ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

THU THẬP DỮ LIỆU SONG NGỮ CHỮ NÔM – CHỮ QUỐC NGỮ

Đồ án môn học: Xử lí ngôn ngữ tự nhiên nâng cao

Chương trình cao học khoá 32

Dưới sự hướng dẫn của: PGS. TS. Đinh Điền
ThS. Lê Thị Thúy Hằng
ThS. Nguyễn Hồng Bửu Long

Thực hiện bởi: Nguyễn Thị Thu Duyên – 22C11005
Đặng Hoàng Minh Triết – 22C11048

TP. HỒ CHÍ MINH – 2022

LỜI CÁM ƠN

Chúng em xin bày tỏ lòng biết ơn chân thành đến PGS. TS. Đinh Điền, ThS. Lê Thị Thúy Hằng và ThS. Nguyễn Hồng Bửu Long đã dành thời gian, kiến thức và sự hỗ trợ để chúng em hoàn thành đồ án môn học này.

Trong suốt quá trình thực hiện đồ án, ThS. Lê Thị Thúy Hằng đã cung cấp cho chúng em những chỉ dẫn quan trọng và những gợi ý giá trị để chúng em có thể tiến hành nghiên cứu và phân tích một cách hiệu quả. Đồng thời, PGS. TS. Đinh Điền và ThS. Nguyễn Hồng Bửu Long đã luôn sẵn sàng lắng nghe và giải đáp những thắc mắc của chúng em, giúp chúng em vượt qua các khó khăn và tiến bộ trong quá trình làm việc.

Chúng tôi cũng xin bày tỏ lòng biết ơn đến các thành viên khác trong lớp học đã chia sẻ ý kiến, góp ý và đóng góp xây dựng vào đồ án của chúng tôi. Đặc biệt là anh Võ Hoài Danh đã giúp chúng tôi trong việc giới thiệu một số ứng dụng thú vị nhằm hỗ trợ cho đồ án này.

Cuối cùng, chúng tôi xin chân thành cảm ơn gia đình, bạn bè và những người thân yêu đã luôn ủng hộ, động viên và tạo điều kiện thuận lợi để chúng tôi hoàn thành đồ án này.

Rất cảm kích vì sự hỗ trợ và đóng góp của mọi người, chúng tôi hy vọng rằng đồ án môn học này sẽ đáp ứng được mong đợi và mang lại giá trị thực tế.

Xin chân thành cảm ơn!

MỤC LỤC

MỞ ĐẦU

Trong thời đại hiện đại của công nghệ thông tin, việc thu thập dữ liệu đang trở nên ngày càng quan trọng và phổ biến trong lĩnh vực nghiên cứu khoa học. Trong ngữ cảnh đó, việc thu thập dữ liệu chữ Nôm và chữ Quốc Ngữ từ các trang web trực tuyến có vai trò đặc biệt quan trọng trong việc bảo tồn và nghiên cứu văn hóa lịch sử Việt Nam.

Hơn thế nữa, việc chuyển ngữ giữa chữ Nôm và chữ Quốc ngữ có thể hỗ trợ các tác vụ như phân tích tài liệu lịch sử, công cụ giáo dục và học ngôn ngữ, lưu trữ và bảo quản kỹ thuật số, tạo siêu dữ liệu cho bộ sưu tập thư pháp và cải thiện hệ thống dịch máy. Với mô hình ghi nhãn chữ Nôm, các nhà nghiên cứu và sử học có thể xác định và phân loại các tài liệu cổ chứa chữ Nôm để phân tích và hiểu nội dung dễ dàng hơn. Người dạy ngôn ngữ và người học có thể sử dụng mô hình để tạo tài liệu học tập, từ điển hoặc công cụ tương tác để học viên thực hành đọc và hiểu các chữ Nôm. Các thư viện, bảo tàng và tổ chức văn hóa có thể sử dụng mô hình này để số hóa và bảo quản các bản thảo hoặc hiện vật chữ Nôm. Đối với các nhà sưu tập và phòng trưng bày nghệ thuật, mô hình này giúp cải thiện khả năng tìm kiếm và hiểu biết về các bộ sưu tập thư pháp của họ. Cuối cùng, tích hợp mô hình ghi nhãn chữ Nôm vào hệ thống dịch máy có thể cải thiện khả năng dịch chính xác và giao tiếp giữa các nhóm ngôn ngữ khác nhau sử dụng chữ Nôm.

Với các tác vụ kể trên, nhu cầu thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ là nhiệm vụ quan trọng và cấp thiết để có thể tiếp tục xây dựng và phát triển các tác vụ trên. Vì lẽ đó, báo cáo này hướng đến nhiệm vụ tập trung vào quá trình thu thập dữ liệu chữ Nôm và chữ Quốc Ngữ từ trang web, với mục tiêu xây dựng một nguồn dữ liệu đáng tin cậy và đầy đủ để phục vụ các nghiên cứu và ứng dụng trong các lĩnh vực này. Trong quá trình nghiên cứu, chúng em đã tìm hiểu và áp dụng các phương pháp và công cụ thích hợp nhằm thu thập dữ liệu một cách hiệu quả và chính xác từ các trang web, đặc biệt là trang web: https://www.scribd.com/lists/21643875/Th%C6%A1-V%C4%83n-Han-Nom.

Mục tiêu chính của báo cáo này là góp phần vào nỗ lực bảo tồn và phát triển nghiên cứu về chữ Nôm và chữ Quốc Ngữ. Bằng cách thu thập dữ liệu đáng tin cậy và dễ truy cập, hy vọng rằng nghiên cứu trong lĩnh vực này sẽ được khám phá sâu hơn và tạo ra những kết quả đáng giá.

Từ khóa: chữ Nôm, chữ Quốc Ngữ, thu thập dữ liệu, khai thác ngữ liệu.

CHƯƠNG 1: GIỚI THIỆU

1.1. Giới thiệu đề tài

Trong thời đại công nghệ thông tin ngày nay, việc thu thập dữ liệu đóng một vai trò vô cùng quan trọng và phổ biến, đặc biệt trong lĩnh vực nghiên cứu ngôn ngữ và văn hóa. Trong bối cảnh đó, đề tài này tập trung vào việc thu thập dữ liệu chữ Nôm và chữ Quốc ngữ tương ứng từ một trang web đáng chú ý về văn học và văn hóa Việt Nam.

Chữ Nôm là một hệ thống chữ viết ngữ tố - âm tiết được sử dụng để biểu diễn tiếng Việt. Đây là bộ chữ do người Việt tạo ra dựa trên chữ Hán, các bộ thủ, âm đọc và từ vựng trong tiếng Việt.

Chữ Nôm đã hình thành và phát triển từ thế kỷ X đến thế kỷ XX. Ban đầu, chữ Nôm thường được sử dụng để ghi lại tên người và địa danh, sau đó dần lan rộng và tiến vào cuộc sống văn hóa của quốc gia. Trong thời kỳ nhà Hồ và nhà Tây Sơn, xuất hiện xu hướng sử dụng chữ Nôm trong văn bản hành chính. Đối với văn học Việt Nam, chữ Nôm mang ý nghĩa đặc biệt quan trọng khi được sử dụng làm công cụ xây dựng nền văn học truyền thống. Trải qua suốt hàng thế kỷ và trở thành một phương tiện quan trọng không những trong văn học, thơ ca mà còn cả trong các lĩnh vực khác như lịch sử, tôn giáo, y học, v.v

Chữ Quốc ngữ là một hệ thống chữ viết tiếng Việt sử dụng bảng chữ cái Latinh và dấu phụ, kết hợp với các chữ cái để biểu thị âm tiết và âm đầu trong tiếng Việt. Hệ thống chữ này được tạo ra dựa trên việc cải tiến bảng chữ cái Latinh và sự kết hợp âm theo quy tắc chính tả của văn tự tiếng Bồ Đào Nha, cùng với một số yếu tố từ tiếng Ý.

Việc tạo ra chữ Quốc ngữ là một sự cải tiến đáng kể trong việc biểu diễn tiếng Việt bằng một hệ thống chữ viết hiệu quả và dễ tiếp cận. Trước đây, việc ghi lại và truyền tải ngôn ngữ Việt thường dựa trên chữ Hán và chữ Nôm, đòi hỏi kiến thức phức tạp và tốn nhiều thời gian để học. Với sự ra đời của chữ Quốc ngữ, việc học và sử dụng tiếng Việt trở nên đơn giản hơn, mở ra nhiều cơ hội tiếp cận kiến thức và giao tiếp với thế giới.

Chữ Quốc ngữ đã trở thành hệ thống chữ viết chính thức và phổ biến nhất trong tiếng Việt từ thế kỷ XX đến nay. Nó đã đóng vai trò quan trọng trong việc phát triển giáo dục, truyền thông, xuất bản và giao tiếp trong cộng đồng quốc tế. Đồng thời, việc sử dụng chữ Quốc ngữ cũng giúp bảo tồn và phát triển văn hóa, văn bản và văn hoá của người Việt Nam, góp phần tạo nên bản sắc và độc đáo của ngôn ngữ và văn hóa Việt.

Như vậy, chữ Nôm là hệ thống chữ viết đặc trưng của Việt Nam, mang trong mình những giá trị văn hóa và lịch sử đặc biệt, trong khi chữ Quốc ngữ là hệ thống chữ viết hiện

đại và phổ biến. Sự đối chiếu và thu thập dữ liệu song song giữa hai hệ thống chữ này từ một trang web đa dạng và phong phú sẽ mang lại những thông tin vô cùng giá trị cho việc khám phá và hiểu sâu hơn về di sản văn hóa của đất nước.

