CMP4030: Advanced Databases
Computer Engineering Department

Semantic Search Engine with Vectorized Database

(Project Proposal)

Team 1

Ahmed Hamdy
Ahmed Aladdin
Asmaa Mohammed
Shehab Khaled

Proposed To

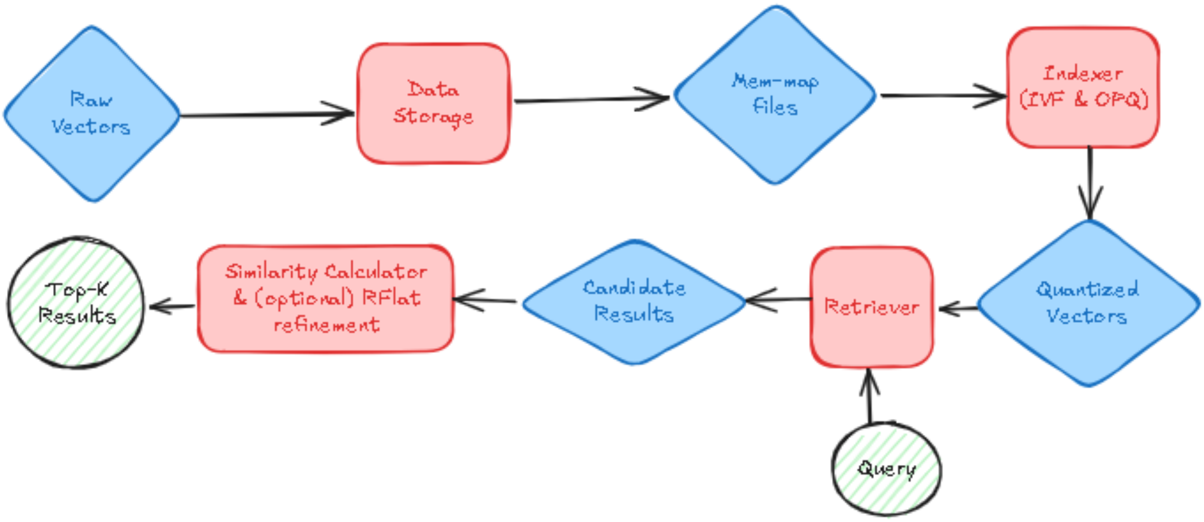
Eng. Abdelrahman Kaseb

Introduction

Using brute force in similarity search is infeasible for large datasets. Instead, we will use **Inverted File with optimized product quantization *(IVF-OPQ)***, one of the most efficient **Approximate Nearest Neighbor *(ANN)*** methods, as it provides the best trade-off for this project's constraints.

Proposed Methodology

Our system will combine **Inverted File (IVF)** for coarse partitioning, **Optimized Product Quantization (OPQ)** for fine-grained encoding, and memory-mapped files for index storage. We may use **Residual Flat refinement *(RFlat)*** as a postprocessing reranking step if there is a significant degradation in the recall.



Indexing Pipeline

Our indexing pipeline will mainly include these two stages:

1. **OPQ transformation:** transform training vectors (OPQ)
2. **Coarse partitioning**: organize vectors to sub-domains (IVF).
3. **Fine-grained encoding:** compress residuals within each cluster (PQ) after applying OPQ rotation.

Retrieval Pipeline

Our retrieval pipeline will consist of the following:

1. **OPQ rotation** → rotated query vector
2. **IVF coarse search** → top coarse centroid IDs with corresponding distances
3. **PQ fine search** → top L candidates (where L>= k)
4. **RFlat reranking (optional)** → final top K results

References

- Vector Indexing Techniques
- Product Quantization in Vector Search
- Optimized Product Quantization (OPQ)