K-means Clustering z wykorzystaniem CUDA

K-means Clustering służy do grupowania podobnych do siebie punktów w zbiorze danych. Wykorzystuje w tym celu prosty algorytm:

1. Losowo wybieramy k punktów - “znaczniki”. Następnie przypisujemy każdemu punktowi jego grupę (1,2…,k). Punkt należy do grupy l, jeśli jego odległość do znacznika l jest najmniejsza spośród k znaczników.
2. Obliczamy centroid każdej grupy (środek masy punktów)
3. Ponownie przypisujemy każdemu punktowi ze zbioru jego grupę, wg. zasady w punkcie 1.
4. Powtarzamy krok 2,3 aż obliczone centroidy w punkcie 2. nie będą się zmieniać.

Do testów wykorzystam własną implementację K-means w C++ opartą na artykule: https://reasonabledeviations.com/2019/10/02/k-means-in-cpp/. Dodatkowo sprawdzę czas działania tego algorytmu w bibliotece Python scikit-learn.

Algorytm można zrównoleglić - obliczanie odległości każdego punktu do centroidu jest niezależne.

Obliczenie centroidu również możemy zrównoleglić używając redukcji.