HỆ THỐNG QUẢN LÝ NGỮ LIỆU SONG NGỮ

I. Tóm tắt. Các chức năng chính của hệ thống gồm:

• Crawl dữ liệu song ngữ tự động từ nguồn web. (theo nguồn dữ liệu, các cặp link URL)

• Quản lý dữ liệu, quản lý người dùng, thống kê khối lượng công việc.

• Phân quyền người dùng thành 3 loại: Quản trị viên (administrator), Người chỉnh sửa (editor). Mỗi loại người dùng sẽ được thiết kế giao diện riêng để phù hợp với chức năng.

• Dóng hàng văn bản song ngữ tự động.

• Chỉnh sửa, dóng hàng văn bản song ngữ thủ công bằng chuyên gia ngôn ngữ.

• Dóng hàng câu song ngữ tự động.

• Chỉnh sửa, dóng hàng câu song ngữ thủ công bằng chuyên gia ngôn ngữ.

• Phát hiện các cặp câu trùng lặp trong CSDL

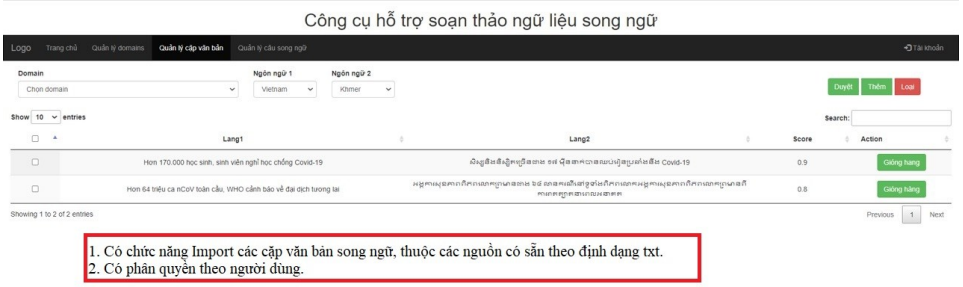
II. Mô tả chi tiết chức năng của hệ thống

1. **Chức năng crawl dữ liệu đơn ngữ tự động.**
   + Gọi tool crawl dữ liệu đơn ngữ.
   + Import (nhập) các nguồn đơn ngữ có sẵn.
   + Export corpus đơn ngữ dưới dạng file text, mỗi câu lưu trên một dòng
2. **Chức năng crawl dữ liệu song ngữ tự động.**
   + Giao điện như hình 1.
   + Quản lý được các Domain song ngữ (cặp URL) dùng để crawl dữ liệu, tránh crawl trùng lặp.
   + **Có trường hợp crawl từ cặp URL**
   + Gọi được tools crawl dữ liệu từ giao diện.
   + Hiển thị được danh sách các văn bản đã crawl được theo Domain.
   + Gọi được Tools gióng hàng văn bản tự động (nhóm cô Hương) để tính Score cho từng cặp văn bản.

(Hình 1 - Giao diện của chức năng crawl văn bản song ngữ)

