Konstrukcja wyszukiwarki semantycznej z bazą oparta na wybranej tematyce z Wikipedii

WŚ

W obecnym świecie, gdzie zasoby informacji rosną w niezwykle szybkim tempie, narzędzia do skutecznego przeszukiwania i wyodrębniania kluczowych informacji z gąszczu danych tekstowych stają się nieocenione. Ten raport opisuje projekt skupiający się na stworzeniu semantycznej wyszukiwarki, która, wykorzystując techniki przetwarzania języka naturalnego (NLP), jest w stanie zrozumieć i uwzględnić znaczenia słów. Projekt wykorzystał cztery popularne techniki konstrukcji wektorów słów: Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), Bag-of-Words (BoW) oraz TF-IDF. Wszystkie one przyniosły zadowalające wyniki, co potwierdza efektywność takiego podejścia w kontekście przetwarzania języka naturalnego. Praca ta stanowi punkt wyjścia do dalszych badań w tej dziedzinie.

1. Wstęp

W erze cyfrowej, gdzie generujemy i konsumujemy ogromne ilości danych tekstowych, skuteczne przeszukiwanie i wyodrębnianie istotnych informacji staje się coraz bardziej istotne. W tym kontekście, przetwarzanie języka naturalnego (NLP), które umożliwia maszynom zrozumienie i manipulowanie językiem ludzkim, odgrywa kluczową rolę. Jednym z istotnych zadań w NLP jest konstrukcja wektorów słów, polegająca na transformacji słów i fraz w numeryczne wektory, które mogą być przetwarzane przez algorytmy uczenia maszynowego.

W moim projekcie skoncentrowałem się na problemie tworzenia semantycznej wyszukiwarki. Semantyczna wyszukiwarka jest zdolna do zrozumienia i uwzględnienia znaczenia słów podczas przeszukiwania zestawu danych tekstowych. Dzięki temu, nawet jeśli konkretne słowo nie występuje w dokumencie, wyszukiwarka może zidentyfikować dokumenty, które są istotne na podstawie semantyki słów, nie tylko ich dosłownego wystąpienia.

Do wykonania projektu wykorzystałem cztery popularne techniki konstrukcji wektorów słów: Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), Bag-of-Words (BoW) oraz TF-IDF.

Wybrałem te konkretne techniki ze względu na ich różnorodność i różne podejścia do problemu. LSA i PCA są technikami redukcji wymiarowości, które starają się uchwycić głębsze, semantyczne znaczenia w danych. Z kolei BoW jest prostszą techniką, reprezentującą dokumenty jako zbiór słów, niezależnie od ich kolejności i kontekstu. Wreszcie, TF-IDF jest techniką, która uwzględnia częstotliwość występowania słów w dokumencie, a także ich rzadkość w całym korpusie.

Moim celem było zrozumienie tych technik, ich potencjalnych zastosowań oraz ocena ich skuteczności w kontekście stworzonej przeze mnie semantycznej wyszukiwarki.

2. Rozwinięcie

Rozwiązanie problemu stworzenia semantycznej wyszukiwarki rozpoczęło się od odpowiedniego przygotowania danych. Zacząłem od pobrania artykułu z Wikipedii jako pliku HTML. Następnie usunąłem z niego niepotrzebne fragmenty i wyodrębniłem czysty tekst artykułu. Usunąłem również białe znaki i przypisy, a następnie przeprowadziłem proces tokenizacji tekstu. W dalszych krokach usunąłem stop words i interpunkcję, co pozwoliło mi na uzyskanie czystego zestawu tokenów do analizy.

Następnie skupiłem się na zastosowaniu różnych metod konstrukcji wektorów słów, które służyły do tworzenia reprezentacji semantycznej dokumentów. Wykorzystałem TF-IDF, LSA, PCA i BoW.

TF-IDF to skrót od częstości występowania słowa pomnożonej przez odwrotność częstości dokumentu. Częstość występowania słowa odnosi się do liczby wystąpień każdego słowa w dokumencie. Odwrotność częstości dokumentu oznacza, że podzielisz każdą z tych liczności słów przez liczbę dokumentów, w których słowo występuje.

Oznacza to, że TF-IDF to metoda oceny ważności słowa w dokumencie. Zasadniczo, im częściej słowo pojawia się w danym dokumencie i im rzadziej pojawia się w innych dokumentach, tym większą ma wagę według metody TF-IDF.

