Laboratorium 8 - sprawozdanie

Krzysztof Wójtowicz

Celem laboratorium było utworzenie wyszukiwarki z zastosowaniem SVD.

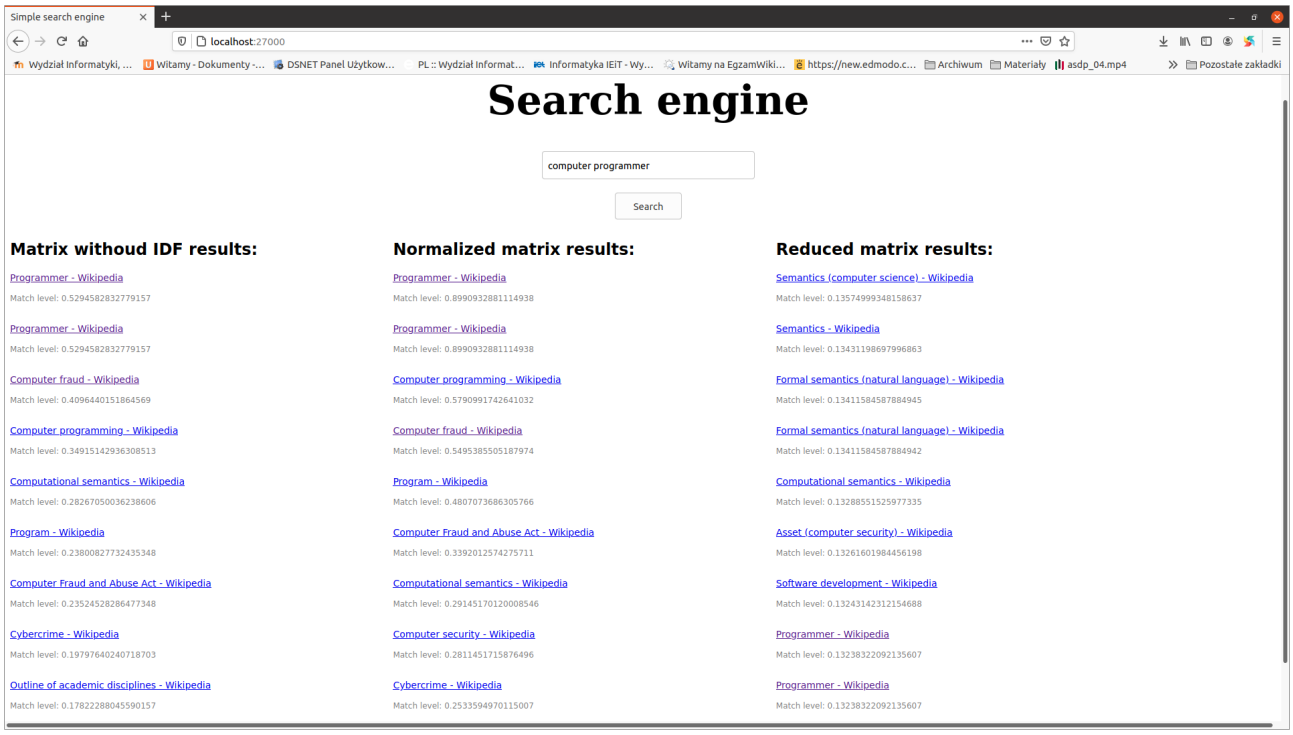
Zbiór dokumentów tekstowych generowany jest dynamicznie przy pomocy własnego web crawlera, ograniczonego jednak do artykułów na angielskiej Wikipedii.
Słownik termów generowany jest na podstawie zawartości artykułów (brane pod uwagę są nagłówki, paragrafy, listy i tabele). Pomijane są słowa, które nie wnoszą informacji semantycznej (stopwords). Pozostałe słowa sprowadzane są do bazy.

