246810 ALGNUMSPR zadanie 3

Michal Jachman

W zadaniu zaimplementowałem algorytm ALS do rekomendacji produktów. Sparsowałem dane z pliku Amazon do bazy danych SQL. Następnie dostosowywałem zapytania do bazy tak, aby uzyskać odpowiednie liczby produktów. Starałem się dobierać danie tak, żeby wybrać użytkowników i produkty, mające najwięcej recenzji. Wynikiem były 3 macierze:

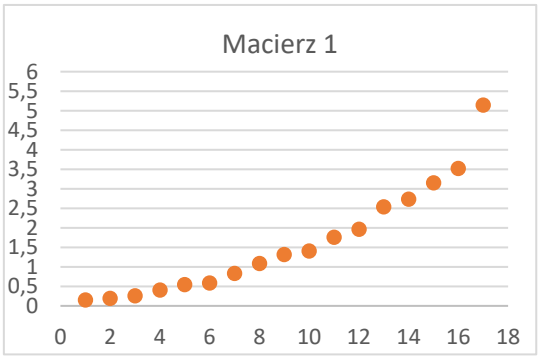
Macierz 1: 44u x 30p (200 recenzji) ok. 15% wypełnienia

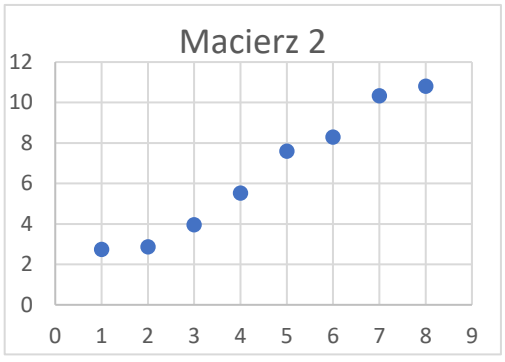
Macierz 2: 273u x 251p (3359 recenzji) ok. 5% wypełnienia

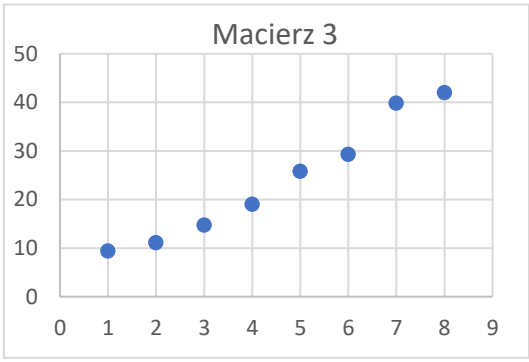
Macierz 3: 194u x 1771p (7759 recenzji) ok. 2% wypełnienia

Podczas testów zakrywałem po 5% danych z każdej macierzy.

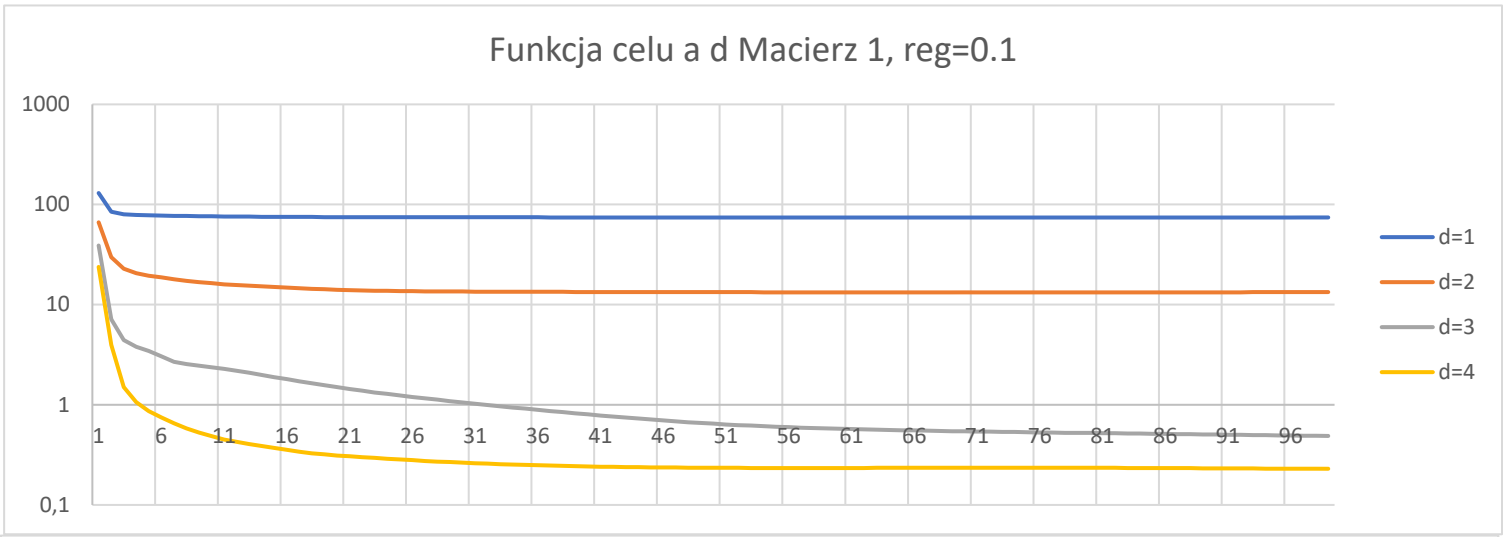
Jak można się spodziewać, im większy parametr d tym większy czas obliczeń algorytmu. Na poniższych wykresach przedstawiłem czas wykonywania algorytmu w sekundach dla każdej z trzech macierzy. Na osi x wartość parametru d. Na osi y czas wykonywania dla 100 iteracji.