LSA to metoda przetwarzania języka naturalnego, która pozwala na analizę relacji między zestawem dokumentów a słowami, które one zawierają. Zasadza się ona na idei, że słowa o podobnym znaczeniu pojawiają się w podobnych kontekstach (tzw. hipoteza dystrybucyjna). W praktyce tworzy się macierz, gdzie wiersze reprezentują unikalne słowa, a kolumny - poszczególne dokumenty. Następnie, za pomocą rozkładu według wartości osobliwych (SVD), zmniejsza się liczbę wierszy, jednocześnie zachowując strukturę podobieństwa między kolumnami.

Analiza głównych składowych (**PCA**) to metoda redukcji wymiarowości, która jest często stosowana do zmniejszenia rozmiaru dużych zbiorów danych. Transformuje ona duży zestaw zmiennych w mniejszy, który nadal zawiera większość informacji z dużego zbioru.

Zmniejszenie liczby zmiennych w zbiorze danych naturalnie prowadzi do pewnej utraty dokładności, ale trikiem w redukcji wymiarowości jest zgodzenie się na nieznaczne pogorszenie dokładności w zamian za uproszczenie. Mniejsze zbiory danych są łatwiejsze do eksploracji i wizualizacji, a także ułatwiają i przyspieszają analizę punktów danych przez algorytmy uczenia maszynowego, eliminując zbędne zmienne do przetworzenia.

Model "bag-of-words" (**BoW**) to uproszczona reprezentacja stosowana w przetwarzaniu języka naturalnego i wyszukiwaniu informacji. W tym modelu tekst (tak jak zdanie lub dokument) jest reprezentowany jako torba (multizbiór) swoich słów, pomijając gramatykę i nawet kolejność słów, ale zachowując ich wielokrotność.

Dla każdej z tych metod, proces wyglądał podobnie. Na podstawie utworzonych tokenów przekształcałem zapytanie do przestrzeni wektorowej, obliczałem podobieństwo kosinusowe, sortowałem indeksy na podstawie wyników podobieństwa kosinusowego i wyświetlałem najlepsze wyniki. To podejście pozwoliło mi na porównanie efektywności poszczególnych metod.

Większość zastosowanych technik konstrukcji wektorów słów zwracała poprawne lub bliskie poprawnym wyniki. Jedynym wyjątkiem była analiza głównych składowych (PCA), gdzie musiałem dostosować liczbę komponentów z 2 do 50. Dla n_components równego 2 wynik nie był w żaden

sposób powiązany z zapytaniem, natomiast zwiększenie tej wartości do 50 dało wynik podobny do uzyskanego przez pozostałe metody.

Dzięki temu podejściu udało mi się z sukcesem zaimplementować semantyczną wyszukiwarkę, która skutecznie odnajdywała odpowiednie dokumenty na podstawie zapytania, niezależnie od zastosowanej metody konstrukcji wektorów słów.

3. Podsumowanie

W trakcie pracy nad tym projektem zaimplementowałem semantyczną wyszukiwarkę z wykorzystaniem różnych metod konstrukcji wektorów słów. Każda z metod - TF-IDF, LSA, PCA i BoW - przyniosła satysfakcjonujące wyniki, co potwierdza skuteczność takiego podejścia w przetwarzaniu języka naturalnego.

Analiza wyników pokazała, że większość zastosowanych metod zwracała poprawne lub bliskie poprawnym wyniki wyszukiwania. Oznacza to, że semantyczna wyszukiwarka działała zgodnie z oczekiwaniami, zwracając dokumenty o największym podobieństwie do wprowadzonego zapytania.

Warto jednak zauważyć, że zakres stosowanych metod nie był wyczerpujący. Istnieje wiele innych podejść do konstrukcji wektorów słów, które mogłyby być zastosowane i przetestowane w ramach tego projektu. Możliwe jest również, że wyniki mogłyby się różnić, gdyby przeprowadzić eksperymenty na większym zestawie danych. Moje obecne wyniki są więc punktem wyjścia do dalszych badań w tej dziedzinie, a nie ostatecznym rozwiązaniem problemu.

4. Bibliografia

- Hobson Lane, Hannes Hapke, Cole Howard - Natural Language Processing in Action Understanding, analyzing, and generating text with Python (2019)
- Radosław Kycia – materiały z zajęć Przetwarzanie Języka Naturalnego
- https://www.nltk.org/api/nltk.tokenize.html
- https://devdocs.io/scikit_learn/
- https://builtin.com/data-science/step-step-explanation-principal-component-analysis
- https://en.wikipedia.org/wiki/Bag-of-words_model