Đề số 01

Phần B (5 điểm): Lab Thực hành (2 bài)
Thời gian làm bài: **45 phút**
Sinh viên đăng nhập và thực hiện nội dung kiểm tra theo yêu cầu trên Canvas.
Submit kết quả bài làm.

Bài 1: MapReduce hoá với Hadoop

Thực hiện bài toán Kmeans trên tập dữ liệu cho sẵn như sau:

Input (đầu vào):

- Dataset point (file dữ liệu: point_3.txt);
- File JAR tương ứng (Kmeans.jar);

Output (đầu ra), yêu cầu submit kết quả 1 file zip gồm:

- File đầu ra: result.txt (sau khi chạy Hadoop Mapreduce hoá với tập dữ liệu trên sử dụng file JAR cung cấp);
- Screenshot quá trình chạy Hadoop trên terminal và tiến trình ở trên WebUI.

Lưu ý: lệnh hadoop jar được chạy với cấu hình sau:

-Dlines=30, -Dmaxloop=50 -Dk=4 -Dthresh=0.0001 -DNumReduceTask=2

Bài 2: Apache Spark (MLLib)

Cho trước những đầu vào sau:

- Dataset dữ liệu giá nhà cho thuê - Housing Price Dataset (Housing.csv);
- File notebook mẫu;

Thực hiện các bước sau và chạy trên Jupyter Notebook:

1. Sử dụng SparkSession để khởi tạo một phiên Spark.
2. Dữ liệu được đọc vào DataFrame từ một tệp CSV.
3. Chuẩn bị dữ liệu (Ví dụ: dữ liệu được chuẩn bị và biến đổi bằng cách sử dụng VectorAssembler để tạo một vector chứa các đặc trưng.)
4. Tập dữ liệu được chia thành tập huấn luyện và tập kiểm tra.
5. Lựa chọn một mô hình để huấn luyện.
6. Mô hình được đánh giá trên tập kiểm tra và dự đoán kết quả.

Yêu cầu submit kết quả là file Jupyter Notebook sau khi đã thực hiện đầy đủ các bước.