HƯỚNG DẪN SỬ DỤNG TOOL VÀ QUÁ TRÌNH LÀM DỮ LIỆU

Ngày 17 tháng 2 năm 2024

Mục lục

1 Yêu cầu và quy định khi tham gia xây dựng bộ dữ liệu

Khi tham gia xây dựng bộ dữ liệu cho nhóm nghiên cứu, yêu cầu những người gán nhãn tuân thủ các quy định và yêu cầu sau đây:

- Nghiêm túc và cẩn thận trong quá trình làm dữ liệu, đảm bảo chất lượng dữ liệu được tạo ra. Dữ liệu chưa đúng sau khi được kiểm tra bởi tác giả và người hướng dẫn sẽ được yêu cầu làm lại hoặc bỏ đi.
- Câu trả lời đặt ra phải có sẵn và hợp lý đối với bảng được cho sẵn.
- Mỗi bảng đặt dữ liệu không quá nhiều cũng không quá ngắn (nằm trong khoảng 10 đến 50 câu).
- Không được sử dụng, lan truyền bộ dữ liệu cho mục đích cá nhân khi chưa liên hệ và được sự cho phép của tác giả và người hướng dẫn.

- Hoàn thành nhiệm vụ theo đúng tiến trình được giao (trường hợp có việc đột xuất hoặc không thể hoàn thành tiến độ sắp tới phải báo cáo, giải trình cho tác giả hoặc người hướng dẫn).

Lưu ý: Tiền thưởng sẽ tính trên số lượng QAs không bị lỗi sau khi được kiểm tra bởi người hướng dẫn và tác giả. Tiền thưởng chỉ được bàn giao khi người gán nhãn hoàn thành được tối thiểu 500 điểm dữ liệu (cặp câu hỏi - câu trả lời) và bộ dữ liệu được hoàn thành với số lượng điểm dữ liệu đã đạt đủ. Trong trường hợp người gán nhãn muốn ứng tiền thì phải liên hệ tác giả hoặc người hướng dẫn để có thêm thông tin.

2 Hướng dẫn sử dụng tool

Tool làm dữ liệu của quá trình làm bộ dữ liệu này có thể truy cập trên web thông qua đường dẫn sau:

`vi-wiki-table-qa-git-optim-tannd-ds.vercel.app`

- **Bước 1.** Tải dữ liệu được giao thông qua drive lưu trữ (sẽ được thông báo trong buổi hướng dẫn gặp mặt người gán nhãn).
- **Bước 2.** Truy cập vào tool và load dữ liệu lên bằng cách kéo thả file dữ liệu vào hoặc chọn file (file sẽ có cấu trúc demo_<x>.json, ví dụ: demo_1.json) và chọn "To Next Step!".
- **Bước 3.** Sau khi giao diện hiện lên thông tin bảng dữ liệu (bên phải giao diện) và khu vực chọn hint - đặt QAs (Questions and Answers) (bên trái giao diện), người gán nhãn tiến hành đặt QAs theo hướng dẫn (sẽ được chi tiết trong mục **3. Hướng dẫn làm dữ liệu**) tuần tự 5 bảng từ Table 0 -> Table 4. (lưu ý: khi làm bước này người gán nhãn cần phải chọn hint - được liệt kê trong vùng ô màu vàng trước khi đặt QAs)
- **Bước 4.** Sau khi đặt dữ liệu xong hết 5 bảng có trong 1 file dữ liệu, chọn vào biểu tượng "Show Confirmed QA Pairs"để hiện những thông tin về các QAs đã làm. Ở đây có thể chỉnh sửa/xóa những QAs bị lỗi. Sau khi hoàn thành file và chỉ khi hoàn thành 5/5 bảng trong file, chọn "Save"để tải dữ liệu sản phẩm về. Dữ liệu sản phẩm sẽ được hướng dẫn lưu trữ bởi người hướng dẫn.
- **Bước 5.** Sau khi làm xong file dữ liệu hiện tại và muốn load dữ liệu mới lên để làm tiếp, chọn "Step 0"để quay về giao diện Load dữ liệu, chọn "Reload"để xóa đi mọi thứ về dữ liệu và QAs cũ (lưu ý: khi chưa hoàn thành xong file hiện tại, người gán nhãn có thể yên tâm thoát trang web của trình duyệt vì dữ liệu vẫn sẽ được lưu trên trình duyệt đó và có thể vào lại web với dữ liệu đang lưu và tiếp tục làm). Sau khi "Reload", người gán nhãn thực hiện lại tuần tự từ **Bước 1**.

