# HƯỚNG DẪN THỰC HIỆN BÀI TẬP LỚN CHO SINH VIÊN

## (Dữ liệu lớn - Học kỳ 2 năm học 2024-2025)

## I. Mục đích, yêu cầu:

- Phát huy tính sáng tạo, vận dụng kiến thức và rèn luyện kỹ năng, kiến thức lập trình của sinh viên để giải quyết một bài toán phân tích dữ liệu lớn trong thực tế.
- Sinh viên thực hành, thiết kế, soạn thảo codes lập trình dựa trên khung Spark, sử dụng ngôn ngữ lập trình R/ Python và các thư viện hỗ trợ để giải quyết một bài toán phân tích dữ liệu lớn có ý nghĩa thực tế.
- Trong trường hợp sinh viên kế thừa codes trên mạng thì cần mô tả rõ phần nào kế thừa và phần nào là codes tự làm, codes phát triển mới.
- Mỗi nhóm sinh viên có 3-4 thành viên.

**II. Chủ đề bài tập lớn gợi ý** (Sinh viên được chọn chủ đề khác tùy theo sở thích nhưng cần được giảng viên duyệt):

Mỗi nhóm sinh viên thảo luận, trao đổi và thống nhất ý tưởng, nội dung làm bài tập lớn của nhóm, ý tưởng phát triển có thể dựa trên lựa chọn đề bài nêu dưới đây (nhưng không bắt buộc, sinh viên được tùy ý thay đổi, tự lựa chọn đề bài khác yêu thích, nhưng cần báo cáo giảng viên duyệt). Sau đó, mỗi nhóm sinh viên phân công, phối hợp cùng nhau thực hiện các nội dung bài tập lớn. Các chủ đề gợi ý:

- Phân tích dữ liệu lớn trong các bài toán có ý nghĩa thực tế như:

+ Xu hướng khách hàng mua sắm hàng hóa khi đi siêu thị

+ Phân tích dịch vụ cung cấp khách hàng dịch vụ viễn thông di động

+ Phân tích xu hướng phân bố tiền lương nhân viên

+ Phân tích doanh thu theo mặt hàng, theo vị trí địa lý

+ Phân tích bản đồ và đặc điểm các dịch bệnh

+ Phân tích đặc điểm tập trung các ca bệnh Covid

+ Phân tích doanh thu công ty điện thoại di động

+ Phân tích xu hướng, bản đồ du lịch

+ Phân tích khách thuê phòng khách sạn, phòng bán chạy, phòng ít khách

+ Phân tích xu hướng khách hàng mua hoặc hoạt động sản xuất, bán hàng của các đại lý bán ô tô, xe máy.

+ …etc...

**- Lưu ý**: Nên chọn tập dữ liệu có kích thước lớn để chứng minh hiệu quả của Spark đối với xử lý dữ liệu lớn. Có thể so sánh hiệu suất mô hình khi lập trình theo cách thông thường với khi lập trình theo mô hình phân tán và xử lý song song.

- Cộng điểm đối với các nhóm thiết lập được cluster xử lý dữ liệu phân tán.

## III. Tham khảo nguồn dữ liệu:

- Các datasets trên Kaggle Website
- Các datasets trên UCI (UC Irvine Machine Learning Repository)
- Các nguồn datasets khác rất phong phú tải về miễn phí từ Internet
- Hugging Face
- …etc...

## IV. Kết quả:

- File MS Word quyển báo cáo bài tập lớn theo đúng mẫu quy định của nhà trường (có file mẫu kèm theo trên LMS).
- Sản phẩm demo hoạt động tốt, đúng yêu cầu bài toán.
- Sinh viên không cần giao nộp mã nguồn (sinh viên được giữ bản quyền), chỉ cần chiếu projector giải thích cách làm, kết quả và giải thích codes, công nghệ, kỹ thuật thể hiện được kết quả sinh viên thực sự đã làm, đã thực hành và học được kiến thức.

## V. Đánh giá:

- Hoàn thành bài tập lớn và các yêu cầu báo cáo tiến độ đúng hạn
- Khối lượng và chất lượng phát triển sản phẩm demo.
- Các bài tập lớn có phạm vi giải quyết bài toán rộng, sử dụng nhiều kỹ thuật lập trình dữ liệu lớn khác nhau sẽ được đánh giá cao hơn.
- Các bài tập lớn cũng được đánh giá cao nếu có nội dung độc đáo, có ý nghĩa thực tế, có nhiều nội dung thực hiện, có độ phức tạp cao, thiết kế logics, codes rõ ràng mạch lạc...
- Điểm sẽ tính theo tỷ lệ % đóng góp của các thành viên (do nhóm sinh viên tự thống nhất) để làm căn cứ chấm điểm từng sinh viên, đảm bảo công bằng (không cào bằng). Đồng thời, điểm cá nhân sinh viên cũng dựa trên kết quả thi vấn đáp khi báo cáo bài tập lớn của nhóm.
- Các nhóm sao chép nội dung, mã nguồn của nhau sẽ bị điểm 0 cho cả nhóm đi sao chép và nhóm cho bạn sao chép.

---