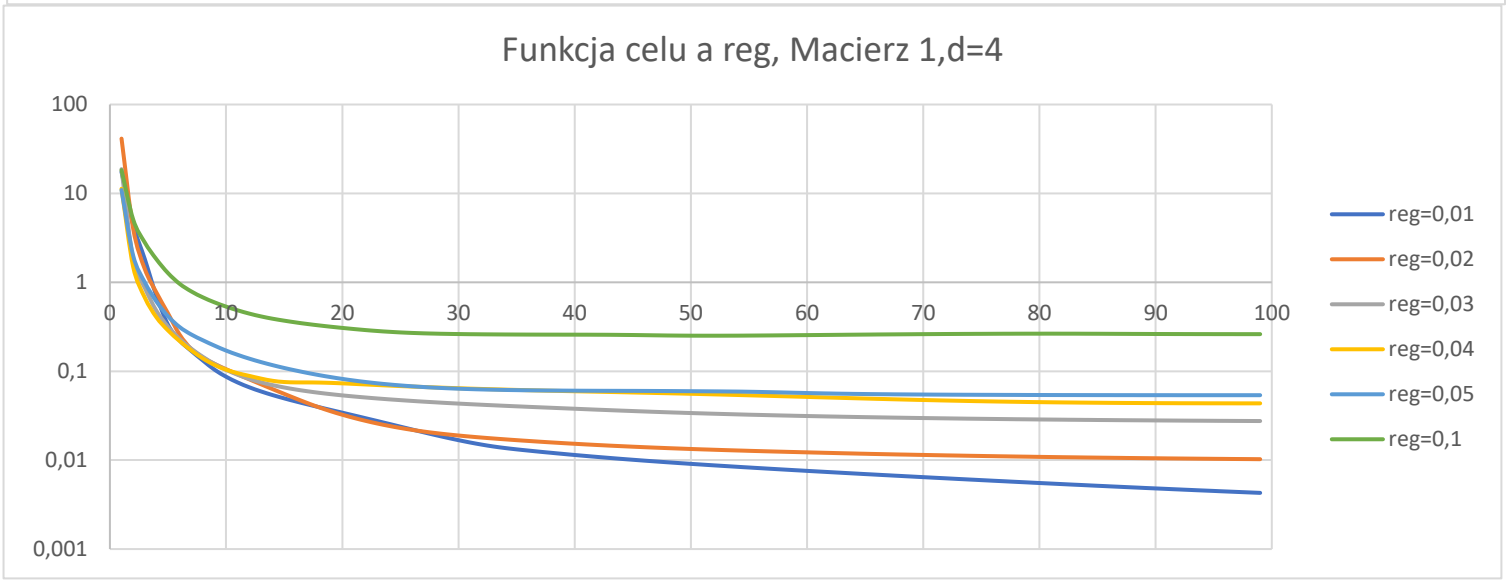






Ponieważ funkcja celu maleje z każdą iteracją, zakładam że algorytm jest poprawny. Im większe d, tym szybciej funkcja maleje i tym mniejszą wartość osiąga. O poprawności implementacji świadczy również porównanie otrzymanych wyników z ocenami testowymi.



