents and interests of the student body at Notre Dame."

This is an example of a factual error based on lack of accurate or current information.

2. q2

3.3 RAG Pipeline

3.4 Insights