Paweł Prus

IDE: Microsoft Visual Studio Code
Procesor: AMD Ryzen 7 6800h
Język: Python 3.10.12
Biblioteki:
aiohttp
aiofiles
asyncio
os
re
shutil
urllib.parse
bs4
numpy
nltk
collections

scipy

flask

Wyszukiwarka

1. Webcrawler (webcrawler.py)

1.1. Webcrawler jest odpowiedzialny za przeszukiwanie internetu w poszukiwaniu treści. Działa na zasadzie przeszukiwania wszerz, pobierając zawartość stron internetowych i indeksując je.

1.2. Całościowe działanie:

1.2.1. Webcrawler pobiera treść stron internetowych, wykorzystując asynchroniczne żądania HTTP przy użyciu biblioteki aiohttp.

1.2.2. Z pobranej treści usuwane są znaczniki HTML, a oczyszczony tekst jest zapisywany do pliku.

1.2.3. Crawler znajduje wszystkie linki w danej treści i dodaje je do kolejki przeszukiwania.

2. Bag of Words (bag_of_words.py)

2.1. Model Bag of Words (BoW) służy do przetwarzania tekstu, w którym analizowany jest jedynie zbiór unikalnych słów występujących w dokumencie, a ich kolejność nie ma znaczenia.

2.2. Całościowe działanie:

2.2.1. Program wczytuje dokumenty z plików tekstowych.

2.2.2. Tworzy unikalny zestaw słów występujących we wszystkich dokumentach.

2.2.3. Dla każdego słowa w zestawie przypisuje unikalny indeks.

2.2.4. Tworzenie macierzy termów: Generuje macierz, w której wiersze reprezentują dokumenty, a kolumny - słowa. Liczba w macierzy oznacza częstotliwość występowania słowa w danym dokumencie

3. Search Engine (search_engine.py)

3.1. Silnik wyszukiwarki używa modelu Bag of Words do porównywania zapytań użytkownika z dokumentami, aby znaleźć pasujące wyniki.

3.2. Całościowe działanie:

3.2.1. Wczytuje wcześniej zaindeksowane dokumenty oraz zestaw słów.

3.2.2. Zapytanie użytkownika jest przekształcane do postaci wektora Bag of Words.

3.2.3. Na podstawie zapytania użytkownika i dokumentów wyznaczane są podobne dokumenty.

3.2.4. Wyniki wyszukiwania są prezentowane w formie szablonu HTML.

4. Wnioski

4.1. Algorytm wyszukiwarki jest skuteczną metodą odnajdywania pasujących wyników w dużych zbiorach danych internetowych.

4.2. Użycie modelu Bag of Words pozwala na efektywne reprezentowanie dokumentów tekstowych i zapytań użytkownika.

4.3. Zastosowanie technik takich jak SVD (Singular Value Decomposition) i IDF (Inverse Document Frequency) może poprawić jakość wyszukiwania, zmniejszając wymiarowość danych i uwzględniając znaczenie poszczególnych słów w dokumencie.