1.2. Động lực

Trong lĩnh vực nghiên cứu văn hóa lịch sử và ngôn ngữ của Việt Nam, việc thu thập dữ liệu chữ Nôm và chữ Quốc ngữ từ các nguồn trực tuyến đóng một vai trò vô cùng quan trọng và hấp dẫn nhằm:

1. *Bảo tồn và khám phá di sản văn hóa*: Chữ Nôm là một hệ thống chữ viết độc đáo của Việt Nam, chứa đựng những giá trị văn hóa lịch sử quý giá. Việc thu thập dữ liệu chữ Nôm từ các trang web giúp bảo tồn và khôi phục những tác phẩm văn học, thơ ca, và kiến thức truyền thống đã được truyền từ đời này sang đời khác. Đồng thời, việc thu thập dữ liệu chữ Quốc ngữ cung cấp một cơ hội phát triển nghiên cứu ngôn ngữ hiện đại và nguồn thông tin văn học đa dạng từ các nguồn trực tuyến.
2. *Nghiên cứu văn hóa và ngôn ngữ đa chiều*: Thu thập dữ liệu chữ Nôm và chữ Quốc ngữ từ trang web cho phép nhà nghiên cứu xây dựng cơ sở dữ liệu đáng tin cậy và đầy đủ về ngôn ngữ và văn hóa của Việt Nam. Từ những dữ liệu thu thập được, nhà nghiên cứu có thể thực hiện phân tích, so sánh, và nghiên cứu đa chiều về diễn đạt, ngữ nghĩa, và sự phát triển của ngôn ngữ và văn hóa trong lịch sử và hiện tại.
3. *Đóng góp vào nghiên cứu khoa học và giáo dục*: Các bài báo cáo và nghiên cứu dựa trên dữ liệu thu thập từ trang web có thể đóng góp vào việc nâng cao nhận thức và hiểu biết về chữ Nôm, chữ Quốc ngữ, và di sản văn hóa của Việt Nam. Những thông tin này không chỉ hỗ trợ các nhà nghiên cứu trong việc làm sáng tỏ các khía cạnh lịch sử và văn hóa, mà còn có thể áp dụng trong giáo dục và truyền thông, góp phần nâng cao chất lượng giáo dục và tăng cường ý thức văn hóa cho cộng đồng.
4. *Khám phá những thách thức và cơ hội mới*: Quá trình thu thập dữ liệu từ trang web có thể đưa ra những thách thức đối với nhà nghiên cứu, như tính chất phong phú và đa dạng của dữ liệu, tính chính xác và đáng tin cậy của nguồn thông tin. Tuy nhiên, đồng thời cũng mở ra cơ hội để phát triển và ứng dụng các công cụ và phương pháp mới trong việc thu thập và xử lý dữ liệu, cải thiện hiệu quả và chất lượng của nghiên cứu.

Tóm lại, việc thu thập dữ liệu chữ Nôm và chữ Quốc ngữ từ trang web không chỉ đơn thuần là công việc nghiên cứu mà còn là một hành trình khám phá và bảo tồn di sản văn hóa của quốc gia. Đây là cơ hội để chúng em đóng góp vào sự phát triển của nghiên cứu

khoa học, giáo dục, và bảo tồn văn hóa, mang lại những giá trị tích cực cho cộng đồng và xã hội.

1.3. Mô tả bài toán

Với ý nghĩa và nhu cầu thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ như hiện nay, nhóm chúng em tập trung vào hai nhiệm vụ chính trong báo cáo này bao gồm xây dựng trình thu thập dữ liệu từ một trang web và chương trình xử lý dữ liệu vừa thu thập thành dạng bảng, mô tả bài toán cụ thể như sau:

Đầu vào:

Trang web https://www.scribd.com/lists/21643875/Th%C6%A1-V%C4%83n-Han-Nom chứa các tệp văn bản chữ Nôm và chữ Quốc ngữ dưới dạng tệp HTML, dạng tệp PDF hoặc các định dạng có thể sao chép thành tập văn bản.

Đầu ra:

- Trình thu thập tự động: Tạo và lưu trữ các tệp văn bản chữ Nôm và chữ Quốc ngữ từ nguồn trên thành các tệp văn bản tương ứng.

- Chương trình xử lý: Tạo một bảng gồm hai cột, một cột chứa văn bản chữ Nôm và một cột chứa văn bản chữ Quốc ngữ từ các tệp văn bản đã thu thập.

Như vậy, kết quả cuối cùng của bài toán này là một tệp excel trong đó có một cột là chữ Nôm, một cột là chữ Quốc Ngữ được trích xuất từ trang web đầu vào. Tệp này có thể sử dụng để nghiên cứu, tìm hiểu và tạo ra các ứng dụng dựa trên tri thức thu được từ chữ Nôm và chữ Quốc ngữ.

1.4. Tổ chức

Báo cáo được tổ chức theo cấu trúc sau: phần đầu tiên giới thiệu về chữ Nôm và chữ Quốc Ngữ, đồng thời tập trung thảo luận về tầm quan trọng của việc thu thập dữ liệu trong việc nghiên cứu và bảo tồn văn hóa lịch sử. Tiếp theo, chúng em trình bày chi tiết về phương pháp và quy trình thu thập dữ liệu từ trang web, bao gồm các công cụ và kỹ thuật sử dụng. Cuối cùng, chúng em trình bày kết quả thu thập dữ liệu và thảo luận về những thách thức và cơ hội trong quá trình thu thập dữ liệu trực tuyến.

CHƯƠNG 2: TỔNG QUAN

2.1. Các công trình liên quan

Việc thu thập dữ liệu chữ Nôm từ các nguồn đáng tin cậy đang là một trong những nhiệm vụ cấp thiết và hàng đầu giúp các nhà nghiên cứu có đủ dữ liệu để thực hiện các nghiên cứu khác dựa trên cơ sở nguồn dữ liệu ban đầu như rút trích đặc trưng, ứng dụng các kỹ thuật học máy, học sâu nhằm tạo ra các ứng dụng hữu ích cho người dùng dựa trên các tri thức được tích góp và ghi chép lại bằng chữ Nôm trong suốt hàng kỷ (từ khoảng thế kỷ thứ 10 đến đầu thế kỷ thứ 20). Một trong số đó phải kể đến là nỗ lực thu thập các tài liệu chữ Nôm viết tay từ hình ảnh của các tài liệu cổ được cung cấp bởi thư viện quốc gia như [1].

Một nỗ lực khác kéo dài khoảng 20 năm của Tổ chức Bảo tồn Chữ Nôm Việt Nam (VNPF) phát sinh từ nỗ lực của người Việt Nam và người Mỹ sau chiến tranh, nhằm bảo tồn và phổ biến nền văn hóa lịch sử của chữ Nôm. Họ thành công mã hóa hơn 18.000 ký tự Hán-Nôm vào Unicode/ISO 10646 và cung cấp phông chữ Nôm Na Tong Light. VNPF đã số hóa và lập danh mục khoảng 4.000 văn bản Hán-Nôm, tạo ra một thư viện kỹ thuật số đầu tiên tại Việt Nam. Ngoài ra, họ cũng phát triển kho lưu trữ di sản chữ Nôm, cung cấp các phiên bản tìm kiếm được của tài liệu quan trọng về mặt văn hóa, lịch sử, thơ ca, tục ngữ và sân khấu của triều đại Lý và Trần. Tính đến đầu năm 2016, một số tác phẩm có sẵn để nghiên cứu bằng Hán-Nôm và Quốc Ngữ trên trang web của VNPF, bao gồm Truyện Kiều, Đại Việt Sử Ký Toàn Thư, thơ của nữ thi sĩ thế kỷ 18 Hồ Xuân Hương, Lục Vân Tiên và Chinh Phụ Ngâm Khúc.

Theo sau đó, với sự phát triển vượt bậc của các mô hình học máy, học sâu, một nỗ lực khác gần đây nhằm thu thập dữ liệu chữ Nôm gồm 2953 trang viết tay thu thập từ Quỹ Bảo tồn Chữ Nôm Việt Nam phục vụ cho hai bài toán phát hiện văn bản và nhận dạng văn bản được báo cáo trong [2] đã ra đời.

2.2. Thách thức

Quá trình thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ nói chung đều phải đối mặt với một số thách thức đáng kể do sự phức tạp của ngôn ngữ chữ Nôm và các vấn đề liên quan đến việc chuyển đổi và bảo tồn dữ liệu. Dưới đây là một số thách thức chính:

Hạn chế về số lượng tư liệu: Dữ liệu chữ Nôm đã khá lâu đời, khiến số lượng tư liệu hiện có rất hạn chế và thưa thớt so với chữ Quốc ngữ. Nhiều tư liệu đã bị mất đi hoặc bị hư hỏng theo thời gian. Điều này làm cho việc thu thập dữ liệu chữ Nôm trở nên khó khăn và tốn nhiều thời gian.

Khó khăn trong việc đọc và hiểu chữ Nôm: Chữ Nôm có hệ thống văn bản và cách diễn đạt khác biệt so với chữ Quốc ngữ, làm cho việc đọc và hiểu nội dung của các tư liệu chữ Nôm trở nên khó khăn đối với người không quen với hệ thống này. Điều này đòi hỏi sự hiểu biết sâu về ngôn ngữ và văn hóa truyền thống.

Rủi ro mất mát và biến dạng dữ liệu: Khi tiến hành số hóa các văn bản chữ Nôm cũ, có nguy cơ mất mát thông tin quan trọng hoặc bị sai sót trong quá trình chuyển đổi sang dạng kỹ thuật số. Điều này có thể xảy ra khi các bản gốc bị hư hỏng hoặc do quá trình quét và nhận dạng ký tự không chính xác.

Đồng bộ hóa và tương thích kỹ thuật: Để hiển thị dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ một cách chính xác trên các nền tảng và thiết bị khác nhau, cần đồng bộ hóa và đảm bảo tính tương thích kỹ thuật giữa hai hệ thống này. Điều này có thể yêu cầu việc phát triển các phông chữ, mã hóa và tiêu chuẩn hóa dữ liệu phù hợp.

Định dạng lưu trữ và tìm kiếm dữ liệu: Tổ chức và lưu trữ dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ đòi hỏi một cơ sở dữ liệu phức tạp và hiệu quả để quản lý và tìm kiếm thông tin dễ dàng. Điều này đòi hỏi sự tinh chỉnh và tích hợp công nghệ thông tin phù hợp.

Phát triển các công cụ hỗ trợ: Các công cụ phần mềm để số hóa, chuyển đổi và hiển thị dữ liệu chữ Nôm và chữ Quốc ngữ cần được phát triển và cập nhật liên tục để đảm bảo tính hiệu quả và tiện lợi trong quá trình thu thập và sử dụng dữ liệu.

Sâu hơn nữa, việc thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ từ các trang web nói riêng còn phải đối mặt với nhiều thách thức khác bao gồm:

Định dạng không đồng nhất: Các trang web thường sử dụng nhiều định dạng khác nhau để hiển thị chữ Nôm, bao gồm Unicode, HTML entities hoặc mã hóa đặc biệt. Điều này tạo ra khó khăn khi cần trích xuất và xử lý dữ liệu, đồng thời yêu cầu phải sử dụng các công cụ phân tích và mã hóa thích hợp.

Lỗi trong mã HTML: Các trang web có thể chứa các lỗi trong mã HTML, chẳng hạn như các thẻ không được đóng đầy đủ, điều này gây ra khó khăn trong quá trình phân tích và trích xuất dữ liệu. Việc sửa lỗi này trước khi thu thập dữ liệu là cần thiết để đảm bảo độ chính xác của dữ liệu thu thập được.

