Władysław Nieć

Zadanie 6 – Wyszukiwarka artykułów z Wikipedii

W celu przygotowania programu pobrałem ponad 59000 artykułów Wikipedii z dziedzin:

- Matematyki
- Fizyki,
- Informatyki
- Astronomii.

Pobieranie artykułów trwało około 9 godzin.

Najpierw pobrałem same tytuły stron, a następnie pobierałem po 10 tekstów całych stron i zapisywałem do bag of words. Po pobraniu przekształciłem pobrane dane do macierzy słów word_matrix. Do jej obsługi utworzyłem słowniki mapujące indeksy słów i indeksy artykułów. Dokonałem zmiany indeksów w macierzy poprzez pomnożenie ich poprzez odpowiednią wartość Inverse Document Frequency. Wektory poziome tak przekształconej macierzy następnie znormalizowałem. Wszystkie te czynności można wykonać samemu, uruchamiając plik setup. Dokonałem kilku prób wyszukiwania na macierzy oryginalnej i jej przybliżeniach, dla rzędów 3, 30, 300 i 1000. Przybliżeń macierzy dokonywałem, rzutując ją na typ csr_matrix dostępny w bibliotece scipy.sparse i wykonując na niej dekompozycję spektralną dostępną w funkcji linalg.svds w tej samej bibliotece. Wyniki wyszukiwania okazały się niezadowalające dla przyblirzeń o niskim rzędzie. Natomiast wyniki dla oryginalnej macierz i tej o rzędzie 1000 okazały się porównywalne. Do porównywania wyszukiwanej frazy z zawartościami artykułów wykorzystywałem podobieństwo cosinus-owe. Moja aplikacja jest napisana całkowicie w języku Python. Wykorzystuję w niej bibliotekę graficzną PySimpleGui. Aby ją uruchomić, należy uruchomić plik main. Uruchamianie aplikacji powinno potrwać kilka sekund. Wyszukiwanie fraz zajmuje podobną ilość czasu. Hasła można wyszukiwać wpisując je w pole tekstowe aplikacji i klikając przycisk search lub wciskając klawisz enter.