Trường Đại học Bách Khoa Hà Nội

Viện Công nghệ Thông Tin và Truyền Thông

Đồ án Tốt nghiệp Đại học

Hệ thống hỗ trợ học tiếng Nhật trực tuyến thông qua podcast

Sinh viên thực hiện Lưu Phương Trang

Người hướng dẫn PGS.TS. Cao Tuấn Dũng

Hà Nội, 06/2018

Trường Đại học Bách Khoa Hà Nội

Viện Công nghệ Thông Tin và Truyền Thông

Đồ án Tốt nghiệp Đại học

Hệ thống hỗ trợ học tiếng Nhật trực tuyến thông qua podcast

Sinh viên thực hiện Lưu Phương Trang

Người hướng dẫn PGS.TS. Cao Tuấn Dũng

Hà Nội, 06/2018

Lời cam kết

Họ và tên sinh viên: Lưu Phương Trang

Điện thoại liên lạc: 0963 955 195 Email: tranglp.1995@gmail.com

Lớp: IS2 – K58 Hệ đào tạo: Việt Nhật

Em – *Lưu Phương Trang* – cam kết Đồ án Tốt nghiệp (ĐATN) là công trình nghiên cứu của bản thân em dưới sự hướng dẫn của *PGS.TS Cao Tuấn Dũng.* Các kết quả nêu trong ĐATN là trung thực, là thành quả của riêng em, không sao chép theo bất kỳ công trình nào khác. Mọi tham khảo trong ĐATN – bao gồm hình ảnh, bảng biểu, số liệu, và các câu từ trích dẫn – đều được ghi rõ ràng và đầy đủ nguồn gốc trong danh mục tài liệu tham khảo. Em xin hoàn toàn chịu trách nhiệm với dù chỉ một sao chép vi phạm quy chế của nhà trường.

Hà Nội, ngày tháng năm

Tác giả ĐATN

Lưu Phương Trang

Lời cảm ơn

Đi qua những năm tháng Bách Khoa, em mới biết tuổi trẻ đáng trân trọng như thế nào. Trân trọng vì những lúc được thầy cô nhiệt tình chỉ bảo, vì những lúc cùng bạn bè trải qua biết bao niềm vui nỗi buồn, vì những lúc khó khăn tưởng chừng như gục ngã nhưng đều đã vượt qua để rồi trưởng thành hơn rất nhiều.

Cảm ơn Bách Khoa! 5 năm, một quãng thời gian có lẽ chẳng đáng gì so với cả cuộc đời nhưng có thể đã là tất cả của tuổi thanh xuân. Em không thể nhớ rõ Bách Khoa đã cho mình bao nhiêu cũng như mình đã cống hiến cho Bách Khoa được gì, chỉ biết rằng tuổi trẻ có Bách Khoa và chắc chắn em sẽ không bao giờ quên điều đó.

Ai đó đã từng nói "Không ai đơn độc đứng trên đỉnh thành công". Tốt nghiệp ra trường đâu đã phải là thành công nhưng nếu chỉ có một mình, em cũng không thể làm được điều đó. Em xin được gửi lời cảm ơn chân thành nhất đến các thầy cô giáo trong trường Đại học Bách Khoa Hà Nội cùng các thầy cô trong Viện Công nghệ Thông tin và Truyền thông và đặc biệt là thầy giáo PGS.TS Cao Tuấn Dũng đã truyền dạy cho em những kiến thức bổ ích, giúp em xây dựng nền móng vững chắc cho sự phát triển trong sự nghiệp và trong cuộc sống sau này.

Cuối cùng là lời cảm ơn đến các bạn trong HEDSPI K58. Cảm ơn vì đã đi cùng nhau những năm tháng đẹp nhất của cuộc đời. Ai rồi cũng có sự lựa chọn riêng, có lối đi riêng, hy vọng tất cả chúng ta đều thành công và vững bước trên con đường mình đã chọn.

Tóm tắt

Ngày nay, học tiếng Nhật đang là nhu cầu của rất nhiều người, bên cạnh việc học ở trường lớp, học qua sách vở thì học qua mạng đang dần trở nên phổ biến. Sau quá trình tìm hiểu, em đã chọn đề tài xây dựng hệ thống hỗ trợ học tiếng Nhật trực tuyến thông qua podcast. Bởi vì các ứng dụng hay trang Web hiện nay chủ yếu dạy về từ vựng, ngữ pháp nhưng lại chưa chú trọng nhiều vào kỹ năng nghe, trong khi đây lại là kỹ năng vô cùng quan trọng của việc học tiếng Nhật.

Tính năng của podcast chính là các nội dung âm thanh và các video clip được tạo ra bởi cá nhân, tổ chức sau đó đăng lên Internet để người dùng nghe và tải xuống. Hệ thống em xây dựng cung cấp phương pháp học tiếng Nhật thông qua nghe bài học dạng audio, video kết hợp xem nội dung dưới dạng văn bản, có kèm giải thích nghĩa của từ, cụm từ quan trọng. Ứng với mỗi bài học đều có phần bài tập điền từ còn thiếu vào chỗ trống để người dùng hiểu bài hơn. Đồng thời đây cũng là một cách học từ vựng hiệu quả vì các từ còn thiếu là những từ quan trọng và mang nhiều ý nghĩa trong bài. Hệ thống cũng hỗ trợ người dùng quản lý thông tin và kết quả quá trình học, làm bài tập.

Dữ liệu bài học vô cùng phong phú vì được tổng hợp từ nhiều nguồn dữ liệu podcast khác nhau và được cập nhật liên tục. Các dữ liệu bài học sau khi thu thập từ các nguồn đều được hệ thống phân loại theo chủ đề, gắn tag phù hợp và tự động thêm thông tin cần thiết trước khi hiển thị cho người dùng học. Do đó, chất lượng bài học cao.

Người dùng có thể truy cập vào trang Web cả trên máy tính và điện thoại với giao diện thân thiện, đẹp mắt và thống nhất. Trang Web cung cấp phương pháp học tiếng Nhật hiệu quả, phù hợp với nhiều người học ở các trình độ khác nhau và có tính ứng dụng thực tế cao.

Mục lục

Danh mục hình vẽ

Danh mục bảng

Danh mục các từ viết tắt

HTML	HyperText Markup Language Ngôn ngữ đánh dấu siêu văn bản
CSS	Cascading Style Sheets Định dạng các phần tử được tạo ra bởi ngôn ngữ đánh dấu
CSDL	Cơ sở dữ liệu
CNTT	Công nghệ thông tin
ĐATN	Đồ án tốt nghiệp

Danh mục thuật ngữ

Podcast	Các bài nghe dạng audio, video
Precision	Độ chính xác
Recall	Độ bao phủ
Crawl	Thu thập dữ liệu
Crawler	Bộ thu thập dữ liệu
Apache	Phần mềm mã nguồn mở cài đặt trên máy chủ Web Server để xử lý các yêu cầu gửi đến máy chủ thông qua giao thức HTTP.
Deploy	Triển khai
API	Application Programming Interface Giao diện lập trình ứng dụng

Chương 1 Giới thiệu đề tài

1.1 Đặt vấn đề

Ngày nay, mối quan hệ giữa Nhật Bản và Việt Nam ngày càng trở lên khăng khít và bền vững. Giữa Việt Nam và Nhật Bản có sự hợp tác chặt chẽ trên nhiều lĩnh vực như: giáo dục, y tế, chăm sóc sức khỏe, xây dựng, chuyển giao công nghệ-thiết bị máy móc và đặc biệt là lĩnh vực Khoa học kỹ thuật.

Do đó nhu cầu học tiếng Nhật hiện nay ngày càng tăng lên. Nhiều trường đại học đã thêm tiếng Nhật vào chương trình học chính thức để sinh viên ra trường ngoài kiến thức chuyên môn còn có thế mạnh là ngôn ngữ tiếng Nhật. Tiêu biểu trong số đó là mô hình dự án HEDSPI liên kết giảng dạy giữa Việt Nam và Nhật Bản của trường Đại học Bách Khoa Hà Nội.

Tuy nhiên không phải ai cũng có cơ hội được học tiếng Nhật tại trường lớp hay các trung tâm uy tín, chất lượng. Do đó, việc tự học cũng dần trở nên phổ biến. Với sự phát triển của công nghệ, đã có rất nhiều phương pháp để tự học tiếng Nhật như đọc báo online, nghe nhạc, xem phim, xem thời sự Nhật.

Là một sinh viên của khoa HEDSPI, em mong muốn phát huy hết kiến thức về CNTT và tiếng Nhật của mình đã học được vào làm ĐATN. Do đó em đã xây dựng hệ thống hỗ trợ học tiếng Nhật trực tuyến thông qua podcast để giúp các bạn đang có nhu cầu học tiếng Nhật có được phương pháp học tập phù hợp, hiệu quả và không mất chi phí. Podcast là các đoạn âm thanh hay các video clip được tạo ra bởi cá nhân, tổ chức sau đó đăng lên Internet để người dùng nghe và tải xuống. Hệ thống của em tổng hợp bài học podcast từ nhiều nguồn khác nhau để hỗ trợ người dùng học tiếng Nhật thông qua việc nghe các bài podcast và xem nội dung bài học dưới dạng văn bản. Hệ thống hướng tới hỗ trợ người học ở các trình độ khác nhau, mang lại nhiều kiến thức và thông tin bổ ích.

1.2 Mục tiêu và phạm vi đề tài

Hiện nay, có rất nhiều trang web, ứng dụng dạy từ vựng, ngữ pháp tiếng Nhật theo các cấp độ khác nhau. Ngoài ra, cũng có một số trang đọc báo online về tin tức của Nhật. Tuy nhiên kỹ năng nghe thì lại chưa có nhiều ứng dụng, trang web chú ý đến. Trong số các trang Web hiện nay hỗ trợ dạy tiếng Nhật thông qua nghe bài học podcast, được nhiều người biết đến nhất là các trang chính thức của Nhật như https://newsinslowjapanese.com, https://www3.nhk.or.jp/news/easy, https://www.erin.ne.jp/jp. Các trang này đều có tính năng nghe audio tiếng Nhật, xem nội dung đoạn audio, giải thích nghĩa một số từ vựng trong bài.

Hệ thống em xây dựng gồm 2 phần là Module Crawler tổng hợp dữ liệu podcast từ nhiều nguồn và phần trang Web dạy tiếng Nhật trực tuyến thông qua bài học podcast. Dữ liệu bài học podcast sau khi thu thập về sẽ chưa được hiển thị ngay cho người dùng thấy mà cần quản trị viên phê duyệt. Quản trị viên sẽ chọn chức năng tự động bổ sung thêm thông tin bài học mà khi thu thập dữ liệu về chưa có, gắn chủ đề và tag cho bài học theo cách kiểu thủ công hoặc tự động. Sau khi bài học đã có đủ thông tin, quản trị viên tiến hành phê duyệt bài học. Tất cả bài học đang hiển trên trang Web đều đã được phê duyệt, do đó có độ chính xác cao.

Trang Web dạy tiếng Nhật ngoài việc cung cấp tính năng cơ bản là nghe audio, video như giống các trang podcast thông thường còn hỗ trợ trải nghiệm học tốt hơn cho người dùng thông qua tính năng làm bài tập thực hành. Người dùng sẽ nghe bài học và điền từ còn thiếu vào chỗ trống, qua đó nâng cao khả năng nghe và giúp nhanh nhớ từ vựng hơn. Mỗi người dùng đều có trang cá nhân riêng để quản lý thông tin của mình, xem lại các bài học đã lưu, theo dõi kết quả quá trình học tập. Đồng thời, hệ thống cũng dựa trên xu hướng lựa chọn bài học mà mỗi người dùng để tự động gợi ý các bài có thể người dùng quan tâm.

Trang Web không những mang lại trải nghiệm học tiếng Nhật tốt hơn cho người dùng mà còn hướng tới hỗ trợ tối đa cho quản trị viên thông qua các tính năng tự động phân loại, tự động gắn tag cho bài học. Do đó giảm bớt công việc cần làm và không yêu cầu quản trị viên biết tiếng Nhật vẫn vận hành được hệ thống.