Đa ngôn ngữ: Một số trang web chứa văn bản chữ Nôm hoặc chữ Quốc ngữ song song với các ngôn ngữ khác, làm cho quá trình trích xuất dữ liệu phức tạp hơn. Cần phải sử dụng các công cụ và phương pháp xử lý ngôn ngữ tự nhiên đa ngôn ngữ để xử lý các văn bản này một cách hiệu quả.

Phân tán dữ liệu: Dữ liệu chữ Nôm có thể được phân tán trên nhiều trang web, việc thu thập và tổng hợp dữ liệu này tốn nhiều công sức và thời gian. Cần phải thiết kế một quy trình thu thập dữ liệu tự động và hiệu quả để thu thập dữ liệu từ nhiều nguồn và kết hợp chúng thành một nguồn dữ liệu toàn diện.

Hạn chế truy cập: Một số trang web có thể có hạn chế về quyền truy cập hoặc tần suất truy cập, điều này có thể dẫn đến việc thu thập dữ liệu chậm chạp hoặc không hoàn toàn. Cần phải xem xét các giải pháp để vượt qua các hạn chế này, bao gồm việc sử dụng các công cụ proxy và điều chỉnh tần suất truy cập.

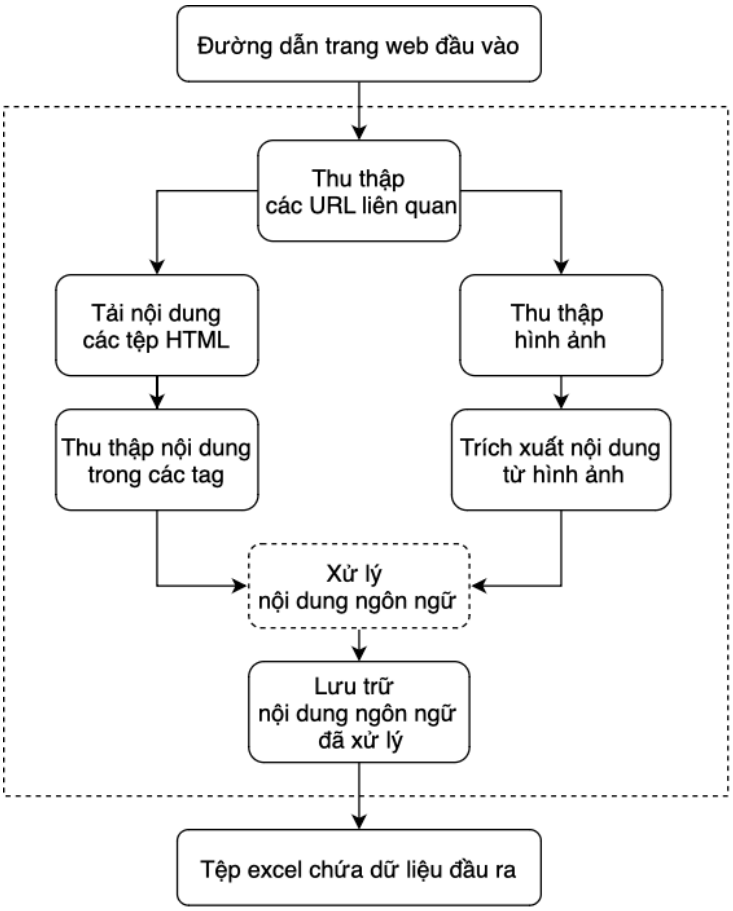
Độ tin cậy và chất lượng dữ liệu: Dữ liệu thu thập từ các trang web có thể không luôn đảm bảo độ tin cậy và chất lượng cao, do sự sai sót hoặc thay đổi trong nội dung trang web. Cần phải áp dụng các phương pháp kiểm tra dữ liệu và xác minh tính chính xác của nó trước khi sử dụng trong nghiên cứu.

Tóm lại, việc thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ từ các trang web đòi hỏi sự cẩn thận và kiên nhẫn trong việc xử lý các thách thức kỹ thuật và quản lý rủi ro liên quan đến bảo mật và chất lượng dữ liệu.

CHƯƠNG 3: PHƯƠNG PHÁP ĐỀ XUẤT

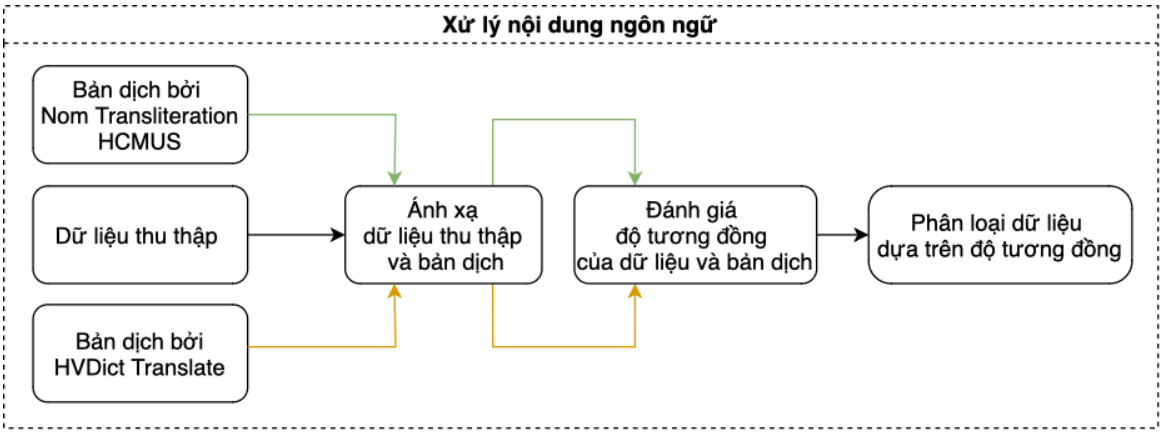
3.1. Sơ đồ chung của hệ thống

Ở báo cáo này, chúng em đề xuất phương pháp thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ từ một trang web (hình 3.1.1) với các bước chính bao gồm: đầu tiên, chúng em thực hiện thu thập các đường dẫn liên quan từ đường dẫn trang web đầu vào, tiếp theo chúng em thực hiện hai tác vụ, một là tải nội dung các tệp HTML, hai là thu thập hình ảnh từ các đường dẫn vừa thu thập được. Đối với các tệp HTML thu thập được, chúng em tiếp tục tiến hành thu thập nội dung cần thiết trong các tag. Đối với các hình ảnh thu thập được, chúng em dự định tiếp tục tiến hành trích xuất nội dung cần thiết từ hình ảnh vừa thu thập trong phần phát triển tiếp theo sau báo cáo này. Các nội dung liên quan sau khi được thu thập sẽ được đưa qua quá trình xử lý nội dung ngôn ngữ (hình 3.1.2). Sau khi xử lý nội dung ngôn ngữ, dữ liệu sẽ được lưu trữ lại và xuất thành tệp excel đầu ra để sử dụng cho các mục đích và tác vụ khác nhau.



Hình 3.1.1. *Sơ đồ chung của hệ thống thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ từ một trang web.*

Nhận thấy để nâng cao tính đúng đắn và độ tin cậy của dữ liệu thu thập, chúng em đề xuất quá trình Xử lý nội dung ngôn ngữ (hình 3.1.2) sau khi thu thập được với các bước sau: ánh xạ dữ liệu vừa thu được bao gồm chữ Nôm và chữ Quốc ngữ với bản dịch chữ Nôm vừa thu thập bởi hai từ điển tin cậy là Nom Transliteration HCMUS và HVDict Translate. Tiếp đến, chúng em đánh giá độ tương đồng của dữ liệu thu thập được với các bản dịch vừa ánh xạ, sau đó phân loại chúng dựa trên độ tương đồng vừa đánh giá được.



Hình 3.1.2. *Sơ đồ mô tả quy trình xử lý nội dung ngôn ngữ được đề cập đến trong sơ đồ chung ở hình 3.1.1*

3.2. Thu thập các đường dẫn (URL) liên quan

Để thu thập danh sách các đường dẫn này và sử dụng chúng cho quá trình tải nội dung, chúng em đã tiến hành quá trình thu thập thông tin và xử lý dữ liệu từ trang web đầu vào. Điều này đảm bảo chúng em có tập hợp đầy đủ và đáng tin cậy các đường dẫn liên quan, giúp cho quá trình nghiên cứu tiếp theo được diễn ra một cách hiệu quả và chính xác. Chúng em hy vọng rằng việc sử dụng các thư viện và công cụ phù hợp sẽ đóng góp vào thành công của nghiên cứu và tạo ra những kết quả đáng giá. Vì vậy, ở giai đoạn thu thập các đường dẫn liên quan này, chúng em đã sử dụng một số thư viện Python quan trọng như requests, BeautifulSoup và re. Thư viện requests được sử dụng để hỗ trợ gửi các yêu cầu HTTP/1.1 một cách dễ dàng và hiệu quả. Để thực hiện việc truy cập vào trang web https://www.scribd.com/lists/21643875/Th%C6%A1-V%C4%83n-Han-Nom, chúng em đã sử dụng thư viện này và kiểm tra tính tồn tại cũng như hoạt động của các đường dẫn.

Tiếp theo, chúng em đã thực hiện phân tích kết quả từ việc truy cập trang web để xác định các đường dẫn liên quan đến các tệp văn bản chữ Nôm và chữ Quốc ngữ. Việc này giúp chúng em tập trung vào những đường dẫn có liên quan đến nghiên cứu của chúng em, cụ thể là các đường dẫn có chứa nội dung là chữ Nôm và chữ Quốc ngữ.

3.3. Tải nội dung các tệp HTML

Trong giai đoạn tiếp theo của nghiên cứu, chúng tôi tiếp tục sử dụng thư viện requests để tải nội dung từ các đường dẫn đã thu thập được trong giai đoạn trước đó. Việc này giúp chúng em thu thập thông tin chi tiết từ các trang web liên quan. Sau khi tải nội dung từ các đường dẫn, chúng tôi tiến hành lưu trữ các tệp HTML này dưới dạng text vào một thư mục

đích. Quá trình lưu trữ dữ liệu này được thực hiện nhằm chuẩn bị cho việc thu thập dữ liệu từ các tag cần thiết trong quá trình phân tích.

