Ton Duc Thang University
Faculty of Information Technology

THUYẾT TRÌNH – QT2

Môn: Xử lý dữ liệu lớn

Thời gian làm bài: 03 tuần

I. Hình thức

- Đề tài được thực hiện theo nhóm **04** – **05** sinh viên.
- Nhóm sinh viên thực hiện các yêu cầu và nộp bài theo hướng dẫn bên dưới.

II. Yêu cầu

Nhóm sinh viên tìm hiểu và cài đặt các thuật toán được yêu cầu trên Google Colab.

a) Câu 1 (4.0 điểm): k-Means trên PySpark DataFrame

- Sinh viên cài đặt thủ công thuật toán k-Means, không sử dụng lớp đối tượng tương ứng trong các thư viện lập trình.
- Phát sinh dữ liệu 2 chiều, lưu trữ bằng DataFrame của PySpark.
- Chạy thuật toán, trực quan hoá kết quả phân cụm sau mỗi bước lặp bằng biểu đồ.

b) Câu 2 (4.0 điểm): CURE trên PySpark DataFrame

- Thực hiện lại câu 1 với thuật toán CURE.

c) Câu 3 (2.0 điểm): Thuyết trình

- Sinh viên viết báo cáo kết quả đề tài theo hình thức thuyết trình. **KHÔNG CÓ MẪU THUYẾT TRÌNH, NHÓM SINH VIÊN TỰ TỔ CHỨC NỘI DUNG**.
- Các thông tin tối thiểu cần có.
    - Danh sách sinh viên: MSSV, Họ tên, Email, Phân công công việc, Mức độ hoàn thành.
    - Tóm tắt cách xử lý từng yêu cầu, nên diễn đạt bằng mã giả/sơ đồ.
    - HẠN CHẾ TỐI ĐA NHÚNG MÃ NGUỒN THÔ VÀO BÀI THUYẾT TRÌNH.
    - Các nội dụng tìm hiểu cần trình bày cô đọng, có ví dụ trực quan.
    - Thuận lợi và khó khăn trong đề tài.

  - Bảng tự đánh giá mức độ hoàn thành các yêu cầu.
  - Tài liệu trích dẫn ghi theo định dạng IEEE.
- Yêu cầu về định dạng: tỷ lệ slide 4x3, hạn chế dùng nền tối/màu sắc vì máy chiếu mờ, đảm bảo khi in bài thuyết trình dạng trắng đen thì các nội dung vẫn rõ ràng.
- Thời lượng tối đa cho phần thuyết trình là **10 phút**.

III. Hướng dẫn nộp bài

- Tạo thư mục với tên theo cú pháp
  <MSSV1>_<MSSV2>_<MSSV3>_<MSSV4>_<MSSV5>
  trong đó gồm:
  - **source.ipynb** → chứa mã nguồn đồ án (giữ lại các kết quả chạy)
  - **source.pdf** → kết xuất pdf của notebook
  - **presentation.pdf** → bài thuyết trình.
- Nén thư mục thành tệp zip và nộp theo deadline.

IV. Quy định

- **Nhóm sinh viên nộp trễ hạn bị 0.0 điểm toàn nhóm.**
- **Sai sót mã số sinh viên nào trong tên tệp nộp bài thì sinh viên tương ứng bị 0.0 điểm.**
- **Thiếu sót các tài liệu được yêu cầu trong tệp nộp bài sẽ bị trừ tối thiểu 50% điểm phần thuyết trình.**
- **Mọi hành vi sao chép code trên mạng, chép bài bạn hoặc cho bạn chép bài nếu bị phát hiện đều sẽ bị điểm 0.0.**
- **Nếu bài làm của sinh viên có dấu hiệu sao chép trên mạng hoặc sao chép nhau, sinh viên sẽ được gọi lên phỏng vấn code riêng để chứng minh bài làm là của mình.**

-- HẾT –