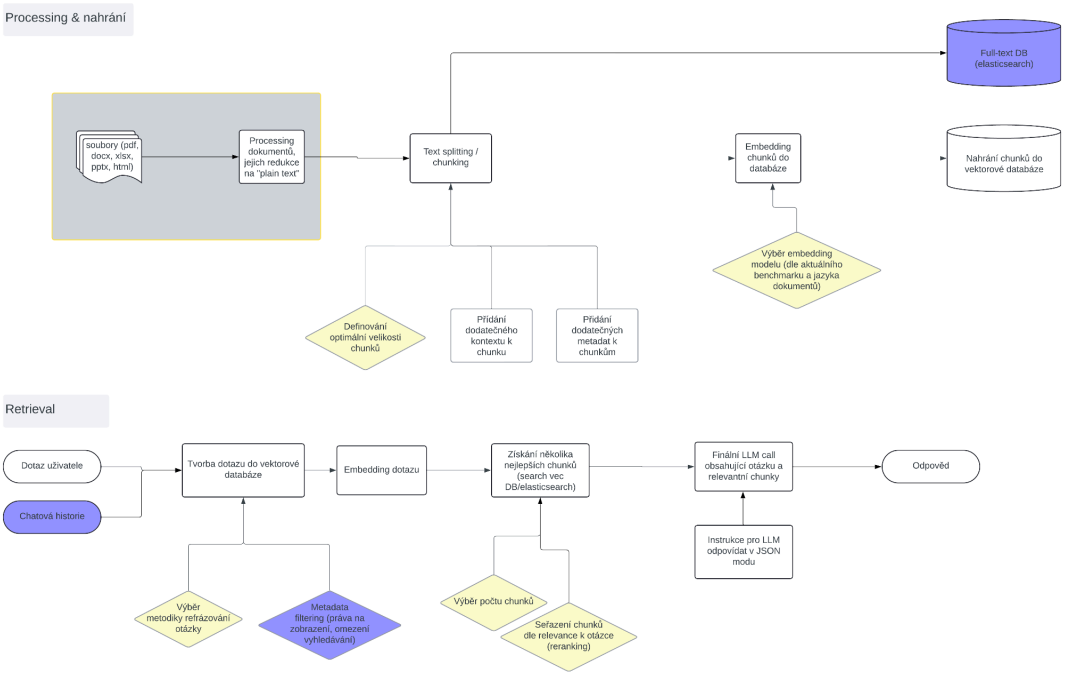
Doprovodný text k zadání bakalářské práce



Úvod

Tato bakalářská práce se zaměřuje na zpracování a retrieval dokumentů pomocí vektorových databází a language models (LLM). Cílem práce je vytvořit efektivní systém pro vyhledávání a zpracování informací z různých typů dokumentů, jako jsou PDF, DOCX, XLSX, PPTX a HTML.

Zpracování a nahrání dokumentů

V první fázi se soubory z různých formátů zpracovávají a redukují na "plain text". Tento krok zahrnuje extrakci textu a jeho přípravu pro další zpracování. Z pohledu bakalářské práce je toto vyřešeno a soubory jsou již redukovány na plain text (dodá Gauss Algorithmic).

Poté následuje text splitting nebo-li chunking, kde se text rozděluje na menší části (chunks), které jsou optimální pro zpracování embedding modely.

Následující kroky zahrnují:

1. **Definování optimální velikosti chunků**: Zajišťuje, aby byly chunky dostatečně malé pro efektivní zpracování, ale dostatečně velké, aby zachovaly kontext (dodá Gauss Algorithmic).
2. **Přidání dodatečného kontextu k chunkům**: Zlepšuje sémantické pochopení jednotlivých chunků (např. zda se jedna o nadpis, název souboru, obsah atd.).

3. **Přidání dodatečných metadat k chunkům**, zejména link na původní dokument, případně pozici textu v něm.

Po přípravě textových chunků jsou tyto chunky vloženy do embedding modelu, který je převádí na vektorové reprezentace. Výběr (dodá Gauss Algorithmic) embedding modelu je proveden na základě aktuálních benchmarků a jazyka dokumentů . Nakonec jsou vektorové reprezentace nahrány do vektorové databáze. Rozšířením může být uložení chunků v originální textové podobě do fulltextové DB (Elasticsearch).

Retrieval

Ve druhé fázi, retrieval, je uživatelův dotaz převeden do vektorové databáze. Tento proces zahrnuje:

1. **Tvorba dotazu do vektorové databáze**: Dotaz je převeden do formy, která je kompatibilní s vektorovou databází.
2. **Embedding dotazu**: Stejně jako u dokumentů je dotaz převeden na vektorovou reprezentaci.
3. **Získání několika nejlepších chunků**: Systém získává nejrelevantnější chunky, které odpovídají dotazu. Rozšířením může být získání chunku i z fulltextové databáze.

Následuje výběr počtu chunků a jejich seřazení podle relevance k dotazu (reranking - předpokládá se použití open-source knihovny, která relevanci řeší). Finální LLM call pak zahrnuje otázku a relevantní chunky, díky čemuž poté LLM poskytne odpověď uživateli..

Volitelné kroky

- **Chatová historie**: Může být využita k obohacení dotazu (doporučeno).
- **Metadata filtering**: Práva na zobrazení a omezení vyhledávání mohou být aplikována na dotazy. Ne každý uživatel může mít přístup ke každému dokumentu.

Závěr

Tato práce se zaměřuje na vytvoření komplexního systému pro zpracování a retrieval dokumentů pomocí moderních technologií NLP a vektorových databází.