|  |  |
| --- | --- |
| Knowledge BAse assistent  7.11.2023 | Abstract  Řešení RestAPI služby, která implementuje funkcionalitu pro potřebu chatbotů založených na externích zdrojích dat. Služba dokáže podporovat více chatbotů založených na různých zdrojích dat. Součástí řešení je i příprava externích dat pro potřeby této služby.  Vladimír Dědourek, Michal Stoklasa |

Obsah

[1 Portál VŠ 2](#_Toc141258336)

[1.1 Cíle řešení 2](#_Toc141258337)

[2 ChatBot 2](#_Toc141258338)

[2.1 Analýza struktury portálu 3](#_Toc141258339)

[2.1.1 Struktura souboru VŠ 3](#_Toc141258340)

[2.1.2 Orientační dotazy studentů 5](#_Toc141258341)

[2.2 Model embeddings 7](#_Toc141258342)

[2.2.1 Strukturální model textu 7](#_Toc141258343)

[2.2.2 Indexování strukturovaného textu 10](#_Toc141258344)

[2.2.3 Příklady operací v DB Qdrant 12](#_Toc141258345)

[2.3 Dotaz odpověď 14](#_Toc141258346)

[2.3.1 Struktura dat v DB 14](#_Toc141258347)

[2.3.2 Konstrukce promptu pro výběr textů na základě otázky 16](#_Toc141258348)

[2.3.2.1 Určení parametrů dotazu 17](#_Toc141258349)

[2.3.2.2 Určení přesné kategorie z uživatelem zadaných dat 18](#_Toc141258350)

[2.3.2.3 Sestavení podmínky pro výběr dat operací search z DB Qdrant 19](#_Toc141258351)

[2.3.2.4 Konstrukce promptu 19](#_Toc141258352)

# Knowledge Base Assistent

## Cíle řešení

# KBAIndex

## Model embeddings



### Indexování strukturovaného textu

Pro indexování se berou texty z tabulky metadat. Tyto texty jsou rozděleny na menší segmenty (chunks) tak, aby maximální velikost nepřekročila maximální počet tokenu (max\_tokens).

Protože pouhým indexováním těchto segmentů (chunks) bez příslušnosti k názvu paragrafu (heading), by mohlo dojít ke ztrátě významové souvislosti, Jsou názvy paragrafu připojeny vždy na začátek jednotlivých segmentů.

Při dělení na segment jsou nedělitelné jednotky celé věty. V případě, že by se věta nevešla do segment, je provedeno dělení na úrovni slov. K rozdělení slov mezi segment nedochází.

Pro každý segment (chunk) jsou ve vektorové databázi přiřazena následující metadata.

# KBAQnA

# RestAPI server

# Přílohy

## Publikování RestAPI serveru

## Vytvoření webové aplikace v Azure

https://learn.microsoft.com/cs-cz/azure/app-service/quickstart-python?tabs=flask%2Cwindows%2Cazure-portal%2Cvscode-deploy%2Cdeploy-instructions-azportal%2Cterminal-bash%2Cdeploy-instructions-zip-azcli

Na webu Azure Portal:

1. Do vyhledávacího panelu v horní části Azure Portal zadejte *App Services*.
2. V nabídce, která se zobrazí pod panelem hledání, vyberte položku s popiskem **App Services** pod nadpisem **Služby** .

## Na stránce **App Services** vyberte **+ Vytvořit** a pak v rozevírací nabídce vyberte **+ Webová aplikace** .

Na stránce **Vytvořit webovou aplikaci** vyplňte formulář následujícím způsobem.

1. **Skupina prostředků** → Vyberte **Vytvořit nový** a použijte název *msdocs-python-webapp-quickstart*.
2. **Název** → *msdocs-python-webapp-quickstart-XYZ* , kde XYZ jsou libovolné tři náhodné znaky. Tento název musí být v rámci služby Azure jedinečný.
3. **Zásobník modulu runtime** → **Pythonu 3.11 (mělo by se shodovat s verzí ve Visual Studiu)**.
4. **Oblast** → Libovolná oblast Azure ve vaší blízkosti.
5. **App Service Plán** → V části **Cenový plán** vyberte **Prozkoumat cenové plány** a vyberte jiný App Service plán.

Plán App Service řídí, kolik prostředků (procesor/paměť) je pro aplikaci k dispozici a náklady na tyto prostředky.  
  
V tomto příkladu vyberte v části **Vývoj/testování** plán **Basic B1** . Za plán Basic B1 se na váš účet Azure budou účtovat malé poplatky, ale doporučuje se pro lepší výkon oproti plánu Free F1.  
  
Až budete hotovi, vyberte **Vybrat** , aby se změny použily

A screenshot of a computer

Description automatically generated

Dále jsem propojil na Github vdedourek2, kde je namapované celé solution PythonApplication.

## Nasazení kódu aplikace do Azure

# Souhrn

Výpočetní čas se převážně spotřebovává při zjištění parametrů dotazu a potom i při generování odpovědi. Vyhledání ve vektorové DB je zpravidla menší než 1 s.

Při zpracování otázky, kde data jsou v mnoha segmentech, nejsou odpovědi kompletní nebo správné. Důvodem je to, že data pro vytvoření vhodného kontextu jsou příliš rozsáhlá a přesahují povolený limit tokenů (viz <https://platform.openai.com/docs/models/overview>).

Ve zdrojových datech chybí informace o studovaných předmětech jednotlivých studijních programů. V případě seznamu studijních předmětů není spolu s názvem uveden i kód předmětu.