Przetwarzanie i analiza dużych zbiorów danych 2020/21

Piotr Wardęcki	234128	234128@edu.p.lodz.pl
Paweł Galewicz	234053	234053@edu.p.lodz.pl
Bartosz Jurczewski	234067	234067@edu.p.lodz.pl

Zadanie 3

1. Cel zadania

Celem zadania była implementacja algorytmu k-średnich z uwzględnieniem dwóch miar - Euklidesowej oraz Manhattan - dla dwóch sposobów generowania centrów - losowego oraz maksymalnie oddalonych od siebie punktów (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 wykonywania zadania niezbędna była instancji *Apache Spark*. Aby ograniczyć liczbę zainstalowanych środowisk skorzystaliśmy z odpowiedniego obrazu dla Dockera [1], który zawierał także *Jupyter Notebook*, *Python* oraz *Miniconda*. Dodatkowo aby ułatwić tworzenie środowiska do kolejnych zadań i między naszymi komputerami skorzystaliśmy z narzędzia *Docker Compose* (nasz plik [2]).

3. Wyniki

Tabela 1. Uzyskane wartości funkcji kosztu

Iteracja	**Metryka euklidesowa**		**Metryka Manhattan**	
	Plik 3b.txt	Plik 3c.txt	Plik 3b.txt	Plik 3b.txt
1	623660345.31	438747790.03	550117.14	1433739.31
2	509862908.30	249803933.63	464869.28	1084488.78
3	485480681.87	194494814.41	470897.38	973431.71
4	463997011.69	169804841.45	483914.41	895934.59
5	460969266.57	156295748.81	489216.07	865128.34
6	460537847.98	149094208.11	487629.67	845846.65
7	460313099.65	142508531.62	483711.92	827219.58
8	460003523.89	132303869.41	475330.77	803590.35
9	459570539.32	117170969.84	474871.24	756039.52
10	459021103.34	108547377.18	457232.92	717332.90
11	458490656.19	102237203.32	447494.39	694587.93
12	457944232.59	98278015.75	450915.01	684444.50
13	457558005.20	95630226.12	451250.37	674574.75
14	457290136.35	93793314.05	451974.60	667409.47
15	457050555.06	92377131.97	451570.36	663556.63
16	456892235.62	91541606.25	452739.01	660162.78
17	456703630.74	91045573.83	453082.73	656041.32
18	456404203.02	90752240.10	450583.67	653036.75
19	456177800.54	90470170.18	450368.75	651112.43
20	455986871.03	90216416.18	449011.36	646481.16

Zmianę względną oznaczyliśmy grecką literą Δ.

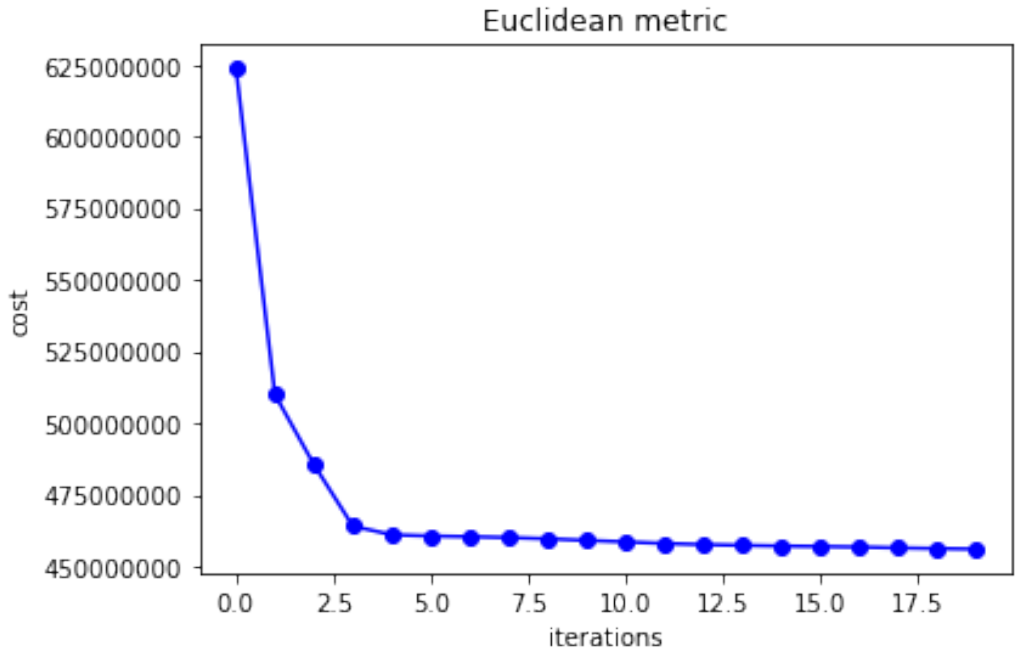
$$\Delta\phi_{3b.txt}\Big(\phi_{3b.txt}(1), \phi_{3b.txt}(10)\Big) = \frac{623660345,31 - 459021103,34}{623660345,31} = 0,2639886329 \approx 26,4\%$$

$$\Delta\phi_{3c.txt}\Big(\phi_{3c.txt}(1), \phi_{3c.txt}(10)\Big) = \frac{438747790,03 - 108547377,18}{438747790,03} = 0,7525973244 \approx 75,26\%$$