Trang Web có giao diện thân thiện, dễ sử dụng, hỗ trợ đa nền tảng, đa người dùng, có thể truy cập qua mạng Internet từ trình duyệt máy tính hoặc điện thoại.

1.3 Định hướng giải pháp

- Để xây dựng trang Web cho phép người học truy cập thông qua trình duyệt với các tính năng như 1.2 thì em đã xây dựng hệ thống gồm phần Backend và Frontend.

  Backend: Là phần xử lý của hệ thống, được viết bằng ngôn ngữ PHP và hỗ trợ của framework Laravel. Em chọn sử dụng framework Laravel bởi vì nó được thiết kế có tính khoa học cao theo mô hình MVC, do phát triển sau các framework khác nên học hỏi và phát triển được thêm rất nhiều chức năng tuyệt với như tích hợp công cụ dòng lệnh Artisan, xây dựng blade template giúp cho việc kết hợp giữa PHP và HTML trở nên đơn giản, sáng sủa.

  Frontend: Hiển thị nội dung trang Web trên trình duyệt sử dụng ngôn ngữ đánh dấu siêu văn bản HTML, kết hợp với CSS để căn chỉnh giao diện, Javascript, Angular Js để tạo hiệu ứng, nhận biết và xử lý hành vi của người dùng trên trang Web.

- Trang Web được hướng tới cung cấp cho nhiều người có thể truy cập ở bất cứ đâu thông qua Internet, do đó cần phải triển khai (deploy) trang Web lên Server. Trong nhiều webserver hỗ trợ việc triển khai trang web thì nổi bật nhất là Phusion Passenger, Apache hay Nginx. Trong đó em lựa chọn sử dụng Apache vì đây là webserver đơn giản và phổ biến nhất trên thế giới nhưng lại có đầy đủ các chức năng như: Virtual hosts, load balancing, access controls…
- Chất lượng bài học đóng vai trò quyết định đến thành công của trang Web. Vì vậy, em đã nghiên cứu phương pháp thu thập dữ liệu để tạo ra kho dữ liệu bài học với số lượng nhiều, nội dung phong phú, mới mẻ. Các phương pháp đó được triển khai trong bộ Crawler tự động lấy dữ liệu. Bộ Crawler được viết bằng ngôn ngữ Ruby, tự động truy cập đến trang cần lấy dữ liệu, phân tích cấu trúc DOM, lấy dữ liệu và lưu vào CSDL.

  Đối với dữ liệu ảnh, audio nếu chỉ lưu đường dẫn trực tiếp của trang lấy dữ liệu thì sau một thời gian, khi đường dẫn đó thay đổi sẽ không thể truy cập được đến các dữ liệu này nữa. Vì vậy em đã tải ảnh, audio lên server trung gian là Cloudinary rồi lưu đường dẫn của Cloudinary vào CSDL.

- ❖ Một số chức năng hỗ trợ người dùng như đánh tag tự động, phân loại bài học theo chủ đề hay điền từ còn thiếu vào chỗ trống đều liên quan đến xử lý ngôn ngữ tự nhiên. Để thực hiện các chức năng này, em đã sử dụng công cụ hỗ trợ là Mecab. Mecab là công cụ mã nguồn mở có chức năng phân tích hình thái từ để tách từ và trả về các thông tin như loại từ, chức năng của từ trong câu, cách đọc… Công cụ này có tính mềm dẻo và ứng dụng cao bởi được xây dựng với phương hướng là tách biệt hoàn toàn chương trình với dữ liệu (từ điển, corpus huấn luyện, các định nghĩa, tham số), do đó dễ dàng cài đặt và sử dụng.

1.4 Bố cục đồ án

Phần còn lại của báo cáo đồ án tốt nghiệp này được tổ chức như sau.

Chương 2 trình bày về hiện trạng một số trang web dạy học tiếng Nhật trực tuyến. Trên cơ sở đó, em đưa ra biểu đồ use case để thấy rõ tương tác của hai tác nhân người dùng và quản trị viên đối với hệ thống và một số yêu cầu phi chức năng.

Chương 3 em giới thiệu về công nghệ để xây dựng trang web mô hình Client-Server, HTML, CSS, MySQL, PHP và framework Laravel. Tiếp theo, là các lý thuyết về thu thập và trích xuất dữ liệu từ Web được coi như cơ sở để xây dựng bộ Crawler tự động lấy dữ liệu. Sau đó em giới thiệu về công cụ tách từ trong tiếng Nhật Mecab được dùng cho các chức năng phân loại bài học theo chủ đề có sẵn và xác định mức độ quan trọng của từ trong văn bản.

Chương 4 giới thiệu về mô hình Client-Server được sử dụng để xây dựng trang Web. Sau đó em đưa ra các thiết kế từ tổng quan đến chi tiết của cả hệ thống để thấy rõ hơn từng thành phần của hệ thống và mối quan hệ giữa các thành phần đó với nhau. Cuối cùng là kết quả đạt được, một số hình ảnh minh họa và kiểm thử cho một vài chức năng quan trọng.

Chương 5 cũng là chương quan trọng nhất giới thiệu về các đóng góp nổi bật của cả hệ thống. Đó là các chức năng liên quan đến xử lý ngôn ngữ tự nhiên, được phát triển dựa trên hỗ trợ tách từ tiếng Nhật của công cụ Mecab như chức năng tự động sinh bài tập điền từ còn thiếu vào chỗ trống, tự động thêm phiên âm cách đọc cho từ Kanji trong tiếng Nhật, phân loại bài học theo chủ đề có sẵn sử dụng thật toán Naive Bayes, gợi ý đánh tag. Ngoài ra hệ thông còn đưa ra gợi ý bài học có thể người dùng quan tâm dựa trên tương tác của người dùng.

Chương 6, chương kết của báo cáo đồ án, để thông qua đó em nhìn lại kết quả mình đã đạt được, tổng kết kiến thức đã học và các kinh nghiệm rút ra. Cuối cùng là một vài hướng phát triển để cải thiện, nâng cao chất lượng dạy học cho trang web nhưng vì thời gian còn hạn chế nên em chưa kịp thực hiện.

Chương 2 Khảo sát và phân tích yêu cầu

Trước tiên khi đi vào xây dựng và triển khai hệ thống, em đã khảo sát hiện trạng các trang web dạy học tiếng Nhật trực tuyến khác để rút ra một số chức năng cần thiết cho trang web. Trên cơ sở đó, em đưa ra biểu đồ use case với hai tác nhân là người dùng và quản trị viên và đặc tả một số use case quan trọng. Cuối cùng là một số yêu cầu phi chức năng của trang Web.

2.1 Khảo sát hiện trạng

Hiện nay có rất nhiều ứng dụng và trang web liên quan đến hỗ trợ học tiếng Nhật như ứng dụng từ điển, trang web luyện thi tiếng Nhật, trang đọc báo online, trang đọc tiểu thuyết Nhật... Tuy nhiên số lượng trang web dạy kỹ năng nghe còn khá ít. Trong số các trang dạy tiếng Nhật thông qua nghe bài học podcast, được biết đến nhiều nhất là một số trang chính thức của Nhật: https://newsinslowjapanese.com, https://www3.nhk.or.jp/news/easy, https://www.nhk.or.jp/lesson/vietnamese. Em khảo sát 2 trang Web được đánh giá cao và có nhiều người sử dụng nhất hiện nay.

Trang https://newsinslowjapanese.com

Hình 1 Giao diện trang News in Slow Japanese

Gồm các chức năng:

- Nghe bài học podcast dưới dạng audio, hiển thị nội dung bài học dạng văn bản tương ứng và một số từ vựng trong bài kèm giải thích nghĩa.
- Có 2 chế độ nghe tốc độ nhanh và chậm.
- Xem danh sách các bài học theo các chủ đề.
- Tìm kiếm bài học theo tên.
- Lưu lại lịch sử xem bài.
- Có tính năng bình luận bài học.

Ưu điểm	Nhược điểm
Nội dung bài học phong phú. Giao diện đẹp, thân thiện với người dùng	Dữ liệu không được cập nhật thường xuyên. Mất phí để tạo tài khoản đăng nhập để sử dụng thêm một số chức năng khác. Chưa có tính năng làm bài tập.

Trang https://www3.nhk.or.jp/news/easy

Hình 2 Giao diện trang NHK news easy

Trang này có các chức năng:

- Nghe bài podcast là các bản tin thời sự dưới dạng audio, hiển thị nội dung dạng text tương ứng.
- Hiển thị giải thích nghĩa của một số từ quan trọng.

- Hiển thị danh sách các bài học podcast theo ngày.
- Có phần khảo sát ý kiến người dùng về nội dung của bài.

Ưu điểm	Nhược điểm
Nội dung được cập nhật thường xuyên theo từng ngày, từng giờ. Nội dung phong phú, chính xác.	Không phân chia bài học theo từng chủ đề. Đường dẫn đến ảnh, audio sau một thời gian sẽ bị xóa đi và không truy cập được nữa. Chưa có tính năng làm bài tập. Không hỗ trợ tạo tài khoản người dùng.

Từ việc khảo sát ưu, nhược điểm của các trang web trên, em đưa ra các tính năng cần của cả hệ thống như sau:

Hệ thống con thu thập dữ liệu bài học

- Hệ thống tự động lấy dữ liệu từ các nguồn tiếng Nhật lưu vào CSDL.
- Để tránh bị mất dữ liệu thì ảnh và audio sẽ được tải lên server trung gian là Cloudinary rồi lấy đường dẫn ở Cloudinary vào CSDL.

Tính năng dành cho người dùng:

- Xem danh sách các bài học theo chủ đề.
- Hiển thị danh sách các bài học mới nhất, được xem nhiều nhất trong tuần.
- Tìm kiếm bài học theo tên. Người dùng nhập tên bài học dạng tiếng Nhật hoặc tiếng Anh vào ô tìm kiếm sẽ thấy danh sách 5 bài học có tên gần sát với nội dung tìm kiếm nhất được hiển thị bên dưới hoặc vào trang tìm kiếm để xem tất cả các kết quả tìm kiếm.
- Gợi ý những bài học có thể người dùng quan tâm. Hệ thống dựa trên xu hướng lựa chọn bài học mà mỗi người để tự động gợi ý các bài phù hợp.
- Nghe bài học podcast, xem nội dung bài học và một số từ vựng trong bài kèm giải thích. Trong khi nghe, người dùng có thể thay đổi âm lượng, tùy chỉnh đến các đoạn muốn nghe hoặc nghe lại bài nhiều lần.
- Các tính năng thích, bình luận và lưu bài học để xem lại. Các tính năng này đều phải đăng nhập từ trước mới thực hiện được.
- Chức năng làm bài tập điền từ còn thiếu vào chỗ trống. Tính năng này cũng yêu cầu đăng nhập. Khi hết thời gian làm bài hoặc khi người dùng chọn nộp bài thì hệ thống tự động hiển thị kết quả thống kê số câu đúng, sai kèm đáp án chi tiết của từng câu. Người dùng có thể làm bài tập nhiều lần cho một bài học.
- Mỗi người dùng đều có trang cá nhân riêng để quản lý thông tin của mình, quản lý danh sách bài học yêu thích, theo dõi kết quả quá trình học tập.

- Gửi góp ý, đánh giá đến nhà phát triển.
- Hỗ trợ hiển thị nội dung trang web bằng 3 ngôn ngữ: Việt, Nhật, Anh.

Tính năng dành cho quản trị viên:

- Quản lý thông tin bài học: Thêm dữ liệu, gắn tag, phân loại, phê duyệt bài học. Tìm kiếm, xem danh sách tất cả bài học và thông tin chi tiết của từng bài.
- Quản lý thông tin người dùng: Xem danh sách người dùng và thông tin chi tiết của từng người, tìm kiếm người dùng, xóa bình luận có nội dung không lành mạnh, khóa tài khoản của người dùng có nhiều bình luận như vậy, bỏ khóa tài khoản.
- Huấn luyện dữ liệu phục vụ cho các chức năng phân loại bài học theo chủ đề và gắn tag tự động.
- Xem nội dung góp ý của người dùng gửi đến.