3. Chức năng quản lý cặp văn bản song ngữ

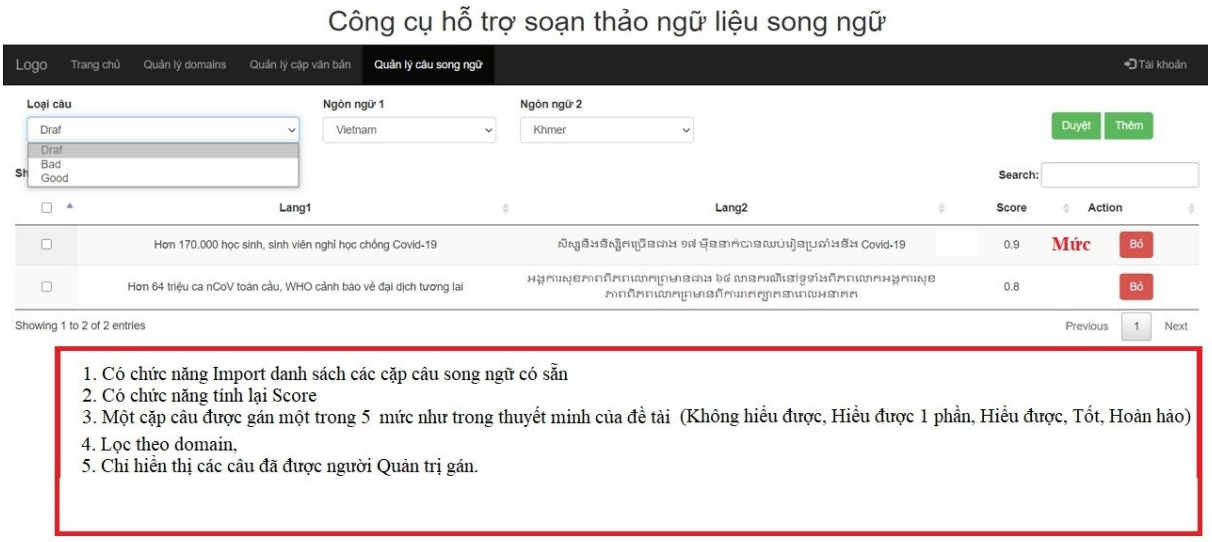
+ Giao diện như hình 2.
+ Quản lý được các danh sách các cặp văn bản song ngữ, lưu được nguồn của văn bản (URL, Nguồn tự do, ...).
+ Hiển thị danh sách cặp văn bản song ngữ theo tài khoản người làm dữ liệu.
+ Cho phép người làm dữ liệu chỉnh sửa, cập nhật văn bản song ngữ (Lưu ý thuận tiện cho trường hợp văn bản dài).
+ Cho phép người làm dữ liệu Import các tệp văn bản (định dạng utf-8 hoặc utf-16) song ngữ từ nguồn có sẵn, đã crawl từ trước, được ghép cặp theo tên file.
+ Cho phép người làm dữ liệu xác nhận từng cặp văn bản song ngữ là Duyệt hoặc Loại bỏ.
+ Khi người dùng đánh dấu 1 cặp văn bản là Duyệt thì gọi đến Tool gióng hàng câu của cô Hương để gióng hàng các câu trong cặp văn bản vừa được chọn. Các câu được gióng hàng tự động được lưu lại.
+ Có khả năng kiểm tra, phát hiện trùng lặp văn bản.
+ Có phân quyền cho người sử dụng.



(Hình 2 - Giao diện của chức năng quản lý cặp văn bản song ngữ)

4. Chức năng quản lý cặp câu song ngữ

+ Giao điện như hình 3.
+ Quản lý được các danh sách các cặp câu song ngữ đã được gióng hàng tự động.
+ Lưu được nguồn của cặp câu (câu được trích ra trong cặp văn bản nào).
+ Hiển thị danh sách cặp câu song ngữ theo tài khoản người làm dữ liệu.
+ Cho phép người làm dữ liệu chỉnh sửa, cập nhật cặp câu song ngữ, sau khi sửa có thể tính lại Score theo Tool của nhóm cô Hương. Cặp câu sau khi sửa cần có kiểm tra cặp câu vừa sửa có trùng với một cặp câu đã tồn tại trong CSDL. Có đánh dấu bằng màu sắc để người dùng dễ quan sát trạng thái câu đã sửa.
+ Cho phép người làm dữ liệu Import các tệp (định dạng utf-8 hoặc utf-16) chứa danh sách các cặp câu song ngữ từ nguồn có sẵn. Tính Score cho các cặp câu được Import. Phát hiện các cặp câu được Import nhưng đã có trong CSDL.
+ Cho phép người làm dữ liệu xác nhận từng cặp câu song ngữ là Duyệt hoặc Loại bỏ.
+ Cho phép người làm dữ liệu đánh giá chất lượng của từng cặp câu song ngữ theo 3 mức: *Tốt, Không tốt, Chưa biết*.
+ Có khả năng kiểm tra, phát hiện trùng lặp cặp câu.
+ Có phân quyền cho người sử dụng



(Hình 3 - Giao diện của chức năng quản lý cặp câu song ngữ)

