MOwNiT - Sprawozdanie z laboratorium 6
Singular Value Decomposition

Weronika Ormaniec
305386

1 Cel ćwiczenia

Celem ćwiczenia była implementacja prostej wyszukiwarki tekstowej z użyciem macierzy *term-by-document* oraz zbadanie jaki wpływ na jej działanie mają przekształcenie *Inveres Document Frequency (IDF)* oraz rozkład *Singular Value Decomposition (SVD)*.

2 Wykonanie ćwiczenia

2.1 Dane

W ćwiczeniu wykorzystano fragment Simple English Wikipedia, który przetworzono z pomocą biblioteki Wiki-Dump Reader. Wybrano 50002 artykuły, które przetworzono w następujący sposób:

- wszystkie litery zamieniono na małe,
- z użyciem wyrażenia regularnego usunięto wszystkie znaki nie będące literami alfabetu łacińskiego lub cyframi,
- usunięto słowa ze stop-listy języka angielskiego zebrane w zbiorze *stopwords* z *nltk.corpus*, czyli słowa najczęściej występujące lub te o małym znaczeniu, np. spójniki,
- wykonano stemming (*nltk.PorterStemmer()*), czyli usunięto końcówki fleksyjne wyrazów.

2.2 Wektor *bag-of-words*

Z przetworzonych artykułów tworzono wektor *bag-of-words*, który zawierał 20 000 najczęściej pojawiających się w artykułach słów. Wszystkie przetworzone artykuły miały początko ponad 200 000 różnych słów, jednak ze względu na niewystarczającą ilość pamięci operacyjnej zdecydowano się na ograniczenie słownika.

2.3 Macierz *term-by-document*

Dla wszystkich plików zliczono liczbę wystąpień każdego słowa z *bag-of-words* i stworzono rzadką macierz *term-by-document*. Następnie, w wersji z *IDF*, przemnożono macierz przez wektor wartości *IDF*, które dla każdego słowa dane jest wzorem $IDF(w) = \log \frac{N}{n_w}$, gdzie N-liczba dokumentów, n_w-liczba dokumentów, w których występuje słowo w. Na sam koniec znormalizowano każdą kolumnę macierzy. Zaimplementowano również wersję bez normalizacji, jednak w ostatecznej aplikacji wykorzystano wersję znormalizowaną, gdyż pozwala ona na nieco szybsze wyszukiwanie, gdy macierz jest już utworzona.

2.4 Przetworzenie zapytania i korelacja

Słowa każdego zapytania przetwarzane są w taki sam sposób jak każdy artykuł. Następnie zapytanie sprowadzane jest do wektora typu *bag-of-words* - pod indeksem odpowiadajcym danemu słowu zapisywana jest liczba jego wystąpień. Wektor zapytania jest normalizowany i wyliczana jest korelacja tego wektora z wektorem każdego pliku, poprzez iloczyn skalarny: $cos\theta_j = q^T A e_j$, gdie q-wektor zapytania, A-macierz *term-by-document*. Ponieważ operowano na wektorach o dodatnich wartościach, pliki najbardziej skorelowane z zapytaniem to te, dla których wyliczona korelacja jest największa.