3.4. Thu thập nội dung trong các tag cần thiết:

Tiếp theo chúng em sử dụng một công cụ phân tích mã nguồn BeautifulSoup trong Python để thu thập nội dung từ các tag HTML cần thiết. Về cơ bản, Beautiful Soup là một gói Python dùng để phân tích các tài liệu HTML và XML, bao gồm cả tài liệu có đánh dấu không hoàn chỉnh (tag soup). Nó tạo ra cây phân tích cho các trang web đã được phân tích, giúp trích xuất dữ liệu từ HTML, phục vụ cho việc thu thập thông tin từ web (web scraping). Công cụ này cho phép chúng em trích xuất thông tin một cách linh hoạt và hiệu quả từ các trang web được lưu trữ trong các tệp HTML.

Mục tiêu của quá trình thu thập dữ liệu là xác định các tag chứa nội dung chữ Nôm và chữ Quốc ngữ trong tệp HTML và thu thập nội dung từ các thẻ này. Các tag này có thể bao gồm các thẻ văn bản chữ Nôm và chữ Quốc ngữ, định dạng, dấu câu, hoặc các thông tin liên quan khác. Việc thu thập dữ liệu từ các tag quan trọng này sẽ đóng vai trò quan trọng trong việc hiểu và phân tích nội dung của các tài liệu, tạo nền tảng cho việc tiếp tục xử lý và phân tích dữ liệu trong các bước sau của nghiên cứu.

Thông thường, các văn bản chữ Nôm được trình bày không thống nhất, một số được trình bày theo chiều dọc, số khác được trình bày theo chiều ngang. Vì vậy, để có thể trích xuất đúng nội dung từ các tệp HTML, chúng em đề xuất sử dụng vị trí góc trên bên trái (top, left) của tất cả các thẻ liên quan. Bằng cách xác định vị trí này, chúng em có thể lọc nội dung từ các thẻ có cùng giá trị vị trí góc trên bên trái. Tiếp đến, chúng em tiến hành lọc sạch dữ liệu bằng một số cách như loại bỏ ký tự lỗi (unprintable), loại bỏ ký tự trùng lắp. Điểm đặc biệt của giai đoạn này là chúng em tiến hành loại bỏ ký tự trùng lắp, bước này rất quan trọng, vì như đã đề cập trước đó, văn bản thu thập được thường bị trộn lẫn giữa chữ Nôm, chữ Quốc ngữ và các thành phần không thực sự có giá trị khác. Vì vậy ngoài việc loại bỏ các thành phần không tiềm năng bằng cách loại bỏ ký tự lỗi, chúng em còn lọc riêng chữ Nôm và chữ Quốc ngữ bằng cách loại bỏ các ký tự trùng lắp khi đối sánh các chuỗi với nhau.

3.5. Thu thập hình ảnh:

Trong giai đoạn này, chúng em sử dụng một số gói thư viện như tqdm, BeautifulSoup, urllib để quản lý tiến trình, thu thập ảnh và tải dữ liệu.

Gói tqdm là một công cụ Python cung cấp thanh tiến trình giúp theo dõi quá trình thực hiện vòng lặp hoặc công việc mất thời gian.

Khi bạn đối mặt với vòng lặp lớn hoặc tác vụ mất nhiều thời gian để hoàn thành, tqdm sẽ thêm một thanh tiến trình vào giao diện dòng lệnh của Python (CLI). Thanh tiến trình này hiển thị thông tin về tiến độ của vòng lặp hoặc công việc, giúp bạn biết tổng số lần lặp, số lần đã hoàn thành và tỷ lệ phần trăm hoàn thành.

Dựa vào tqdm, người dùng có thể dễ dàng theo dõi và quản lý tiến độ các tác vụ mất thời gian, từ đó tăng hiệu suất và tiện lợi trong việc thực hiện các ứng dụng Python có liên quan đến vòng lặp hoặc xử lý dữ liệu lớn.

Gói urllib trong Python cung cấp bộ công cụ làm việc với các URL và hoạt động liên quan đến mạng. Gói này bao gồm ba mô-đun chính: urllib.request để thực hiện yêu cầu HTTP và tải dữ liệu từ URL, urllib.parse để phân tích và xây dựng các thành phần của URL như giao thức, domain, đường dẫn và tham số truy vấn, và urllib.error để xử lý các lỗi mạng.

Nhờ gói urllib, người dùng có thể dễ dàng làm việc với dữ liệu trên mạng và thực hiện các tác vụ liên quan đến web trong ứng dụng Python của họ. Điều này làm cho gói urllib trở thành một công cụ hữu ích trong việc xây dựng các ứng dụng mạng trong Python.

Chúng em đã sử dụng các thư viện đã giới thiệu để triển khai việc truy cập và tải xuống các hình ảnh từ các đường dẫn đã xác định, và sau đó lưu trữ chúng vào các thư mục tương ứng. Quá trình này bao gồm các bước sau:

Kiểm tra tính hợp lệ của các đường dẫn đầu vào để đảm bảo chúng có thể được truy cập.

Xác định các hình ảnh có sẵn trong đường dẫn đang xét để biết chính xác số lượng hình ảnh cần tải xuống.

Thực hiện tải xuống các hình ảnh từ các đường dẫn đã xác định và lưu trữ chúng vào các thư mục tương ứng.

Qua việc sử dụng các thư viện phù hợp, chúng em có thể dễ dàng tự động thực hiện quy trình này, giúp tiết kiệm thời gian và công sức so với việc thực hiện thủ công. Điều này đảm bảo tính chính xác và hiệu quả trong việc quản lý và lưu trữ các hình ảnh từ các nguồn dữ liệu khác nhau.

3.6. Xử lý nội dung ngôn ngữ

3.6.1. Lưu trữ nội dung thành các file txt

Trong giai đoạn này, chúng em tiếp tục tiến hành thu thập nội dung cần thiết trong các thẻ của tệp HTML và lưu chúng lại dưới dạng tệp txt để dễ xử lý ở bước tiếp theo. Sau khi thu thập xong, chúng em thực hiện đánh giá và đối sánh lại một lần nữa các tệp txt để xác định danh sách các tệp thực sự có chứa nội dung đang cần thu thập. Về chi tiết cài đặt, chúng em sử dụng các thông số top, left trong style của một thẻ để xác định kí tự có cùng trên một dòng hay không. Bằng cách sử dụng start_top như một con trỏ để so sánh với thuộc tính top của các thẻ nó chạy qua. Lưu các thông số của một hàng theo cấu trúc sau:

- Top: Thông số top được lấy trong style của kí tự cuối cùng trong hang
- Left: Thông số left được lấy trong style của kí tự cuối cùng trong hang
- start_top: Thông số được lấy làm chuẩn để mốc xác định các kí tự cùng hang
- Text: Nội dụng thu thập trong các thẻ được nối với nhau và được xác định là cùng hàng theo quy luật mà chúng em đã đề xuất

3.6.2. Dùng Nom Transliteration HCMUS và HVDict Translate API dịch làm chuẩn

Việc thu thập dữ liệu từ các trang web đặt ra một thách thức lớn về độ tin cậy, do đó, để đảm bảo tính chính xác của dữ liệu thu thập được, chúng tôi đã quyết định sử dụng thêm các bản dịch từ chữ Nôm sang chữ Quốc ngữ đáng tin cậy. Trong nỗ lực này, chúng tôi đã sử dụng hai API quan trọng là Nom Transliteration HCMUS từ Đại học Khoa học Tự nhiên TP.HCM và HVDict Translate.

Trong báo cáo này, chúng tôi đã áp dụng API Nom Transliteration để dịch chữ Nôm sang chữ Quốc ngữ một cách chính xác và tin cậy. Sau đó, chúng tôi so sánh kết quả dịch thu được với các văn bản chữ Quốc ngữ tương ứng từ các trang web đã thu thập được. Việc này giúp đảm bảo rằng dữ liệu sử dụng trong nghiên cứu của chúng tôi đáng tin cậy và chính xác.

3.6.3. Ánh xạ các bản dịch và dữ liệu thu thập được

Kết hợp kết quả dịch chữ Nôm từ hai API trên và dữ liệu chữ Quốc ngữ đã thu thập từ trang web để tạo các cặp câu song ngữ.

3.6.4. Phân loại dữ liệu dựa trên độ tương đồng so với bản dịch:

Về cơ bản, việc đánh giá độ tương đồng thường dựa trên các phương pháp đo độ tương đồng như TF-IDF, Cosine Similarity, hay khoảng cách Levenshtein. Trong đó, TF-IDF,

Cosine Similarity và Khoảng cách Levenshtein là ba khái niệm quan trọng trong xử lý ngôn ngữ tự nhiên và phân tích dữ liệu. Cụ thể:

TF-IDF (Term Frequency-Inverse Document Frequency) là một phương pháp tính toán trọng số của các từ trong một văn bản so với một tập các văn bản. Nó đo lường tần suất xuất hiện của một từ trong văn bản (TF) và đồng thời điều chỉnh bằng độ quan trọng của từ đó trong tập văn bản (IDF). TF-IDF thường được sử dụng trong các tác vụ như tìm kiếm thông tin, tóm tắt văn bản, phân loại văn bản và gom nhóm dữ liệu.

Cosine Similarity là một phép đo độ tương đồng giữa hai véc-tơ trong không gian đa chiều. Nó đo lường góc giữa hai véc-tơ và cho biết mức độ hướng giống nhau của chúng. Giá trị Cosine Similarity nằm trong khoảng [-1, 1], trong đó 1 đại diện cho sự giống nhau hoàn toàn, 0 đại diện cho sự không tương đồng và -1 đại diện cho sự đối nghịch hoàn toàn. Cosine Similarity thường được sử dụng trong các bài toán gom nhóm (clustering) và tìm kiếm thông tin để so sánh mức độ tương đồng giữa các văn bản, tài liệu hoặc tập dữ liệu.

Khoảng cách Levenshtein, còn gọi là khoảng cách sửa đổi, đo lường số lượng các thao tác chỉnh sửa (chèn, xóa hoặc thay thế ký tự) cần thiết để biến đổi một chuỗi thành chuỗi khác. Nó được sử dụng để so sánh độ giống nhau giữa hai chuỗi, đặc biệt là khi xử lý dữ liệu chuỗi có thể xuất hiện các sai sót hoặc thay đổi nhỏ. Khoảng cách Levenshtein thường được áp dụng trong các tác vụ như gợi ý từ điển, phân loại văn bản gần giống và so sánh chuỗi dữ liệu không chuẩn.

Như vây, trong báo cáo này, chúng em cân nhắc sử dụng khoảng cách Levenshtein, là một phép đo độ tương đồng giữa hai chuỗi, cụ thể là dữ liệu thu thập được lần lượt với các bản dịch từ hai từ điển vừa nêu trên. Ngoài ra, do số lượng ký tự ở các câu là khác nhau, điều này làm khó khăn so sánh độ tương đồng giữa các câu. Vì thế, chúng em đề xuất chuyển đổi khoảng cách Levenshtein vừa tính được trên các câu thành tỉ lệ phần trăm để có cái nhìn tổng quan hơn về dữ liệu.

