Project

1 Mô tả

Đồ án này sẽ dựa trên cuộc thi Learning Agency Lab - Automated Essay Scoring 2.0. Nhóm của bạn sẽ thực hiện việc xây dựng và đào tạo một mô hình chấm điểm bài luận của học sinh. Mô hình này có thể giúp xây dựng một hệ thống chấm bài tự động cho học sinh, giúp rút ngắn thời gian chấm điểm của giáo viên và nhanh chóng có phản hồi về điểm số cho học sinh.

Mục tiêu của đồ án này giúp các bạn thực hành việc xử lý trên tập dữ liệu văn bản. Do vậy, nhằm đơn giản hóa bài toán, các bạn cần thực hiện việc xây dựng mô hình dự đoán điểm số của bài luận (từ 1-6) dựa trên bài nội dung bài luận được đưa vào.

Các công việc yêu cầu cụ thể được trình bày dưới đây.

2 Công việc cụ thể

2.1 EDA (Exploratory Data Analysis)

Phân tích Khám phá Dữ liệu giúp chúng ta có cái nhìn đầu tiên về dữ liệu. Trong phần này, các bạn cần thực hiện việc phân tích và thống kê về dữ liệu. Một số phân tích gợi ý: thống kê độ dài của câu hỏi, độ dài của đoạn văn, phân bố của label,....

2.2 Thử nghiệm trên mô hình tự xây dựng

Sinh viên được yêu cầu, sử dụng dataset đã được cung cấp, và tiến hành huấn luyện cùng đánh giá hiệu năng của mô hình. Các nhóm cần tham dự cuộc thi trên kaggle và thực hiện việc submmit lên hệ thống cuộc thi. Kết quả hiệu năng của mô hình được tính dựa trên điểm số trên hệ thống.

2.3 Thử nghiệm trên mô hình ngôn ngữ lớn (Điểm cộng)

Hiện nay, các mô hình ngôn ngữ lớn đang rất phát triển và đạt được nhiều hiệu quả đáng kể trong các tác vụ liên quan đến xử lý ngôn ngữ tự nhiên. Do vậy, các nhóm có thể thử nghiệm việc sử dụng các kĩ thuật tận dụng sức mạnh của mô hình ngôn ngữ lớn như Prompt engineering để giải quyết bài toán này.

Lưu ý: Đây chỉ là phần điểm cộng. Các nhóm có thể lựa chọn việc thực hiện hoặc không.

3 Các nội dung cần nộp

1. Source code: gồm các folder con $Q1, Q2, Q3$ chứa code theo từng yêu cầu 2.1, 2.2, 2.3. Lưu ý, code phải chạy được trên nền tảng Google Colab và kèm file readme để cấu hình nếu cần thiết cho từng phần.
2. Báo cáo: yêu cầu cần đầy đủ các thành phần sau đây:
   (a) Bìa
   (b) Mục lục
   (c) Tự đánh giá: Tự đánh giá về mức độ hoàn thành của từng yêu cầu. Bảng phân công công việc của từng thành viên và mức độ hoàn thành của từng thành viên

(d) Nội dung chính: Chia thành từng mục theo từng yêu cầu của đồ án. Bao gồm:

i. EDA (Exploratory Data Analysis): thể hiện bằng các biểu đồ, các bảng. Mô tả ý nghĩa và giải thích và nhận xét các số liệu.

ii. Thử nghiệm trên mô hình tự xây dựng: Mô tả từng thành phần của mô hình và thực nghiệm độ hiệu quả của mô hình trên bộ dataset đã cho và đưa ra nhận xét.

iii. Thử nghiệm trên mô hình ngôn ngữ lớn(nếu có): Mô tả mô hình. Báo cáo kết quả của thực nghiệm của mô hình trên bộ dataset đã cho và đưa ra nhận xét.

3. Một đoạn video ngắn (15-20p): Các nhóm cần chuẩn bị slide và trình bày về đồ án của nhóm mình. Nếu dung lượng video quá nặng, các nhóm có thể lưu lên youtube hoặc google drive sau đó chèn link vào báo cáo

4. Slide

4 Lưu ý

- Đồ án sẽ được đánh giá trên tỉ lệ giữa khối lượng công việc được trình bày trong báo cáo và số lượng thành viên. Báo cáo cần được trình bày chỉn chu và chi tiết vì 50% điểm đồ án đến từ báo cáo.

- Mỗi nhóm gồm 4 - 6 thành viên. Không chấp nhận các nhóm lẻ. Nhóm có ít hơn / nhiều hơn số thành viên quy định cần gửi email cho GVLT và GVTH cho đến tối đa 2 tuần sau khi công bố đồ án. Lưu ý, khi tham gia cuộc thi trên Kaggle cần bắt đầu tên đội với cú pháp sau: `FIT-HCMUS-<GroupID>`. Các nhóm đăng ký nhóm và lấy mã nhóm tại đây: Danh sách nhóm. Trong trường hợp, nhóm tạo nhiều tài khoản thì có thể thêm thông tin sau tiền tố trên.

- Bất cứ hình thức đạo văn, sao chép có chủ đích dù là vô tình hay cố tình đều sẽ bị 0 điểm toàn môn học. Trong trường hợp cần thiết, GVLT và GVTH có quyền yêu cầu SV/nhóm SV thực hiện vấn đáp để đưa ra quyết định cuối cùng.