2.2 Tổng quan chức năng

2.2.1 Biểu đồ use case tổng quan

Biểu đồ use case tổng quan được biểu diễn như sau:

Hình 3 Biểu đồ use case tổng quan

Mô tả biểu đồ use case tổng quan

- Tác nhân: người dùng và quản trị viên.
- Vai trò: Người dùng sử dụng các dịch vụ, tính năng mà trang web cung cấp. Quản trị viên quản lý thông tin liên quan đến dữ liệu bài học và thông tin người dùng.

Với tác nhân là người dùng: Người dùng không cần đăng nhập mà vẫn có thể sử dụng tất cả các chức năng ngoại trừ xem trang cá nhân và một số chức năng nhỏ của use case xem chi tiết nội dung chi tiết của bài học.

Tên use case	Mô tả tóm tắt
Xem danh sách bài học theo thể loại	Dữ liệu bài học được chia theo 4 thể loại cũng là dữ liệu được lấy 4 nguồn khác nhau. Đó là tiếng Nhật dành cho người mới bắt đầu, tiếng Nhật thời sự, tiếng Nhật đời sống và tiếng Nhật Ted Talk. Người dùng chọn vào từng thể loại để xem danh sách các bài học thuộc thể loại đó.
Xem danh sách bài học theo chủ đề	Các bài học cũng được chia theo 17 chủ đề: văn hóa, thể thao, sự kiện, khoa học, giáo dục… Người dùng chọn vào từng chủ đề để xem danh sách các bài học thuộc chủ đề đó.
Xem chi tiết nội dung bài học	Người dùng nghe podcast và xem nội dung podcast dưới dạng văn bản tương ứng. Các tính năng thích, lưu bài học, bình luận, và làm bài tập yêu cầu phải đăng nhập mới thực hiện được. Use case này được giải thích cụ thể hơn trong phần 2.2.2
Tìm kiếm bài học	Người dùng nhập tên bài học bằng tiếng Nhật hoặc tiếng Anh vào ô input ở màn hình chính sẽ thấy danh sách 5 bài học phù hợp nhất được hiển bị ngay bên dưới. Khi ấn vào nút có biểu tượng tìm kiếm ở cạnh ô input sẽ di chuyển đến màn hình tìm kiếm để xem tất cả kết quả phù hợp.
Xem trang cá nhân	Chức năng này yêu cầu người dùng phải đăng nhập trước. Người dùng xem được danh sách bài học đã lưu, lịch sử hoạt động, lịch sử làm bài tập, thay đổi thông tin cá nhân. Use case này được giải thích cụ thể hơn trong phần 2.2.2

Bảng 1 Mô tả use case tổng quan với tác nhân người dùng

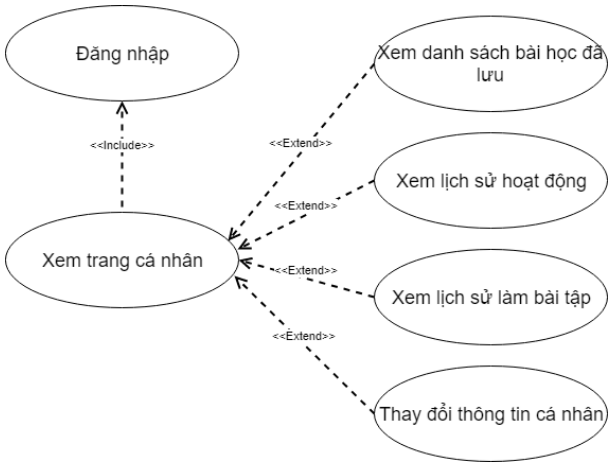
Với tác nhân là quản trị viên: Tất cả các chức năng của quản trị viên đều phải đăng nhập mới thực hiện được.

Tên use case	Mô tả tóm tắt
Quản lý thông tin bài học	Quản trị viên có thể xem danh sách bài học, thêm dữ liệu còn thiếu cho bài học, phân loại chủ đề, gắn tag cho bài học, xóa chủ đề và tag của bài học, phê duyệt bài học.
Quản lý thông tin người dùng	Quản trị viên có thể xem danh sách người dùng, tìm kiếm người dùng, xem thông tin về người dùng và khóa tài khoản người dùng nếu thấy có nhiều bình luận có nội dung không lành mạnh, bỏ khóa tài khoản.
Huấn luyện dữ liệu	Quản trị viên huấn luyện dữ liệu phục vụ cho việc đánh tag và phân loại bài học theo chủ đề.

Bảng 2 Mô tả use case tổng quan với tác nhân quản trị viên

2.2.2 Biểu đồ use case phân rã

Use case phân rã "Xem trang cá nhân"



Hình 4 Biểu đồ use case phân rã "Xem trang cá nhân"

Tác nhân: Người dùng.

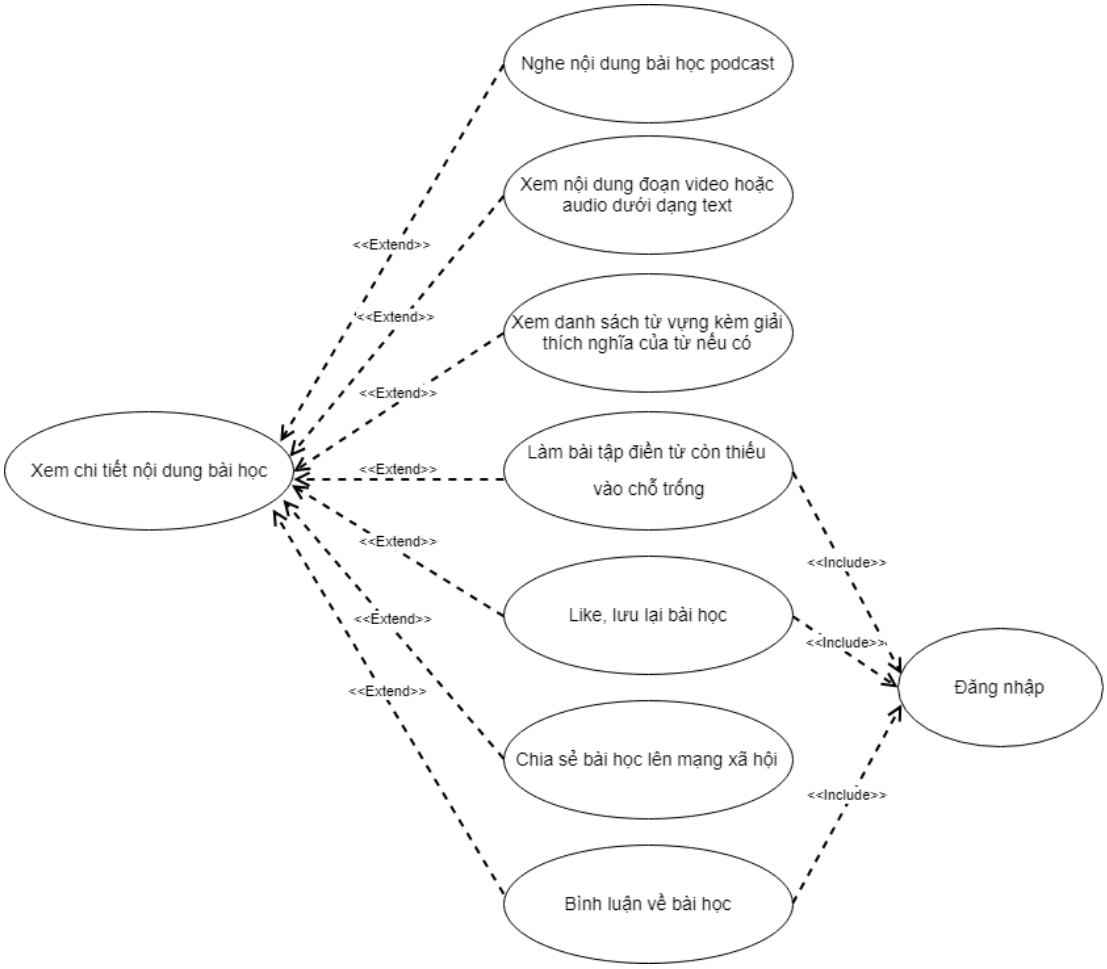
Mô tả use case:

Tên use case	Mô tả tóm tắt
Xem lịch sử hoạt động	Người dùng xem được danh sách các bài học mà mình đã có tương tác: xem, thích, lưu bài học, làm bài tập.
Xem lịch sử làm bài tập	Hiển thị danh sách kết quả tất cả các lần bài tập theo thứ tự thời gian làm bài dưới dạng thống kê về tổng số câu hỏi, số câu trả lời đúng, số câu trả lời sai, số câu không trả lời.

Thay đổi thông tin cá nhân	Người dùng thay đổi các thông tin như avatar, tên hiển thị và mật khẩu đăng nhập.

Bảng 3 Mô tả use case phân rã "Xem trang cá nhân"

Use case phân rã "Xem chi tiết nội dung bài học"



Hình 5 Biểu đồ use case phân rã "Xem chi tiết nội dung bài học"

Tác nhân: Người dùng.

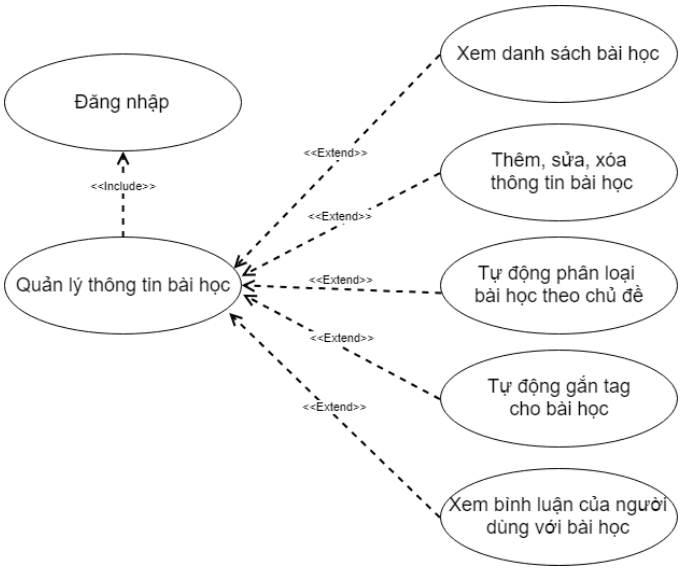
Mô tả use case:

Tên use case	Mô tả tóm tắt
Nghe nội dung bài học podcast	Người dùng tùy chỉnh âm lượng, tua đến thời gian mình muốn nghe và có thể nghe lại nhiều lần.
Bình luận về bài học	Phải đăng nhập để thực hiện chức năng bình luận. Sau khi viết bình luận người dùng có thể chỉnh sửa hoặc xóa bình luận của mình.

Làm bài tập điền từ vào chỗ trống	+ Chức năng này yêu cầu phải đăng nhập mới sử dụng được. + Trong khi làm bài tập, một số thông tin không liên quan đến bài tập như các bình luận của người dùng sẽ bị ẩn đi. + Đến khi hết thời gian của đoạn video hoặc khi người dùng ấn nút kết thúc thì kết quả tự động được hiển thị. + Kết quả là thống kê tổng số câu, số câu đúng, số câu sai, số câu chưa làm và chỉ rõ đáp án từng câu cùng câu trả lời của người dùng đã làm để người dùng tiện so sánh, đối chiếu. + Một bài tập có thể làm nhiều lần, các từ ẩn đi của mỗi lần làm bài sẽ khác so với lần trước.

Bảng 4 Mô tả use case phân rã "Xem chi tiết nội dung bài học"

Use case phân rã "Quản lý thông tin bài học"



Hình 6 Biểu đồ use case phân rã "Quản lý thông tin bài học"

Tác nhân: Quản trị viên.