Để so khớp và đánh giá mức độ tin cậy của dữ liệu thu thập được chúng em chia thành 4 cấp độ như sau:

- Với các câu có độ tương đồng giữa dữ liệu thu thập được với các bản dịch lớn hơn hoặc bằng 75%, chúng được đánh giá là đáng tin cậy (Reliable)

- Với các câu có độ tương đồng trong khoảng từ 60% đến 75%, chúng được đánh giá là có thể tin cậy được (Trustable)
- Với các câu có độ tương đồng trong khoảng từ 50% đến 60%, chúng được đánh giá là cần xác minh lại (Needs_verification)
- Với các câu có độ tương đồng còn lại, chúng được đánh giá là không biết (Unknown)

3.7. Lưu trữ nội dung ngôn ngữ đã xử lý dưới dạng Excel:

Để lưu trữ nội dung ngôn ngữ đã xử lý, chúng em dùng thư viện openpyxl. Đây là một thư viện Python mạnh mẽ và dễ sử dụng khi việc làm việc với các tệp Excel (.xlsx). Nó cung cấp các tính năng cho phép đọc, ghi và xử lý dữ liệu trong các tệp Excel một cách linh hoạt và tiện lợi. Ngoài các tính năng cơ bản như đọc và ghi dữ liệu, Openpyxl còn hỗ trợ tính năng mạnh mẽ khác giúp tối ưu hóa công việc với tệp Excel như cho phép thêm và xóa các ô, hàng và cột trong tệp Excel một cách dễ dàng, giúp điều chỉnh cấu trúc bảng một cách linh hoạt. Openpyxl cũng hỗ trợ đa luồng (multithreading) và tiến trình (multiprocessing), cho phép xử lý đồng thời nhiều tệp Excel mà không gây tắc nghẽn. Điều này rất hữu ích khi chúng ta có các tệp lớn hoặc muốn tăng hiệu suất xử lý dữ liệu về sau.

Chúng em tiến hành tạo các tệp excel tương ứng cho từng tác phẩm văn học được thu thập (ở đây là 35 tác phẩm văn học kinh điển được thu thập từ trang web https://www.scribd.com/lists/21643875/Th%C6%A1-V%C4%83n-Han-Nom). Trong đó, mỗi tệp excel có 6 cột bao gồm: Top, Left, Start_top, Text, Translated by Document, Translated by Dictionary. Lưu trữ các cặp câu song ngữ và thông tin liên quan vào một tệp Excel để dễ dàng quản lý và sử dụng cho các nghiên cứu và ứng dụng tiếp theo.

3.7.1 Tách dữ liệu thu thập từ file Text

Sử dụng tệp txt đã lấy được từ các bước trên, chia các thuộc tính đang được cách nhau bằng dấu ";" sau đó truyền vào từng cột tương ứng trong Excel.

3.7.2 Xử lý dữ liệu chữ lấy được

Riêng với dữ liệu text chúng em sẽ chia chung thành 2 kiểu cấu trúc sau:

- Diễn giải Nôm tự trên cùng một dòng với chữ Quốc ngữ (dòng sẽ bao gồm Nôm tự và chữ Quốc ngữ)
- Diễn giải Nôm tự ở các dòng khác nhau (một dòng sẽ bao gồm hoàn toàn chữ Nôm hoặc có rất ít chữ Quốc ngữ, và dòng Nôm tự sẽ được giải nghĩa bằng chữ Quốc ngữ nhưng ở dòng khác trong văn bản)

Chúng em sẽ tiến hành xử lý 2 kiểu cấu trúc theo 2 cách:

- Nếu cùng một dòng chứa đồng thời Nôm tự và chữ Quốc ngữ. Sử dụng HVDict để xác định xem kí tự đó có là Nôm tự hay không, sau đó tách riêng chữ Nôm và chữ quốc ngữ ra. Tiếp tục mang chữ Nôm tự đi dịch bằng cách gọi API để dịch chữ Nôm và so sánh kết quả với chữ Quốc ngữ cùng dòng
- Nếu việc diễn giải chữ Nôm sẽ ở một dòng khác (dạng thơ), tức là trong một dòng sẽ chứa phần lớn là chữ Nôm. Sử dụng từ điển Nom Transliteration HCMUS để dịch và lấy kết quả so khớp với dòng chữ quốc ngữ có trong bài. Đôi khi sẽ có những trường hợp dòng Nôm tự được giải nghĩa bởi nhiều dòng chữ quốc ngữ ta sẽ linh hoạt so sánh chữ quốc ngữ có nằm trong kết quả lấy được việc dịch chữ Nôm hay không, sau đó ghép các dòng lại với nhau (lấy kết quả dịch làm mốc)

Lưu ý: Việc nhận định một dòng sẽ sử dụng cách xử lý 1 hay 2 sẽ dựa vào tổng số kí tự Nôm chứa trong dòng đó với tổng số từ quốc ngữ

3.7.3 Loại bỏ các kí tự vô nghĩa

Sử dụng các hàm để xác định các từ vô nghĩa (không thể in ra được màn hình).

3.7.4 So khớp bản dịch bởi các API từ điển với bản dịch có trong dữ liệu

Sử dụng các hàm nói trên để so độ tương đồng giữa và ở từng mức độ tương đồng sẽ lưu bản dịch của trang web với từng màu khác nhau tương ứng với từng mức độ tương đồng với bản dịch bởi API từ điển (trong báo cáo này, chúng em lấy bản dịch bởi Nom Transliteration HCMUS làm mốc). Việc này giúp tăng độ tin cậy đối với dữ liệu mà chúng em thu thập được. Ngoài ra, với đề xuất này, chúng em hi vọng có thể giảm bớt thời gian kiểm tra lại dữ liệu thu được.

3.7.5 Lưu vào tệp Excel

Sau khi xử lý dữ liệu qua các bước sẽ lưu lại bằng tệp Excel với cấu trúc như sau:

- *Top:* Thông số top được lấy trong style của kí tự cuối cùng trong hàng
- *Left:* Thông số left được lấy trong style của kí tự cuối cùng trong hàng
- *start_top:* Thông số được lấy làm chuẩn để mốc xác định các kí tự cùng hàng
- *Text:* Nôm tự thu được sau khi tách

- *Translated by Document:* Chữ Quốc ngữ diễn giải Nôm tự có trong trang web đang thu thập cùng với các màu sắc tương ứng với độ tương đồng giữa dữ liệu thu thập được so với bản dịch bằng từ điển.
- *Translated by Dictionary:* Nội dụng được diễn giải Nôm tự dựa theo bản dịch từ tool Nom Transliteration HCMUS

CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

4.1. Mô tả nguồn dữ liệu

Trong khuôn khổ báo cáo này, chúng em thu thập dữ liệu từ trang web Scribd. Đây là một nền tảng trực tuyến phổ biến, cho phép người dùng chia sẻ và truy cập hàng triệu tài liệu, sách, bài viết, tạp chí và nhiều tài liệu văn bản khác từ khắp nơi trên thế giới. Một trong những điểm hấp dẫn của Scribd là sự đa dạng của nội dung mà nó cung cấp. Người dùng có thể tìm thấy và đọc các tác phẩm từ nhiều lĩnh vực khác nhau như văn học, khoa học, lịch sử, công nghệ, nghệ thuật, thể thao và nhiều chủ đề khác.

Bên cạnh đó, đáng chú ý là trang web này hiện tại đang cung cấp một khối lượng đáng kể các văn bản song ngữ chữ Nôm và chữ Quốc ngữ. Cụ thể, cho đến thời điểm nghiên cứu báo cáo này của chúng em, trang web cung cấp 35 tác phẩm văn học kinh điển với khoảng 3585 trang tại đường dẫn https://www.scribd.com/lists/21643875/Th%C6%A1-V%C4%83n-Han-Nom. Điều này mở ra một cơ hội vô cùng to lớn trong nỗ lực thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ.

4.2. Kết quả thực nghiệm

4.2.1. Thu thập các đường dẫn liên quan

Đối với tác vụ thu thập các đường dẫn liên quan từ trang web đầu vào, chúng em đã hoàn thành tốt tác vụ này khi có thể thu thập hoàn toàn các đường dẫn liên quan, cụ thể là thu thập đủ 35 đường dẫn dẫn đến 35 tác phẩm văn học có trong đường dẫn đầu vào. Các đường dẫn này bao gồm:

1. https://www.scribd.com/document/354356904/Thập-Giới-Co-Hồn-Quốc-Ngữ-Văn-Le-Thanh-Tong
2. https://www.scribd.com/document/351003106/Thien-Nam-Ngữ-Lục-Thơ-Nom
3. https://www.scribd.com/document/354358928/Cung-Oan-Ngam-Khuc-Nguyễn-Gia-Thiều-Le-Văn-Đặng-Diễn-Nom
4. https://www.scribd.com/document/354358921/Chinh-Phụ-Ngam-Diễn-Ca-Nguyễn-Văn-Xuan
5. https://www.scribd.com/document/354359286/Chinh-Phụ-Ngam-Han-Việt

