DESCRIPTION OF COURSEWORK

Course Code	SWE404
Course Name	Big Data Analytics
Lecturer	Lu Yang
Academic Session	2020/04
Assessment Title	Assignment 4

A. Introduction/ Situation/ Background Information

This assignment evaluates the students' understanding of Lecture 10, which are mainly about MLlib and clustering algorithms.

B. Course Learning Outcomes (CLO) covered

At the end of this assessment, students are able to:

CLO 1 Describe the concept and understand broad knowledge of Big Data.

CLO 2 Demonstrate appropriate knowledge of Hadoop, Spark, Storm and Map Reduce framework and apply them to build a VM-based environment.

CLO 3 Integrate knowledge and understanding of the basic principles, techniques and methodologies of organizing and searching Big Data, and apply them to create value with business insight.

CLO 4 Identify the awareness of the wide applicability of Big Data for real-world practical purposes.

CLO 5 Demonstrate the need to continually follow Big Data development trends.

C. University Policy on Academic Misconduct

1. Academic misconduct is a serious offense in Xiamen University Malaysia. It can be defined as any of the following:
   i. **Plagiarism** is submitting or presenting someone else's work, words, ideas, data or information as your own intentionally or unintentionally. This includes incorporating

published and unpublished material, whether in manuscript, printed or electronic form into your work without acknowledging the source (the person and the work).

ii. **Collusion** is two or more people collaborating on a piece of work (in part or whole) which is intended to be wholly individual and passed it off as own individual work.

iii. **Cheating** is an act of dishonesty or fraud in order to gain an unfair advantage in an assessment. This includes using or attempting to use, or assisting another to use materials that are prohibited or inappropriate, commissioning work from a third party, falsifying data, or breaching any examination rules.

2. All the assessment submitted must be the outcome of the student. Any form of academic misconduct is a serious offense which will be penalised by being given a zero mark for the entire assessment in question or part of the assessment in question. If there is more than one guilty party as in the case of collusion, both you and your collusion partner(s) will be subjected to the same penalty.

D. Instruction to Students

This assignment is an **individual** assignment. Each student should submit Jupyter Notebook file on Moodle, named as "Assignment4_YourStudentID.ipynb".

The deadline is **18:00, 29th June**. Overdue penalty will be given to the assignment that is submitted after the deadline.

* Your codes will be sent to a **Plagiarism** detection system for duplication checking. Please write your codes independently. (Modify your code if you copy some fragment from the Internet because your classmates may copy the same fragment.)

E. Evaluation Breakdown

No.	Component Title	Percentage (%)
1.	Assignment 4	100
	TOTAL	**100**

F. Task(s)

You are required to implement k-means algorithm in PySpark. Please be noted that all the data flow should be conducted by RDD or Spark DataFrame. You should have at least the following components:

(a) A data normalization function norm_data(X) that normalizes the training and test data. (10 marks)

(b) A distance calculation function cal_dist(x1, x2) that calculate the distance between x1 and x2 and return the value of the distance. (10 marks)

(c) Centroids initialization function centroid_init(data, k) that selects k initial centroids from the data. (10 marks)

(d) Closest centroid searching function get_closest(x, centroids) that returns the index of the centroid that is closest to the data point x. (10 marks)

(e) Centroid updating function centroid_update(X, index) that returns the updated centroid by the index of the closest centroid of each data point. (10 marks)

(f) The k-means main function kmeans(data, k, max_iter, tol) that returns the k centroids. (10 marks)

(g) The prediction function predict(X, centroids) to use the trained centroid to predict the data point X and return the predicted labels. (10 marks)

(h) The evaluation function evaluate(predictions, labels) that returns the Silhouette Coefficient of the clustering result. (10 marks)

The function arguments can be different based on your design. Finally, compare your implementation with the k-means in MLlib and try to analyze the difference. (20 marks)

There are two bonus questions:

(i) Implement k-means++ in centroid initialization. (10 marks)

(j) Implement NMI for evaluation. (10 marks)

The training dataset:

https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass/pendigits

The test dataset:

https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass/pendigits.t