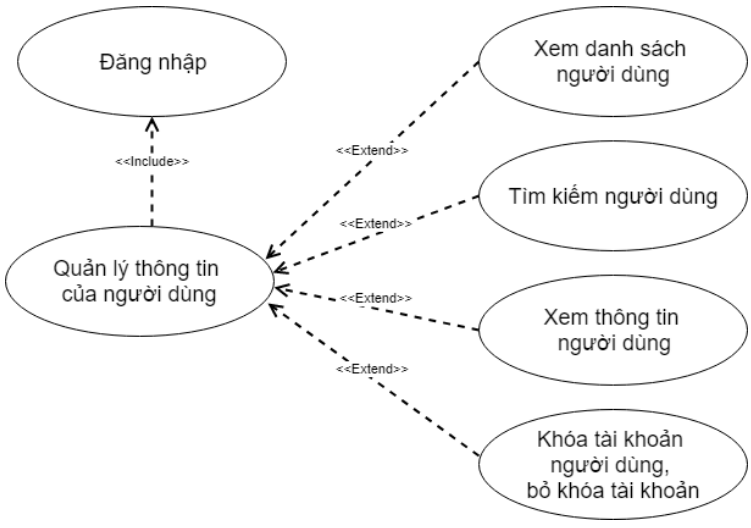
Mô tả use case:

Tên use case	**Mô tả tóm tắt**
Xem danh sách bài học	+ Hệ thống biển thị ra danh sách tất cả các bài học được sắp xếp theo thứ tự thời gian. + Quản trị viên có thể chọn chỉ xem danh sách bài theo nguồn lấy dữ liệu, theo chủ đề hoặc theo trạng thái của bài (có 2 trạng thái là đã phê duyệt và chưa phê duyệt)

	+ Có thể sort danh sách bài học theo các các trường: Id, tên bài học, thời gian.
Thêm, sửa, xóa thông tin bài học	+ Dữ liệu khi lấy từ các nguồn dữ liệu về chưa có phiên âm cách đọc chữ Kanji trong tiếng Nhật. Do đó, quản trị viện cần tự động thêm dữ liệu này cho bài. + Hệ thống đưa ra gợi ý danh sách các chủ đề theo thứ tự giảm dần của mức độ phù hợp. Quản trị viên dựa vào đó, lựa chọn chủ để phù hợp cho bài. + Hệ thống đưa ra gợi ý tag là những từ quan trọng xuất hiện trong bài. Quản trị viên dựa vào đó để gắn tag cho bài. + Xem, chỉnh sửa, xóa các thông tin của bài học. + Sau khi bài học có đủ dữ liệu, quản trị viên phê duyệt bài học để hiển thị cho người dùng thấy.
Tự động phân loại chủ đề cho bài học	Chức năng này để tự động phân loại chủ đề cho các bài học vừa lấy dữ liệu về, chưa được phân loại.
Tự động gắn tag cho bài học	Chức năng này để tự động gắn tag cho các bài học vừa lấy dữ liệu về, chưa được gắn tag.

Bảng 5 Mô tả use case phân rã "Quản lý thông tin bài học"

Use case phân rã "Quản lý thông tin của người dùng"



Hình 7 Biểu đồ use case phân rã "Quản lý thông tin người dùng"

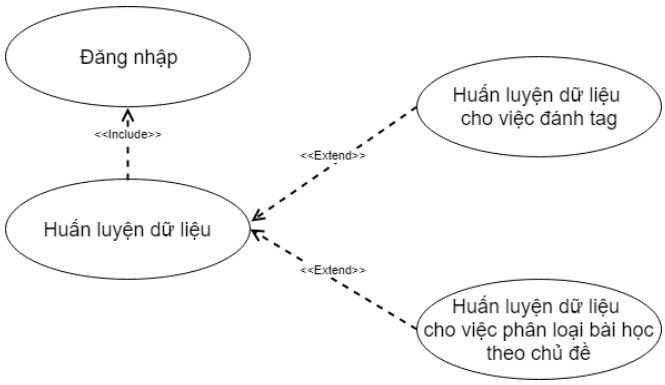
Tác nhân: Quản trị viên.

Mô tả use case:

Tên use case	Mô tả tóm tắt
Xem danh sách người dùng	+ Hệ thống hiển thị danh sách tất cả người dùng được sắp xếp theo thứ tự thời gian tạo tài khoản. + Có thể chọn xem danh sách người dùng theo trạng thái. Có 2 trạng thái là đang hoạt động hoặc đã bị quản trị viên block.
Tìm kiếm người dùng	Quản trị viên tìm kiếm người dùng theo tên hoặc theo ID của người dùng.
Xem thông tin người dùng	Quản trị viên chỉ xem được các thông tin liên quan đến người dùng địa chỉ email, tên đăng nhập, các bình luận, không chỉnh sửa được các thông tin này.
Khóa tài khoản người dùng, bỏ khóa tài khoản	+ Khóa tài khoản của người dùng có nhiều bình luận với nội dung không lành mạnh và gửi mail thông báo cho người dùng biết. + Quản trị viên cũng có thể bỏ khóa các tài khoản đã bị khóa.

Bảng 6 Mô tả use case phân rã "Quản lý thông tin người dùng"

Use case phân rã "Huấn luyện dữ liệu"



Hình 8 Biểu đồ use case phân rã "Huấn luyện dữ liệu"

Tác nhân: Quản trị viên.

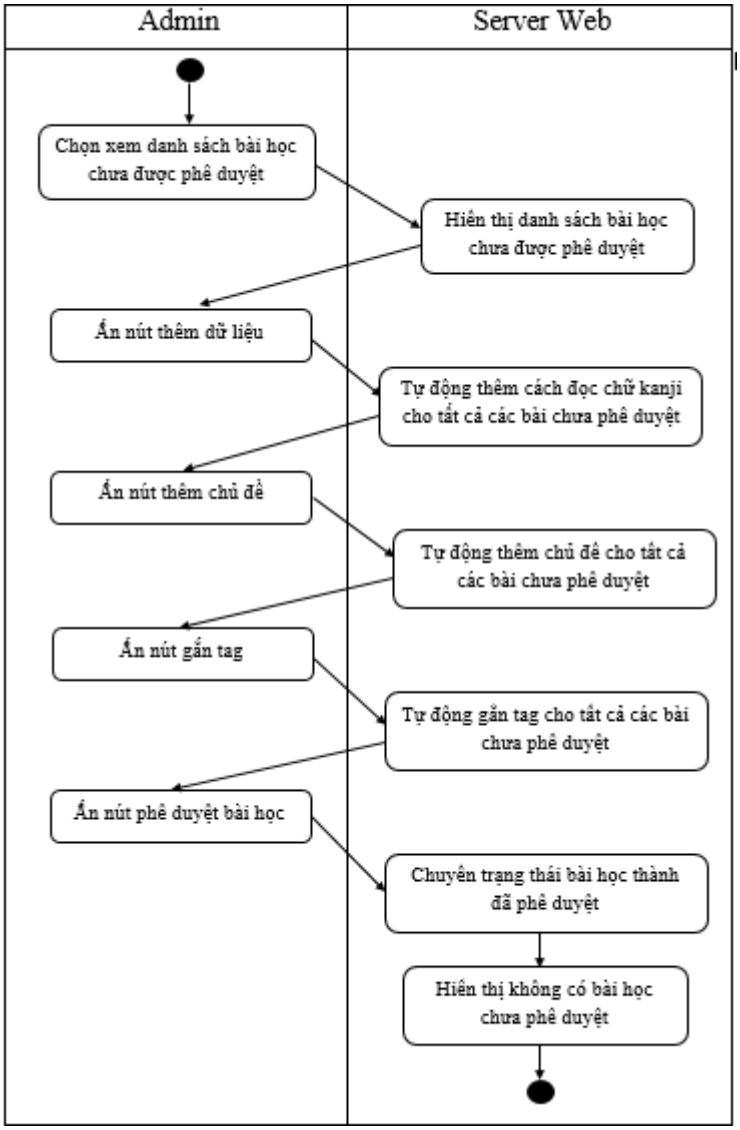
Mô tả use case: Quản trị viên có thể chọn huyến luyện dữ liệu cho việc đánh tag và dữ liệu cho việc phân loại bài học theo chủ đề để cập nhật các dữ liệu đã phân tích được vào CSDL, giúp chức năng gợi ý đánh tag và tự động phân loại đưa ra kết quả chính xác nhất.

2.2.3 Quy trình nghiệp vụ

Quy trình nghiệp vụ "Phê duyệt tất cả bài học mới được lấy dữ liệu"

Tác nhân: Quản trị viên

Mô tả: Đối với các bài học vừa được lấy từ các nguồn dữ liệu về, quản trị viên không cần thiết phải vào từng bài để thêm thông tin, phê duyệt bài học. Thay vào đó, có thể chọn chức năng tự động thêm các thông tin như phiên âm cách đọc chữ Kanji, tự động phân loại chủ đề và gắn tag cho tất cả bài cùng lúc rồi phê duyệt tất cả các bài đó. Quy trình thực hiện như sau:



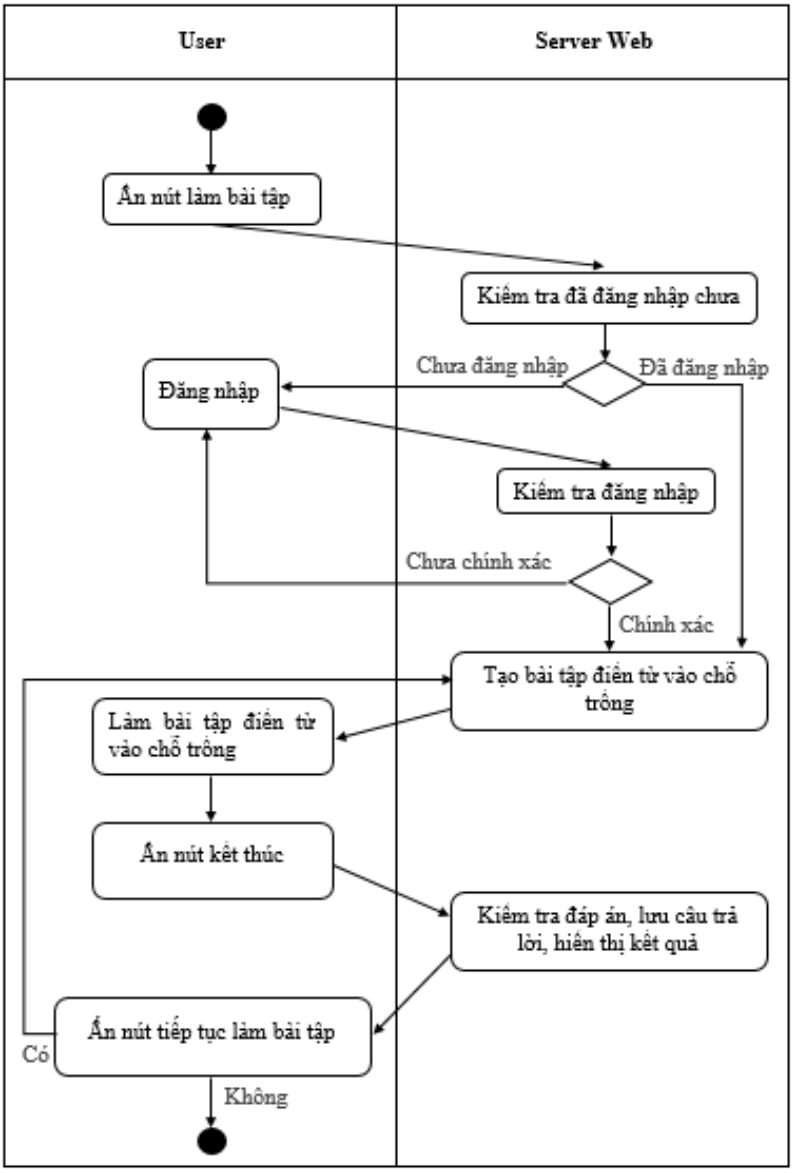
Hình 9 Quy trình nghiệp vụ "Phê duyệt tất cả bài học mới được lấy dữ liệu"

2.3 Đặc tả chức năng

2.3.1 Đặc tả use case “Làm bài tập điền từ vào chỗ trống”

Tác nhân: Người dùng	Tiền điều kiện: Phải đăng nhập
Dữ liệu vào: Câu trả lời của người dùng	Dữ liệu ra: Đáp án
Mô tả tóm tắt: Người dùng làm bài tập điền từ vào chỗ trống do hệ thống tự động sinh ra. Sau khi làm xong, hệ thống tự động hiển thị kết quả.	