6. https://www.scribd.com/document/351002527/Ma-Thanh-Tan-Truyện-Lý-Hồng-Phượng
7. https://www.scribd.com/document/354356908/Bạch-Van-Am-Quốc-Ngữ-Nguyễn-Bỉnh-Khiem
8. https://www.scribd.com/document/354356906/Bai-Tựa-Đại-Nam-Quốc-Ngữ
9. https://www.scribd.com/document/354361314/Minh-Tam-Bửu-Giam-Diễn-Nom
10. https://www.scribd.com/document/354361116/Nữ-Tử-Tu-Tri-Diễn-Nom
11. https://www.scribd.com/document/354361115/Mộng-Lien-Đinh-Thi-Thảo-Nguyễn-Gia-Tuyển
12. https://www.scribd.com/document/354359991/Giao-Tử-Phu-Mạc-Đĩnh-Chi
13. https://www.scribd.com/document/354359995/Hồ-Xuan-Hương-Thi-Tập-Le-Văn-Đặng-Diễn-Nom
14. https://www.scribd.com/document/354358926/Cac-Bai-Tuyệt-Cu-Trong-Quốc-Am-Thi-Tập-Nguyễn-Đinh-Hoa-Le-Văn-Đặng
15. https://www.scribd.com/document/354358920/Ba-Huyện-Thanh-Quan-Lam-Trau-Diễn-Nom
16. https://www.scribd.com/document/354358923/Bai-Hat-Đĩ-Diễn-Nom
17. https://www.scribd.com/document/354359285/Đi-Chợ-Tinh-Tiền-Diễn-Nom
18. https://www.scribd.com/document/354359289/Di-Chuc-Văn-Nguyễn-Khuyến
19. https://www.scribd.com/document/352233420/1958-Giai-Nhan-Kỳ-Ngộ-Phan-Chau-Trinh
20. https://www.scribd.com/document/365582497/1547-Truyền-Kỳ-Mạn-Lục-Nguyễn-Dữ
21. https://www.scribd.com/document/365802429/1950-Cung-Oan-Ngam-Khuc-Dẫn-Giải-Va-Chu-Thich-Ton-Thất-Lương
22. https://www.scribd.com/document/365802849/1952-Gia-Huấn-Ca-Nguyễn-Trai-Đinh-Gia-Thuyết-Chu-Thich
23. https://www.scribd.com/document/350234561/1898-Quảng-Tập-Viem-Văn-Edmond-Nordemann
24. https://www.scribd.com/document/368541567/Tổng-Tập-Văn-Học-Nom-Thơ-Nom-Han-Luật-Nguyễn-Ta-Nhi
25. https://www.scribd.com/document/352647408/Le-Triều-Nguyễn-Tướng-Cong-Gia-Huấn-Ca-Maurice-Durand

26. https://www.scribd.com/document/352647934/Cung-Oan-Thi-Maurice-Durand
27. https://www.scribd.com/document/352647926/Thien-Nam-Toan-Quốc-Diễn-Am-Maurice-Durand
28. https://www.scribd.com/document/352647937/Truyện-Vua-Le-Thai-Tổ-Maurice-Durand
29. https://www.scribd.com/document/352647415/Đại-Nam-Quốc-Tuý-Maurice-Durand
30. https://www.scribd.com/document/352648389/Thanh-Tổ-Kệ-Diễn-Quốc-Am-Maurice-Durand
31. https://www.scribd.com/document/352649175/Văn-Tế-Tướng-Sĩ-Trận-Vong-Maurice-Durand
32. https://www.scribd.com/document/352649182/Bạch-Van-Thi-Tập-Maurice-Durand
33. https://www.scribd.com/document/352649184/Phương-Hoa-Tan-Truyện-Maurice-Durand
34. https://www.scribd.com/document/329053271/1958-Thien-Nam-Ngữ-Lục-Tập-1-Nguyễn-Lương-Ngọc-Đinh-Gia-Khanh
35. https://www.scribd.com/document/352647412/Phật-Thuyết-Đại-Thanh-Mạt-Kiếp-Chan-Kinh-Maurice-Durand

4.2.2. Tải nội dung các tệp HTML

Về cơ bản, tệp HTML là cốt lõi của bất kỳ trang web nào, xác định cấu trúc và nội dung của trang, đồng thời cung cấp khả năng tương tác và tích hợp với các ngôn ngữ và công nghệ khác để tạo ra trải nghiệm web đa dạng và phong phú. Với nỗ lực của mình, chúng em đã tải về được 35 tệp HTML từ tất cả 35 đường dẫn liên quan đã thu thập được ở bước trên. chứa nội dung trang web. Sau đó, chúng em tiến hành kiểm tra và đánh giá nội dung hữu ích, nghĩa là nội dung tệp cần có chứa chữ Nôm và chữ Quốc ngữ cần thu thập, trong 35 tệp HTML vừa tải được, có 15 tệp có chứa nội dung hữu ích. Các tệp còn lại không chứa nội dung cần thiết một phần vì trang web yêu cầu đăng nhập, một phần vì chế độ bảo mật. Cụ thể, trang web yêu cầu đăng nhập để truy cập nội dung hoặc sử dụng API để hiển thị dữ liệu. Ngoài ra, một số đường dẫn chứa hình ảnh cũng là một trong những nguyên do khiến tệp html tải về không chứa nội dung chữ Nôm và chữ Quốc ngữ, mặc dù phần hiển thị trên trang web đó là đầy đủ.

4.2.3. Thu thập nội dung trong các tag và lưu thành tệp txt

Các nội dung trong các thẻ của trang web sẽ được lưu lại thành các thuộc tính được viết cách nhau bằng dấu ";" và lưu thành một hàng trong tệp txt. Sau bước này, chúng em thu thập được tổng cộng có 15 tệp chứa nội dung chữ Nôm và chữ Quốc ngữ như mong đợi bao gồm:

1. Bà Huyện Thanh Quan Làm Trâu Diễn Nôm
2. Bạch Vân Am Quốc Ngữ Nguyễn Bỉnh Khiêm
3. Bài Tựa Đại Nam Quốc Ngữ
4. Các Bài Tuyệt Cú Trong Quốc Am Thi Tập Nguyễn Đình Hòa Lê Văn Đặng
5. Chinh Phụ Ngâm Diễn Ca Nguyễn Văn Xuân
6. Chinh Phụ Ngâm Hán Việt
7. Cung Oán Ngâm Khúc Nguyễn Gia Thiều Lê Văn Đặng Diễn Nôm
8. Đi Chợ Tính Tiền Diễn Nôm
9. Di Chúc Văn Nguyễn Khuyến
10. Giao Tử Phu Mạc Đĩnh Chi
11. Hồ Xuân Hương Thi Tập Lê Văn Đặng Diễn Nôm
12. Minh Tâm Bửu Giám Diễn Nôm
13. Mộng Liên Đình Thi Thảo Nguyễn Gia Tuyền
14. Nử Tử Tu Tri Diễn Nôm
15. Thập Giới Cô Hồn Quốc Ngữ Văn Lê Thánh Tông

```
662;1326;0;Chinh Phụ Ngâm Diễn Ca 婦吟 歌
795;1303;662;i Điễễn âm段氏點音
900;None;795;Nguồn: Chinh Phuêt 1950, tr. 143-167.
994;663;900;Phụ Chú:
1091;None;994;ả Nguyễn Văn Xuân (1921-2007), một trong những tác phẩm ra đời để lại
1185;663;1091;cho ông
1279;663;1185;ều kỷ niệất, đó là trườợp của cuốn biên khảo Chinh phụ  ngâm diễ n
1385;663;1291.25;âm tân khúc.
1465;663;1385;Lng sách của một gia đình nghèo, có xuất
1558;None;1465;xứ từ
1650;663;1558;gia đình bà Chúa Nhấhành Thái; lễch chữ Nôm có
1745;663;1650.5;cuốn
1852;663;1745;Chinh phụ  ngâm diễ n âm tân khúc, có tựa của nhà xuất bản (năm Gia Long thứ
1931;663;1852;14, 1815),
2025;663;1931;nguyên tự (tựa) của dịch giả. Ông cho r ằng cuốộc tếđộc lập
2117;663;2025;mà cũng
2211;663;2117;là một bốốn Chinh phụ  ngâm b ị  khảo của hoả Hoàng Xuân Hãn
2304;663;2224.125;)
2396;663;2304;– là một kỳthư về ngành nghiên cứu văn học theo văn bản học.
2490;663;2396.96875;Và ông đã xác quyết dịch giảđích thực là Phan Huy Ích và tên tác phẩm là
2597;663;2490.9375;Chinh phụ  diễ n âm tân khúc chứ không phải là Chinh phụ  ngâm...
2676;640;2597; Điều này đã gây ra sự tranh cãi không ít trên văn đàn miền Nam, thời trước 1975.
2768;663;2676.875;Trong thập niên 1960 có lúc anh đượ c mờ i dạy giờ  môn chữ  Nôm tại Đại Học Huế
2870;663;2770.98828125;và phát hiện trong một tủsách tư nhân bản văn chữ  Nôm của Phan Huy Ích dịch
2972;663;2872.953125;Chinh Phụ Ngâm của Đặng Trần Côn; bản dịch này trước kia được xem như của
3076;663;2974.90625;bà Đoàn ThịĐiểm. Khoảng 1920 nhiều học giảđã cho rằng bản dịch là của
3178;663;3076;Phan Huy Ích, nhưng chỉcó văn bản quốc ngữ . Phát hiện của Nguyễn Văn Xuân
3280;663;3178;đã giúp thêm bằng chứ ng cho học giảHoàng Xuân Hãn, ngườ i đãẳng định,
3379;663;3280;từ  khoảng 1950, dịch phẩm Chinh Phụ Ngâm thông dụng hiện nay là của Phan
3480;663;3381.9765625;Huy Ích.
3705;747;3480;Tập 征婦吟備緑Chinh Phụ Ngâm Bị Lục (1902)
3834;750;3706.9375;隆和?藏板jêu Tàng Bản
3951;750;3840.25;có ghi 3 dòng, phía sau trang bìa:
4070;835;3951;青池仁睦鄧陳先生琨著
4204;835;4070;文江中富段夫人點演音
4338;835;4204;神溪桐峯承書
4454;729;4338;Thanh trì nhân mục Đặng Trần tiên sanh Côn trứ
4559;729;4456.625;Văn giang trung phú Đoàn phu nhân Điểm diễn âm
4658;729;4559;Thần khê Đồng Phong Thừa thư .
4861;663;4659.5;1
4981;663;4861;??坦涇干?Thuở  trời đất nổi cơn gió bụi,
5113;663;4986.4375;客?紅?Khách má hồng nhiều nỗi truân chiên.
547;663;5119.4921875;撐箕瀋瀋層?Xanh kia thăm thẳm tầng trên,
680;663;552.375;為埃??朱?餒尼Vì ai gây dự ng cho nên nỗi này ?
797;663;686.71875;5
917;663;797;?長城??月Trống Trườ ng Thành lung lay bóng nguyệt,
1051;663;921.875;?甘泉?瞲式?Khói Cam Tuyền mờ  mịt thứ c mây.
1185;663;1056.71875;?吝鎌寶??Chín lần gươm báu chống tay,
1317;663;1189.875;姅?傳檄定?出征ửa đêm truyền hịch định ngày xuất chinh.
```

