Universitas Indonesia Library >> UI - Skripsi Membership

Building a Retrieval Module for a Retrieval-Augmented Generation (RAG) System For Data Discovery Queries = Membangun Modul Retrieval untuk Sistem Retrieval-Augmented Generation (RAG) untuk Menjawab Pertanyaan Data Discovery

David Alexander, author

Deskripsi Lengkap: https://lib.ui.ac.id/detail?id=9999920568186&lokasi=lokal

---

Abstrak

Data discovery is a problem where analysts spend more time looking for relevant data than analyzing it. To address this issue, solutions such as Aurum and Ver have been created to capture relationships between multiple data sources. However, these solutions only look at the data and do not capture any context humans might have in mind when producing or consuming data. To solve the issue of extracting and storing information from users, two problems need to be solved: getting data creators to document their data and creating a solution for storing and retrieving relevant content regarding a user query about data. This paper focuses on the second part of the problem, where a system to store and retrieve relevant content about data is created, which will be created as a retrieval-augmented generation (RAG) system. A retrieval module is created using a vector store index where documents are turned into vector embedding. When a query comes, the same model is used to turn the query into a vector embedding. The similarity of query and document embedding are compared and the top-k most similar document to the query is returned. Multiple embedding models are chosen and evaluated based on index-building times, hit rate, mean reciprocal rank (MRR), and query times. Smaller, retrieval-focused, and correctly languaged models like `bge-small-en-v1.5` are recommended for efficient index-building and query performance while offering competitive hit rates and MRR scores. Larger models do not necessarily offer better retrieval quality. Using a multilingual model in a case where only one language is needed produces bad results, making model specialization crucial for optimizing performance.

.....Data discovery adalah masalah di mana analis menghabiskan lebih banyak waktu untuk mencari data yang relevan daripada menganalisisnya. Untuk mengatasi masalah ini, solusi seperti Aurum dan Ver telah dibuat untuk menangkap hubungan antara berbagai sumber data. Namun, solusi ini hanya melihat data dan tidak menangkap konteks yang mungkin dimiliki manusia saat memproduksi atau mengonsumsi data. Untuk menyelesaikan masalah ekstraksi dan penyimpanan informasi dari pengguna, dua masalah perlu diselesaikan: membuat pembuat data mendokumentasikan data dan menciptakan solusi untuk menyimpan dan mengambil konten yang relevan terkait dengan pertanyaan pengguna tentang data. Makalah ini berfokus pada bagian kedua dari masalah tersebut, yaitu menciptakan sistem untuk menyimpan dan mengambil konten yang relevan tentang data, yang akan dibuat sebagai sistem retrieval-augmented generation (RAG). Modul retrieval dibuat menggunakan indeks penyimpanan vektor di mana dokumen diubah menjadi embedding vektor. Ketika sebuah pertanyaan datang, model yang sama digunakan untuk mengubah pertanyaan menjadi embedding vektor. Kesamaan antara embedding pertanyaan dan dokumen dibandingkan dan dokumen paling mirip ter-k dengan pertanyaan dikembalikan. Beberapa model embedding dipilih dan dievaluasi berdasarkan waktu pembuatan indeks, hit rate, mean reciprocal rank (MRR), dan waktu pengambilan data. Model yang lebih kecil, berfokus pada pengambilan, dan berbahasa yang benar seperti bge-small-en-v1.5 direkomendasikan untuk pembuatan indeks dan kinerja pertanyaan yang efisien sambil

menawarkan hit rate dan skor MRR yang kompetitif. Model yang lebih besar tidak selalu menawarkan kualitas pengambilan yang lebih baik. Menggunakan model multibahasa dalam kasus di mana hanya satu bahasa yang dibutuhkan menghasilkan hasil yang buruk, membuat spesialisasi model sangat penting untuk mengoptimalkan kinerja.