3 Hướng dẫn làm dữ liệu

Dựa trên thông tin của bảng dữ liệu hiện ở bên phải giao diện tool, người gán nhãn sẽ tiến hành đặt các cặp QAs liên quan đến bảng. Để đảm bảo cho quá trình này được đồng nhất giữa các người gán nhãn và đảm bảo sự đa dạng của việc đặt câu hỏi, nhóm chúng tôi tạo ra 11 Hints (từ khóa sẽ xuất hiện trong câu hỏi hoặc yêu cầu câu hỏi phải thỏa mãn) như sau:

- **"Cái gì, cái nào, công ty nào,... hoặc những từ đồng nghĩa (What)"**
- **"Làm thế nào, như thế nào, bằng cách nào,... hoặc những từ đồng nghĩa (How)"**
- **"Ở đâu, trong thành phố nào, ở nước nào,... hoặc những từ đồng nghĩa (Where)"**
- **"Ai, người nào,... hoặc những từ đồng nghĩa (Who)"**
- **"Vì sao, tại sao,... hoặc những từ đồng nghĩa (Why)"**
- **"Khi nào, thời gian nào, ... hoặc những từ đồng nghĩa (When)"**
- **"Sử dụng hỏi kết hợp giữa các ô, các hàng, các cột"**
- **"Sử dụng yêu cầu liệt kê, sắp xếp"**
- **"Sử dụng tính toán"**
- **"Câu hỏi không thể trả lời từ kiến thức trong bảng"**

Sau đây là những ví dụ đặc trưng của từng hint trong quá trình làm dữ liệu. Trong quá trình làm dữ liệu nếu như có thể đặt ra được những trường hợp QAs mới, hãy note lại và liên hệ vào nhóm annotator ***Lưu ý: Những ví dụ dưới chỉ là những ví dụ minh họa về ý nghĩa của hint. Trong thực tế quá trình làm dữ liệu sẽ phải dựa trên dữ liệu có trong bảng***

3.1 Cái gì, cái nào, công ty nào,... hoặc những từ đồng nghĩa (What)

Ví dụ minh họa:

Question: Sản phẩm Nokia Lumia 520 (521) sử dụng ***màn hình gì/nào***?
Answer: ***4,0"IPS LCD 480 x 800 px***
****
Question: Thống đốc Suzuki Naomichi là thống đốc của ***đô đạo phủ huyện nào***?
Answer: ***Hokkaidō***

3.2 Làm thế nào, như thế nào, bằng cách nào,... hoặc những từ đồng nghĩa (How)

Đa số dựa trên các ô chứa nhiều thông tin (nhiều chữ)

3.3 Ở đâu, trong thành phố nào, ở nước nào,... hoặc những từ đồng nghĩa (Where)

Ví dụ minh họa:

Question: Lãnh chúa ABC có lâu đài tọa lạc ***ở đâu***?
Answer: ***Anh***
****
Question: Thống đốc Suzuki Naomichi là thống đốc ***ở đô đạo phủ huyện nào***?
Answer: ***Hokkaidō***

3.4 Ai, người nào,... hoặc những từ đồng nghĩa (Who)

Ví dụ minh họa:

Question: ***Người nào*** là người phát minh ra TableQA Annotation tool ?
Answer: ***Đỗ Nhật Tân***
****
Question: ***Ai*** là thống đốc ở đô đạo phủ huyện Hokkaidō?
Answer: ***Suzuki Naomichi***