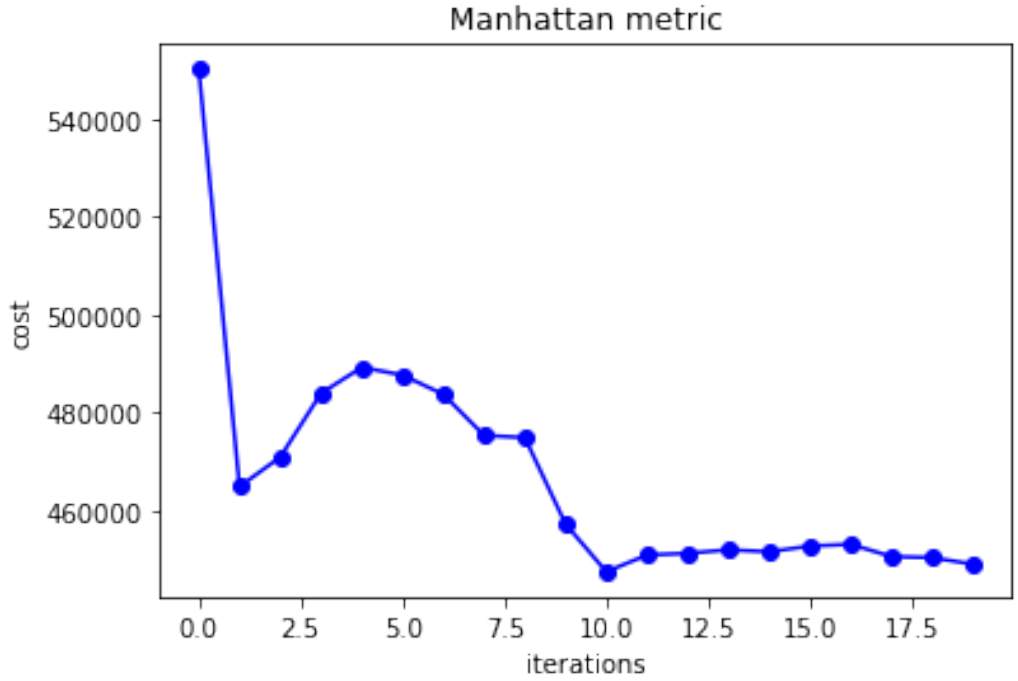
$$\Delta\psi_{3b.txt}\Big(\psi_{3b.txt}(1), \psi_{3b.txt}(10)\Big) = \frac{550117,14 - 457232,92}{550117,14} = 0,1688444392 \approx 16,88\%$$

