HƯỚNG DẪN THỰC HIỆN

bài tập thực hành, đồ án giữa kỳ và cuối kỳ

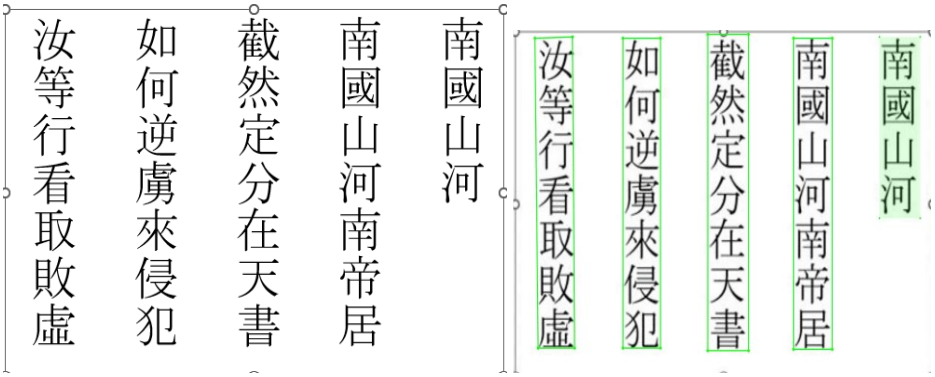
Môn: Nhập môn XLNNTN

Để hướng đến các ứng dụng thực tế sau này (sẽ sử dụng các LLM, công nghệ AI, ML trên các kho ngữ liệu đặc thù của tiếng Việt tại các đơn vị ở Việt Nam), nội dung bài tập thực hành, đồ án giữa kỳ và đồ án cuối kỳ sẽ như sau:

1. Bài tập thực hành (thực hiện cá nhân, chiếm 30% điểm): GVTH sẽ hướng dẫn SV cách viết một số chương trình công cụ (trong sheet “tools”) mà trong đó chủ yếu sử dụng các hàm, API, thư viện có sẵn cũng như các kỹ thuật lập trình cá nhân của mình để rút trích nội dung văn bản (dạng ảnh, text) từ file pdf hay html nhằm xây dựng nên kho ngữ liệu.
2. Đồ án giữa kỳ (thực hiện cá nhân, chiếm 40% điểm): SV sẽ sử dụng các chương trình hay công cụ (từ phần thực hành) để tiến hành xây dựng tự động nên một kho ngữ liệu song song (sheet “parallel corpus”) hay “ngữ liệu đơn ngữ” (sheet “mono corpus”) hay “tự điển/từ điển” (sheet “dictionary”) từ nguồn ngữ liệu thô sẽ cung cấp cho từng bạn SV. Đối với ngữ liệu song song, SV cần viết chương trình công cụ gióng hàng (alignment) giữa các trường thông tin (cột trong file excel) sao cho trùng khớp (matching) về nội dung.
3. Đồ án cuối kỳ (thực hiện theo nhóm 3-4 SV, chiếm 30% điểm): Nhóm SV sẽ chọn một đồ án (trong sheet “MODEL” và sheet này đang được bổ sung thêm). Nhóm SV sẽ huấn luyện model của nhóm dựa trên các kho ngữ liệu được các bạn trong lớp đã xây dựng trong các đồ án giữa kỳ.

Với kho ngữ liệu song song, mỗi SV cần xây dựng nên 1 kho có kích thước tối thiểu 40k chữ. Chẳng hạn: với hình bài thơ “Nam quốc sơn hà” dưới đây:

a.Từ ảnh văn bản (bài thơ) bên trái, sẽ có công cụ (thư viện) tự động để chúng ta xác định từng cột (bounding box) kèm theo tọa độ (4 góc: từ điểm góc trên bên trái, góc trên bên phải, góc dưới bên phải, góc dưới bên trái):



[(341, 8), (379, 8), (379, 149), (341, 149)]