3.5 Vì sao, tại sao,... hoặc những từ đồng nghĩa (Why)

Đa số dựa trên các ô chứa nhiều thông tin (nhiều chữ)

3.6 Khi nào, thời gian nào, ... hoặc những từ đồng nghĩa (When)

Question: Tòa nhà cao nhất nước Mỹ được xây dựng trong ***thời gian nào***?
Answer: ***1900***

3.7 Sử dụng hỏi kết hợp giữa các ô, các hàng, các cột

Question: Tòa nhà nào trong danh sách nằm ở ***Trung Quốc*** và có chiều cao trên ***100m***?
Answer: ***Tòa nhà A***

3.8 Sử dụng yêu cầu liệt kê, sắp xếp

Question: ***Sắp xếp*** các tòa nhà theo thứ tự tăng dần chiều cao.
Answer: ***Tòa nhà A, Tòa nhà E, Tòa nhà C, Tòa nhà B, Tòa nhà D***

3.9 Sử dụng tính toán

Có 1 số lưu ý khi các bạn sử dụng hint này như sau:

- Các phép toán được sử dụng trong quá trình làm dữ liệu là (cộng, trừ, nhân, chia, trung bình cộng, lớn nhất, nhỏ nhất, lớn nhì, lớn ba, đếm) Trong trường hợp các bạn nghĩ ra được các phép toán mới, các bạn phải hỏi vào nhóm annotator để được xác thực khả năng được làm hay không nhé, các phép toán mới nếu hợp lệ sẽ giúp cho bộ dữ liệu trở nên đa dạng.
- Có thể sử dụng tính toán đối với các đối tượng thời gian (khoảng thời gian, ...), không gian (khoảng cách, tổng khoảng cách,...), điểm số (tổng điểm, trung bình, ...), ...

Question: ***Có tổng cộng bao nhiêu*** các tòa nhà cao trên 100m?
Answer: ***5***
****
Question: ***Trung bình*** các tòa nhà trong danh sách là bao nhiêu?
Answer: ***100.4m***

3.10 Câu hỏi không thể trả lời từ kiến thức trong bảng

Đối với Hint này, chỉ có 1 lưu ý là câu hỏi phải liên quan tới bảng. Ví dụ bảng dữ liệu đó là dữ liệu liên quan đến sinh học, khi sử dụng hint này phải sử dụng câu hỏi liên quan tới sinh học, nếu sử dụng câu hỏi liên quan đến văn học sẽ được cho là không hợp lệ.

Sau đây là các trường hợp được sử dụng trong quá trình sử dụng hint này:

- Có thể sử dụng các câu hỏi liên quan đến ô trống trong bảng (ô không chứa dữ liệu)
- Sử dụng câu hỏi không thỏa mãn bất kỳ điều kiện trong bảng
- Sử dụng câu hỏi đối lập/đối nghịch với ngữ cảnh của bảng dữ liệu
- Thay đổi bối cảnh (thực thể, địa điểm, thời gian...) trong bảng

Câu trả lời được cố định bởi đáp án ***"Null"***, không thay đổi đáp án.

3.11 Câu hỏi Yes/No

Đây là dạng câu hỏi kiểm chứng với đáp án sẽ là Yes hoặc No dựa vào câu hỏi và bối cảnh của bảng dữ liệu. Tuy nhiên, loại câu hỏi này trong tiếng Việt sẽ có sự đa dạng lớn về kiểu trả lời.
Các lưu ý ở hint này sẽ là tự nhiên trong việc đặt câu hỏi, TRÁNH việc sử dụng từ ngữ ở văn phong nói trong việc trả lời ("Không bé ơi", "OK") và trả lời súc tích ngắn gọn với 1 từ duy nhất (trường hợp nào phải trả lời bằng 2 từ các bạn hãy liên hệ và trao đổi với nhóm).
Sau đây là những ví dụ có thể được sử dụng trong quá trình làm dữ liệu:

- Ông A mất năm 1900 đúng không? - (Đúng/Không)
- Tòa nhà A có phải ở New York không? - (Phải/Không)
- Việt Nam đã từng có dịch Covid chưa? - (Rồi/Chưa)
- Ông A có sở hữu B không? - (Có/Không)
- ...

