Use-case

Návrh projektu: Implementace lokálního jazykového modelu pro potřeby KÚ Středočeského kraje

Vypracoval: Adam Seifert

Pozice: Uchazeč o pozici Specialista pro vývoj interních AI modelů (Junior)

Úvod

Tento dokument popisuje technický a procesní postup, jakým bych postupoval při zavádění interního AI modelu na Krajském úřadě. Hlavním cílem je vytvořit systém, který pomůže úředníkům s prací nad citlivými dokumenty (smlouvy, usnesení, metodiky), aniž by jakákoliv data opustila interní síť úřadu. Jako vývojář se v návrhu zaměřuji především na využití prověřených open-source nástrojů a technologii RAG (Retrieval-Augmented Generation).

---

1. Bodový postup řešení

Můj postup by se skládal z následujících kroků:

1. **Příprava lokálního prostředí:** Instalace základního operačního systému a zprovoznění ovladačů pro grafické karty (NVIDIA CUDA), které jsou pro běh AI nezbytné.
2. **Výběr a testování modelu:** Stažení open-source modelů (např. **Llama 3.1 8B** nebo **Mistral**) a jejich testování v lokálním prostředí pomocí nástrojů jako **Ollama** nebo **vLLM** pro ověření rychlosti odpovědí.
3. **Příprava datové pipeline (Ingestce):** Vytvoření skriptů v Pythonu, které „přečtou“ dokumenty z vybraných složek, rozdělí je na menší části (chunks) a převedou je do formátu, kterému AI rozumí (vektorizace).
4. **Vytvoření RAG systému:** Propojení modelu s vektorovou databází. To zajistí, že se model nebude „vymýšlet“, ale bude odpovídat pouze na základě dokumentů KÚ.
5. **Vývoj jednoduchého webového rozhraní:** Vytvoření interní aplikace (např. v **Streamlit** nebo **Gradio**), kde budou úředníci moci pokládat dotazy.
6. **Ladění a zpětná vazba:** V úzké spolupráci se zkušenějšími kolegy z IT a vybranými uživateli z odborů budu ladit přesnost odpovědí.

---

2. Co budu k projektu potřebovat (znalosti a technologie)

Abych mohl projekt úspěšně realizovat, budu pracovat s následujícím stackem:

Technologie a Software:

- **Jazyk Python:** Hlavní nástroj pro veškerý vývoj, práci s daty a propojení modelů.
- **Knihovny LangChain nebo LlamaIndex:** Jsou to standardní nástroje pro stavbu AI aplikací. Pomáhají s „lepením“ modelu, databáze a uživatelského vstupu dohromady.
- **Vektorová databáze (např. Qdrant nebo ChromaDB):** Slouží k ukládání obsahu dokumentů v číselné podobě pro rychlé vyhledávání.
- **Docker:** Pro snadné nasazení a izolaci aplikací v rámci serveru.

Infrastruktura (Hardware na KÚ):

- **GPU (Grafická karta):** Pro juniorní start a testování stačí výkonnější spotřebitelská karta (např. NVIDIA RTX 4090), ale pro ostrý provoz pro více lidí bude potřeba serverová karta (např. NVIDIA A100 nebo L40S) s dostatečnou pamětí (VRAM), aby se tam model „vešel“.
- **Server:** Alespoň 64 GB RAM a moderní vícejádrový procesor pro obsluhu datových toků.
- **Síť:** Úplné odříznutí od internetu. Veškeré knihovny a modely se stáhnou jednou a poté bude server fungovat v izolovaném režimu.

---

3. Konkrétní use-case: Rychlá rešerše v usneseních Rady kraje

Problém: Úředník potřebuje zjistit, jak se v minulosti rozhodovalo o dotacích na opravy památek v určitém regionu. Musí prohledávat stovky PDF souborů, což trvá hodiny.

Řešení: Vytvořím „AI Archiváře“. Úředník se zeptá: *„Jaké dotace na opravy kostelů jsme schválili v roce 2023 na Kutnohorsku?“*

1. Systém prohledá vektorovou databázi všech usnesení.
2. Najde relevantní odstavce, kde se mluví o „Kutné Hoře“, „opravách“ a „církevních stavbách“.
3. Tyto odstavce předhodí modelu společně s dotazem.
4. Model vypíše: *„V roce 2023 byly schváleny tři dotace: 1. Oprava střechy v obci X (500 tis. Kč), 2. Fasáda kostela v obci Y (1 mil. Kč)... Zdroj: Usnesení č. 456/2023.“*

Python

```
# Technical example: Initializing a local embedding model
# This runs locally on our GPU without external API calls.
from langchain_community.embeddings import HuggingFaceBgeEmbeddings

model_name = "BAAI/bge-small-en-v1.5" # Can be replaced with a Czech-
optimized model
encode_kwargs = {'normalize_embeddings': True}

# Setting up the embedding function for our local vector store
embed_model = HuggingFaceBgeEmbeddings(
    model_name=model_name,
```

```
    model_kwargs={'device': 'cuda'}, # Using local GPU
    encode_kwargs=encode_kwargs
)
```

---

4. Tři hlavní rizika a jejich řešení

1. **Model si vymýšlí (Halucinace):**
    - *Riziko:* AI odpoví sebejistě, ale nepravdivě.
    - *Řešení:* Nastavím systém tak, aby model mohl čerpat **pouze** z nahraných dokumentů (tzv. Temperature 0). Pokud informaci v dokumentech nenajde, musí odpovědět: „Omlouvám se, ale v interních podkladech jsem tuto informaci nenašel.“
2. **Hardware nestíhá (Vysoká latence):**
    - *Riziko:* Na odpověď se čeká minutu, což úředníky odradí od používání.
    - *Řešení:* Použiji techniku zvanou **kvantizace** (zmenšení modelu). Tím se model zrychlí a zabere méně paměti na grafické kartě při zachování 95 % přesnosti.
3. **Bezpečnost a přístupová práva:**
    - *Riziko:* Úředník z jednoho odboru se skrze AI dostane k citlivým datům jiného odboru, ke kterým nemá mít přístup.
    - *Řešení:* Do vektorové databáze uložím ke každému kousku textu i informaci o tom, z jakého odboru pochází. AI pak bude filtrovat výsledky podle toho, kdo je zrovna přihlášený (např. přes LDAP/Active Directory).

---

5. Časový rámec (Odhad juniora)

Tento plán počítá s tím, že se budu muset s některými systémy KÚ seznámit za běhu.

- **1. měsíc:** Seznámení se s IT infrastrukturou KÚ, instalace serveru, rozchození prvního modelu „nanečisto“.
- **2. měsíc:** Sběr dat z prvního odboru (např. Legislativa) a testování úspěšnosti vyhledávání v jejich dokumentech.
- **3. měsíc:** Vývoj jednoduchého webového rozhraní pro testovací skupinu uživatelů.
- **4.–5. měsíc:** Ladění systému na základě reálných dotazů, opravování chyb, práce na zabezpečení.
- **6. měsíc:** Oficiální spuštění pro první odbor a příprava na rozšíření pro zbytek úřadu.

Závěr

Mým cílem není vytvořit „vševědoucí AI“, ale praktického pomocníka, který ušetří kolegům na úřadě nudnou práci s dohledáváním informací. Jako junior se budu soustředit na stabilitu, bezpečnost dat a postupný rozvoj systému pod dohledem zkušenějších kolegů. Věřím, že on-premise řešení je pro instituci typu Krajského úřadu jedinou správnou cestou, jak využít AI a přitom neriskovat únik citlivých informací.