**SẮP XẾP DỮ LIỆU LỚN**

Môn học: cấu trúc dữ liệu và giải thuật

TA: Nguyễn Thái Vũ

Email: vunguyenthai73@gmail.com

1. **Giới thiệu**

* Mục tiêu: rèn luyện cho sinh viên khả năng sắp xếp file dữ liệu không fit trên RAM.
* Thời gian: 3 tuần.
* Sinh viên làm việc nhóm (2 - 3 người). Không được phép là cá nhân, do bài tập ngoài rèn luyện kỹ năng lập trình, còn rèn luyện khả năng làm việc nhóm.

1. **Bài toán**
   1. **Phát biểu bài toán**

* Bộ dữ liệu: <https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews>. Sinh viên vào đường link vào download file “Books\_rating.csv”.
* Nhiệm vụ là sắp xếp theo “The Id of Book” từ nhỏ đến lớn. Trường hợp hai rating có “book id” bằng nhau, thì cứ sắp xếp ngẫu nhiên.
* Kết quả sắp xếp được lưu vào file “sorted\_books\_rating.csv”
  1. **Gợi ý**
* Sinh viên dùng ý tưởng thiết kế thuật toán Merge sort hoặc B Tree hoặc dùng bất kỳ các thiết kế nào sinh viên tìm hiểu.
* Dữ liệu khoảng 3GB, để đảm bảo sinh viên có thao tác sai thì vẫn không làm ảnh hưởng đến máy tính. Tuy nhiên, sinh viên cần xử lý dữ liệu như việc dữ liệu không fit trên RAM.
* Gợi ý cho việc chạy code:
  + Tạo dữ liệu kiểm thử file nhỏ.
  + Chạy thử trên file nhỏ.
  + Nếu chạy trên file nhỏ thành công, thì chạy trên toàn bộ dữ liệu.

1. **Đánh giá**

* Các yếu tố đánh giá:
* Làm việc nhóm (10%)
* Thiết kế kiến trúc (30%)
* Áp dụng vào bài toán sắp xếp (30%)
* Báo cáo kết quả, độ phức tạp (30%)
* Điểm cộng: dùng các công cụ quản lý source code và làm việc nhóm. Ví dụ: github, … (5%)
* Báo cáo gồm các nội dung:
  + Giới thiệu.
  + Thành viên, đóng góp, mức độ hoàn thành công việc.
  + Mô tả kiến trúc và thuật toán được áp dụng vào bài toán.
  + Kết luận .

1. **Lưu ý**

* Ngôn ngữ lập trình: C/C++. Sinh viên được phép dùng các thư viện, công cụ hỗ trợ bằng ngôn ngữ C/C++
* Sinh viên có thể tạo nhiều file source code, file thư viện, …
* Các nhóm có thể thảo luận với nhau và tham khảo các ý kiến từ nhiều nguồn. Nhưng code phải là của bạn. Mọi hành vi gian lận đều sẽ xử phạt điểm 0.
* Nội dung nộp gồm:
  + Folder <MSSV1>\_<MSSV2>\_<MSSV3>:
    - File report.pdf
    - Folder source code.
  + Nén lại và nộp file \*.zip