ĐỒ ÁN THỰC HÀNH ỨNG DỤNG PHÂN TÁN 2023

Xây dựng công cụ crowdsourcing để gán nhãn dữ liệu

v1.0 (03/03/2023)

Mô tả

Là công cụ hỗ trợ gán nhãn dữ liệu phục vụ cho các dự án trí tuệ nhân tạo cho nhiều bài toán khác nhau:

- Phân loại văn bản:
  - Mẫu dữ liệu là một văn bản
  - Cho một văn bản, người gán nhãn sẽ phân loại văn bản đầu vào thuộc các loại nào đó
- Hỏi đáp:
  - Mẫu dữ liệu là một cặp câu hỏi / văn bản
  - Cho một văn bản và một câu hỏi, người gán nhãn sẽ đưa ra quan hệ đúng sai của câu hỏi và đoạn văn đó. Văn bản có phải là trả lời cho câu hỏi.
- Dịch máy:
  - Mẫu dữ liệu là một văn bản ở một ngôn ngữ
  - Cho một văn bản ở một ngôn ngữ, người gán nhãn sẽ đưa ra bản dịch của văn bản đó ở ngôn ngữ yêu cầu
- Gán nhãn thực thể:
  - Mẫu dữ liệu là một văn bản
  - Cho một văn bản, người gán nhãn sẽ chọn các tên thực thể trong văn bản đó.
- Gán nhãn cặp văn bản (đồng nghĩa):
  - Mẫu dữ liệu là một cặp văn bản
  - Cho một cặp văn bản, người gán nhãn sẽ đưa ra nhãn tương ứng về tác vụ.
- Gán nhãn câu trả lời của cặp câu hỏi và văn bản:
  - Cho một câu hỏi, một đoạn văn, người gán nhãn sẽ đưa ra câu trả lời chính xác cho đoạn văn
- Tìm câu hỏi đồng nghĩa:
  - Cho một câu hỏi, người gán nhãn sẽ đưa ra danh sách các câu hỏi đồng nghĩa

Các vai trò người dùng:

- Quản trị hệ thống
- Người dùng quản lý
- Người gán nhãn:
  - Cấp 1
  - Cấp 2
  - …

Một số định nghĩa:

- Tác vụ là gì: Là một nhiệm vụ gán nhãn dữ liệu, có nhiều loại tác vụ khác nhau.
- Mẫu dữ liệu: Là một mẫu dữ liệu cần gán nhãn. Các mẫu dữ liệu của các tác vụ khác nhau có thông tin khác nhau.
- Có 3 loại tác vụ chính:
    - Text classification: Gán nhãn mẫu dữ liệu vào trong một tập nhãn được quy định trước (bởi người quản lý)
    - Text generation: Người gán nhãn nghĩ các nội dung tương ứng với tác vụ.
    - Sequence labeling: Người gán nhãn gán nhãn các đoạn trong văn bản đầu vào

Các chức năng chính:

Người dùng quản lý

- **Tạo dự án**: Đặt tên dự án, mỗi dự án sẽ gán một loại nhãn dữ liệu. Mỗi dự án sẽ gán nhãn 1 hoặc nhiều tác vụ. Định nghĩa các nhãn để gán nhãn dự án.
- **Nhập dữ liệu / Import dữ liệu**: Đưa các mẫu dữ liệu cần gán nhãn vào dự án
    - Từng mẫu đơn lẻ
    - Từ file: JSONL, CSV, …
- **Mời/Thêm các người dùng** để gán nhãn dữ liệu với loại tương ứng: Cấp 1, Cấp 2, Cấp 3.
- **Xem/hiệu chỉnh/tìm kiếm dữ liệu** đã được gán nhãn bởi các người dùng.
- **Xem thống kê** các nhãn dữ liệu được gán nhãn.
- **Phân công gán nhãn**: có thể phân công ngẫu nhiên, gán nhãn toàn bộ tuần tự, hoặc phân công chỉ định theo ID của mẫu dữ liệu trong dự án
- **Import / Export**:
    - Import dữ liệu từ API cung cấp
    - Export dữ liệu đã được gán nhãn

Người gán nhãn:

- **Hiển thị mẫu dữ liệu để gán nhãn**. Mỗi mẫu dữ liệu có thể có nhiều người gán nhãn khác nhau. Quản lý có thể quy định số lượng tối đa người gán nhãn cho 1 dự án.
- Sau khi được mời và chấp nhận lời mời thì có thể thực hiện được nhiệm vụ gán nhãn, có thể xem danh sách các dự án mình đang tham gia.
- Sau khi gán nhãn có thể xem được mình đã gán nhãn được bao nhiêu mẫu dữ liệu, danh sách mẫu dữ liệu đã gán: xem số lượng đã thực hiện và chi tiết.

Người gán nhãn cấp 2, 3:

- Review các nhãn dữ liệu được gán bởi các người dùng
- Revise bằng cách gán nhãn lại

Các chức năng chung:

- Đăng ký tài khoản
- Đăng nhập / đăng xuất / đổi mật khẩu
- Quản lý người dùng của quản trị

Tham khảo: **doccano**

Yêu cầu:

Cho phép sử dụng các framework.

Các bài toán gán nhãu tham khảo để làm đồ án. Giáo viên sẽ cung cấp dữ liệu mẫu

- Gán nhãn loại câu hỏi trên dữ liệu TREC57 Phân loại câu hỏi:
  - Tác vụ Text Classification: Phân loại câu hỏi vào trong 1 tập nhãn
- Gán nhãn hỏi đáp trên bộ dữ liệu Hỏi đáp
  - Tác vụ Text Classification: Phân loại câu hỏi – câu trả lời vào 1 tập nhãn
- Gán nhãn đồng nghĩa
  - Tác vụ Text Classification: Phân loại 1 cặp câu vào 1 tập nhãn
- Tìm câu hỏi đồng nghĩa:
  - Tác vụ Text Generation: Nhập dữ liệu văn bản cho một tác vụ

Dùng bộ dữ liệu khác.

Công nghệ:

- Web Server / Server side script: PHP / Python (Bốc thăm)
- Database: MySQL, Elasic search / Mongo
- HTML, AJAX, Jquery, CSS, Javascript