Hình 4.2.3.1. *Hình ảnh tệp txt sau khi được lưu từ tác phẩm Chinh Phụ Ngâm Diễn Ca - Nguyễn Văn Xuân*

```
531;928;0;ẠCH VÂN AM QUỐC NGỮ
649;1724;533.625;NGUYỄN BỈNH KHIÊM
768;663;650.5;大正六年 未科進士及第第 甲第 名程國公阮秉謙詩集公待少時
909;663;768;本縣知縣遊行見公出對云六七功不如子程狀元對曰二千石若
1049;663;909;公及長名望箸居官致仕作白云庵詩集?興詩百首用國音長短律
1168;928;1049;[1]
1300;928;1174.5;吝矧?戈?戈Lẩn thẩn ngày qua tháng qua,
1434;1032;1308.234375;Một phen xuân tớ i, một phen già.
1567;928;1441.71875;愛憂???印渃Ái ưu vặc vặc trăng in nướ c,
1702;928;1575.1875;名利淩淩???lợ i lâng lâng gió thổi hoa.
1836;928;1709.21875;案Án sách hãy còn án sách cũ,
1969;928;1839.5;渃?伴貝渃?? Nướ c non bạn với nướ c non nhà.
2089;928;1976.6171875;[2]
2222;928;2095.5;?整篆庫来台Giàu chễnh chện khó lai thai,
2355;928;2228.46875;運轉流通??埃ận chuyển lưu thông há của ai.
2489;928;2362.21875;Vũng nọ ghe khi làm bãi cát,
2631;928;2494.75;耒?固課?丸thuở  lút hòn dai.
2770;928;2638.25;坤頑買別升時降mớ i biết thăng thờ i giáng
2905;928;2777.8125;大ại dột nào hay tiểu có đài.
3038;928;2912.1875;匏屈包饒時吏撮Đã khuất bao nhiêu thì lại duỗi,
3173;928;3046.40625;道?差Đạo tr ờ i lồng lộng chẳng hề sai.
3298;928;3180.9609375;[3]
3432;928;3304.5;???????Giàu ba bữa, khó hai niêu, ?
3565;928;3437.875;安分時欣?每調?An phận thì hơn hết mọi điều,
3698;928;3572.4375;?旺茶梅唏???Khát uống trà mai hơi ngút ngút, ?
3832;928;3703.625;焠掑軒月?囂囂?Sốt k ề hiên nguyệt gió hiu hiu .
3973;928;3839.7265625;幀?Giang san tám bức là tranh vẻ, ?
4111;928;3980.4375;???務意錦繞?Hoa cỏ tư mùa ấy gấm thêu,
4245;928;4118.86328125;?且?????式?Thong thả hôm khuya nằm sớ m thức, ?
4378;1032;4252.865234375;萬匏隊德?堯Muôn vàn đã đội đức tr ờ i Nghiêu.
4498;928;4385.84375;[4]
4632;928;4504.5;????Giàu cơm thịt, khó cơm rau,
4765;928;4635.0;安分羅仙路沛求phận là tiên, lọ phải cầu.
4899;928;4772.796875;?旺茶椿唏??ớ m uống trà thông hơi ngút ngút,
5033;928;4904.75;?掑軒月?楼楼k ề hiên nguyệt tỏ làu làu.
5175;928;5039.640625;?Vun thông tướ i cúc ba thằng mọn,
547;928;5182.6875;?焐?ử lửa hâm trà một mụ hầu.
667;928;554.90234375;[5]
806;928;673.5;???重?埃認Giàu sang ngườ i tr ọng khó ai nhìn,
940;928;807.625;汆胞腰為几呂?ấy dạ yêu vì k ẻ lỡ  nhèn?
1073;928;947.796875;課?酉嘲嘲拱朗ở  khó, dẫu chào, chào cũng lảng,
1208;928;1080.1875;欺時淆chẳng hỏi, hỏi thì quen.
1347;928;1215.59375;hiềm dan díu điều làm bạn,
```

Hình 4.2.3.2. *Hình ảnh tệp txt sau khi được lưu từ tác phẩm Bạch Vân Am Quốc Ngữ - Nguyễn Bỉnh Khiêm*

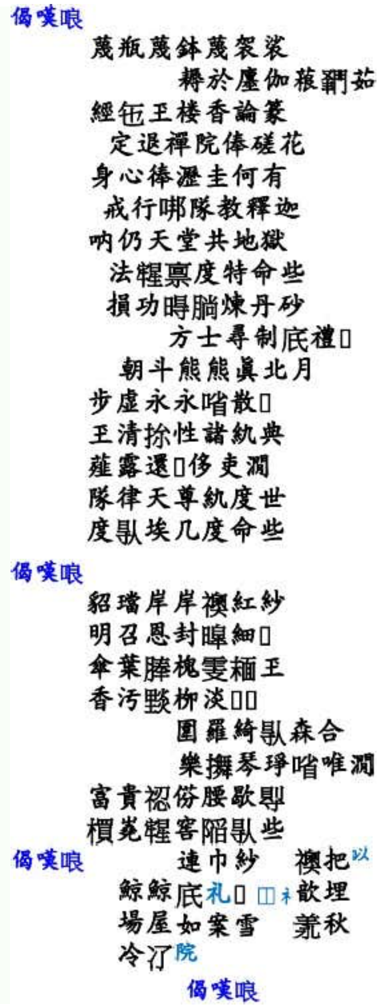
4.2.4. Thu thập hình ảnh

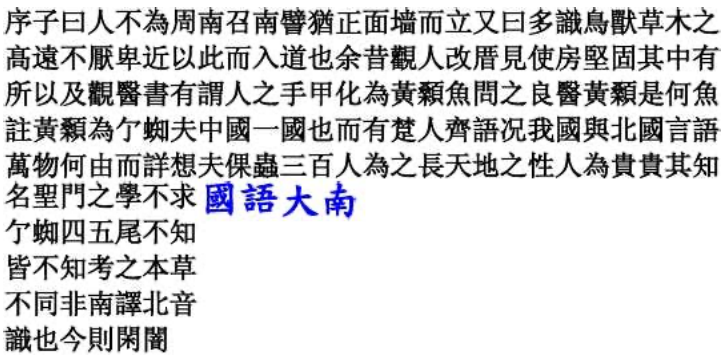
Tính đến thời điểm hiện tại, kết quả của nghiên cứu đã ghi nhận thành công trong việc thu thập tổng cộng 56 hình ảnh có định dạng .jpg từ những trang web có liên quan. Đây là một thành tựu quan trọng, vì những hình ảnh này đóng vai trò là dữ liệu đầu vào then chốt cho những tác vụ thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ từ hình ảnh. Đồng thời, những hình ảnh này sẽ chịu trách nhiệm chủ yếu trong việc hỗ trợ và cung cấp nguồn dữ liệu phong phú cho việc xây dựng các mô hình thị giác máy tính trong tương lai.

馬成新傳造錄演歌詩
MÃ THÀNH TÂN TRUYỆN TẠO LỤC DIỄN CA THI
Lý Hồng Phượng phiên âm

說断課希洪王
Thuyết đoạn thuở vua Hồng Vương
治位天下罘方順和
Trị vì thiên hạ bốn phương thuận hoà
人民楽業歐歌
Nhân dân lạc nghiệp âu ca
㐌㐌 通且茹茹罒咹
Nơi nơi thong thả, nhà nhà làm ăn
固得戶馬䏧能
Có người họ Mã tên Năng,
忠君愛国职升冲䏧
Trung quân ái quốc chức thăng trong đời.
福汝享祿茹歪
Phước nhờ hưởng lộc nhà trời,
添生男子迮䏧坤當
Thêm sinh nam tử dưới đời khôn đương.
㐌咍經史文章
Đã hay kinh sử văn chương
武㞾揀習覔唐老通
Võ thời đánh tập trăm đường làu thông.

1





Hình 4.2.4.1 *Một số hình ảnh chứa chữ Nôm và chữ Quốc ngữ thu thập được từ các đường dẫn liên quan*

4.2.5. Xử lý ngôn ngữ và lưu thành tệp excel

Top	Left	Start_top	Text	Translated by Document	Translated by Dictionary
917	663	797	長城月	trống thành nguyệt, --in--[Original] Trống Trườ ng Thành lung lay bóng nguyệt,	trường thành nguyệt
1051	663	921.875	甘泉礎式	cam tuyền mịt thứ --in--[Original] Khói Cam Tuyền mờ mịt thứ c mấy.	cam tuyền mịt thức
1185	663	1056.71875	杏麟實	lần gươm báu --in--[Original] Chín lần gươm báu chống tay,	lần gươm báu
1317	663	1189.875	抨傳檄定出征	truyền hịch định xuất chinh. --in--[Original] ửa đêm truyền hịch định ngày xuất chinh.	nửa truyền hịch định xuất chinh
1434	663	1329.796875	9		
1575	663	1434	泝清平	nướ c thanh bình --in--[Original] Nướ c thanh bình ba trăm năm cũ,	nước thanh bình
1735	663	1580.6875	襖戎挥官武自尼	áo nhung quan vũ tứ --in--[Original] Áo nhung trao quan vũ tứ nay.	áo nhung huy quan vũ tứ này
1868	663	1738.75	使塘	sứ --in--[Original] Sứ trời sớm giục đườ ng mây,	sứ đường
2003	663	1874.74609375	詫	sá --in--[Original] Phép công là trọng, niềm tây sá nào.	sá
2119	663	2008.75	13		
2238	663	2119	塘扔弓箭	đeo cung tiễn, --in--[Original] Đườ ng giong ruổi lưng đeo cung tiễn,	đường đeo cung tiễn
2372	663	2243.4375	餞迭鈑妻琴	tiễn đưa bận thê noa. --in--[Original] Buổi tiễn đưa lòng bận thê noa.	tiễn đưa bịn thê noa
2506	663	2377.75	旗嗜脸脸	cờ tiếng xa --in--[Original] Bóng cờ tiếng trống xa xa,	cờ tiếng xa xa
2639	None	2511.4375	愁醝愁房	sầu oán phòng. --in--[Original] Sầu lên ngọn ải, oán ra cử a phòng.	sầu ải oán phòng
2755	663	2645.6875	17		
2874	663	2755	払豪傑	chàng hào kiệt, --in--[Original] Chàng tuổi trẻ vốn dòng hào kiệt,	chàng hào kiệt
3008	663	2880.7109375	攝筆硯踱役刀弓	xếp bút nghiên theo việc đao cung. --in--[Original] Xếp bút nghiên theo việc đao cung.	nếp bút nghiên theo việc đao cung
3142	663	3013.9375	城連儻献庭	thành liền mong hiến bệ --in--[Original] Thành liền mong hiến bệ rồng,	thành liền mong hiến bệ
3276	663	3148.703125	賊	giặc --in--[Original] Thước gươm đã quyết chẳng dung giặc trời.	giặc
3392	663	3288.40625	21		
3511	663	3392	志膠馭	chí da ngư --in--[Original] Chí làm trai dặm nghìn da ngư a,	chí da ngựa
3667	663	3516.8125	毛	mao. --in--[Original] Gieo Thái Sơn nhẹ nữ a/nữ a hồng mao.	mao
3826	663	3673.59375	啫茄扔幅袍	giã nhà đeo bào, --in--[Original] Giã nhà đeo bứ c chiến bào,	giã nhà đeo bức bào
3962	663	3831.6875	檐棶溝秋	roi cầu thu. --in--[Original] Thét roi cầu Vị, ào ào gió thu.	roi cầu vị thu
4078	663	3966.875	25		
4196	663	4078	外頭穩泝如	ngoài đầu cầu nước như --in--[Original] Ngoài đầu cầu nước trong như lọc,	ngòi đầu cầu nước như
4330	663	4202.40625	塘辺棶莽	bên cầu còn --in--[Original] Đườ ng bên cầu cỏ mọc còn non.	đường bên cầu còn
4463	663	4336.6171875	迭払弋弋	chàng đặc đặc --in--[Original] Đưa chàng lòng đặc đặc buồn,	đưa chàng đằng đặc
4597	663	4469.359375	步坤平馭水坤平船	bộ khôn khôn bằng bằng ngư thuyền. --in--[Original] Bộ khôn bằng ngư a, thủy khôn bằng thuyền.	bộ khôn bằng ngựa thuỷ khôn bằng thuyền
4714	663	4609.91015625	29		
4833	663	4714	Nướ c có chảy mà phiền chẳng rã,		
4967	663	4839.59375	固麻匙	mà dạ --in--[Original] Cóc ở thơm mà dạ chẳng khuây.	có mà dạ
5100	663	4973.375	叼叼吏扮	dặn dặn, lại cầm --in--[Original] Dặn rồi dặn, lại cầm tay,	dặn dặn lại cầm
548	663	5106.8515625	膠沒跑跑史	một lại --in--[Original] Bước đi một bước c giây giây lại dừ ng	đá một tri tri lại
664	663	553.25	33		

