Annotation Guideline

GIỚI THIỆU

Hướng dẫn gán nhãn này có sự tham gia chỉnh sửa và thực hiện của các thành viên đến từ nhóm 5:

- 22022530 Nguyễn Nhật Tân
- 22022508 Ngô Việt Anh
- 22022644 Nguyễn Tiến Dũng

Bài toán hướng đến là phân loại thể loại của một bài báo dựa trên tiêu đề:

- Đầu vào: Tiêu đề một bài báo
- Đầu ra: Thể loại của bài báo đó (giáo dục, giải trí, sức khỏe,...)

1. Mục đích

- Trình bày quy trình gán nhãn (annotation) cho tập dữ liệu tiêu đề bài báo.
- Trình bày các vấn đề mà các annotators gặp phải và cách xử lý trong quá trình gán nhãn cho loại dữ liệu này.
- Mục tiêu cuối là cho ra một bộ dữ liệu đã gán nhãn phục vụ cho việc huấn luyện mô hình.

2. Quy tắc gán nhãn

- Đọc và hiểu rõ tiêu đề bài báo trước khi chú thích.
- Chọn thể loại báo phù hợp dựa trên nội dung chính của tiêu đề.
- Tránh tác động của thông tin nằm ngoài tiêu đề, chỉ tập trung vào tiêu đề bài báo.

3. Các loại nhãn

- Kinh tế: các bài báo về kinh tế, tài chính, thị trường, doanh nghiệp, quản lý, cổ phiếu, khởi nghiệp và các sự kiện xu hướng kinh doanh.
- Giáo dục: các thông tin về giảng dạy, học tập, chính sách giáo dục, sự phát triển và thay đổi trong lĩnh vực giáo dục.
- Xe: các bài báo về ô tô, xe máy, công nghệ xe, sự kiện trong ngành công nghiệp ô tô và các thông tin liên quan đến lĩnh vực xe cộ.
- Sức khỏe: các nghiên cứu và xu hướng trong lĩnh vực y tế như bệnh tật, phòng ngừa bệnh, chăm sóc sức khỏe, dược phẩm và các vấn đề y tế công cộng.

- Công nghệ - Game: các bài báo về công nghệ, game, ứng dụng di động, xu hướng công nghệ và các thông tin liên quan đến ngành công nghệ và game.

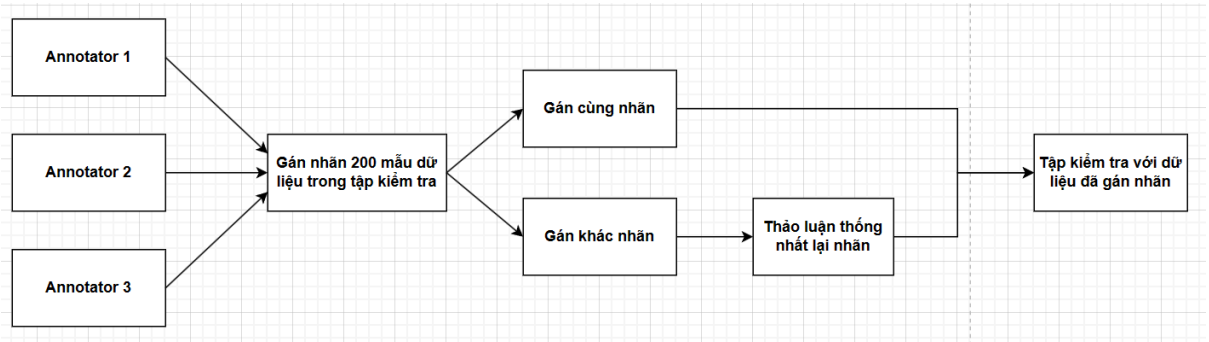
4. Ví dụ

Tiêu đề bài báo	Thể loại
Thủ tướng quy định trường hợp EVN được điều chỉnh giá bán lẻ điện bình quân	Kinh tế
9 thí sinh bị đình chỉ ngay đợt thi Đánh giá năng lực đầu tiên của ĐHQGHN	Giáo dục
Ô tô gầm cao cỡ nhỏ, dưới 600 triệu: Hyundai Venue giảm giá, đấu Kia Sonet	Xe
Bác sĩ viết đơn thuốc cẩu thả khiến bệnh nhân ngộ độc do uống quá liều	Sức khỏe
Bộ xử lý Apple M1, M2 và M3 dính lỗ hổng bảo mật không thể khắc phục	Công nghệ - game

5. Quy trình gán nhãn

- Trong bài toán này có một điểm lợi đó là các bài báo được viết và đăng tải lên đã luôn đi kèm thể loại, đồng nghĩa với việc khi thu thập dữ liệu ta hoàn toàn có thể thu thập cả tiêu đề và thể loại thay vì chỉ thu thập được tiêu đề rồi sau đó mới gán nhãn.
- Tuy nhiên để đảm bảo quy trình xử lý dữ liệu vẫn có phần gán nhãn thì chúng ta vẫn sẽ gán nhãn tay cho phần dữ liệu trong tập dữ liệu kiểm tra với khoảng 100 mẫu.
- Mỗi người sẽ tự gán nhãn cho 100 mẫu trong tập kiểm tra này rồi so sánh tỉ lệ gán nhãn giống nhau của cả 3 người và đánh dấu các tiêu đề mà các thành viên gán nhãn khác nhau.
- Các tiêu đề có kết quả gán nhãn khác nhau cần phải được thảo luận thống nhất lại nhãn được gán và xem lý do gây sự gán lệch.
- Ngoài cách gán nhãn tay ra thì có thể dùng một số thuật toán nhỏ gán tự động thông qua một số từ tiêu biểu của thể loại đó.

  (VD: có từ là tên hãng xe như “toyota”, “honda” => Thể loại xe)



6. Đánh giá kết quả

- Kết quả sau cùng: sai 4/200 mẫu, đạt accuracy 98%.

id	Title	Dũng	Tân	Việt Anh	tag thống nhất
625	Xe Trung Quốc 'nhái' Land Rover một thời tại Việt Nam rớt giá thê thảm	xe	xe	kinh tế	xe
717	Mitsubishi Xpander HEV 2024 thêm động cơ điện, trang bị 'xịn', giá hơn 600 triệu đồng	xe	xe	kinh tế	xe
1165	Các yếu tố giúp Samsung duy trì vị trí dẫn đầu trên thị trường bộ nhớ flash	công nghệ - game	cong-nghe-game	kinh tế	cong-nghe-game
1228	Doanh số iPhone giảm mạnh 24% tại Trung Quốc	kinh tế	cong-nghe-game	kinh tế	cong-nghe-game

- Rút ra số lượng các mẫu gán lệch và suy ra tỉ lệ gây bối rối.

7. Kết luận

- Điểm mạnh: vì được cả 3 thành viên cùng gán nhãn nên sẽ đảm bảo được sự chính xác về nhãn sau cùng so với nhãn thật của bài báo dù có các tiêu đề có thể gây nhầm lẫn.
- Điểm yếu: vì số lượng mẫu được gán nhãn tay chỉ là 100 nên có thể tỉ lệ tiêu đề báo gây bối rối được rút ra sẽ có xu hướng thấp.
- Với tổng cộng 1000 mẫu trong đó 900 mẫu trong tập huấn luyện được gán nhãn sẵn do đặc thù bài toán thì việc gán nhãn tay cho 100 mẫu trong tập kiểm tra là tương đối phù hợp cho việc trải nghiệm bước làm gán nhãn cũng như rút ra một vài thông số.