Mateusz Budnik 171927

Clustering using the K-mean algorithm

Finding clusters (collections of data points aggregates together because of certain similarities) using the k-mean algorithm.

Algorithm:

1. Specify the number of clusters
2. For a given set of N elements, choose k elements and make them cluster centers
3. Calculate the distance of each object from the center of the clusters
4. Based on the distance, assign the object to the nearest cluster
5. Recalculate cluster centers
6. Go to step 3 or finish

Sample solution (pseudocode)

```
Int main(){
  Int k = …
  load data
  choose randomly k cluster centers
  For( steps count){
    GPU thread————————————————————————————————
      assign data point to appropriate GPU thread based on calculated id
      calculate the distance
      assign point to cluster
    ——————————————————————————————————————————
    synchronize()
    calculate new cluster centers
  }
}
```

Data

To test the implementation of the algorithm the weather data will be used. Observations will be clustered to groups such as “sunny weather”, “rainy weather”, “windy weather” and so on. Each observation consist information such as air pressure, air temperature, average wind direction and some more. There are 1587257 rows in the dataset.

The dataset is availabile at this link: https://www.kaggle.com/julianjose/minute-weather