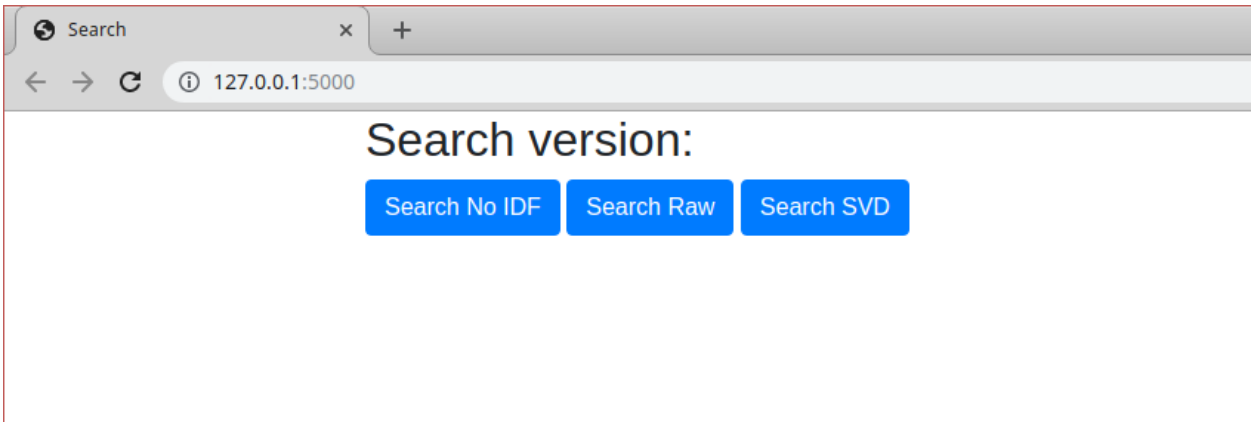
2.5 SVD

Do rozkładu SVD wykorzystano TruncatedSVD z pakietu scikit-learn. Wykonano redukcję wymiarów, a konkretnie metodą *transform* wygenerowano z macierzy *term-by-document* macierz, która odpowiada za przekształcenie przestrzeni słów w pewną przestrzeń liniową o mniejszym wymiarze. W momencie przetwarzania zapytania, jest ono mapowane na tę mniej wymiarową przestrzeń. Następnie wyliczany jest przybliżony wektor korelacji dzięki atrybutowi *TruncatedSVD.components*. Dzięki tej metodzie możemy wykonać mniej pojedyńczych mnożeń oraz w trakcie analizy bazować bardziej na znaczeniu grup słów zamiast na konkretnych słowach. Przetestowano mapowanie na przestrzeń 50, 100 i 250 wymiarową.

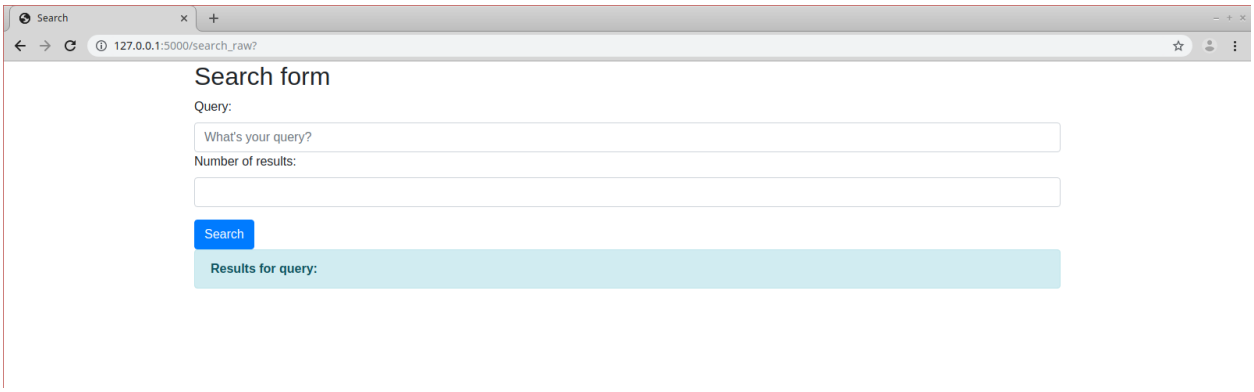
3 Aplikacja

Do zaimplementowania prostej aplikacji webowej wykorzystano framework Flask. Po uruchomieniu aplikacji można wybrać jedną z 3 wersji wyszukiwarki:

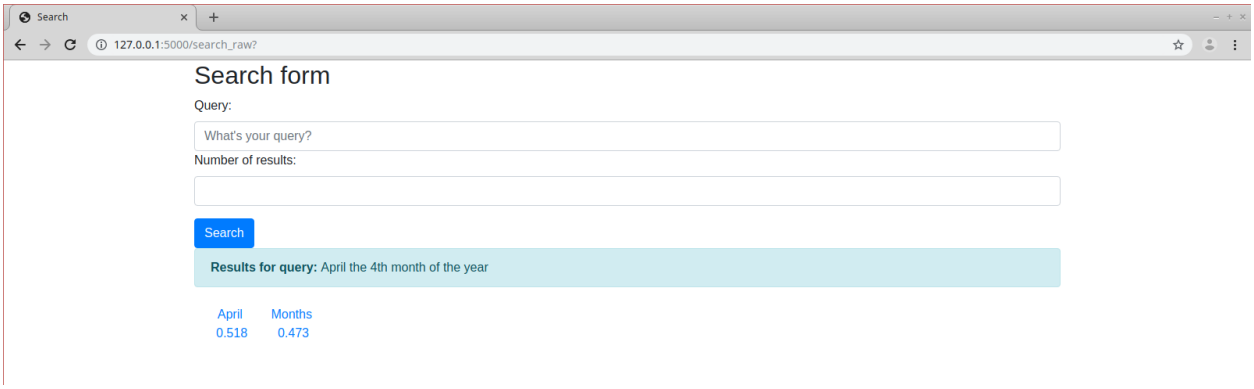
- Search No IDF - surowa wersja wyszukiwarki bez IDF,
- Search Raw - surowa wersja wyszukiwarki z IDF,
- Search SVD - wyszukiwarka z IDF i SVD.



Wszystkie wersje operują na 50 002 artykułach, a potrzebne do obliczeń macierze zostały wyliczone wcześniej i w momencie uruchamiania aplikacji są jedynie wczytywane z pliku. Online wyliczany jest tylko wektor zapytania.



