Jakub Kruczek

Metody Obliczeniowe w Nauce i Technice

Laboratorium 6

Singular Value Decomposition

Sprawozdanie

1. **Opis**

A. **Dane**
Dane zostały przetworzone przez “Porter Stemming Algorithm”.

B. **Bag of words**
Zbiór jest słownikiem i posiada budowę: *“słowo”: <liczba_wystąpień>*

C. **Term by document**
Jest to “Set”, który posiada wszystkie słowa.

D. **Inverse document frequency i normalizacja**
Budowa: słownik *“słowo”: <wartość_idf>*.

E. **SVD i low-rank approximation**
Użyto wbudowanej funkcji w sciPy.

2. Przykładowe czasy działania:

```
/usr/bin/python3.4 /home/kruczjak/PycharmProjects/lab6-mownit/main.py
Czytanie i parsowanie
[####################]100.0% 999/999 time: 3343.4486389160156ms
Przetwarzanie inverse document frequency
[####################]100.0% 999/999 time: 3902.397871017456ms
Tworzenie macierzy A
[####################]100.0% 999/999 time: 5013.051986694336ms
SVD
Czas: 7924.123287200928ms
Gotowe. Pelny czas: 20185.07194519043ms
Wyszukaj: google

Normalizowanie
[####################]100.0% 999/999 time: 204819.69165802002ms
Dokladnosc: 0.0278708513156
Inne wyniki: [{0.011676458806956435}, {0.011676458806956435}, {0.027867243742592501}, {0.0278708
Uzywajac SVD:
[####################]100.0% 999/999 time: 11247.866868972778msDokladnosc: 0.000146391751467
Inne wyniki: [{0.00014211679236797402}, {0.00014217531524095471}, {0.00014639122605252129}, {0.0

Process finished with exit code 0
```

3. Wnioski

- Najdłuższym etapem działania algorytmu było obliczanie miary prawdopodobieństwa i normalizacja wektorów cech bag-of-words i wektora powstałego z zapytania do długości 1.
- Najkrótsze (oprócz wczytywania i parsowania plików) było mnożenie bag-of-words przez wyliczane *inverse document frequency.*
- Poprzez zastosowanie SVD w celu usunięcia szumów z macierzy A, udało się o wiele skrócić czas wyliczania prawdopodobieństwa.
- Dzięki normalizacji IDF zmniejszył się wpływ słów znajdujących się w większości tekstów na wpływ wyszukiwania (szczególnie przydatne, gdy zapytanie składa się z wielu słów.