Luồng sự kiện chính:

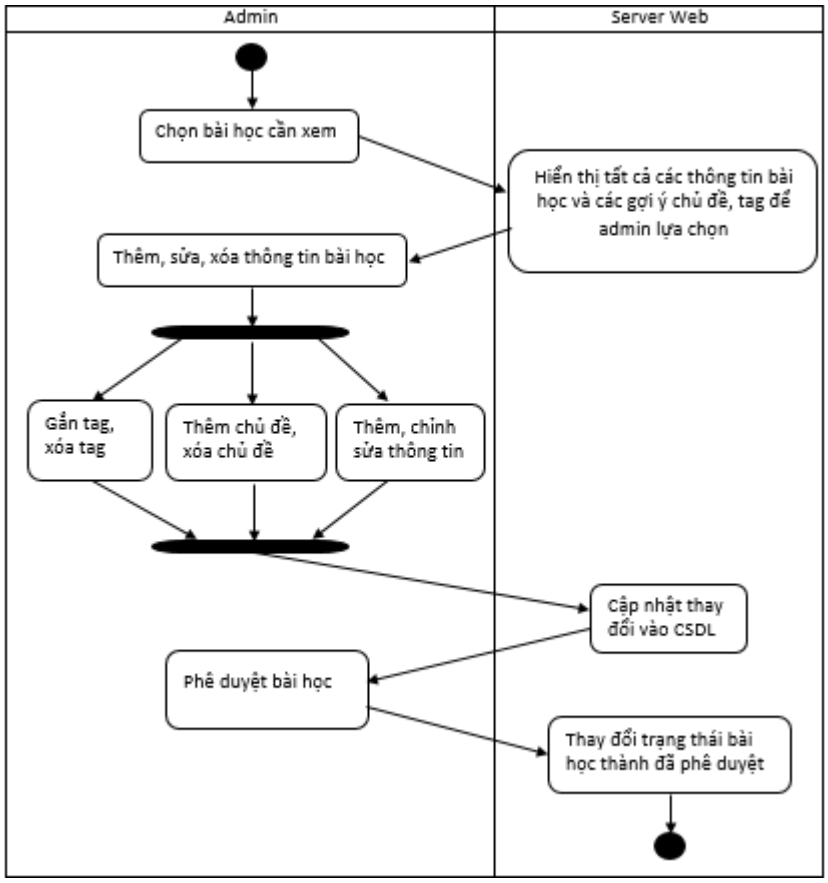


Hình 10 Luồng sự kiện use case “Làm bài tập điền từ vào chỗ trống”

2.3.2 Đặc tả use case “Thêm, sửa, xóa thông tin bài học”

Tác nhân: Quản trị viên	Tiền điều kiện: Phải đăng nhập
Dữ liệu vào: Lựa chọn chủ đề, lựa chọn tag cho bài học	Dữ liệu ra: Bài học đã được phê duyệt
Mô tả tóm tắt: Quản trị viên lựa chọn chủ đề và tag từ gợi ý của hệ thống để gắn cho bài học. Sau khi bài học có đầy đủ thông tin thì quản trị viên phê duyệt bài học để hiển thị cho người dùng học.	

Luồng sự kiện:



Hình 11 Luồng sự kiện use case “Thêm, sửa, xóa thông tin bài học”

2.4 Yêu cầu phi chức năng

Bên cạnh các yêu cầu về chức năng, trang Web còn đáp ứng các yêu cầu phi chức năng như sau:

- Tính dễ dùng: Trang web có giao diện đẹp, thân thiện với người dùng, có sự thống nhất về màu sắc, vị trí, kích thước của các phần tử trong một trang và giữa các trang với nhau. Hỗ trợ hiển thị bằng ba ngôn ngữ: Việt, Nhật, Anh, phù hợp với mọi người dùng ở các trình độ khác nhau, dù đã biết tiếng Nhật hay chưa biết đều có thể dùng được.
- Yêu cầu hiệu năng: Thời gian tải trang Web trong khoảng 1.35-2.8s. Trang Web hoạt động bình thường với số lượng người dùng hiện tại là gần 200 người.
- Yêu cầu linh động: Trang web đã được đẩy lên server, có tên miền cụ thể nên người dùng có thể truy cập mọi lúc, mọi nơi.
- Độ tin cậy: Trang Web đã có sự dụng thử của một số người trong một thời gian và em đã sửa hết các lỗi phát sinh trong thời gian dùng thử.
- Yêu cầu tương thích: Trang web tương thích với tất cả các trình duyệt hiện nay như Chrome, Firefox, Safari ở cả máy tính và điện thoại.
- Yêu cầu riêng tư, an toàn: Chỉ người dùng mới có thể xem, chỉnh sửa thông tin cá nhân của mình.

Trong chương này, em đã tìm hiểu đưa rút ra một số yêu cầu cần thiết cho trang web. Các chức năng của trang web cũng được biểu diễn qua biểu đồ use case với hai tác nhân là người dùng và quản trị viên. Tiếp theo, ở chương 3 em sẽ trình bày về công nghệ sử dụng và các thuật toán áp dụng cho trang web.

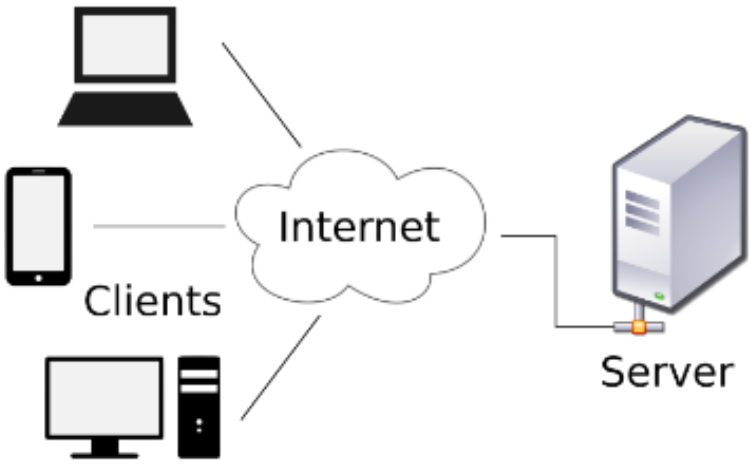
Chương 3 Công nghệ sử dụng

Từ việc khảo sát và phân tích yêu cầu, tiếp theo em sẽ giới thiệu về một số lý thuyết và công nghệ sử dụng. Đó là công nghệ lập trình Web, phương pháp thu thập và trích xuất dữ liệu từ Web, công cụ Mecab tách từ trong tiếng Nhật. Công cụ Mecab được sử dụng để hỗ trợ cho các chức năng liên quan đến xử lý ngôn ngữ tự nhiên mà thuật toán và cơ sở lý thuyết của các chức năng đó cũng được trình bày cụ thể trong hai phần cuối của chương này là phân loại bài học theo chủ đề có sẵn và xác định mức độ quan trọng của từ trong văn bản.

3.1 Công nghệ lập trình web

3.1.1 Giới thiệu mô hình Client-Server

Mô hình client-server là một mô hình nổi tiếng trong mạng máy tính, được áp dụng rất rộng rãi và là mô hình của nhiều trang web hiện có. Trong mô hình này có nhiều Client và một Server liên lạc với nhau qua hệ thống mô hình mạng. Khi client yêu cầu một thông tin gì đó thì Client sẽ gửi yêu cầu đến cho Server, Server xử lý các yêu cầu từ Client rồi phản hồi những thông tin mà Client cần.



Hình 12 Mô hình hệ phân tán Client-Server

Thuật ngữ server được dùng cho những chương trình thi hành như một dịch vụ trên toàn mạng. Các chương trình server này chấp nhận tất cả các yêu cầu hợp lệ đến từ mọi nơi trên mạng, sau đó nó thi hành dịch vụ và trả kết quả về máy yêu cầu. Một chương trình được coi là client khi nó gửi các yêu cầu tới máy có chương trình server và chờ đợi câu trả lời từ server. Chương trình server và client nói chuyện với nhau bằng các thông điệp (messages) thông qua một cổng truyền thông liên tác IPC (Interprocess Communication).

Để một chương trình server và một chương trình client có thể giao tiếp được với nhau thì giữa chúng phải có một chuẩn để nói chuyện, chuẩn này được gọi là giao thức. Nếu một chương trình client nào đó muốn yêu cầu lấy thông tin từ server thì nó phải tuân theo giao thức mà server đó đưa ra.

Các giao thức chuẩn (ở tầng mạng và vận chuyển) được sử dụng rộng rãi nhất hiện nay như: giao thức TCP/IP, giao thức SNA của IBM, OSI, ISDN, X.25 hoặc giao thức LAN-to-LAN NetBIOS. Một máy tính chứa chương trình server được coi là một máy chủ hay máy phục vụ (server) và máy chứa chương trình client được coi là máy khách (client). Mô hình mạng trên đó có các máy chủ và máy khách giao tiếp với nhau theo 1 hoặc nhiều dịch vụ được gọi là mô hình Client-Server.

3.1.2 HTML, CSS, Javascript, Angular JS

HTML (viết tắt cho HyperText Markup Language, hay là "Ngôn ngữ Đánh dấu Siêu văn bản") là một ngôn ngữ đánh dấu được thiết kế ra để tạo nên các trang web với các mẩu thông tin được trình bày trên World Wide Web.

CSS: Trong tin học, các tập tin định kiểu theo tầng – dịch từ tiếng Anh là Cascading Style Sheets (CSS) – được dùng để miêu tả cách trình bày các tài liệu viết bằng ngôn ngữ HTML và XHTML. Ngoài ra ngôn ngữ định kiểu theo tầng cũng có thể dùng cho XML, SVG, XUL. Các đặc điểm kỹ thuật của CSS được duy trì bởi World Wide Web Consortium (W3C). Thay vì đặt các thẻ quy định kiểu dáng cho văn bản HTML ngay trong nội dung của nó, bạn nên sử dụng CSS.

JavaScript là một ngôn ngữ lập trình dựa trên nguyên mẫu với cú pháp phát triển từ C. Giống như C, JavaScript có khái niệm từ khóa, do đó, JavaScript gần như không thể được mở rộng. Cũng giống như C, JavaScript không có bộ xử lý xuất/nhập

(input/output) riêng. Trong khi C sử dụng thư viện xuất/nhập chuẩn, JavaScript dựa vào phần mềm ngôn ngữ được gắn vào để thực hiện xuất/nhập.

Angular JS là framework viết bằng Javascript. Cho phép người dùng sử dụng HTML như ngôn ngữ mẫu và có thể mở rộng cú pháp của HTML để diễn đạt các thành phần ứng dụng một cách rõ ràng hơn. Hai tính năng cốt lõi là Data Binding và Dependency injection của Angularr JS đã hỗ trợ sẵn các chức năng cơ bản để giảm bớt việc code cho người dùng. Angular JS hiện đã hoạt động được trên tất cả trình duyệt.

3.1.3 MySQL

MySQL là hệ quản trị cơ sở dữ liệu tự do nguồn mở phổ biến nhất thế giới và được các nhà phát triển rất ưa chuộng trong quá trình phát triển ứng dụng. Vì MySQL là cơ sở dữ liệu tốc độ cao, ổn định và dễ sử dụng, có tính khả chuyển, hoạt động trên nhiều hệ điều hành cung cấp một hệ thống lớn các hàm tiện ích rất mạnh. Với tốc độ và tính bảo mật cao, MySQL rất thích hợp cho các ứng dụng có truy cập CSDL trên internet. MySQL miễn phí hoàn toàn cho nên bạn có thể tải về MySQL từ trang chủ. Nó có nhiều phiên bản cho các hệ điều hành khác nhau: phiên bản Win32 cho các hệ điều hành dòng Windows, Linux, Mac OS X, Unix, FreeBSD, NetBSD, Novell NetWare, SGI Irix, Solaris, SunOS,...

MySQL là một trong những ví dụ rất cơ bản về Hệ Quản trị Cơ sở dữ liệu quan hệ sử dụng Ngôn ngữ truy vấn có cấu trúc (SQL).