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

3.1. Wykresy dla losowo wygenerowanych punktów

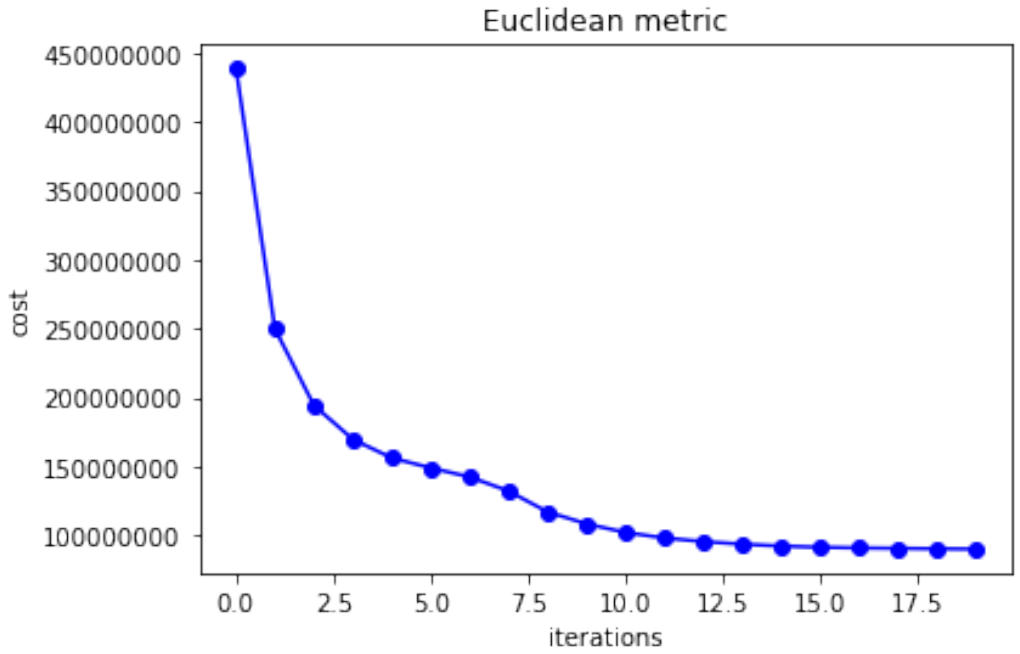


Rysunek 1. Metryka Euklidesowa dla losowych punktów - 3b.txt

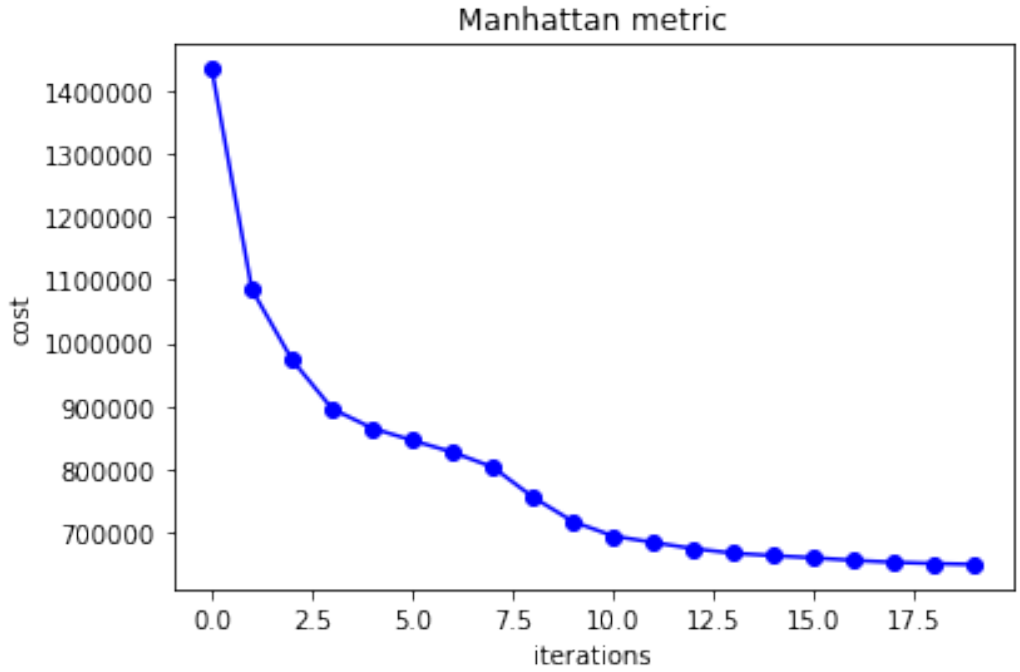


Rysunek 2. Metryka Manhattan dla losowych punktów - 3b.txt

3.2. Wykresy dla punktów najbardziej oddalonych od siebie



Rysunek 3. Metryka Euklidesowa dla punktów oddalonych najdalej do siebie - 3c.txt



Rysunek 4. Metryka Manhattan dla punktów oddalonych najdalej do siebie - 3c.txt

4. Wnioski

- Przy centrach wybieranych z najbardziej oddalonych od siebie punktów zmiana kosztu w kolejnych iteracjach jest znacząco większa niż w przypadku punktów losowych. Jest to spowodowane tym, że w pierwszych kilku iteracjach centra będą znajdowały się na "krańcach" zbioru danych, dlatego odległości od punktów są bardzo duże.
- Dla losowo wybranych centrów metryka Euklidesowa wydaje się być bardziej optymalna. Kosz pierwszej iteracji jest mniejszy, niż koszt 20 iteracji przy najbardziej oddalonych punktach.
- Odwrotnie jest w przypadku metryki Manhattan. Tutaj lepszą metodą wydaje się branie najbardziej oddalonych punktów.
- Przypadek 2 jest jedynym, w który występuje wzrost kosztu w kilku kolejnych iteracjach. Anomalia ta może wynikać z niekorzystnego doboru losowych punktów wybranych na centra.
- *Apache Spark* jest szczególnie przydatny do równoległego przetwarzania rozproszonych danych za pomocą algorytmów iteracyjnych.

Bibliografia

[1] *Jupyter Notebook Python, Spark Stack* `https://hub.docker.com/r/jupyter/pyspark-notebook`

[2] *Plik Docker Compose do zadania 2* `https://github.com/jurczewski/PiADZD/blob/master/zad2/docker-compose.yml`