

RMIT Classification: Trusted

```

mapper.py
Input: point file
c ← LoadClusters()
For each point p in point file
    n ← NearestClusterID(clusters c, point p)
    p ← ExtendPoint(point p)
    emit(clusterid n, point p)

reducer.py
Input: clusterid n, points [p1, p2, ...]
s ← InitPointSum ()
For each point p ∈ [p1, p2, ...] do
    s ← s + p
m ← ComputeCentroid (point s)
emit(clusterid n, point p)
    
```

Lloyd's Algorithm

1. Select k random point $\{S_1, S_2, \dots, S_k\}$ as seeds.
2. **Until clusters converge:**
 - Assign each point x_i to the cluster c_j such that $d(x_i, S_j)$ is minimal
 - For each point x_i , find nearest centroid c_j , assign the point x_i to cluster
 - Refine the seeds to the centroid of each cluster
 - For each cluster, the centroid is updated on each dimension k :

$$c_j^{(t+1)} = \frac{1}{|X_j^{(t)}|} \sum_{x \in X_j^{(t)}} x$$

If not converged

5/15/2022 Big Data Processing

```

#!/bin/bash
i=1
while :
do
    hadoop jar ../../../../hadoop-streaming-3.1.4.jar \
        -D mapred.reduce.tasks=1 \
        -D mapred.text.key.partitionner.options=-k1 \
        -file centroids.txt \
        -file ./mapper.py \
        -mapper ./mapper.py \
        -file ./reducer.py \
        -reducer ./reducer.py \
        -input /testMapReduce/dataset.txt \
        -output /testMapReduce/mapreduce-output$i \
        -partitioner org.apache.hadoop.mapred.lib.KeyFieldBasedPartitioner

    rm -f centroids1.txt
    hadoop fs -copyToLocal /testMapReduce/mapreduce-output$i/part-00000 centroids1.txt
    seeiftrue="python reader.py"

    if [ $seeiftrue = 1 ]
    then
        rm centroids.txt
        hadoop fs -copyToLocal /testMapReduce/mapreduce-output$i/part-00000 centroids.txt
        break
    else
        rm centroids.txt
        hadoop fs -copyToLocal /testMapReduce/mapreduce-output$i/part-00000 centroids.txt
    fi
    i=$((i+1))
done
    
```

Is it okay if the reduce task number is not 1?

Current locations of k centroids are saved in this file.

What is the role of $\$i$ here?

Converged? - whether the new and old locations of each centroid are same or not

output final locations of k centroids

Update locations of the k centroids