Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra kybernetiky

Vektorový model pro vyhledávání informací

Numerické metody
KMA/NM

David Žahour
19.3.2023
A19B0334P

1 Úvod

Cílem této práce bylo vytvořit program, který umožní načítání souborů v HTML formátu a následné vektorové vyjádření těchto souborů pomocí metody IDF. Tento program také umožňuje načítání vyhledávacích dotazů a poskytuje seznam dokumentů relevantních k těmto dotazům.

2 SentenceTransformers pro zlepšení vyhledávání

Pro zlepšení výsledků vyhledávání jsme v této práci využili knihovnu SentenceTransformers. Tato knihovna umožňuje vytváření embedingů významu textových dokumentů a jejich následné porovnání, což zajišťuje větší přesnost a relevantnost výsledků vyhledávání. Jednou z výhod tohoto přístupu je, že umožňuje lepší porovnávání podobnosti mezi texty, i když jsou psány odlišnými způsoby nebo používají jiné slovní spojení.

3 Funkce

V programu, který načítá soubory v HTML formátu a umožňuje vyhledávání v dokumentech, se využívá knihovna SentenceTransformers pro vytváření významových vektorů dokumentů a dotazů a porovnávání těchto vektorů pomocí kosinové podobnosti. Dále se využívá metoda IDF (inverzního dokumentového frekvence) pro vektorové vyjádření každého dokumentu a dotazu, a to pomocí metody transform z knihovny TfidfVectorizer. Nakonec se sčítají oba kosinové výsledky a výsledné skóre určuje relevanci každého dokumentu k danému dotazu. Seznam nejrelevantnějších dokumentů je řazen podle tohoto skóre a program vrátí výsledky vyhledávání.