4 Những trường hợp được cho là không hợp lệ trong quá trình làm dữ liệu

Đối với những trường hợp dữ liệu không hợp lệ, dữ liệu đó sẽ bị loại bỏ và được người hướng dẫn nhắc nhở. Nếu tái phạm nhiều lần sẽ bị loại khỏi quá trình đóng góp dữ liệu theo quy định.

4.1 Câu hỏi không nghiêm túc

Ví dụ: Q: 1 cộng 1 là 2, 2 thêm 2 là 4, trung tâm châu đại phúc là?

4.2 Sai chính tả

Trường hợp bảng dữ liệu bị sai chính tả sẽ không bị tính là lỗi, và người làm dữ liệu có thể copy từ đó qua answer.

4.3 Không đúng, đầy đủ cấu trúc của 1 câu hỏi, hỏi chưa rõ ràng

Ví dụ: Q: Năm có nhiều cuộc thi bị hoãn nhất? -> Sửa lại: Năm nào có nhiều cuộc thi bị hoãn nhất? / Năm có nhiều cuộc thi bị hoãn nhất là năm nào?

4.4 Spam cấu trúc câu hỏi

Trường hợp file dữ liệu có quá nhiều cấu trúc câu hỏi bị spam (sử dụng lặp lại nhiều lần giữa các bảng) sẽ bị cảnh báo làm lại và hủy luôn file dữ liệu bị spam đó (file dữ liệu có bảng vi phạm lỗi spam).
Ví dụ:

- Q: Năm sinh của ông A là năm nào?
- Q: Năm sinh của ông B là năm nào?
- Q: Năm sinh của ông C là năm nào?
- Q: Năm sinh của ông D là năm nào?

4.5 Câu trả lời tự biên tự diễn dựa trên kiến thức cá nhân/bản thân và không có trong bảng dữ liệu

Chỉ sử dụng kiến thức có trong bảng, mà không được sử dụng kiến thức sẵn có của cá nhân mà không có trong bảng để trả lời.

4.6 Spam câu hỏi yes/no

Qua quan sát làm dữ liệu từ nhóm annotator cũ, chúng mình nhận ra việc đặt câu hỏi yes/no (Ví dụ: Có phải tòa nhà ABC nằm ở thành phố D không?) chiếm khá nhiều bởi sự dễ trong lúc đặt câu hỏi và trả lời. Có 1 lưu ý ở đây là các bạn hãy làm 1 cách tự nhiên và thoải mái với tinh thần không spam về 1 hint cụ thể. Nhóm mình giới hạn việc đặt hint yes/no với tỉ lệ 10%, tuy nhiên sẽ có trường hợp các bạn làm với tâm thế không spam nhưng vẫn dính trên 10%, bọn mình sẽ phân tích và sẽ xét xử sau đối với trường hợp đó (có thể vẫn sẽ confirm hoặc không confirm và xóa đi những dữ liệu thừa). Các bạn cũng lưu ý đối với các trường hợp những hint khác, nếu nhóm mình phát hiện spam (>15%), nhóm mình cũng sẽ loại bỏ những câu dư thừa.

5 Phần kết

Trong quá trình làm dữ liệu, các góp ý và phát hiện ra được vấn đề mới (vấn đề về tool, về có loại câu hỏi mới nhưng không có hint, guideline thiếu chặt chẽ...) trong quá trình làm sẽ giúp ích bọn mình rất nhiều trong quá trình hoàn thiện bài nghiên cứu. Nếu có thể, các bạn có thể liên hệ bằng cách inbox riêng hoặc nhắn tin thông báo vào nhóm annotator, nhóm mình sẽ rất biết ơn đối với những góp ý. Xin cảm ơn các bạn đã đọc qua guideline.