KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện thoại: (08) 38.354.266 – Fax: (08) 38.350.096

Khoa Học Web

PROJECT 3

PHÂN TÍCH DỮ LIỆU WEB

Biên soạn:
Lê Ngọc Thành

1. Nội dung

Áp dụng các mô hình học máy để phân tích dữ liệu thu thập từ web.

2. Yêu cầu

Project được thực hiện theo nhóm. Thời gian và cách thức nộp, xem trên Moodle.
Nội dung cần nộp:

- Báo cáo trình bày trong file .doc/.docx/pdf chứa:
  - Thông tin nhóm: tên nhóm, mssv…
  - Mức độ hoàn thành tổng thể của mỗi yêu cầu.
  - Mức độ hoàn thành của từng thành viên.
  - Mô tả theo các yêu cầu trong mục 3.
    Khuyến khích trình bày đơn giản, có hình minh họa.
- Souce code kèm hướng dẫn chạy nếu thực hiện trong môi trường khác Jupyter Notebook hoặc python gốc.
- Dataset nếu có điều chỉnh so với Project 1 và 2 thì cần mô tả thêm.
- Ngôn ngữ lập trình bắt buộc: Python
  - Cho phép sử dụng các thư viện đã được giới thiệu trong lý thuyết.

3. Yêu cầu chi tiết

Từ dữ liệu trực quan hóa ở project 2, nhóm tiến hành triển khai các mô hình học máy để rút ra được các kết luận.

Cụ thể trong project này, nhóm được yêu cầu thực hiện các nhiệm vụ sau:

- Đặt ra các câu hỏi cần áp dụng mô hình học máy để giải quyết. Tập trung chính vào các bài toán phân lớp dữ liệu.
- Số lượng bài toán cần tối thiểu bằng số thành viên của nhóm.
- Mô tả bài toán chi tiết, liên hệ với dữ liệu đã trực quan trong project 2.
- Chọn lựa mô hình các mô hình học máy cho từng bài toán, lý giải tại sao chọn các mô hình đó. Khuyến khích mỗi bài toán chạy nhiều mô hình khác nhau để so sánh đánh giá.
- Mô tả các thức huấn luyện, cách phân chia tập dữ liệu, các độ đo đánh giá chất lượng của mô hình.
- Chạy thực thi thuật toán sử dụng các thư viện. Mô tả từng bước ý nghĩa của từng bước trong báo cáo.
- Trình bày kết quả chạy trên các bộ kiểm thử (test set), trực quan hóa kết quả và nhận xét trên kết quả.
- Kết luận liệu mô hình có giúp trả lời câu hỏi được đưa ra ban đầu không? Tại sao?

4. Những giới hạn

- Không được lấy các code và dữ liệu có sẵn để chạy mà phải chạy trên dữ liệu đã thu thập từ web.
- Một số thư viện như numpy, pandas, seaborn, matplotlib, sklearn, pytorch nên sử dụng.

5. Đánh giá

- Các tiêu chí đánh giá:
    1. Đặt ra các vấn đề cần giải quyết (10%)
    2. Mô tả dữ liệu liên quan (5%)
    3. Chọn lựa, giải thích tính phù hợp của các mô hình học máy trên dữ liệu và bài toán nêu ra (10%)
    4. Thực hiện huấn luyện và mô tả chi tiết các thuật toán được triển khai để huấn luyện (25%)
    5. Mô tả cách phân chia dữ liệu huấn luyện và kiểm thử, lý giải cách phân chia và chứng tỏ kết quả không quá phụ thuộc vào cách phân chia đó. (10%)
    6. Giải thích các độ đo để đánh giá mô hình (5%)
    7. Phân tích và trực quan hóa kết quả thu được, lý giải các điểm quan trọng (25%)
    8. Kết luận về vấn đề nêu ra ban đầu (10%)

6. Qui định

- Bài không có báo cáo và code sẽ không chấm.
- Thành viên không tham gia sẽ không có điểm.
- Các nguồn tài liệu tham khảo (nếu có) cần ghi đầy đủ trong báo cáo ở mục *Tài liệu tham khảo*. Lưu ý cần phân biệt giữa tham khảo và đạo văn.
- Đặt tên thư mục bài làm là MSSV1_MSSV2_MSSV03_..., với MSSV là mã số sinh viên, nén toàn bộ bài nộp thành 1 tập tin trước khi nộp. Nếu kích thước >20MB thì upload lên server ngoài như Google Drive, …, nộp link và giữ link public ít nhất trong 1 năm.
- Bài giống nhau sẽ 0 điểm môn học.

7. Liên hệ

Mọi thắc mắc trong quá trình thực hiện vui lòng gửi mail về lnthanh@fit.hcmus.edu.vn