234073
Numer indeksu
Aleksandra Kowalczyk
Imię i nazwisko

234102
Numer indeksu
Zbigniew Nowacki
Imię i nazwisko

234106
Numer indeksu
Karol Podlewski
Imię i nazwisko

Kierunek Informatyka Stosowana
Stopień II
Specjalizacja Data Science
Semestr 2

Data oddania 18 listopada 2020

Przetwarzanie i analiza dużych zbiorów danych

Zadanie 3

1 Cel zadania

Celem zadania była implementacja algorytmu k-średnich z uwzględnieniem dwóch miar - euklidesowej oraz Manhattan - dla dwóch rozmieszeń centrów skupień - losowego oraz maksymalnie od siebie oddalonych centroidów według odległości euklidesowej.

Dla każdej iteracji należało obliczyć funkcje kosztu $\phi(i)$ oraz $\psi(i)$, wygenerować wykresy oraz obliczyć procentową zmianę kosztu po 10 iteracjach algorytmu dla obydwu miar odległości z wskazaniem, które z dwóch początkowych rozmieszeń skupień pozwoliło uzyskać lepsze rezultaty.

Miara euklidesowa:

$$\text{odległość: } ||a-b|| = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2} \tag{1}$$

$$\text{koszt: } \phi = \sum_{x \epsilon X} \min_{c \epsilon C} ||x-c||^2 \tag{2}$$

Miara Manhattan:

$$\text{odległość: } |a-b| = \sum_{i=1}^{d} |a_i - b_i| \tag{3}$$

$$\text{koszt: } \psi = \sum_{x \epsilon X} \min_{c \epsilon C} |x-c| \tag{4}$$

2 Opis implementacji

Do implementacji zadania wykorzystano język Python wraz z API PySpark, które umożliwiło skorzystanie z możliwości języka Apache Spark w pythonowym kodzie. Wszystkie funkcje (odległości, kosztu) zostały zaimplementowane w programie, nie korzystano z bibliotek.

3 Uzyskane wyniki

Analiza zajęła około 7 minut i 20 sekund.

Tabela 1: Uzyskane wartości funkcji kosztu

Iteracja	Metryka euklidesowa		Metryka Manhattan	
	Plik 3b.txt	Plik 3c.txt	Plik 3b.txt	Plik 3c.txt
1	623 660 345,30	438 747 790,02	550 117,14	1 433 739,31
2	509 862 908,29	249 803 933,62	464 661,07	1 084 488,77
3	485 480 681,87	194 494 814,40	471 200,04	973 431,71
4	463 997 011,68	169 804 841,45	484 160,69	895 934,59
5	460 969 266,57	156 295 748,80	489 251,72	865 128,33
6	460 537 847,98	149 094 208,10	487 564,74	845 846,64
7	460 313 099,65	142 508 531,61	483 404,05	827 219,58
8	460 003 523,88	132 303 869,40	475 365,34	803 590,34
9	459 570 539,31	117 170 969,83	474 924,05	756 039,51
10	459 021 103,34	108 547 377,17	457 233,64	717 332,90
11	458 490 656,19	102 237 203,31	447 495,09	694 587,92
12	457 944 232,58	98 278 015,74	451 004,30	684 444,50
13	457 558 005,19	95 630 226,12	451 222,09	674 574,74
14	457 290 136,35	93 793 314,05	451 973,84	667 409,46
15	457 050 555,05	92 377 131,96	451 585,35	663 556,62
16	456 892 235,61	91 541 606,25	452 756,64	660 162,77
17	456 703 630,73	91 045 573,83	452 893,79	656 041,32
18	456 404 203,01	90 752 240,10	450 382,23	653 036,75
19	456 177 800,54	90 470 170,18	450 023,96	651 112,42
20	455 986 871,02	90 216 416,17	448 929,47	649 689,01

$$\text{Zmiana względna } \phi_{3b.txt}\Big(\phi_{3b.txt}(1), \phi_{3b.txt}(10)\Big) = \frac{623660345,3 - 459021103,34}{623660345,3} = 0,2639886329 \approx 26,4\%$$