5. **Chức quản lý kho ngữ liệu đơn ngữ**
    - Cho phép crawler dữ liệu đơn ngữ từ các nguồn link các trang web theo các ngôn ngữ trong đề tài (có thể mở rộng cho các ngôn ngữ khác)
    - Có khả năng quản lý dữ liệu đơn ngữ trong CSDL (thêm sửa xóa): Có lưu trữ dữ liệu đơn ngữ (lấy ở nguồn các trang web nào, thời gian ngày giờ crawler), xuất dữ liệu đơn ngữ theo file, thống kê dữ liệu đơn ngữ (lấy ở đâu, số lượng kích thước bao nhiêu, ngày giờ tạo)
    - Cho phép import (nhập) dữ liệu đơn ngữ từ các nguồn dữ liệu khác để bổ sung vào kho ngữ liệu đơn ngữ.
6. **Chức năng thống kê, báo cáo**
    - Cho phép thống kê báo cáo kho ngữ liệu song ngữ theo các miền dữ liệu crawler được, số lượng kích thước bao nhiêu.
    - Cho phép thống kê kho ngữ liệu đơn ngữ theo các miền dữ liệu crawler được, số lượng kích thước bao nhiêu.
7. **Chức năng quản trị hệ thống**

    - Quyền quản trị hệ thống: Quản lý các user, phân quyền cho các user, cấu hình hệ thống.

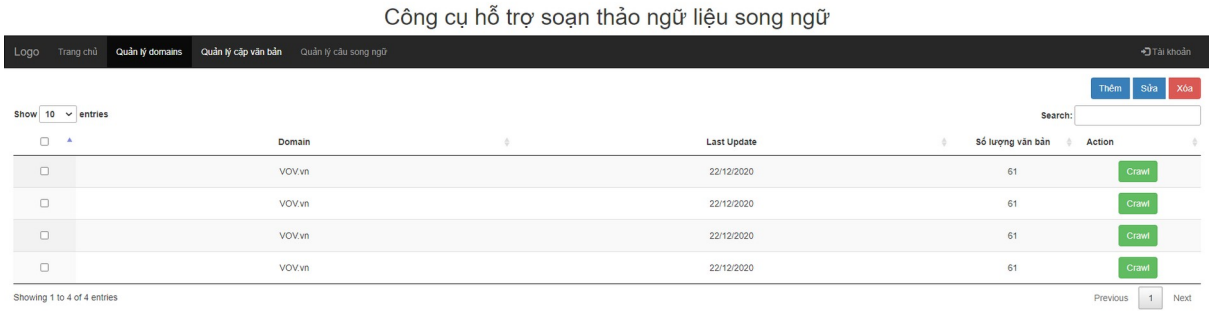
    - Quyền quản lý dự án: User quản lý dự án phụ trách theo cặp ngôn ngữ, thực hiện các chức Crawl dữ liệu, Quản lý dữ liệu (import, export), phân công làm dữ liệu, thống kê báo cao.

    - Quyền chuyên gia ngôn ngữ (Người làm dữ liệu): Chỉ gắn nhãn dữ liệu đối với cặp văn bản, cặp câu song ngữ. Không được quyền Import, Export dữ liệu.

Để tránh trùng lặp, thuận tiện cho người làm dữ liệu cần quản lý các dữ liệu sau:

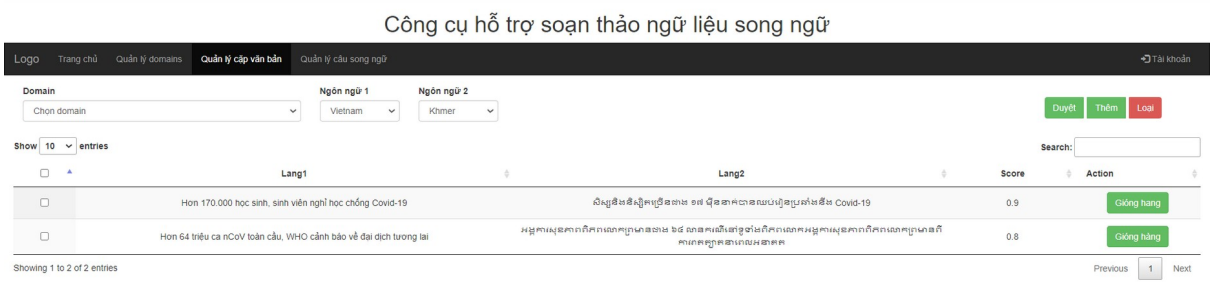
1. Quản lý (thêm, xóa, sửa, quét (crawl)) các tên miền (host) chứa dữ liệu song ngữ, là các tên miền gốc như vov.vn. Các trang web này là đầu vào cho công cụ kiểu bitextor vừa crawl dữ liệu đồng thời gióng hàng văn bản để tạo ra cặp văn bản song ngữ thô (draft) để người biên tập duyệt.

