Sprawozdanie 3 – Jan Górski

Metody Obliczeniowe w Nauce i Technice
Laboratorium 7
Singular Value Decomposition

Wybranym przeze mnie zbiorem dokumentów są 2613 książek, artykułów udostępnionych w ramach projektu Gutenberg. Pobrane zostały przy pomocy biblioteki dostępnej bilioteki gutenberg. Kod do pobierania dokumentów znajduje sie w pliku *download.py.*

Wyznaczony zbiór termów jest to iloczyn pierwszych 5000 słów (wcześniej zmodyfikowanch) z dokumentów oraz zbioru słów ze słownika *words.txt.* Operacje na słowach znajdują sie w pliku *word_checking.py*. Natomias tworzenie zbioru termów w pliku *do_matrix.py* a sam zbiór zapisany jest w *pliku union.txt.*

Częstośćwystępowania słów został policzony również w pliku *do_matrix.py* a sama częstość występowania danych słów zostałą zapisana w pliku *matrix.txt*, natomiast lista wszystkich dokumentów została zapisana w pliku *name.txt.*

W pliku *preapare.py* znajdują sie wszystkie operacje na macierzy tj mnożenie przez IDF (inverse document frequency) oraz normalizacja macierzy . W tym pliku obliczane są również różne macierze SVD przy pomocy funckji biblioteki Scipy dla rzadkich macierzy.

Obsługa zapytań znajduje sie w pliku *finder.py* tam z wykorzystaniem policzonych wcześniej pliku odbywa sie wyszukiwanie.

Wyniki z usuwaniem szumu oraz bez usuwania znacząco sie różniom(tym bardziej im mniejsza jest wartość k), często fragment wzięty z tekstu nie był w pierwszej 10 najlepszych wyników dla sposobu bez usuwania szumu. Najlepsze wyniki zapytań otrzymywałem dla k rzędu 1400
Jeśli nie użyje IDF to znajduje się pewna mała pula dokumentów która w większości przypadków zajmuje 10 najwyższych miejsc. Natomiast jeśli użyje się IDF to problem ten nie występuje. Program z usuwaniem szumu działa wolniej, jest to spowodowane ładowaniem większego pliku do pamięci oraz tym że macierz po SVD nie jest tak samo rzadka jak ta bez usuwania.