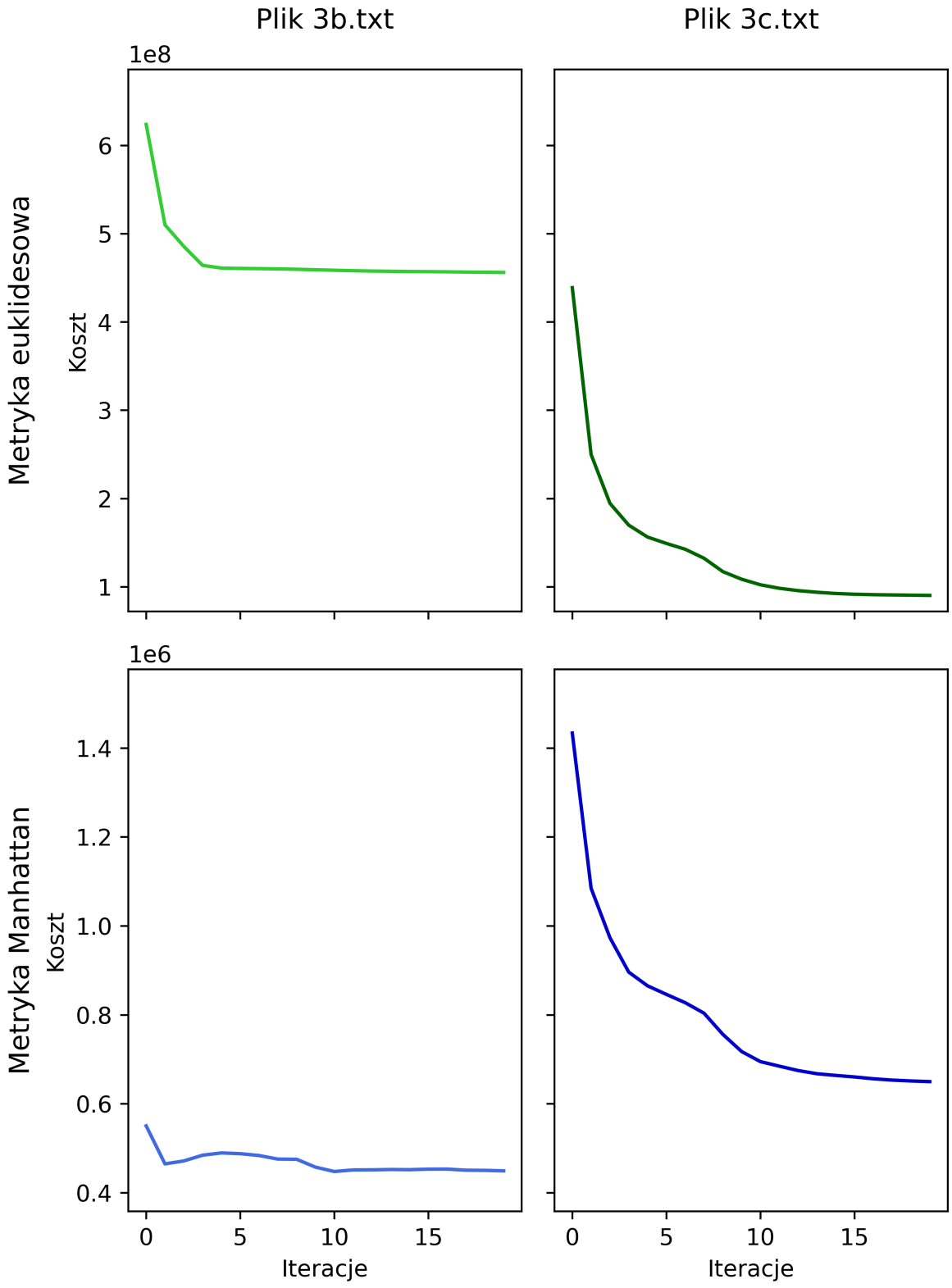
$$\text{Zmiana względna } \phi_{3c.txt}\Big(\phi_{3c.txt}(1), \phi_{3c.txt}(10)\Big) = \frac{438747790,02 - 108547377,17}{438747790,02} = 0,7525973244 \approx 75,26\%$$

$$\text{Zmiana względna } \psi_{3b.txt}\Big(\psi_{3b.txt}(1), \psi_{3b.txt}(10)\Big) = \frac{550117,14 - 457233,64}{550117,14} = 0,1688431304 \approx 16,88\%$$

$$\text{Zmiana względna } \psi_{3c.txt}\Big(\psi_{3c.txt}(1), \psi_{3c.txt}(10)\Big) = \frac{1433739,31 - 717332,90}{1433739,31} = 0,4996768973 \approx 49,97\%$$

Tabela 2: Procentowa zmiana kosztu po 10 iteracjach

Metryka	Rozmieszczenie centroidów	Zmiana kosztu
Metryka euklidesowa	Plik 3b.txt	26,4%
	Plik 3c.txt	75,26%
Metryka Manhattan	Plik 3b.txt	16,88%
	Plik 3c.txt	49,97%



Rysunek 1: Wykresy funkcji kosztu

4 Analiza wyników

Wykresy funkcji (rysunek 3) celu dla identycznych początkowych skupień posiadają podobny kształt, co widać szczególnie dobrze w przypadku pliku `3c.txt`, gdzie spadek kosztu robi się mniej gwałtowny po 4 iteracji, by delikatnie przyśpieszyć między 8 a 10 iteracją, po czym wykres zaczyna się widocznie wypłaszczać (co dokładniej można zobaczyć w tabeli 1).

W przypadku pliku `3b.txt` kształt funkcji nie jest już tak zbliżony, głównie ze względu na fakt, że dla metryki Manhattan wartość funkcji celu nie zawsze maleje (wzrost można zaobserwować w iteracjach 3-5, 12-14 oraz 16-17). Niezależnie od metryki, kształt funkcji zaczyna się dużo wcześniej wypłaszczać niż w przypadku drugiego początkowego rozłożenia centroidów.

Wniosek, że dla badanych początkowych środków skupień wykresy szybciej zaczynają się wypłaszczać w przypadku gdy centroidy są losowe (plik `3b.txt`) widać także w tabeli 2. Zmiana kosztu jest prawie 3-krotnie większa dla maksymalnie oddalonych cenroidów. Wartości te są też większe dla metryki euklidesowej.