שמאטע (Rag in Yiddish)

http://exodao.net

SCHMATE (pronounced SHMAH-teh) : Extending re-Isearch with vector datatypes for embeddings.

May 2024 Draft 1.2

Re-Isearch is a 100% open source novel multimodal search and retrieval engine using mathematical models and algorithms different from the all-too-common inverted index. It is a kind of hybrid between full-text, XML, object and graph noSQL-db that natively ingests a wide range of document types and formats. It has been open-sourced through a grant from NLnet/NGI-Zero Search. See our talk from FOSDEM '22: A lightning intro to re-Isearch

It has polymorphic object types allowing for search as text but also as a large number of datatypes including: numerical, computed, range, date, time, geo, boolean etc. as well as a number of hashes including several phonetic. These datatypes have their own, for their individual datatypes, storage and retrieval algorithms (including relevant ranking and similarity methods). Project Schmate intends to extend re-Isearch with a flat vector datatype tuned for embeddings.

This new datatype is intended to provide a powerful alternative to popular vector databases like FAISS for Dense Passage Retrieval (DPR) in, among other domains, Retrieval Augmented Generation (RAG) applied to popular open source context constrained large language models. LLaMa/LLaMa2/LLaMa3 have, for example, a relatively small 2k, resp. generally 4k and 8k context while Mistral-7B has a context length of 32k, This is still too small to be able to include the whole local or updated corpus but only some bits (passages). RAG (Retrieval Augmented Generation) is a means to try to maneuver out of this constraint but because of the fixed unit of retrieval in typical vector databases used for DPR (Dense Passage Retrieval) they demand a prior segmentation of content into size constrained blobs or passages.

This is where re-Isearch and its proposed new datatypes and extensions enter. Since re-Isearch has a fully dynamic unit of retrieval, definable at search time or by heuristic, its simplifies the creation and maintenance of DPR systems and provides a significant advantage for, among other applications, RAG. There are, of course, a myriad of other uses.

While the current re-Isearch can be used for vector search of embeddings without an internal datatype optimized for the function using a bi-encoder architecture to encode queries and nodes into the desired dense vectors its performance is sub-optimal on this specific task to some of the vector engines such as Meta's FAISS with its dedicated vector store. The herein proposed datatype should provide significant performance improvements at least en-par with, for example, FAISS or uSearch.

It's one thing to have a good search but it must also be economically feasible, efficient and sustainable. Re-Isearch has been designed to compile and run efficiently on pretty much any platform, even small embedded systems. By constraining computational requirement we can leverage existing hardware and adhere to our aim to eliminate the need for costly data centers and reduce the environmental impact for search.

While we would delight in developers joining our ExoDAO moon-shot, the project is focused on re-Isearch and is fully de-coupled from a dependence upon participation and 100% useable in and of itself.

And, as always: ***100% self-contained (no hidden API demands to some cloud service) true open source with Apache 2.0 license on our software components and CC BY 4.0 on documentation.***

Contact Edward C. Zimmermann. Jakob-Klar-Str. 8, D-80796 Munich