Voice-Interactive Semantic Search Interface with Vector Databases

Martin Bulín[1], Adam Frémund[2]

1 Introduction

Semantic searching offers significant advantages over full-text search, particularly because it allows users to formulate queries in natural language without needing to know the precise indexed key phrases. By using vector databases that store and index data as high-dimensional vectors, we can search through large datasets in real-time. In this work, we present a custom web-based interface for state-of-the-art semantic search on arbitrary textual data. Additionally, we integrate our in-house speech technologies - ASR and TTS (Švec et al. (2022)) to enhance user interaction. The interface supports two modes: 1) Searching based on retrieval-augmented generation (RAG) with an LLM generating answers in a chat-like format, and 2) raw semantic matching with indexed data. In both modes, the original PDF file is shown and the exact source of the retrieved information is provided.

2 Methodology

The source of factual information consists of a collection of PDF documents. Following the recipe in Figure 1, these documents are parsed into individual contexts (chunks). Various methods for this parsing exist, including fixed-length or paragraph-based approaches. In this work, however, we experiment with a technique known as semantic chunking (Chase (2022)).

Once the contexts are extracted, we use the Czech SentenceTransformer (Reimers et al. (2019)) model `fav-kky/FERNET-C5` to encode them into vectors. Along with each vector, we retain the source file and position of each item to facilitate later sourcing. The data are then indexed into a vector database, with all processing performed offline up to this point.

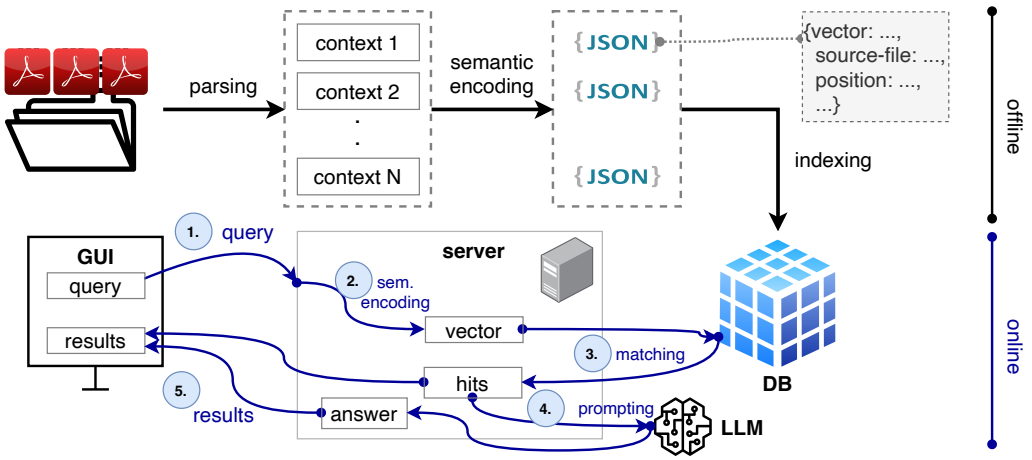

Figure 1: Pipeline of data offline processing and the online interaction with the interface.

[1] PhD student of Cybernetics, focused on Artificial Intelligence, e-mail: bulinm@kky.zcu.cz
[2] PhD student of Cybernetics, focused on Artificial Intelligence, e-mail: afremund@kky.zcu.cz

3 The Application

The online searching flow is illustrated at the bottom part of Figure 1, with the `GUI` block detailed in Figure 2. Here, in the top-left corner, users can select the collection they wish to search (the interface supports multiple collections from various domains). The top control panel allows users to activate RAG mode (enabling chatbot-like answers) and choose to use TTS to read the answers aloud. ASR can be activated by pressing `CTRL+Enter` at any time. The left part displays the user's queries and, if activated, the LLM's answers, allowing for up to three follow-up queries per search. The search results are sorted by cosine similarity score at the top of the left pane. The right side of the page features a PDF viewer, highlighting the matching context within the original file and, if available, the link to this file online.

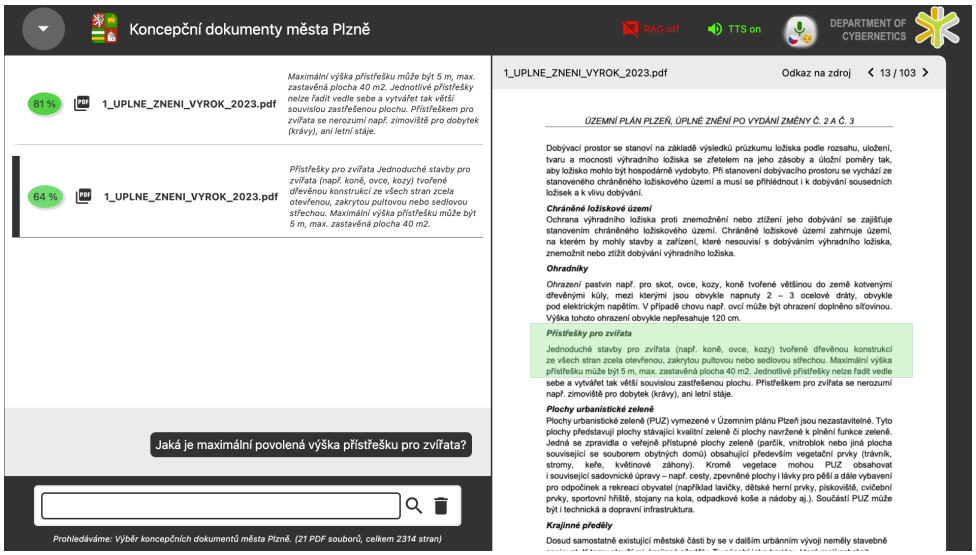

Figure 2: Graphical visualization of the semantic search interface.

The application is built on React with static generation and a Python backend server. The procedure is general and can be applied to various tasks. Currently, our demo includes: 1) Regulation rules of the University of West Bohemia, 2) Collection of laws of the Czech Republic, and 3) Conceptual documents of the city of Pilsen. The system currently supports searching in Czech and English documents. In the future, we plan to extend the interface to multiple modalities, enabling semantic searching in audio, video, and images.

Acknowledgement

This work was supported by the UWB grant, project No. SGS-2022-017.

References

Chase, H. (2022) *LangChain*. `https://github.com/langchain-ai/langchain`. Released [2022-10-17]

Švec J. at al. (2022). Multi-modal communication system for mobile robot. IFAC-PapersOnLine, 55(4), 133-138.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.