Bài thực hành số 3

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 phần mềm tùy ý để gom cụm cho các tập tin Clustering1-NoClass-Train.arff và Clustering2-NoClass-Train.arff như sau:

- Sử dụng thuật toán K-Means để thực nghiệm nhiều lần sau đó cho biết số cụm của mỗi tập tin có khả năng là bao nhiêu?
- Cho biết đặc điểm của từng cụm được phát hiện.
- Hai tập tin Clustering1-NoClass-Test.arff và Clustering2-NoClass-Test.arff chứa các điểm dữ liệu (tương ứng với 2 tập tin huấn luyện). Sử dụng kết quả đã huấn luyện để cho biết mỗi điểm trong các tập tin này thuộc về cụm nào?

Câu 2:

Sử dụng tập tin lưu kết quả thi TNPTTH 2020 của TpHCM trong tuần trước:

- 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 của tổ hợp).
- Sử dụng K-Mean để thử gom cụm và tìm số cụm có khả năng trong tập dữ liệu này.
- Cho biết đặc điểm từng cụm.

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

Câu 3:

Tập tin car.cvs.zip (trên LMS) chứa dữ liệu về các xe ô tô được bán trong cả năm 2020. Sau khi giải nén tập tin này: sẽ có 1 tập tin **dạng text (.csv) và sử dụng bộ mã UTF-8**
Các thông tin trên 1 dòng của tập tin này bao gồm

- car_model:Hiệu xe
- km: Số km đã sử dụng
- imp_exp: Xe nhập khẩu?
- km_1:Số km đã sử dụng (có thể bỏ vùng này)
- imp_exp_1:Xe nhập khẩu? (có thể bỏ vùng này)
- car_type: Loại xe
- out_color: màu sơn
- in_color: màu nội thất
- door_num: số cửa
- seat_num: số chỗ ngồi
- new_old: xe cũ /mới
- car_year: năm lưu hành
- title:Mô tả xe
- price: Giá bán
- area: Khu vực
- poster_name: Người rao
- poster_add: ĐC liên hệ
- poster_tel: Số phone liên hệ

Yêu cầu:

- Chuyển dữ liệu trong tập tin này sang mã Unicode
- Dùng 1 ứng dụng tùy ý để xóa bỏ các thông tin không cần thiết, riêng tư để chuẩn bị gom cụm.
- Chuẩn hóa các vùng dữ liệu còn lại.
- Dùng thuật toán gom cụm và cho biết số cụm có khả năng trong tập tin này là bao nhiêu cụm? cho biết đặc điểm từng cụm.

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 3.
- Tìm hiểu công cụ Gom cụm K-means được cài đặt trên Python và sau đó trình bày chương trình để thực hiện thuật toán gom cụm này bằng Python với tập tin Clustering1-NoClass-Train.arff và dự báo các cụm của từng từng điểm dữ liệu trong tập tin Clustering1-NoClass-Test.arff

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.