Wyniki zapytania prezentowane są w postaci linków do k znaleznionych artykułów wraz z otrzymaną wartością korelacji.



4 Wyniki

Poniżej zaprezentowano działanie aplikacji dla zapytań: "Months", "queens and kings of England", "Gravitational waves", "Euclidean algorithm". Badano wszystkie trzy tryby działania aplikacji. *SVD* przetestowano w 3 wersjach.

4.1 Wersja bez IDF

Results for query: Months

Months	Month	Hebrew calendar	Cruiserweight	Islamic calendar	Central European Summer Time
0.577	0.427	0.31	0.308	0.271	0.267

Köppen climate classification	Endometrium	Eyelash	Pronghorn
0.252	0.243	0.229	0.218

Results for query: queens and kings of England

List of English monarchs	England	Queen	Elizabeth I of England	King	Queens	Eleanor of Provence	Canute
0.592	0.553	0.547	0.537	0.536	0.535	0.535	0.53

Edred of England	Regent
0.515	0.505

Results for query: Gravitational waves

Wave (physics)	Wave	List of wave topics	Longitudinal wave	Sine wave	Wavelength	Ocean surface wave
0.637	0.623	0.612	0.591	0.576	0.575	0.552

Surface wave	Gravity wave	Radio wave
0.51	0.479	0.476

Results for query: Euclidean algorithm

Euclidian algorithm	Complexity theory	Las Vegas algorithm	Genetic algorithms	Monte-Carlo algorithm	Las-Vegas algorithm
0.707	0.444	0.397	0.354	0.316	0.316

Genetic algorithm	Key (cryptography)	Exponentiation by squaring	Non-Euclidean geometry
0.298	0.278	0.255	0.244

4.2 Wersja z IDF

Results for query: Months

Months	Month	Cruiserweight	Hebrew calendar	Central European Summer Time	Islamic calendar	June
0.947	0.411	0.32	0.287	0.285	0.246	0.237

Köppen climate classification	Eyelash	Pronghorn
0.227	0.213	0.198

Results for query: queens and kings of England

Queen regnant	Queen Regnant	King	England	Queens	List of English monarchs	Queen	Queen consort
0.546	0.546	0.541	0.524	0.517	0.477	0.477	0.449

Queen Consort	Queen Elizabeth, The Queen Mother
0.449	0.448

Results for query: Gravitational waves

Wave (physics)	Wave	Longitudinal wave	List of wave topics	New Wave	New wave	Ocean surface wave	Waves
0.64	0.632	0.59	0.573	0.568	0.568	0.568	0.546

Wavelength	Gravity
0.525	0.524

Results for query: Euclidean algorithm

Euclidian algorithm	Complexity theory	Genetic algorithms	Las Vegas algorithm	Las-Vegas algorithm	Monte-Carlo algorithm
0.985	0.557	0.542	0.506	0.477	0.455

Exponentiation by squaring	Genetic algorithm	Non-Euclidean geometry	Decidability theory
0.399	0.381	0.362	0.346

Można zauważyć, że wyniki zapytania zawierają dokładnie te słowa, co w zapytaniu. Tak samo wewnątrz tych artykułów dokładnie te frazy powtarzają się kilkukrotnie. Nie zaobserwowano szczególnych różnic w zwróconych artykułach pomiędzy wersją z *IDF* i bez *IDF*, jednak w porównaniu z poprzednią wersją wzrosła wartość otrzymanych korelacji.

4.3 Wersja z IDF i SVD 50

Results for query: Months

2013	Day	2010	2012	2011	2014	2001	2004	1992	1998
0.013	0.012	0.012	0.012	0.012	0.011	0.011	0.01	0.01	0.01

Results for query: queens and kings of England

History of England	List of English monarchs	England	1272	1689	Tudor dynasty	930s	Glorious Revolution
0.247	0.244	0.241	0.202	0.201	0.199	0.192	0.187

William III of England	Henry VI of England
0.186	0.186

Results for query: Gravitational waves

Chemical element	2008 Atlantic hurricane season	Hurricane Florence (2006)	2006 Atlantic hurricane season
0.014	0.013	0.013	0.013

Storm history of Hurricane Wilma	2007 Atlantic hurricane season	Tropical cyclone	Storm history of Hurricane Ivan	Chemistry
0.013	0.013	0.013	0.012	0.012

Storm history of Hurricane Katrina
0.012

Results for query: Euclidean algorithm

Computer	Computer science	Computer program	Computer numbering formats	List of words about computers
0.013	0.013	0.013	0.012	0.012

Personal computer	Software	Computer jargon	Malware	Number
0.012	0.011	0.011	0.011	0.011

Wyniki są skorelowane znaczeniowo z zapytaniem, ale utracono możliwość znalezienia konkretnego hasła. Zostało wychwycone znaczenie miesiąca jako okresu czasu, dlatego wynikiem zapytania są konkretne lata. Wyniki zapytania o królowe i królów Anglii są ogólniejsze niż bez zastosowania SVD. Dla zapytania o fale grawitacyjne znaleziono powiązanie między falą a zjawiskami pogodowymi (prawodpodobnie dlatego, że często występowały blisko słowa woda) oraz między grawitacją i chemią (prawdopodobnie poprzez słowo fizyka). Algorytm Euklidesa został wrzucony do jednej kategorii z prawdopodobnie wszystkim związanym z informatyką.