- Thêm là bổ sung tên miền mới vào danh sách.
- Xóa là loại bỏ hẳn khỏi danh sách tên miền.
- Sửa là chức năng sửa tên miền.
- Quét/Quét lại là chức năng gọi công cụ vi-bitextor (tương tự bitextor) để quét và tìm cặp văn bản song ngữ, đưa vào CSDL song ngữ thô nếu chưa có.



2. Quản lý (thêm, duyệt, loại bỏ) cặp văn bản song ngữ. Khi cặp văn bản được duyệt, nó sẽ được gióng hàng câu để tạo ra tập cặp câu thô (draft). Khi loại, cặp văn bản sẽ được đánh dấu không đạt để khi crawl lại sẽ không phải loại thêm một lần nữa cặp văn bản này.

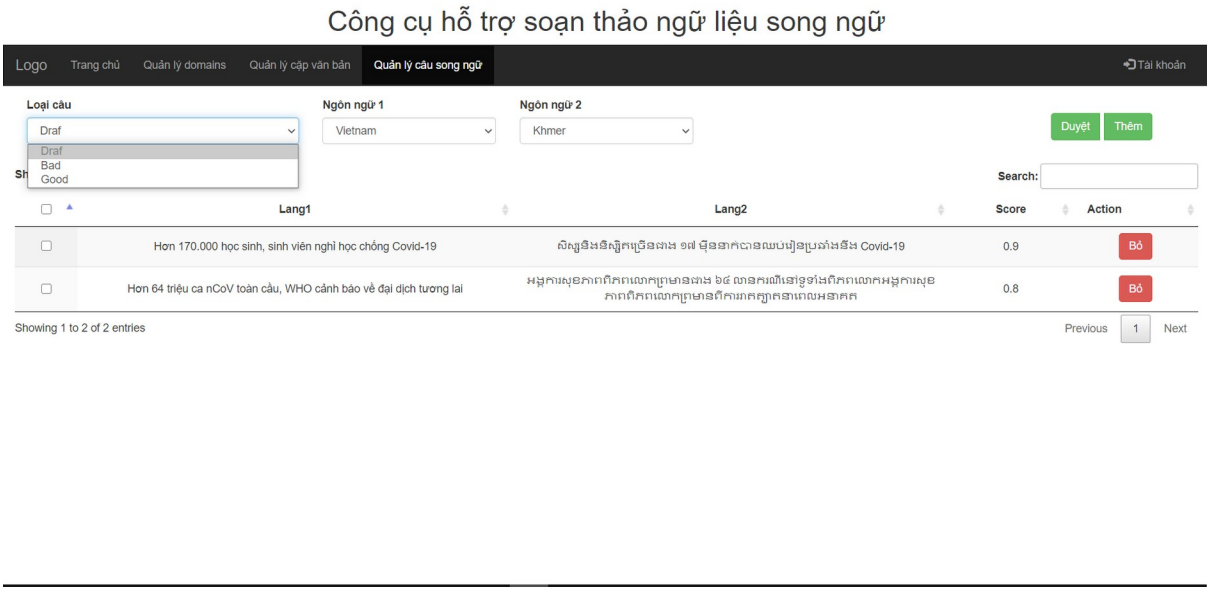
- Thêm là thêm cặp urls chứa hai văn bản mà người đưa vào. Ghi thêm thông tin người đưa vào và thời gian.
- Duyệt là chức năng gọi công cụ gióng hàng văn bản (tương tự vecalign, hunalign) để tìm các cặp câu thô (draft).
- Loại là chức năng đánh dấu cặp văn bản bị loại, sau này sẽ không cần xử lý lại.



