**Bài tập 2 – Khai thác tập phổ biến và luật kết hợp**

**Mining frequent Itemsets and Association Rules**

Môn học: Khai thác dữ liệu và ứng dụng - 19KHDL

*Lecturer: Lê Ngọc Thành*

*TA: Nguyễn Thái Vũ*

*Ghi chú: Tài liệu này biên soạn dựa trên một số bài tập của CE Department- Sharif University of Technology.*

1. **Giới thiệu**

* Mục tiêu của bài tập: sinh viên làm quen với các thuật toán về khai thác tập phổ biến và luật kết hợp.
* Thời gian: 2 tuần (chi tiết xem trên moodle)
* Dữ liệu được cho trong file churn.txt
* Sinh viên đọc file mô tả về dữ liệu trong file churn\_description.pdf để hiểu hơn về dữ liệu.

1. **Yêu cầu bài tập**

* Sinh viên sử dụng ngôn ngữ lập trình Python. Sinh viên được phép sử dụng các thư viện core Python, Numpy, Matplotlib, Pandas. Các thư viện như Sklearn không được phép.
* Lưu ý, các bạn có thể tham khảo bất kỳ đâu, tuy nhiên code phải là của các bạn. Các hình thức sao chép mà không ghi mã nguồn, hoặc sao chép giữa các sinh viên sẽ bị 0 điểm môn học.
* Tiền xử lý dữ liệu (20%):
  + Nêu rõ mục đích, lý do, ảnh hưởng của biện pháp tiền xử lý.
  + Ví dụ: đối với các biến có hệ số tương quan cao, ta có thể bỏ đi thuộc tính. Hoặc bỏ đi các cột có tỷ lệ dữ liệu bị thiếu cao.
* Khai thác tập phổ biến và luật kết hợp (80%):
  + Giả sử bạn là nhà phân tích dữ liệu. Mục tiêu của bạn, là muốn khám phá ra các tập phổ biến/tập luật có ích cho việc ra quyết định của doanh nghiệp (xem thêm phần 6.3 Which Patterns Are Interesting? –Pattern Evaluation Methods, sách Data Mining. Concepts and Techniques, 3rd Edition để hiểu rõ hơn thế nào là tập luật có ích).
  + Hãy tự đưa ra 3 câu hỏi và sử dụng các thuật toán phân tích và khai thác tập phổ biến và luật kết hợp (Apriori, FP-growth,…) , để tìm ra câu trả lời tương ứng.
  + Lưu ý, 3 câu hỏi trên, cần đáp ứng sơ bộ các tiêu chí:
    - Bạn muốn biết gì thêm từ dữ liệu?
    - Các tập phổ biến/tập luật có ý nghĩa gì?
    - Các tập phổ biến/tập luật đã trả lời cho câu hỏi của bạn?

1. **Yêu cầu báo cáo**

* Sinh viên đặt tên thư mục là <MSSV> và nén lại thành <MSSV>.zip hoặc <MSSV>.rar
* Sinh viên tổ chức gồm 2 thư mục:
  + Thư mục code: chứa các file code tương ứng. Lưu ý, tên file, hàm, biến, ... cần được đặt có ý nghĩa (comment nếu cần thiết).
  + Thư mục report: chứa file report.pdf, mô tả lại các quá trình:
    - Tiền xử lý dữ liệu: nêu rõ mục đích, lý do, ảnh hưởng của biện pháp tiền xử lý.
    - Khai thác tập phổ biến và luật kết hợp: mô tả 3 câu trả lời tương ứng.
* Đánh giá:
  + Hoàn thành yêu cầu bài tập: 70%
  + Báo cáo tốt, giải thích rõ ràng: 30%

1. Tài liệu tham khảo

* Slide lý thuyết.
* Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition.
* File churn\_description.pdf

Nếu có câu hỏi các bạn có thể gửi lên nhóm Zalo hoặc email trực tiếp vunguyenthai73@gmail.com