4.4 Wersja z IDF i SVD 100

Results for query: Months

2013	Treviso	Mede, Lombardy	Macerata	Chieti	Belluno	Libourne	Vicenza	Indre River	Province of Macerata
0.017	0.016	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015

Results for query: queens and kings of England

List of English monarchs	England	History of England	Tudor dynasty	Henry II	House of Plantagenet	1689
0.41	0.343	0.343	0.315	0.305	0.296	0.294

Henry VI of England	1272	Glorious Revolution
0.286	0.281	0.278

Results for query: Gravitational waves

Electricity	Power station	2006 Atlantic hurricane season	Tropical cyclone	2008 Atlantic hurricane season	Speed of light
0.025	0.025	0.025	0.024	0.024	0.024

Hurricane Isaac (2006)	2005 Atlantic hurricane season	1986 Atlantic hurricane season	2007 Atlantic hurricane season
0.024	0.023	0.023	0.023

Results for query: Euclidean algorithm

Computer science	Computer	Computer program	Chipset - southbridge	Computer system	Computers	Computer programs
0.021	0.02	0.018	0.017	0.017	0.017	0.017

Software application	Computer Programs	Programming
0.017	0.017	0.017

Zapytanie o miesiąc wydaje się nie mieć powiązania z wynikami, jednak jest to na tyle ogólne pojęcie, że mogło tak się zdarzyć. Pozostałe wyniki prezentują się podobnie, jak w wersji poprzedniej.

4.5 Wersja z IDF i SVD 200

Results for query: Months

Common year	Year	Gregorian calendar	New Years Day	Hebrew calendar	Leap year	June	Holiday
0.03	0.03	0.029	0.029	0.028	0.027	0.027	0.027

Islamic calendar	1705
0.027	0.027

Results for query: queens and kings of England

List of English monarchs	England	Tudor dynasty	History of England	Elizabeth I of England	Edward IV of England
0.489	0.421	0.405	0.403	0.348	0.342

House of Plantagenet	Henry II	Edward V of England	Henry VI of England
0.337	0.334	0.332	0.325

Results for query: Gravitational waves

Speed of light	Wave–particle duality	Light	Čerenkov radiation	Physics	String theory	Albert Einstein	Electromagnetism
0.061	0.06	0.058	0.057	0.054	0.053	0.053	0.053

A Brief History of Time	Radio wave
0.053	0.052

Results for query: Euclidean algorithm

Symmetric-key algorithm	Blowfish (cipher)	Key generation	Twofish	Key schedule	Cypher	Key space	Key size
0.04	0.04	0.04	0.039	0.037	0.037	0.037	0.036

Meet-in-the-middle attack	Cryptography
0.036	0.036

Dla tej wersji otrzymano wyjątkowo interesujące wyniki. Zapytanie o miesiąc dostaje wynik znaczeniowo powiązany z okresem czasu. Zapytanie o fale grawitacyjne jest powiązane ściśle z fizyką. Algorytm Euklidesa nie jest powiązany ani dokładnie z arykułem o algorytmie Euklidesa, ani z całą informatyką, a jedynie z działem informatyki, gdzie jest bardzo często używany, czyli z kryptografią.

4.6 Wnioski

Poprzez konstrukcję macierzy *term-by-document* można zbudować całkiem sprawnie działającą wyszukiwarkę tekstową.

Przytoczone przykłady nie demonstrują zbyt dobrze działania *IDF* jednak jest ono widoczne, gdy stworzymy zapytanie złożone ze słowa potocznego i bardziej specyficznego np. "Garfield the cat". Dla tego zapytania, wśród 20 zwróconych artykułów, zanotowano pojawienie się 2 artykułów powiązanych ze słowem "Garfield" dla wersj bez *IDF* i 5 dla wersj z *IDF*.

Wersję wyszukiwarki bez SVD lepiej używać, gdy zależy nam na bardzo konkretnych wynikach, bezpośrednio powiązanych słowami z zapytaniem. Wersja z SVD zwróci luźniej powiązane wyniki. O tym jak słabe będzie to powiązanie decyduje wymiar przestrzeni, na którą mapujemy przestrzeń słów - im mniejszy wymiar, tym luźniejsze powiązanie znaczeniowe.