Wyszukiwarka składa się z dwóch programów:

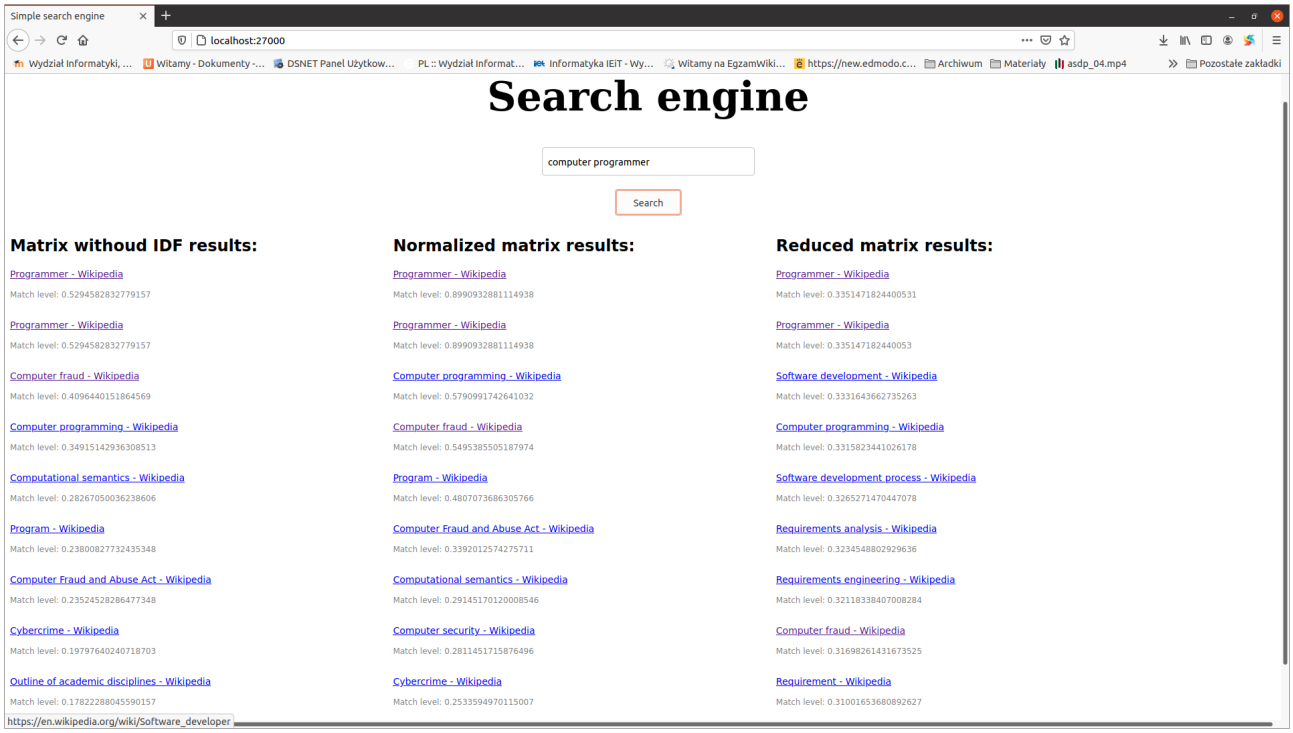
- backand.py - odpowiedzialny za przygotowanie zbioru *term-by-document matrix* w postaci bez zastosowania IDF, po zastosowaniu IDF i normalizacji oraz po redukcji szumów; wygenerowane dane przechowywane są na dysku w postaci *pickle;*
- app.py - aplikacja odpowiedzialna za stronę kliencką, odczytuje dane wygenerowane przez backend, pobiera zapytanie od użytkownika i zwraca najlepsze dopasowania.

Przykładowe wyszukiwania dla różnych wartości k przy redukcji szumów z użyciem SVD:

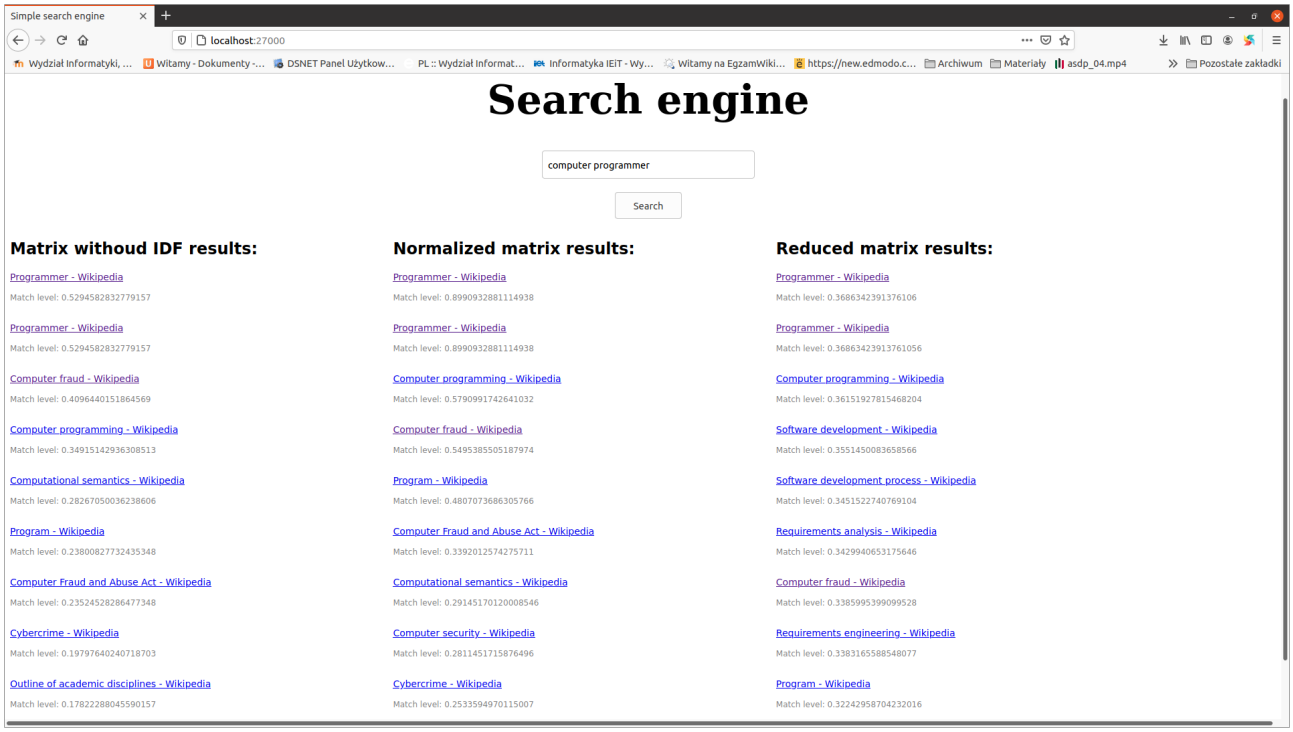
k = 10



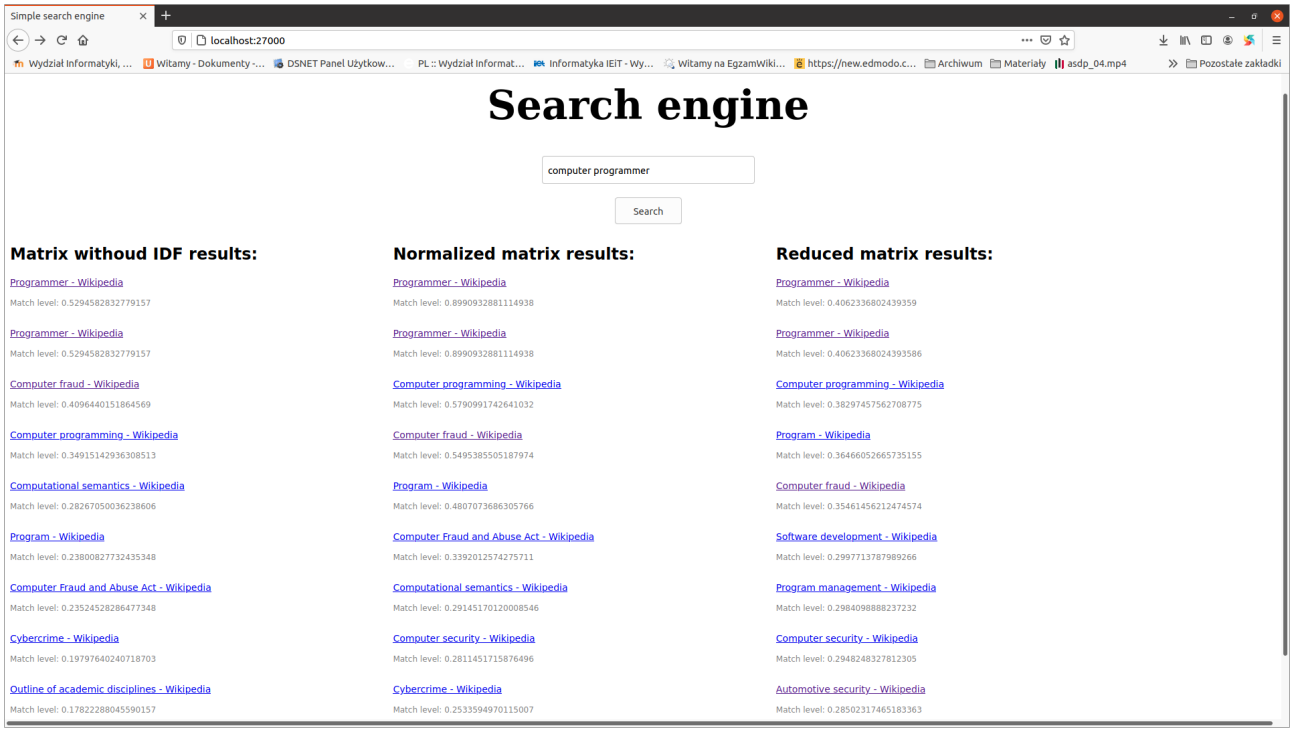
k = 30


Simple search engine
localhost:27000
Search engine
computer programmer
Search
Matrix withoud IDF results:
Programmer - Wikipedia
Match level: 0.5294582832779157
Programmer - Wikipedia
Match level: 0.5294582832779157
Computer fraud - Wikipedia
Match level: 0.4096440151864569
Computer programming - Wikipedia
Match level: 0.34915142936308513
Computational semantics - Wikipedia
Match level: 0.28267050036238606
Program - Wikipedia
Match level: 0.23800827732435348
Computer Fraud and Abuse Act - Wikipedia
Match level: 0.23524528286477348
Cybercrime - Wikipedia
Match level: 0.19797640240718703
Outline of academic disciplines - Wikipedia
Match level: 0.17822288045590157
Normalized matrix results:
Programmer - Wikipedia
Match level: 0.8990932881114938
Programmer - Wikipedia
Match level: 0.8990932881114938
Computer programming - Wikipedia
Match level: 0.5790991742641032
Computer fraud - Wikipedia
Match level: 0.5495385505187974
Program - Wikipedia
Match level: 0.4807073686305766
Computer Fraud and Abuse Act - Wikipedia
Match level: 0.3392012574275711
Computational semantics - Wikipedia
Match level: 0.29145170120008546
Computer security - Wikipedia
Match level: 0.2811451715876496
Cybercrime - Wikipedia
Match level: 0.2533594970115007
Reduced matrix results:
Programmer - Wikipedia
Match level: 0.3351471824400531
Programmer - Wikipedia
Match level: 0.335147182440053
Software development - Wikipedia
Match level: 0.3331643662735263
Computer programming - Wikipedia
Match level: 0.3315823441026178
Software development process - Wikipedia
Match level: 0.3265271470447078
Requirements analysis - Wikipedia
Match level: 0.3234548802929636
Requirements engineering - Wikipedia
Match level: 0.32118338407008284
Computer fraud - Wikipedia
Match level: 0.31698261431673525
Requirement - Wikipedia
Match level: 0.31001653680892627
https://en.wikipedia.org/wiki/Software_developer