3. Quản lý (thêm, duyệt, sửa và duyệt, loại bỏ) các cặp câu

- Thêm là người dùng tự thêm cặp câu bằng tay.
- Duyệt là người dùng đánh dấu cặp câu là tốt.
- Sửa và duyệt chức năng người dùng thêm cặp câu mới, nhưng liên kết đến cặp câu gốc.

- Loại bỏ là chức năng đánh dấu cặp câu đó không đạt yêu cầu, để sau này có tìm thêm được cặp câu đó thì cũng loại luôn.



Các lưu ý gồm:

1. *Đối với tài khoản **Đánh giá viên**, được quyền xem lại những cặp câu đã làm của một **người làm dữ liệu** nào đó (cộng tác viên) trong một khoảng thời gian nào đó, để đánh giá chất lượng các cặp câu, tiến độ công việc. Do đó, đánh giá viên cần được thêm tính năng lọc các cặp câu đã được đánh giá theo người làm dữ liệu (Cộng tác viên), theo ngày làm dữ liệu (các cặp câu từ **ngày nào** → **ngày nào**), điều kiện lọc có thể là kết hợp cả hai điều kiện*

trên để: Chỉnh sửa đánh giá, đưa các cặp câu đánh giá chưa đúng trở về trạng thái chưa được đánh giá.

2. *Khi Import danh sách các cặp câu từ 1 file dữ liệu cần phải cho người import cung cấp thêm thông tin **Cặp ngôn ngữ**, **miền ngôn ngữ** (là nội dung ghi nguồn gốc dữ liệu như: VOV, Tedtalk, vv), lưu vết user nào Import dữ liệu đó, có thể chỉ cho phép Admin, **Đánh giá viên** được quyền Import.*
3. ***Người Quản trị** có thể lọc dữ liệu đã làm theo các tiêu chuẩn: Cộng tác viên, miền dữ liệu, theo mức độ chất lượng (Score).*
4. *Quản trị viên có thể Import dữ liệu, xóa dữ liệu do mình đã Import.*
5. *Có Chức năng hiển thị thông tin khối lượng công việc của từng user.*
6. *Khi người dùng sửa cặp câu thì tính lại score*
7. ***Các file text có định dạng UTF-8 hoặc UTF-1**6*

Yêu cầu:

1. Mỗi tài khoản Cộng tác viên, Đánh giá viên được gắn với 1 cặp ngôn ngữ duy nhất (Việt-Lào; Việt-Khmer; Việt-Trung). Do đó Cộng tác viên không cần chọn cặp ngôn ngữ khi làm dữ liệu.

2. Ở cửa sổ Văn bản, thêm chức năng để Cộng tác viên tự nhập hoặc copy/paste 2 văn bản song ngữ, chỉnh sửa. Sau đó, Cộng tác viên gọi Tool gióng hàng tự động để sinh ra các cặp câu đã được gióng tự động --> bổ sung các cặp câu này vào kho cặp câu song ngữ chưa được đánh giá.

3. Người quản trị có thể xem được các thông tin (theo các cặp ngôn ngữ): Số cặp câu chưa được đánh giá, số cặp câu đã đánh giá, số cặp câu đã đánh giá theo từng user.

4. Các Cộng tác viên có thể xem được tiến độ của mình như số cặp câu mình đã đánh giá.

5. Đối với tài khoản Đánh giá viên, cần được quyền xem lại những cặp câu đã làm của một người làm dữ liệu nào đó (cộng tác viên) trong một khoảng thời gian nào đó, để đánh giá chất lượng các cặp câu, tiến độ công việc. Do đó, đánh giá viên cần được thêm tính năng lọc các cặp câu đã được đánh giá theo người làm dữ liệu (Cộng tác viên), theo ngày làm dữ liệu (các cặp câu từ ngày nào → ngày nào), điều kiện lọc có thể là kết hợp cả hai điều kiện trên để: Chỉnh sửa đánh giá, đưa các cặp câu đánh giá chưa đúng trở về trạng thái chưa được đánh giá.

6. Người Quản trị có thể lọc dữ liệu đã làm theo các tiêu chuẩn: Cộng tác viên, miền dữ liệu, theo mức độ chất lượng (Score).