MySQL được sử dụng cho việc bổ trợ PHP, Perl, và nhiều ngôn ngữ khác, nó làm nơi lưu trữ những thông tin trên các trang web viết bằng PHP hay Perl,...

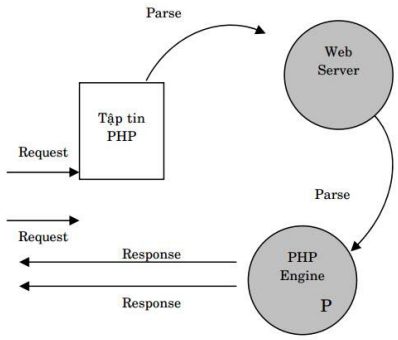
3.1.4 Ngôn ngữ lập trình PHP và Framework Laravel

PHP - viết tắt hồi quy của **"Hypertext Preprocessor"**, là một ngôn ngữ lập trình kịch bản được chạy ở phía server nhằm sinh ra mã HTML trên client. PHP đã trải qua rất nhiều phiên bản và được tối ưu hóa cho các ứng dụng web, với cách viết mã rõ rãng, tốc độ nhanh, dễ học nên PHP đã trở thành một ngôn ngữ lập trình web rất phổ biến và được ưa chuộng.

PHP chạy trên môi trường Webserver và lưu trữ dữ liệu thông qua hệ quản trị cơ sở dữ liệu nên PHP thường đi kèm với Apache, MySQL và hệ điều hành Linux (LAMP).

- Apache là một phần mềm web server có nhiệm vụ tiếp nhận request từ trình duyệt người dùng sau đó chuyển giao cho PHP xử lý và gửi trả lại cho trình duyệt.
- MySQL cũng tương tự như các hệ quản trị cơ sở dữ liệu khác (Postgress, Oracle, SQL server...) đóng vai trò là nơi lưu trữ và truy vấn dữ liệu.
- Linux: Hệ điều hành mã nguồn mở được sử dụng rất rộng rãi cho các webserver. Thông thường các phiên bản được sử dụng nhiều nhất là RedHat Enterprise Linux, Ubuntu...

Khi người sử dụng gọi trang PHP, Web Server sẽ triệu gọi PHP Engine để thông dịch trang PHP và trả kết quả cho người dùng như hình bên dưới.



Hình 13 Quy trình hoạt động của PHP

Laravel là một công cụ mã nguồn mở, một framework ra đời khá muộn so với các framework khác nhưng lại nhanh chóng được đón nhận và hiện đang có cộng đồng người sử dụng nhiều nhất. Sở dĩ Laravel có được các thành công như vậy bởi nó ra đợi sau nên đã kết hợp được ý tưởng tốt của các framework khác và cung cấp nhiều tính năng tuyệt vời, hữu ích cho quá trình phát triển nhanh ứng dụng.

Laravel sử dụng kiến trúc MVC, có tính bảo mật cao, một số tính năng đã được dựng sẵn, hỗ trợ blade template… Laravel không chỉ giúp các lập trình viên code ít hơn, phát triển ứng dụng nhanh hơn mà còn làm ứng dụng dễ hiểu và dễ bảo trì hơn.

3.2 Thu thập và trích xuất dữ liệu từ Web

3.2.1 Phương pháp trích xuất dữ liệu từ Web

Để bóc tách ra các thông tin cần thiết từ một trang HTML có 2 phương pháp chính:

Phương pháp 1 : Không sử dụng cấu trúc trang HTML.

Tư tưởng của phương pháp này là crawl toàn bộ nội dung text của trang web về, sau đó trích xuất dữ liệu sử dụng chính đặc thù của dữ liệu đó.

Một ví dụ điển hình của phương pháp này là trích xuất giá tiền của một sản phẩm. Đặc điểm chung của giá tiền là đằng trước có chứa các từ khóa như “giá”, “price”… và đằng sau sẽ là “VND”, “đồng” …

Ưu điểm của phương pháp này là có thể trích xuất thông tin từ bất kỳ trang Web nào, bởi vì không cần quan tâm tới cấu trúc của trang Web. Tuy nhiên nhược điểm của nó là không biết được trang Web hiện đang thu thập có đáng tin cậy hay không, và quan trọng hơn là xác suất để trích xuất chính xác thông tin là thấp, bởi vì nó phụ thuộc nhiều vào kết quả so khớp.

Phương pháp 2 : Sử dụng cấu trúc của trang HTML.

Tư tưởng của phương pháp này là trước hết phân tích cấu trúc trang HTML bằng cách phân tích cấu trúc cây DOM của nó, sau đó sẽ tiếp cận từng nút con cần thiết để trích rút thông tin.

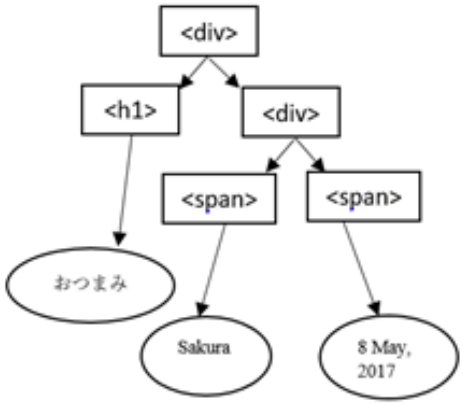
Nhược điểm của phương pháp này là người lập trình cần phải phân tích trước cấu trúc của trang Web. Nếu muốn thêm vào một nguồn mới, thì người lập trình phải viết thêm code. Hơn nữa, cấu trúc của nguồn là thay đổi thường xuyên, do đó người lập trình cũng cần phải update code thường xuyên.

Thế nhưng ưu điểm của nó là độ chính xác của thông tin trích xuất được có thể lên tới 100%. Bởi vì nó không dựa vào một tập luật so khớp, mà dựa trên cấu trúc DOM của chính trang HTML.

Trong đồ án này, em đã tiến hày trích xuất dữ liệu dựa trên phương pháp thứ 2. Bởi vì yêu cầu thông tin về các bài học cần phải chính xác tuyệt đối, hơn nữa nguồn dữ liệu cũng phải có độ tin cậy cao và cấu trúc HTML của các trang em muốn lấy dữ liệu thường ít khi thay đổi. Trong phần tiếp theo em sẽ trình bày về phương pháp trích xuất dữ liệu dựa trên phương pháp thứ 2.

3.2.2 Cây DOM biểu diễn văn bản HTML và truy xuất dữ liệu bằng CSS selector

DOM là một cách biểu diễn của văn bản HTML nói riêng và XML nói chung ở dạng cây. Mỗi tag trong HTML sẽ tương ứng với một node trong DOM. Và mỗi node trong DOM có thể được xác định bởi một CSS selector. Một CSS selector có thể xác định một hoặc nhiều node trong DOM [4]. Dưới đây là ví dụ về đoạn HTML và cây DOM tương ứng.

Văn bản HTML	Cây DOM
 <h1 id= ‘post-title’>おつまみ</h1> Sakura 8 May, 2017 	

Để lấy dữ liệu, CSS selector sẽ tìm đến các thẻ chứa nội dung mong muốn. Các thẻ được phân biệt với nhau bằng tên các class hoặc id. Một số cú pháp CSS Selector hay được sử dụng:

Ví dụ	Mô tả ví dụ
#content	Chọn tất cả các node có id = “content”
.postmeta	Chọn node có class = “postmeta”
div p	Chọn tất cả các node p trong node div
div>p	Chọn tất cả các node p mà có cha của nó là node div.
div+p	Chọn tất cả các node p mà được đặt ngay sau node div

Bảng 7 Cú pháp CSS Selector thường được sử dụng

3.2.3 Lưu trữ hình ảnh, audio lên server trung gian Cloudinary

Cloudinary một cloud-based service cung cấp giải pháp quản lý hình ảnh bao gồm upload, lưu trữ, thao tác, tối ưu hóa. Cloudinary cung cấp các APIs toàn diện và màn hình quản lý giúp dễ dàng tích hợp vào các trang web và ứng dụng di động.

Cloudinary cũng cung cấp Ruby Gem là **gem cloudinary** để dễ dàng trong việc tương tác với ứng dụng viết bằng ngôn ngữ Ruby. Vì vậy, trong khi thu thập và trích xuất dữ liệu, sau khi lấy được đường dẫn ảnh, audio, hệ thống Crawler sẽ gọi hàm upload trong **gem cloudinary** để upload ảnh, audio từ đường dẫn lên server Cloudinary và trả về đường dẫn của ảnh, audio đó ở Cloudinary.

Để sử dụng Cloudinary, cần đăng ký tài khoản rồi lấy các thông tin như Cloud name, API Key, API Secret để cấu hình các thông tin cần thiết cho việc sử dụng dịch vụ Cloudinary. Với các tài khoản miễn phí, Cloudinary cung cấp 20 GB dung lượng lưu trữ. Đối với trang web hiện tại, con số này đủ để đáp ứng các nhu cầu lưu trữ dữ liệu.

3.3 Công cụ Mecab xử lý ngôn ngữ tiếng Nhật

Mecab xử lý vắn bản đầu vào tiếng Nhật, phân tích và tách chúng thành những từ có nghĩa. Không những thế, Mecab còn đưa ra các thông tin liên quan đến từ như từ loại, tác dụng của từ trong câu, cách viết, cách phát âm của từ. [3]

Với từ “食べる” có nghĩa là ăn, từ này đang ở thể thông thường. Giống như tiếng Anh, các từ trong tiếng Nhật được chia thành nhiều thể. Công cụ Mecab vẫn nhận biết được và trả về thể thông thường của từ là “食べる”

食べます	Thể lịch sự	食べません	Thể phủ định lịch sự
食べました	Thể quá khứ lịch sự	食べませんでした	Thể phủ định quá khứ lịch sự
食べた	Thể quá khứ	食べなかった	Thể phủ định quá khứ

食べて	Thể て	食べれば	Thể điều kiện

Bảng 8 Các thể của từ trong tiếng Nhật

Với văn bản đầu vào: “たばこを辞める” (Tôi bỏ thuốc lá) thì kết quả đầu ra là:

たばこ	名詞,一般,*,*,*,*, たばこ, タバコ, タバコ
を	助詞,格助詞,一般,*,*,*,を,を,を
辞める	動詞,自立,*,*,一段,未然形,やめる,ヤメル,ヤメル

Bảng 9 Kết quả phân tích câu sử dụng công cụ Mecab

Các thông tin liên quan đến mỗi từ được cách nhau bởi dấu “,”. Kí tự “*” nghĩa là từ đang không có thông tin đó. Ví dụ với từ “たばこ” (thuốc lá), kết quả phân tích được là:

Phần 1: Chức năng chính của từ	名詞 danh từ
Phần 2: Thông tin phụ của từ	一般 danh từ thông thường
Phần 3: Thông tin phụ của từ	* (không có)
Phần 4: Thông tin phụ của từ	* (không có)
Phần 5: Kiểu biến đối	* (không có)
Phần 6: Dạng khác của từ	* (không có)
Phần 7: Từ ở thể từ điển	タバコ
Phần 8: Cách đọc	タバコ
Phần 9: Cách phát âm	タバコ

Bảng 10 Giải thích kết quả phân tích câu sử dụng công cụ Mecab

Nhờ công cụ Mecab em đã phân tích được nội dung của bài học thành các từ có nghĩa và biết được tác dụng của từ trong câu. Từ đó phục vụ cho việc tự động sinh bài tập điền từ vào chỗ trống và bài toán tự động đánh tag, phân loại bài học theo chủ đề.

