YẾU CẦU CỦA ĐỒ ÁN GIỮA KỲ

Do có nhiều đồ án GK khác nhau, vì vậy các yêu cầu cũng có sự khác nhau, tuy nhiên cũng có một số yêu cầu chung như sau:

A. YÊU CẦU CHUNG:

1. Tất cả các xử lý hãy cố gắng thực hiện tự động ở mức càng cao càng tốt. Đây là một tiêu chí để đánh giá đồ án.
2. T chỉ ràng buộc về tên tác phẩm ngữ liệu thô (vd: “Đại Việt Quốc sử diễn ca”; hay “Đại Nam thực lục tập x”, …) chứ không bắt buộc phải dùng ngữ liệu thô mà T cung cấp. Các bạn có thể tìm bản khác để thuận tiện hơn.
3. T cũng không ràng buộc phải sử dụng các công cụ, giải thuật, phương pháp hay mô hình T cung cấp. T chỉ gợi ý, tư vấn mà thôi. Ví dụ: để OCR chữ Hán, thì có một số công cụ tốt sau: KanDianGuJi, ChatGPT, Gemini, Google Vision, .. còn để OCR chữ Nôm thì theo T biết thì có API của NomNaOCR và của Lab T (tools.clc.hcmus.edu.vn, kimtudien.com.vn hay kimhannom.vn).
4. Để đánh giá độ chính xác của các model do các bạn lựa chọn/tự xây dựng, T cung cấp mỗi dạng ngữ liệu một golden dataset nhỏ để đánh giá.
5. Cách đánh giá ngữ liệu các bạn: bên T sẽ sử dụng công cụ để đánh giá tự động kho ngữ liệu do nhóm các bạn xây dựng, vd công cụ đo độ chính xác của việc dóng hàng (có bị sai ko), về kiểm lỗi ngôn ngữ (chính tả, tách câu,..), …
6. Đầu ra cần tồn tại dưới dạng XML trong đó có các thẻ để lưu metadata, như: tên tác phẩm, tác giả, ngôn ngữ, niên đại, … Ngoài ra, mỗi câu cần có thẻ XML chứa <sentence_ID> gồm 14 ký tự (đã giải thích trong file SinoNom_OCR_Transliteration Alignment.pdf) để sau này, có thể truy xuất ngược biết câu đó thuộc chương nào, tập nào, bộ nào, tác phẩm tên gì, do ai viết, thời đại nào, …. Ngoài ra, Sentence_ID và metadata sau này sẽ giúp người sử dụng lọc tìm ra các tác phẩm mình mong muốn một cádh nhanh chóng và tự động.

B. YÊU CẦU RIÊNG:

Mỗi loại ngữ liệu khác nhau, ngoài các yêu cầu chung nói trên, còn cần thỏa các yêu cầu riêng như sau:

1. Với ngữ liệu song ngữ Hán-Việt (đầu vào là ảnh/txt): cần dóng hàng dịch nghĩa văn bản giữa câu Hán và Việt (dùng Dic-based hay BERT align,. ..). Định dạng ngữ liệu: 2 dạng [a] XML với thẻ như quy định chung và thêm thẻ chỉ ngôn ngữ (C hay V nghĩa là Chinese hay Vietnamese) cho từng cặp câu đã được dóng. Vd: <STC_ID=”ABC_001.001.01”><C>…… </C><V> ……..</V></STC_ID>> và [b] dạng Excel (như trong file SinoNom_OCR_Transliteration Alignment.pdf đã hướng dẫn) để T dễ nhìn và dễ dò xác suất một cách thủ công.
2. Với ngữ liệu đơn ngữ Hán (đầu vào là ảnh): cần OCR ảnh văn bản Hán rồi dóng hàng giữa kq ảnh (bbox) và character OCR ra được. Định dạng gồm XML (vì là

đơn ngữ, nên không cần thẻ ngôn ngữ C hay V) và Excel như trên. Mỗi câu cũng có cần có stc_id như trên.

3. Với ngữ liệu đơn Việt (đầu vào là ảnh): cần OCR ảnh văn bản Việt rồi dóng hàng giữa kq ảnh (bbox) và character OCR ra được. Định dạng gồm XML (vì là đơn ngữ, nên không cần thẻ ngôn ngữ C hay V) và Excel như trên. Mỗi câu cũng có cần có stc_id như trên.
4. Với ngữ liệu đơn ngữ Hán (đầu vào là txt): cần gán nhãn thực thể (NER: Named Entity Recognition) bằng model có sẵn hay tự xây dựng. Bộ nhãn của NER cần có ít nhất các thẻ cơ bản như: PER, LOC, ORG, TITLE, TME và NUM. Định dạng gồm XML có các thể NER và Excel như trên. Mỗi câu cũng có cần có stc_id như trên.
5. Với ngữ liệu đơn ngữ Việt (đầu vào là txt): cần gán nhãn thực thể (NER: Named Entity Recognition) bằng model có sẵn hay tự xây dựng. Bộ nhãn của NER cần có ít nhất các thẻ cơ bản như: PER, LOC, ORG, TITLE, TME và NUM. Định dạng gồm XML có các thể NER và Excel như trên. Mỗi câu cũng có cần có stc_id như trên.