k = 50

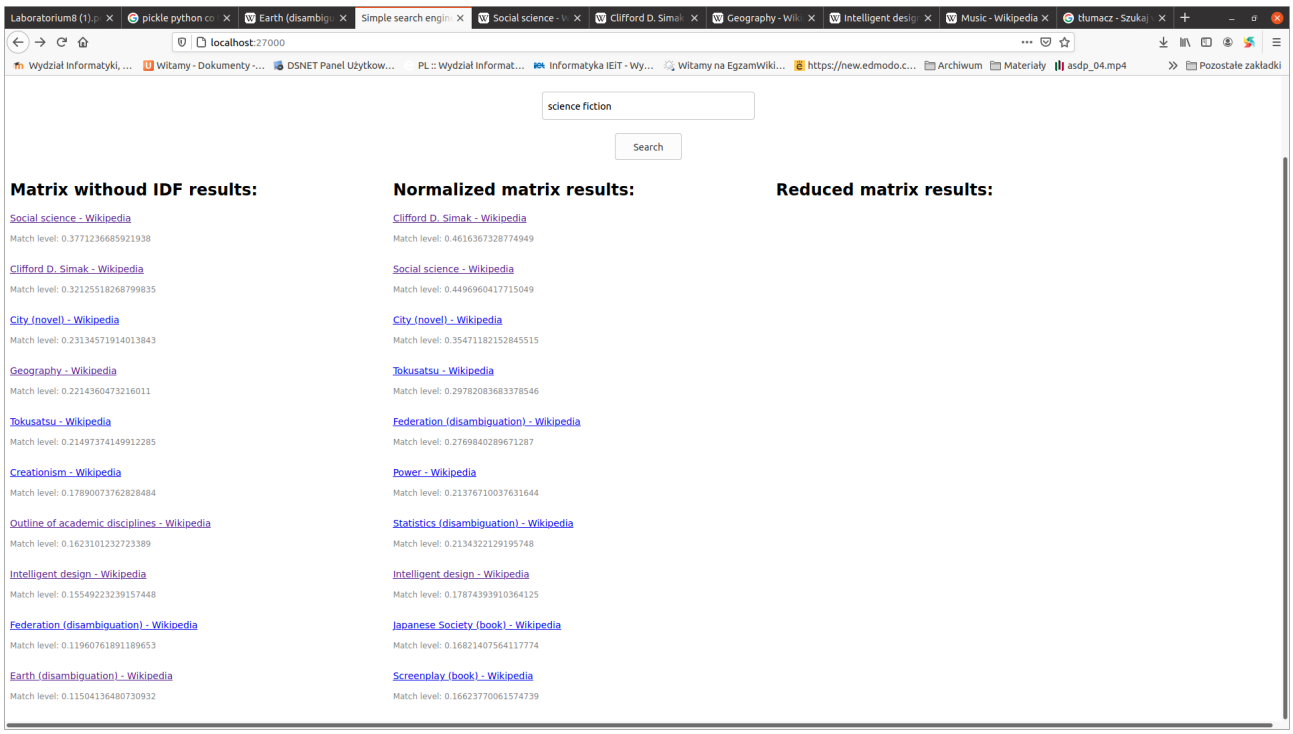

Simple search engine
localhost:27000
Search engine
computer programmer
Search
Matrix withoud IDF results:
Programmer - Wikipedia
Match level: 0.5294582832779157
Programmer - Wikipedia
Match level: 0.5294582832779157
Computer fraud - Wikipedia
Match level: 0.4096440151864569
Computer programming - Wikipedia
Match level: 0.34915142936308513
Computational semantics - Wikipedia
Match level: 0.28267050036238606
Program - Wikipedia
Match level: 0.23800827732435348
Computer Fraud and Abuse Act - Wikipedia
Match level: 0.23524528286477348
Cybercrime - Wikipedia
Match level: 0.19797640240718703
Outline of academic disciplines - Wikipedia
Match level: 0.17822288045590157
Normalized matrix results:
Programmer - Wikipedia
Match level: 0.8990932881114938
Programmer - Wikipedia
Match level: 0.8990932881114938
Computer programming - Wikipedia
Match level: 0.5790991742641032
Computer fraud - Wikipedia
Match level: 0.5495385505187974
Program - Wikipedia
Match level: 0.4807073686305766
Computer Fraud and Abuse Act - Wikipedia
Match level: 0.3392012574275711
Computational semantics - Wikipedia
Match level: 0.29145170120008546
Computer security - Wikipedia
Match level: 0.2811451715876496
Cybercrime - Wikipedia
Match level: 0.2533594970115007
Reduced matrix results:
Programmer - Wikipedia
Match level: 0.3686342391376106
Programmer - Wikipedia
Match level: 0.36863423913761056
Computer programming - Wikipedia
Match level: 0.36151927815468204
Software development - Wikipedia
Match level: 0.3551450083658566
Software development process - Wikipedia
Match level: 0.3451522740769104
Requirements analysis - Wikipedia
Match level: 0.3429940653175646
Computer fraud - Wikipedia
Match level: 0.3385995399099528
Requirements engineering - Wikipedia
Match level: 0.3383165588548077
Program - Wikipedia
Match level: 0.3224295870423201 6