3.4 Phân loại bài học theo các chủ đề có sẵn sử dụng thuật toán Naive Bayes

3.4.1 Phát biểu bài toán và mô hình phân loại tổng quát

Bài toán phân loại bài học theo chủ đề có thể được phát biểu như sau: Cho trước một tập văn bản **D={D1,D2,D3,…,Dn}** và tập chủ đề được định nghĩa **C={C1,C2,…,Cn}.**

Nhiệm vụ của bài toán này là gán lớp Di thuộc về Cj đã được định nghĩa. Hay nói cách khác, mục tiêu của bài toán này là đi tìm hàm f:

f: DxC -> Boolean

$$f(d,C) = \begin{cases} true \\ false \end{cases}$$

Có rất nhiều hướng tiếp cận bài toán đã được nghiên cứu như: tiếp cận bài toán phân loại dựa trên lý thuyết đồ thị, cách tiếp cận sử dụng lý thuyết tập thô, cách tiếp cận thống kê… Tuy nhiên tất cả các phương pháp trên đều dựa vào các phương pháp chung là máy học đó là: học không có giám sát, học có giám sát... Trong đó:

- Học không có giám sát là học với tập dữ liệu huấn luyện ban đầu hoàn toàn chưa được gắn nhãn. Học không có giám sát là phương pháp học sử dụng cho lớp bài toán gom cụm, phân cụm.
- Học có giám sát là học với tập dữ liệu huấn luyện ban đầu hoàn toàn được gắn nhãn từ trước. Học có giám sát là phương pháp học sử dụng cho lớp bài toán phân lớp, phân loại.

Điểm khác nhau giữa hai cách học này là đối với học không giám sát, trước khi phân cụm không cần biết có bao nhiêu cụm, và các cụm đó là gì. Còn đối với học có giám sát thì cần phải biết cụ thể các lớp cần phân loại.

Dữ liệu của trang web được lấy từ các nguồn khác nhau, trong đó dữ liệu bài học của nguồn https://newsinslowjapanese.com đã được phân loại theo 17 chủ đề như: văn hóa, thể thao, du lịch, động vật… Bởi vậy, em đã chọn cách học có giám sát và sử dụng dữ liệu của nguồn này để làm dữ liệu huấn luyện, phục vụ bài toán phân loại cho những bài học chưa có chủ đề.

Cụ thể, thuật toán phân loại mà em chọn là Naive Bayes bởi vì nó đơn giản, dễ cài đặt, chạy nhanh, kết quả phân loại được tương đối chính xác, CSDL độc lập so với chương trình nên dễ dàng cập nhật.

3.4.2 Thuật toán phân loại Naive Bayes

Đây là thuật toán được xem là đơn giản nhất trong việc phân loại. Bộ phân lớp Bayes có thể dự báo các xác suất là thành viên của lớp, chẳng hạn xác suất mẫu cho trước thuộc về một lớp xác định với giả định các thuộc tính là độc lập nhau (độc lập điều kiện lớp)

Thuật toán Naive Bayes dựa trên việc tính xác suất có điều kiện.

$$\mathbf{P(Y|X)} = \frac{\mathbf{P(X|Y)} * \boldsymbol{P(Y)}}{\boldsymbol{P(X)}}$$

Trong đó:

- Y đại diện một giả thuyết, giả thuyết này được suy luật khi có được chứng cứ mới X.
- P(X) là xác suất X xảy ra
- P(Y) là xác xuất Y xảy ra
- P(X|Y) là xác suất X xảy ra khi Y xảy ra (xác suất có điều kiện, khả năng của X khi Y đúng)
- P(Y|X) xác xuất hậu nghiệm của Y nếu biết X

Áp dụng trong bài toán phân loại, các dữ liệu cần có:

- D: Tập dữ liệu huấn luyện đã được vector hóa dưới dạng vector X = (X_1, X_2, X_3,…, X_n). Các thuộc tính X_1, X_2, X_3, …, X_n độc lập xác suất đôi một với nhau.
- C_i: Tập các tài liệu của D thuộc lớp C_i với i = {1, 2, 3,…}

Với tài liệu mới X^{new} = (X_1, X_2,…, X_n) cần phân lớp, định lý Bayes [6] là:

$$\mathbf{P}(\boldsymbol{C_i}|\mathbf{X}) = \frac{\boldsymbol{P(X|C_i)} * \boldsymbol{P(C_i)}}{\boldsymbol{P(X)}}$$

Theo tính chất độc lập điều kiện:

$$\mathbf{P(X|Ci)} = \prod_{k=1}^{n} \boldsymbol{P(X_k|C_i)} = \mathbf{P(X_1 \mid C_i)* P(X_2 \mid C_i)* P(X_3 \mid C_i)\ldots* P(X_n \mid C_i)}$$

Vì P(X) có giá trị là như nhau đối với mỗi lớp nên luật phân lớp cho tài liệu mới X^{new}={ X_1, X_2,…, X_n} là:

$$\text{Max}\left(P(C_i) * \prod_{k=1}^{n} P(X_k | C_i)\right)$$

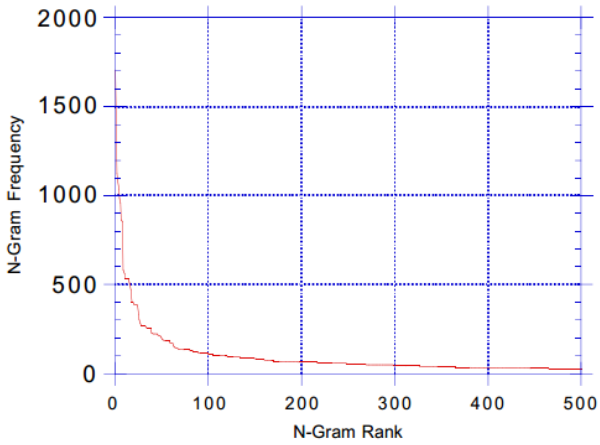
Trong đó:

- $P(C_i)$: Được tính dựa trên tần suất xuất hiện tài liệu trong tập huấn luyện.
- $P(X_k | C_i)$ Được xác định bởi giá trị tập thuộc tính trong quá trình huấn luyện.

3.5 Xác định mức độ quan trọng của từ sử dụng thuật toán TF-IDF

Xử lý ngôn ngữ là một kĩ thuật quan trọng nhằm giúp máy tính hiểu được ngôn ngữ của con người, qua đó hướng dẫn máy tính thực hiện và giúp đỡ con người trong những công việc có liên quan đến ngôn ngữ.

Tuy nhiên trong ngôn ngữ luôn có những từ có xuất hiện với tần suất nhiều hơn các từ khác. Một trong những phát biểu nổi tiếng nhất Zipf's law phát biểu về vấn đề này là "The nth most common word in a human language text occurs with a frequency inversely proportional to n" [5], dịch sang Tiếng Việt là "Từ quan trọng thứ n trong một văn bản ngôn ngữ của con người xuất hiện với một tần số nghịch với n." Để trực quan hóa phát biểu đó, dưới đây là biểu đồ về mối quan hệ giữa tần suất xuất hiện và độ quan trọng của từ.



Hình 14 Biểu đồ mối quan hệ giữa tần suất xuất hiện và độ quan trọng của từ

Biểu đồ trên cho thấy luôn có một tập các từ mà tần số xuất hiện, sử dụng nhiều hơn các từ khác nhưng mức độ quan trọng lại không cao, điều này đúng trong bất kì ngôn

ngữ nào. Chính vì vậy cần có một phương pháp để làm mịn đường cong tần số trên hay là việc cân bằng mức độ quan trọng giữa các từ.

Một trong những kĩ thuật được sử dụng đó là TF-IDF, viết tắt của Term Frequency - Inverse Document Frequency. TF-IDF là trọng số của một từ trong văn bản thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản. [7]

TF (Term Frequency) dùng để ước lượng tần suất xuất hiện của từ trong văn bản. Với từ t nằm trong văn bản d thì giá trị TF(t, d) được xác định là:

$$TF(t, d) = \frac{\text{Số lần xuất hiện của từ t trong văn bản d}}{\text{Tổng số từ trong văn bản d}}$$

IDF (Inverse Document Frequency) dùng để ước lượng mức độ quan trọng của từ đó. Khi tính tần số xuất hiện TF thì các từ đều được coi là quan trọng như nhau. Giá trị IDF sẽ làm giảm mức độ quan trọng của những từ xuất hiện nhiều nhưng lại không mang nhiều ý nghĩa. Với văn bản d nằm trong tập hợp các văn bản D thì IDF(t, D) xác định bởi:

$$IDF(t, D) = log\_e(\frac{\text{Tổng số văn bản trong tập mẫu D}}{\text{Số văn bản có chứa từ t}})$$

Mức độ quan trọng của từ t trong văn bản d mà văn bản d nằm trong tập văn bản D được xác định là:

$$TF\text{-}IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

Ở chương 3, em đã giới thiệu về tất cả công nghệ và thuật toán được áp dụng cho cả hệ thống. Trong đó, công nghệ thu thập và trích xuất dữ liệu sử dụng để xây dựng bộ Crawler tự động lấy dữ liệu. Công cụ Mecab để tách văn bản tiếng Nhật thành từng từ, được sử dụng cho các chức năng phân loại bài học theo chủ đề có sẵn áp dụng thuật toán Naive Bayes và xác định mức độ quan trọng của từ thông qua thuật toán TF-IDF. Sau khi đã xác định công nghệ sử dụng, tiếp theo ở chương 4 em đi vào phát triển và triển khai ứng dụng.

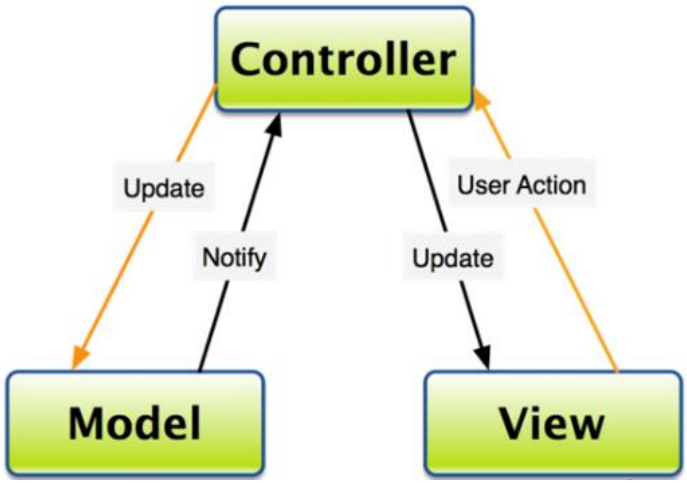
Chương 4 Phát triển và triển khai ứng dụng

Chương 3 đã thảo luận về công nghệ sử dụng và các cơ sở lý thuyết làm nền tảng cho việc triển khai ứng dụng. Do đó trong chương này, em giới thiệu về kiến trúc phần mềm được sử dụng để xây dựng trang Web đó là kiến trúc MVC. Sau đó là các thiết kế từ tổng quan đến chi tiết, em mô tả các thành phần trong hệ thống và mối quan hệ giữa các thành phần đó với nhau. Tiếp đến, em đưa ra kết quả đạt được, một số hình ảnh minh họa và kiểm thử cho một vài chức năng quan trọng.

4.1 Thiết kế kiến trúc

4.1.1 Lựa chọn kiến trúc phần mềm

Trang web được thiết kế theo kiến trúc phần mềm MVC, viết tắt của Model-View-Controller, mỗi thành phần có một nhiệm vụ riêng biệt và độc lập với các thành phần khác. Cách thức hoạt động của mô hình MVC



Hình 15 Mô hình MVC

- Model: Quản lý dữ liệu hệ thống và thao tác với các dữ liệu đó. Chứa các hàm, các phương thức truy vấn trực tiếp với CSDL.
- View: Đảm nhận việc hiển thị thông tin, tương tác với người dùng, nơi chứa tất cả các đối tượng GUI như textbox, images...
- Controller: Đóng vài trò trung gian giữa Model và View với nhiệm vụ tiếp nhận yêu cầu từ Client, xử lý yêu cầu, gọi các hàm, phương thức trong Model để lấy dữ liệu, gửi dữ liệu qua View tương ứng rồi trả kết quả về cho Client.

Mô hình này thể hiện tính chuyên nghiệp trong lập trình, phân tích thiết kế. Do được chia thành các thành phần độc lập nên giúp phát triển ứng dụng nhanh, đơn giản, sản phẩm dễ nâng cấp, bảo trì.