Hình 4.2.5.1 *Tệp excel chứa dữ liệu thu thập được từ tác phẩm Chinh Phụ Ngâm Diễn Ca - Nguyễn Văn Xuân*

Như vậy sau khi thực hiện các thao tác vừa mô tả trên, chúng em đã thành công thu thập được các tệp excel với tổng cộng 1455 dòng bao gồm 2 cột, một cột chữ Nôm và một cột chữ Quốc ngữ cùng các thông số khác. Trong 1455 dòng dữ liệu này, chúng em đã thu thập được khoảng 8617 Nôm tự. Đây là một con số đáng kể, đóng vai trò vô cùng quan trọng đối với công tác thu thập dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ vốn đang trong tình trạng khan hiếm dữ liệu như hiện nay.

Tên dữ liệu	Số lượng
Đường dẫn chứa dữ liệu song ngữ chữ Nôm và chữ Quốc ngữ	35
Hình ảnh	56
Tệp HTML	35
Tệp văn bản (txt)	35

Số dòng dữ liệu song ngữ thu thập được	Khoảng 1455 dòng
Số Nôm tự	Khoảng 8617 Nôm tự

Bảng 4.2 *Bảng kết quả thực nghiệm*

CHƯƠNG 5: THẢO LUẬN

5.1. Lợi ích của phương pháp đề xuất

Phương pháp đề xuất khi thu thập dữ liệu song ngữ (bao gồm chữ Nôm và chữ Quốc ngữ) mang lại nhiều lợi ích đáng kể. Đầu tiên, nó *đảm bảo tính chính xác và đầy đủ của dữ liệu* thông qua việc tự động thu thập từ các trang web, giảm thiểu sai sót trong quá trình sao chép thủ công. Thứ hai, phương pháp này *tiết kiệm thời gian và công sức* cho nhà nghiên cứu, giúp họ tập trung vào phân tích dữ liệu và nâng cao chất lượng nghiên cứu. Thứ ba, phương pháp *cho phép thu thập dữ liệu đa ngôn ngữ*, bao gồm cả chữ Nôm và chữ Quốc ngữ, mở rộng phạm vi nghiên cứu và đáp ứng nhu cầu đa dạng của dự án. Thứ tư, việc sử dụng phương pháp này *đảm bảo tính đồng nhất của dữ liệu* thông qua việc chuyển đổi dữ liệu thành định dạng chuẩn, tạo điều kiện thuận lợi cho việc tiếp cận và phân tích dữ liệu trong tương lai. Thứ năm, phương pháp đề xuất *linh hoạt và có thể tùy chỉnh* cho các yêu cầu nghiên cứu cụ thể. Cuối cùng, phương pháp này *cho phép tổng hợp lượng lớn dữ liệu* từ nhiều trang web khác nhau, cung cấp nguồn tài nguyên đáng kể cho quá trình nghiên cứu và phân tích dữ liệu song ngữ. Tóm lại, phương pháp đề xuất này mang lại nhiều lợi ích to lớn, từ đảm bảo tính chính xác và đầy đủ của dữ liệu đến khả năng tiết kiệm thời gian và tăng cường tính linh hoạt, giúp nâng cao chất lượng và hiệu quả của quá trình nghiên cứu và phân tích dữ liệu song ngữ.

5.2. Hạn chế của đề tài

5.2.1. Khó khăn đến từ dữ liệu

- *Về mặt bản quyền:* Trang web cung cấp dịch vụ đọc trực tuyến, nên cần yêu cầu đăng nhập tài khoản để đọc được bản đầy đủ
- *Về cấu trúc trang web:* Trang web có cấu trúc khá phức tạp, các thông tin văn bản được tách riêng ra thành từng kí tự, từng thẻ tag khác nhau không theo một cấu trúc nào
- *Về cấu trúc trình bày dữ liệu:* Cấu trúc trình bày các thông tin ở trang web không theo một format nhất định, nhất là các Nôm tự được viết theo cả chiều ngang và chiều dọc, cùng dòng cũng những khác dòng, diễn giải từng kí tự hay là giải thích theo từng dòng. Vậy nên rất khó có thể tìm ra một quy luật chung cho việc crawl dữ liệu để phù hợp với tất cả các kiểu cấu trúc văn bản trong dữ liệu nguồn này.

- *Về Nôm tự:* Chính chữ Nôm cũng là một trong nhưng khó khăn khi crawl dữ liệu vì đôi khi có những kí tự chữ Nôm không thể biểu diễn được, các kí tự không in ra được sẽ được mặc định thành các kí tự vô nghĩa, gây cản trở việc thu thập dữ liệu

5.2.2. Hạn chế của phương pháp

- Phương vấn còn chưa thể tối ưu cho tất cả các trường hợp, định dạng của các trang web khác nhau
- Chưa thể lấy được hết những kí tự không xác định hoặc không thể in ra màn hình
- Thời gian thu thập dữ liệu chưa được tối ưu

5.3. Hướng phát triển

Cho đến thời điểm hiện tại, chúng em đã tải được tổng cộng 56 ảnh định dạng .jpg từ các trang web liên quan. Những dữ liệu hình ảnh này sẽ đóng vai trò quan trọng trong quá trình phát triển tiếp theo của dự án này. Chúng em dự định sử dụng những hình ảnh này để trích xuất văn bản bằng các mô hình thị giác máy tính.

Kế hoạch tiếp theo của chúng tôi là áp dụng các thuật toán nhận dạng văn bản và trích xuất thông tin từ những hình ảnh này. Quá trình này có thể bao gồm việc sử dụng kỹ thuật OCR (Optical Character Recognition) để chuyển đổi dữ liệu hình ảnh sang dạng văn bản có thể đọc được. Sau đó, chúng tôi sẽ xử lý và làm sạch dữ liệu để đảm bảo tính chính xác và độ tin cậy.

Sau khi trích xuất được dữ liệu văn bản từ hình ảnh, chúng tôi sẽ tiếp tục nghiên cứu và phát triển các mô hình thị giác máy tính phức tạp hơn. Điều này nhằm nâng cao hiệu suất và độ chính xác của việc trích xuất thông tin từ những hình ảnh phức tạp trong tương lai.

Mục tiêu cuối cùng của chúng tôi là tạo ra một hệ thống trích xuất thông tin tự động, linh hoạt và mạnh mẽ từ những hình ảnh được thu thập từ các trang web. Những thành tựu trong dự án này có thể đóng góp quan trọng vào việc giải quyết các vấn đề thực tiễn và nhu cầu trong lĩnh vực phân tích dữ liệu và tổ chức thông tin.

KẾT LUẬN

Trong quá trình nghiên cứu và thu thập dữ liệu, chúng em đã thành công thu thập tổng cộng 1455 dòng dữ liệu và lưu trữ lại thành các tệp Excel. Dữ liệu bao gồm thông tin quan trọng về chữ Nôm và chữ quốc ngữ cùng với các thông số liên quan. Tổng quan, chúng ta đã thu thập được khoảng trên 8000 chữ Nôm tự và 56 hình ảnh.

Số lượng lớn chữ Nôm thu thập được cung cấp một nguồn tài nguyên quý giá cho nghiên cứu về ngôn ngữ và văn hóa dân tộc. Chúng ta có thể sử dụng dữ liệu này để nghiên cứu và phân tích các đặc điểm ngữ pháp, từ vựng và cấu trúc của chữ Nôm, từ đó hiểu sâu hơn về ngôn ngữ truyền thống của dân tộc Việt Nam.

Sự hiện diện của 56 hình ảnh cũng là một thành tựu quan trọng trong việc nghiên cứu chữ Nôm. Những hình ảnh này có thể là các bản thảo, tài liệu lịch sử hoặc tài liệu nghiên cứu liên quan đến chữ Nôm, có thể cung cấp thông tin thêm về việc sử dụng và phát triển chữ Nôm trong quá khứ.

Tổng kết lại, việc thu thập dữ liệu từ các tệp Excel này đã mở ra những cơ hội mới để tìm hiểu và nghiên cứu về chữ Nôm và ngôn ngữ Việt Nam. Đây là một bước quan trọng trong việc bảo tồn và phát triển di sản văn hóa của dân tộc, đồng thời hỗ trợ trong việc nghiên cứu và khai thác sâu hơn về ngôn ngữ và văn hóa của quốc gia.

Tài liệu tham khảo

[1] Van Phan, Truyen, Bilan Zhu, and Masaki Nakagawa. "Collecting handwritten nom character patterns from historical document pages." 2012 10th IAPR International Workshop on Document Analysis Systems. IEEE, 2012.

[2] Dang, Hoang-Quan, et al. "NomNaOCR: The First Dataset for Optical Character Recognition on Han-Nom Script." 2022 RIVF International Conference on Computing and Communication Technologies (RIVF). IEEE, 2022.