Bài thực hành số 4

Lưu ý: Chủ đề trong diễn đàn nhóm tạo ra dùng để nộp các bài tập, các bài báo cáo, các kết quả thảo luận,.. trong suốt học kỳ. Các nhóm có thể tham khảo nội dung của các nhóm khác nhưng tuyệt đối không sao chép cho các bài kết quả của nhóm mình (đề cao tính trung thực).

Trong các buổi làm việc: yêu cầu làm việc theo nhóm, các nhóm có thể trao đổi, hỗ trợ nhau khi cần thiết. Nhóm có thể mang máy tính cá nhân cùng làm việc.

Câu 1:

- Sử dụng tập tin GroceryStore-AssociateRules.txt để thử nghiệm thuật toán Luật Kết Hợp.
- Sinh viên tự chọn các giá trị ngưỡng!

Câu 2:

- Sử dụng tập tin kết quả thi TNTHPT 2020 của TpHCM để thực hiện các yêu cầu sau:
- Tạo tập tin KHTN chỉ chứa các thí sinh dự thi tốt nghiệp các môn khối tự nhiên.
- Xóa bỏ các vùng không cần thiết, đánh số lại vùng SBD nhằm bảo đảm không lộ lọt thông tin riêng tư.
- Với mỗi thí sinh: chỉ giữ lại điểm thi 6 môn (bao gồm T-AV-V và 3 môn Lý-Hóa-Sinh).
- Sử dụng ứng dụng bất kỳ để thay đổi nội dung 6 vùng trên theo qui tắc sau: Nếu điểm môn thi đó>8 thì thay thế bằng 1 ngược lại thay thế bằng 0
- Dùng thuật toán Luật Kết Hợp để phát hiện sự liên quan giữa các môn học.

Kết quả câu 1 và câu 2 sẽ nộp trong buổi thực hành.

Câu 3:

Tập tin RestaurantDataset.csv chứa thông tin khoảng trên 100.000 hiệu ăn/nhà hàng/cửa hàng bán thức ăn/… tại thành phố NewYork. Dữ liệu có 3 cột. Cột đầu cho biết khu vực của cửa hàng (có 5 khu vực) ví dụ Manhattan/Brooklyn, cột 2 cho biết lĩnh vực/phân loại nhà hàng (ví dụ nhà hàng Pháp, nhà hàng hải sản, nhà hàng Thái,…) và cột thứ 3 là 1 ký tự đánh giá chất lượng của cửa hàng (có 5 ký tự A/B/C/P/Z)

Yêu cầu:

- Tìm cách chuyển đổi dữ liệu gốc này để thực hiện thuật toán Luật Kết Hợp.
- Ví dụ sau khi thực hiện thuật toán này, có thể phát hiện 1 số luật thú vị sau:

[Japanese] => [MANHATTAN] (Conf: 60.3%, Supp: 1.9%)
Phần lớn nhà hàng Nhật ở Newyork nằm ở khu vực Manhattan.

[B,Japanese] => [MANHATTAN] (Conf: 58.4%, Supp: 0.6%)
Phần lớn nhà hàng Nhật có đánh giá chất lượng B nằm ở khu Manhattan.

[STATENISLAND] => [A] (Conf: 62.5%, Supp: 2.2%)
Các nhà hàng ở khu vực StateIsland được đánh giá loại A.

Câu 4: Phần thu hoạch

Phân công thực hiện, thực hiện ở nhà, thời hạn 1 tuần và gửi vào chủ đề trong diễn đàn.

- Báo cáo các kết quả của câu 2.
- Thuật toán FP-Growth cũng là một cải tiến của Luật Kết Hợp. Thử sử dụng thuật toán này (bằng bất kỳ ứng dụng nào) để thực hiện lại yêu cầu câu 3.

Lưu ý: Nhóm có thể dùng bất kỳ ứng dụng để thực hiện báo cáo (Power Point, Word, Clip/Video,…). Phần thu hoạch này cần nộp lên diễn đàn trước buổi học lý thuyết kế tiếp: GV có thể mời 1 nhóm lên trình bày.