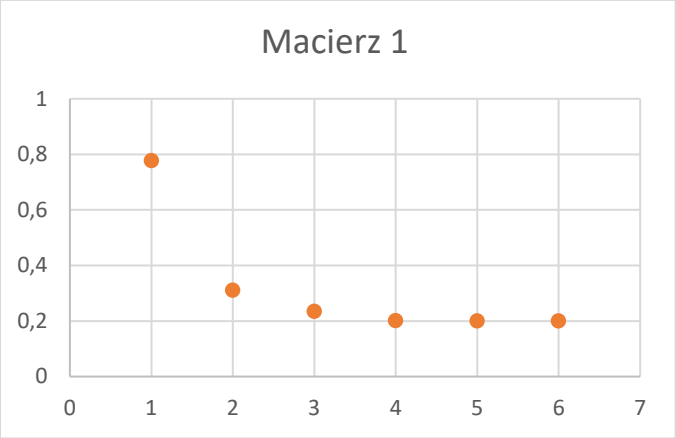


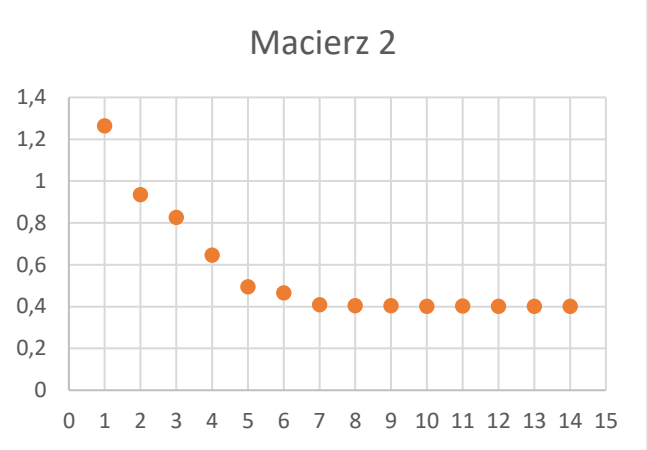
Przykładowe wartości obliczonych ocen dla małej macierzy, przy d=4,reg=0.1 i stopniu zakrycia=5%.

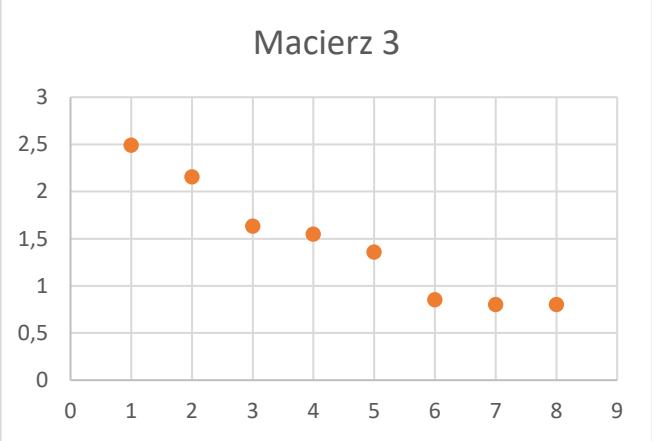
Obliczona ocena	Prawdziwa wartość	różnica
4,994671	5	2,84E-05
5,018134	5	0,000329
5,027862	5	0,000776
1,971629	2	0,000805
4,967409	5	0,001062
4,965616	5	0,001182
1,964779	2	0,001241
4,042087	4	0,001771
4,941527	5	0,003419
4,941527	5	0,003419
5,070345	5	0,004948
3,089874	3	0,008077
3,903364	4	0,009339
5,168996	5	0,02856
5,203243	5	0,041308
4,750523	5	0,062239
4,731822	5	0,07192
4,724638	5	0,075824
4,716435	5	0,080409
4,697084	5	0,091758
3,677921	4	0,103735
4,381197	4	0,145311
4,591445	5	0,166917
3,558666	4	0,194776
5,444366	5	0,197461
4,504141	5	0,245876
4,42758	5	0,327665
4,42758	5	0,327665
4,423362	5	0,332512
4,660129	4	0,435771
4,710571	4	0,504911
4,721742	4	0,520911
2,054337	3	0,894278
4,035514	5	0,930233

Przy zakryciu co 20 elementu dla każdej macierzy, błąd średniokwadratowy prezentuje się następująco. Dla małej macierzy optymalne d wynosi 4. Powyżej tej wartości, błąd średniokwadratowy oscyluje w granicach 0,2.

Dla macierzy 2 optymalne d to 6, większe d. Błąd średniokwadratowy zatrzymuje się w okolicy 0,4. Dla macierzy 3 optymalne d to 8. Dalej błąd oscylował w okolicy 0,9. Na wykresach pokazane są wartości przy przyjęciu liczby iteracji=100 i reg=0.1 i stopniu zakrycia=5%.







Podsumowując, im większa macierz, tym mniejszy stopień wypełnienia udaje się uzyskać. Co za tym idzie, wyniki rekomendacji są mniej dokładne i Algorytm osiąga swoje maksimum precyzji przy większych d.