



# **Voice-Interactive Semantic Search Interface with Vector Databases**

Martin Bulín<sup>1</sup>, Adam Frémund<sup>2</sup>

### **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 highdimensional 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 retrievalaugmented 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.

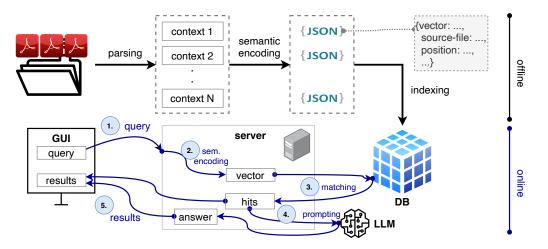


Figure 1: Pipeline of data offline processing and the online interaction with the interface.

<sup>&</sup>lt;sup>1</sup> PhD student of Cybernetics, focused on Artificial Intelligence, e-mail: bulinm@kky.zcu.cz

<sup>&</sup>lt;sup>2</sup> 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.

Koncepční dokumenty	města Plzně		🜓 TTS on	<sup>ام</sup> 🚱	EPARTMENT OF	*
(813) 🖾 1_UPLNE_ZNENI_VYROK_2023.pdf	Maximální výška přistřešku může byl 5 m, max. zastavěná plocha 40 m2. Jednotlivé přístřešky nelze řadit vedle sebe a vytvářet tak větší souvislou zastřešneno plochu. Přístřeškem pro zvířata se nerozumí např. zimoviště pro dobylek (krávy), ani letní stálje.	1_UPLNE_ZNENI_VYROK_2023.pdf		Odkaz na zd	iroj 🕻 13 / 103	>
		ÚZEMNÍ PLÁN PLZEŘ, ÚPLN	IÉ ZNĚNÍ PO VYDÁ	ÁNÍ ZMĚNY Č. 2 A	Č. 3	
64 % 🖾 1_UPLNE_ZNENI_VYROK_2023.pdf	Přístřešky pro zvřlata Jednoduché stavby pro zvřlata (např. koně, ovce, kozy) tvořené dřevěnou konstrukcí ze všech stran zcela otevřenou, zkatýrou puřítovou enbo sedlovou střechou. Maximální výška přístřešku může být 5 m, max. zastavěné plocha 40 m2.	Dobývací prostor se stanoví na základ Vysledků průzkrumu tobkaka podle rozahu, ukdení, hran a monosti vývřahního bůžka se zřetkom ma jaho zásdov jučkný prostoru se vychází za aby ložisko mohlo být hospodárně vydobýto. Při stanovení dobývacího prostoru se vychází za stanoveného chrálného bůžislového území a musí se přihlédnout i k obbývání sousedních kožete k skvu dobývaní.				
		Chránofe ložistová území Ochrane výmtachho ložiska proti znemožnéní nebo zližení jeho dobývání se zajištýje stanovením chránéhoho ložiskového uzemí. Chránnéh oblakové území zahrnuje území, na kterém by mohy stavý za zližení, tiste nesovské s dobýváním výhradního ložiska, znemožní nebo zliži dobývání výhradního ložiska. Ohradníhy				
		Ohrazení pastvín např. pro skot, ovce, kzoy, kond tvořené věšlisnou do zamě kotvenými ditervánými kůty, mazů teterými jsou obvykén nagenuty 2 – 3 ocebové držu, obvyké pod elektrickým napělitím. V případě chovu např. ovce může být ohrazení doplného siťovnou. Výkka lohoto ohrazení udvykén enpřesatvija 120 cm.				
		Přístřešky pro zvířata Jednoduché stavby pro zvířata (např. koně, ovce, kozy) tvolené dřevěnou konstrukci zv všech stran zcela devřenou, zakrydou půtlovou nebo sedlovou střechou. Nacimální výška přístřeške může být 5 mar. zastavěné půchat do filo. Jednotlevě přístřešky nebe řatá vede				
			ešenou plochu. Přís	enou plochu. Přístřeškem pro zvířata se nerozumí ní stále.		
Jaká je maximální povolená výška přístřešku pro zvířata?		Plocity urbanicisie is aknet Plochy urbanicisie a sknet pickty pleditivný ji lacity stávující kvalimili zelené (3 plocity narkzené k piňeli funkce zelené. Jedná se zpravádla o velpeň plintupné pickty narkzené (patřík, urbanicki nebo jiná plocha související se souborem odyvých domů) obsahující pledeválm vegetační právy (třahník, storom, teně, k kelitova zálkovy), Kronen vegetačen mároku PUL2 obsahovat pro obpočnek a retvrasi obyvátě (narkžel kalitik, cividení) prová, postorní hřítěk, storyna k kolitik, dostáve kole a nátědový ji), Soudistí PUL2 může				
		být i technická a dopravní infrastruktura.				
Prohledáváme: Výběr koncepčních dokumentů města Pizr	A (21 BDE soubor) collear 2214 strant	Krajinné předěly				
romedavame. vyser koncepenien dokumenta mesta Pizi	te. (277 b) bobbord, centern 2314 strain	Dosud samostatně existující městské čá				

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.