[(261, 9), (297, 9), (297, 251), (261, 251)]

[(181, 6), (219, 6), (219, 252), (181, 252)]

[(102, 9), (137, 9), (137, 250), (102, 250)]

[(20, 9), (55, 8), (57, 250), (23, 250)]

Các bạn sẽ điền tọa độ từng box vào cột “Image box“ (mỗi box 1 hàng) trong file excel như hình dưới:

Image box

[(341, 8), (379, 8), (379, 149), (341, 149)]
[(261, 9), (297, 9), (297, 251), (261, 251)]
[(181, 6), (219, 6), (219, 252), (181, 252)]
[(102, 9), (137, 9), (137, 250), (102, 250)]
[(20, 9), (55, 8), (57, 250), (23, 250)]

b. Trong phần thực hành, các bạn sẽ viết chương trình (sử dụng thư viện OCR và một số kỹ thuật khác của các bạn) để biến ảnh văn bản trên thành ký tự Hán Nôm và dịch chữ Quốc ngữ thành chữ Hán Nôm. Sau đó, các bạn sẽ viết chương trình để đối chiếu giữa 2 kết quả này (dựa trên tự điển Chữ Quốc ngữ => Nôm và tự điển các chữ Hán Nôm tương đồng về mặt hình thái) để chọn ra chữ Hán Nôm đúng. Kết quả cuối cùng như hình dưới (trong sheet “example”):

ID	Image box	SinoNom char	Âm Hán Việt	Nghĩa thuần Việt
LBPv.023.001.01	[(341, 8), (379, 8), (379, 149), (341, 149)]	南國山河	Nam quốc sơn hà	Sông núi nước Nam
LBPv.023.001.02	[(261, 9), (297, 9), (297, 251), (261, 251)]	南國山河南帝居	Nam quốc sơn hà nam đế cư	Sông núi nước Nam vua Nam ở
LBPv.023.001.03	[(181, 6), (219, 6), (219, 252), (181, 252)]	截然定分在天書	Tiệt nhiên định phận tại thiên thư	Cương giới đã ghi rành rành trong sách Trời
LBPv.023.001.04	[(102, 9), (137, 9), (137, 250), (102, 250)]	如何逆虜來侵犯	Như hà nghịch lỗ lai xâm phạm	Cớ sao lũ giặc bạo ngược kia dám tới xâm phạm
LBPv.023.001.05	[(20, 9), (55, 8), (57, 250), (23, 250)]	汝等行看取敗虛	Thủ đẳng hành khan thủ bại hư	Chúng bay hãy chờ xem, thế nào cũng chuốc lấy bại vong

Tổng số các chữ trong ví dụ này (trong 4 cột tương ứng: ImageBox, SinoNomChar, ÂmHánViệt và NghĩaThuầnViệt) là: 32+32+32+43 = 139 chữ.

Kho ngữ liệu song song có thể vắng mặt cột “nghĩa thuần Việt” (tùy loại ngữ liệu thô cung cấp cho các bạn).

Cột ID (mã số): gồm 12 ký tự với ý nghĩa như sau (từ trái sang phải): **DSGk.fff.ppp.ss**

D: Domain: chỉ lĩnh vực, như: văn học (L), lịch sử (H), y học (M), tôn giáo (R), …

S: Subdomain: chỉ lĩnh vực con, như: văn học – nhóm B (LB), tôn giáo Công giáo (RC),...

G: Genre: chỉ thể loại: văn vần (P), văn xuôi (T), ..

k: Kind: loại ngữ liệu thô (có cột thuần Việt “v”)

fff: #số hiệu file ngữ liệu thô trong lĩnh vực con đó

ppp: #số trang trong file ngữ liệu thô đó

ss: #số câu hay số box trong trang ngữ liệu thô đó (số này do các bạn đánh số tự động, còn các ký tự khác sẽ cung cấp sẵn cho các bạn khi giao ngữ liệu thô.Vd: **LBPv.023.001.**ss).