BÀI TẬP THỰC HÀNH

TẬP PHỔ BIẾN & LUẬT KẾT HỢP

1 Qui định

- **Thời gian làm bài**: 1 tuần (Xem deadline trong link nộp bài trên moodle) .
- **Nộp bài**: Tập tin nén (.zip hoặc .rar) đặt tên theo MSSV. Trong đó bao gồm:
  - Tập tin báo cáo: Đặt tên theo MSSV và lưu với một trong các định dạng: doc, docx, pdf.
  - Các tập tin dữ liệu theo yêu cầu của bài tập.
- *Các hành vi sử dụng toàn bộ/một phần bài làm của người khác sẽ bị 0 điểm cho toàn bộ phần thực hành.*

2 Nội dung bài tập

Bài 1 – Apriori & FP-Growth

Cho cơ sở dữ liệu giao dịch như sau:

Mã giao dịch	Các hạng mục
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
5	Jam, Soda, Chips, Milk
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

a) Với minsup = 35% và minconf = 80%, hãy áp dụng thuật toán Apriori và FP-Growth để tìm tất cả các tập phổ biến. So sánh kết quả của 2 thuật toán.

b) Với kết quả thu được ở câu a), liệt kê tập phổ biến tối đại, tập phổ biến đóng.

c) Từ các tập phổ biến thu được ở câu a), liệt kê tất cả các luật kết hợp có dạng $\{item1, item2\} \rightarrow item3$ thỏa ngưỡng minsup và minconf đã cho.

Dữ liệu nộp lại:

- Đáp án cho các yêu cầu a, b, c: Trình bày vào tập tin báo cáo.

Bài 2 – Thực hành với công cụ WEKA

Cho tập dữ liệu về vùng phân bố một số loài thực vật ở các bang của nước Mỹ và Canada trong 3 tập tin như sau:

- **plants.data**: liệt kê tên của loài thực vật và các vùng phân bố tương ứng. Mỗi loài thực vật được liệt kê trên một dòng.
- **Stateabbr.txt**: mô tả tên của các bang có trong dữ liệu và tên viết tắt tương ứng.

Yêu cầu:

a) Hãy chuyển dữ liệu trong tập tin plants.data từ dạng giao dịch sang dạng nhị phân như sau:
   - Cột đầu tiên là tên các loài cây, các cột tiếp theo tương ứng với các vùng phân bố.
   - Dòng đầu tiên liệt kê tên viết tắt của các vùng phân bố. Lưu ý bỏ cột đầu tiên.
   - Các dòng tiếp theo: mỗi dòng tương ứng với một loài thực vật.
   - Giá trị các ô: nhận một trong 2 giá trị "y" hoặc "n". Trong đó, "y" cho biết loài thực vật có xuất hiện trong vùng phân bố tương ứng và "n" là không xuất hiện.
   - Lưu lại tập tin theo định dạng csv với tên plants.csv.

Gợi ý: sinh viên có thể dùng các chức năng trong Microsoft Excel như "**text to column**", "**copy và paste transpose**" (chuyển dòng thành cột), hàm **IF**, copy công thức… để hỗ trợ việc chuyển đổi.

Lưu ý: Trong trường hợp vùng phân bố không xuất hiện trong tập tin mô tả mà xuất hiện trong tập dữ liệu thì sinh viên thêm cột cho vùng đó.

b) Sử dụng công cụ WEKA và trả lời câu hỏi sau:
   - Có tất cả bao nhiêu loài thực vật.
   - Có tất cả bao nhiêu vùng phân bố.
   - Số loài thực vật trên mỗi vùng phân bố.
   - Vùng phân bố có ít loài cây nhất, cho biết số lượng tương ứng.
   - Vùng phân bố có nhiều loài cây nhất, cho biết số lượng tương ứng.
   - Trung bình một vùng phân bố có bao nhiêu loài cây.

c) Điều chỉnh nội dung của tập tin để khai thác mẫu phổ biến như sau:
   - Thực hiện thay thế toàn bộ giá trị "n" thành "?"
   - Cột đầu tiên (tên loài thực vật), không cần thiết cho việc khai thác nên hãy xóa nó đi.
   - Lưu kết quả dưới định dạng arff, với tên là plants.arff.

d) Khai thác tập phổ biến:

Sử dụng thuật toán Apriori trong Weka để khai thác tất cả các tập phổ biến với $minsup = 0.1$. Trình bày kết quả định lượng trong tập tin báo cáo như sau:

Kích thước	Số lượng tập phổ biến
1 hạng mục	17
2 hạng mục	30
…	…
n hạng mục	…

Danh sách tất cả các tập phổ biến thỏa yêu cầu được lưu trong tập tin **F1.doc** (Sinh viên copy đoạn chính giữa 2 dòng "**Generated sets of large itemsets**" và "**Best rules found**")

e) Khai thác luật kết hợp

- Sử dụng thuật toán FP-Growth trong Weka để khai thác luật kết hợp
- Chỉ liệt kê những luật sinh ra từ tập phổ biến có kích thước lớn nhất và có độ tin cậy từ 0.95 trở lên.
- Trình bày kết quả định lượng trong tập tin báo cáo như sau:

Tập hạng mục phổ biến	Số lượng luật
$ak = y, ca = y, bd = y$	2
…	…

Danh sách tất cả các luật kết hợp thỏa yêu cầu được lưu trong tập tin **AR.doc**(.pdf hoặc .docx)

Dữ liệu nộp lại

- Tập tin dữ liệu: plants.csv, plants.arff
- Tập tin báo cáo
- Các tập tin khác: F1.doc, AR.doc