k = 100



Najlepsze rezultaty uzyskano dla k w przedziale 30 - 50, przy wielkości zbioru dokumentów wynoszącej 1000, jednak bez znajomości wszystkich dokumentów indeksowanych przez wyszukiwarkę trudno jest ocenić adekwatność wyników wyszukiwania.
Używanie SVD ma jednak znaczący wpływ na szybkość wyszukiwania - zredukowana macierz nie jest już macierzą rzadką, przez co operacje na niej zajmują dużo więcej czasu.

Porównując wyniki wyszukiwania z zastosowanie IDF i bez niego zauważono, że wyniki są bardziej odpowiednie w przypadku złożonych zapytań. Poniżej przedstawiono wyniki wyszukiwania dla frazy "science fiction". Bez zastosowania IDF rezultaty zostały zdominowane przez dokumenty zawierające słowo "science" w dużej ilości, natomiast w przypadku zastosowania IDF można zauważyć, że wyniki bardziej odpowiadają całej frazie "science fiction".


Laboratorium8 (1).p
pickle python co
Earth (disambigu
Simple search engin
Social science - W
Clifford D. Simak
Geography - Wiki
Intelligent desig
Music - Wikipedia
tłumacz - Szukaj
localhost:27000
Wydział Informatyki, ...
Witamy - Dokumenty -...
DSNET Panel Użytkow...
PL :: Wydział Informat...
Informatyka IEiT - Wy...
Witamy na EgzamWiki...
https://new.edmodo.c...
Archiwum
Materiały
asdp_04.mp4
Pozostałe zakładki
science fiction
Search
Matrix withoud IDF results:
Social science - Wikipedia
Match level: 0.3771236685921938
Clifford D. Simak - Wikipedia
Match level: 0.32125518268799835
City (novel) - Wikipedia
Match level: 0.23134571914013843
Geography - Wikipedia
Match level: 0.22143604732160 11
Tokusatsu - Wikipedia
Match level: 0.21497374149912285
Creationism - Wikipedia
Match level: 0.17890073762828484
Outline of academic disciplines - Wikipedia
Match level: 0.1623101232723389
Intelligent design - Wikipedia
Match level: 0.15549223239157448
Federation (disambiguation) - Wikipedia
Match level: 0.11960761891189653
Earth (disambiguation) - Wikipedia
Match level: 0.11504136480730932
Normalized matrix results:
Clifford D. Simak - Wikipedia
Match level: 0.4616367328774949
Social science - Wikipedia
Match level: 0.4496960417715049
City (novel) - Wikipedia
Match level: 0.35471182152845515
Tokusatsu - Wikipedia
Match level: 0.29782083683378546
Federation (disambiguation) - Wikipedia
Match level: 0.2769840289671287
Power - Wikipedia
Match level: 0.213767100376316 44
Statistics (disambiguation) - Wikipedia
Match level: 0.2134322129195748
Intelligent design - Wikipedia
Match level: 0.17874393910364125
Japanese Society (book) - Wikipedia
Match level: 0.16821407564117774
Screenplay (book) - Wikipedia
Match level: 0.16623770061574739
Reduced matrix results: