# DSEID ZG522 Big Data Systems

## Assignment II

## Hadoop Map-Reduce Framework

**Note:**

- *This is a take-home assignment to be carried out by each learner individually and independently*.
- *There are two programming exercises - requiring only one dataset to be used – on Hadoop MapReduce framework.*
- *The learner may use any programming language as the programming interface, while Java or Python is recommended.*
- *You may consult / discuss with other learners peripheral aspects such as the environment but not on solving the specific problems in terms of design or implementation.*

**End of Note.**

### Learning Outcomes:

- Ability to write programs that perform data processing using the MapReduce programming paradigm
- Secondary outcome is getting familiarity with custom implementation of machine learning algorithms for clustering and classification techniques

### Problem Context:

Assume that you are working as analyst for “MyMall”, a supermarket chain. The mall has collected some interesting characteristics of customers who had visited

the mall earlier. (Refer the attached Mall_Customers.csv file for the same). The marketing management is planning a campaign to increase the sales of a new product. Before that they want to analyze the segmentation of existing customers so that they can have the clearer picture about the customer categories.

**Exercise 1:**

Write a MapReduce program that will take the mall dataset as input and produce the clusters representing customer segments using k-means clustering. You may have to do some preprocessing of the data as required. You may write your own custom implementation of k-means matching the problem statement and dataset. The program should output the cluster number, centroid used and number of records belonging to that cluster.

**Exercise 2:**

As an outcome of Exercise 1, you will obtain clusters in the given dataset. Apply appropriate labels to those clusters. Explain your logic behind nomenclature of the clusters. Show the sample dataset.

**Exercise 3:**

As an outcome of Exercise 2, you will obtain a labelled dataset. Using this labelled dataset, write a MapReduce program that will predict the customer segment for the unknown customer record using the k-nearest neighbor classification technique.

You may write your own custom implementation matching the classification requirement given above. .

Your program should output the confusion matrix and accuracy rate obtained based upon your implementation.

Given an unknown customer record, your program should output the 3 customer records that are similar to him/her and using majority voting should predict the category for customer.

References


1) Hadoop Lab sheets and records. (*Available on the course website.*)

2) Your text book for data mining (for K-means and K-NN)

**3)** (for K-means): Oracle Blogs.

https://blogs.oracle.com/datascience/introduction-to-k-means-clustering

**4) (for k-nn) :**

http://www.math.le.ac.uk/people/ag153/homepage/KNN/OliverKNN_Talk.pdf

5) Kaggle customer segmentation example


==========END==========