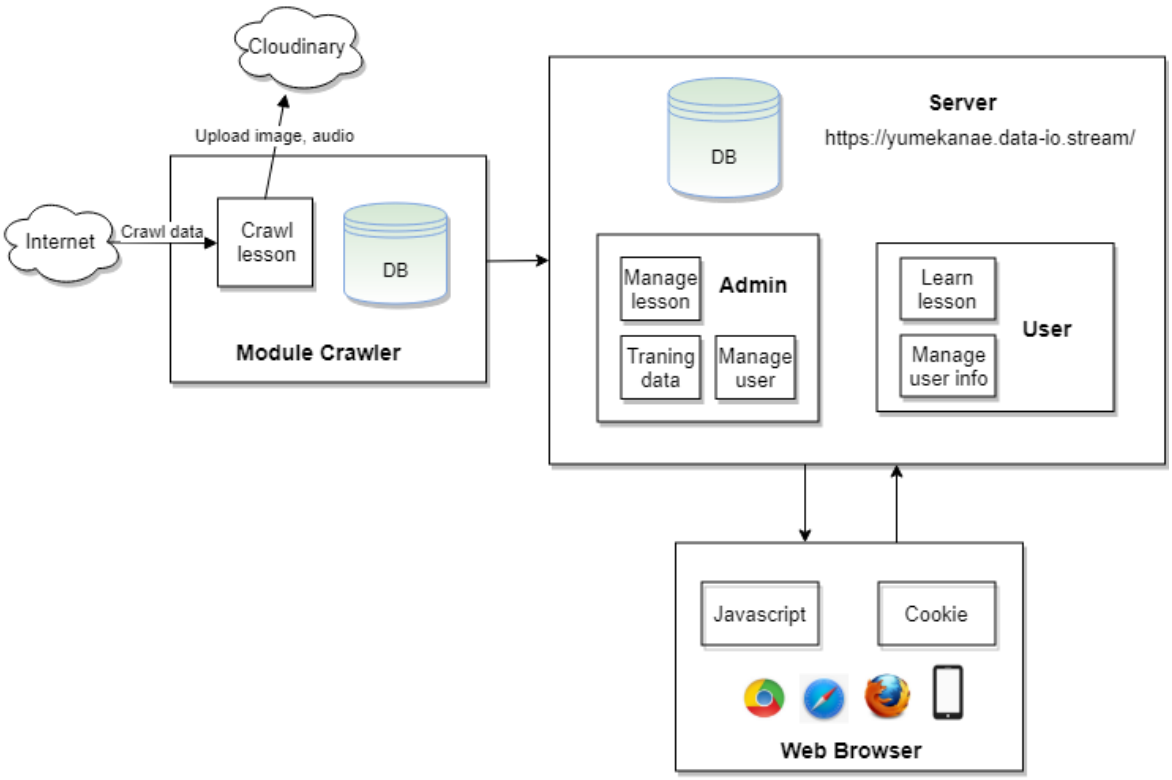
Trang web em xây dựng tuân thủ theo đúng kiến trúc MVC gồm 3 phần Model, Controller, View.

- Model: Gồm các thành phần để quản lý và thao tác với CSDL như: class Category, class Lesson, class Topic, class Tag, class User…
- Controller: Gồm các file tương ứng với các file trong phần Model để xử lý yêu cầu từ Client như CategoryController, LessonController, TopicController, TagController …
- View: Ứng với mỗi Controller có forder chứa các file hiển thị nội dung được Controller tương ứng gửi đến. Giao diện trang Web được thiết kế hiển thị cân xứng, đẹp mắt trên cả trình duyệt máy tính và trình duyệt điện thoại.

4.1.2 Thiết kế tổng quan

Hệ thống gồm Module Crawler và WebServer và giao diện phía người dùng.

Mô hình tổng quan của hệ thống được biểu diễn như trong hình dưới đây:



Hình 16 Thiết kế tổng quan của hệ thống

Module Crawler là 1 module độc lập dùng để thu thập dữ liệu bài học từ trên Internet lưu vào CSDL, được người quản trị chạy riêng trên máy tính. Dữ liệu ảnh và audio sẽ được tải lên server trung gian là Cloudinary rồi lưu đường dẫn ở Cloudinary vào CSLD để tránh mất dữ liệu.

WebServer cung cấp các chức năng của hệ thống dành cho quản trị viên và người dùng.

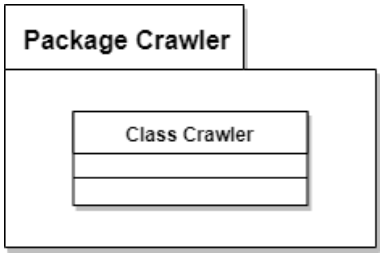
- Quản trị viên quản lý tất cả dữ liệu bài học như thêm, sửa, xóa bài học, tự động gắn tag, phân loại chủ đề cho bài, phê duyệt bài học... Ngoài ra, quản trị viên còn nắm được các thông tin liên quan người dùng như tổng số người dùng, quản lý bình luận đối với mỗi bài học và có thể huấn luyện dữ liệu phục vụ các chức năng phân loại bài học theo chủ đề và gắn tag tự động.
- Người dùng có các chức năng học bài, làm bài tập và quản lý thông tin cá nhân của mình như xem danh sách các bài học đã lưu, xem lịch sử hoạt động, xem kết quả quá trình làm bài tập, thay đổi thông tin cá nhân.

Trang Web có thể hiển thị trên tất cả các trình duyệt hiện nay như Chrome, Firefox, Safari và cả trình duyệt điện thoại với giao diện đẹp mắt, thân thiện với người dùng.

Trên trình duyệt, sử dụng Cookie để lưu thông tin đăng nhập, ngôn ngữ hiển thị hiện tại của trang Web. Javascript để xử lý hành vi của người dùng như nhập thông tin vào form đăng nhập, viết bình luận… rồi gửi các thông tin đó đến Server để xử lý và lưu vào CSDL.

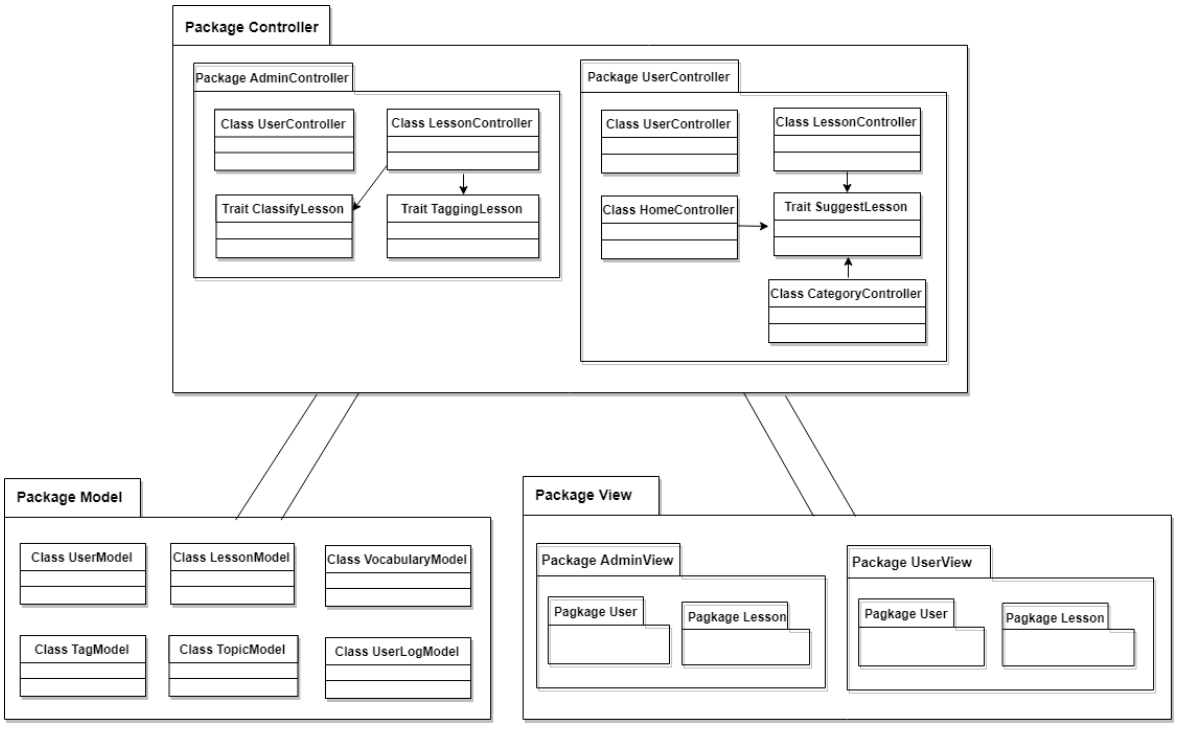
4.1.3 Thiết kế chi tiết gói

Module Crawler được tạo nên từ gói Crawler.



Hình 17 Thiết kế chi tiết gói Crawler

WebServer được thiết kế theo mô hình MVC ứng với 3 gói là gói Model, gói View và gói Controller.



Hình 18 Thiết kế chi tiết gói Controller

Gói Controller gồm 2 gói con là AdminController và UserController. Các yêu cầu từ Client muốn truy cập vào gói này đều vào phải đi qua middleware CheckAminLogin

để kiểm tra chỉ các tài khoản của quản trị viên mới có quyền truy cập. Lớp LessonController kế thừa Trait ClassifyLesson và Trait TaggingLesson để sử dụng các phương thức xử lý tính năng loại bài học theo chủ đề và gắn tag cho bài học. Gói UserController xử lý các yêu cầu từ người dùng. Các lớp LessonController, HomeController, CategoryController có tính năng gợi ý bài học cho người dùng nên cần kế thừa Trait SuggestLesson. (Trait có thể được hiểu như một Class, được PHP 5.4 trở lên hỗ trợ để khắc phục nhược điểm đơn kế thừa trong PHP)

Mỗi lớp trong gói Model tương ứng với một bảng trong CSDL. Mỗi lớp bao gồm thuộc tính và phương thức để định nghĩa kiểu liên kết giữa bảng ứng với lớp đó và các bảng khác trong CSDL. Ngoài ra, mỗi lớp còn chứa các phương thức truy vấn dữ liệu để các lớp trong gói Model sử dụng.

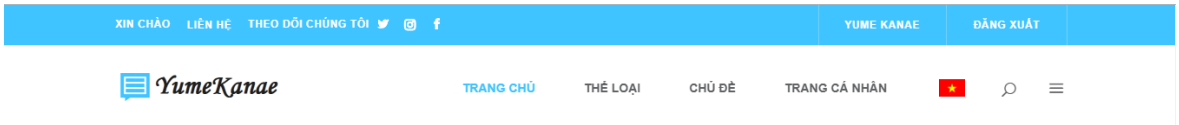
Gói View cũng gồm 2 gói con là AdminView và UserView. Ứng với mỗi lớp trong gói Controller sẽ có một gói View tương ứng, chứa các file HTML để hiển thị dữ liệu trên trình duyệt.

4.2 Thiết kế chi tiết

4.2.1 Thiết kế giao diện

Trang Web của em hướng tới hiển thị trên mọi màn hình với kích thước khác nhau và cả màn hình Web mobile. Hệ thống đã được thử nghiệm hiển thị tốt ở một số độ phân giải như 1366x768, 1280x720, 800x600 trên máy tính và tỷ lệ màn hình 16:9 trên điện thoại.

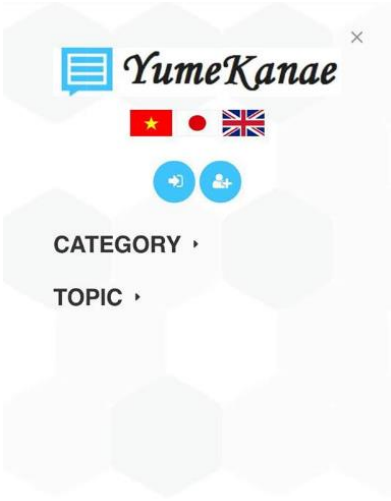
Giao diện trang Web được thiết kế thống nhất ở mọi màn hình, với màu đặc trưng là màu xanh. Ở mỗi trang đều có header menu để người dùng chọn xem bài học theo thể loại, xem bài học theo chủ đề, xem trang cá nhân, thay đổi ngôn ngữ, đến trang tìm kiếm bài học. Ngoài ra ở header menu còn có phần liên kết với trang giới thiệu ở các mạng xã hội như Facebook, Instagram…



Hình 19 Giao diện header menu

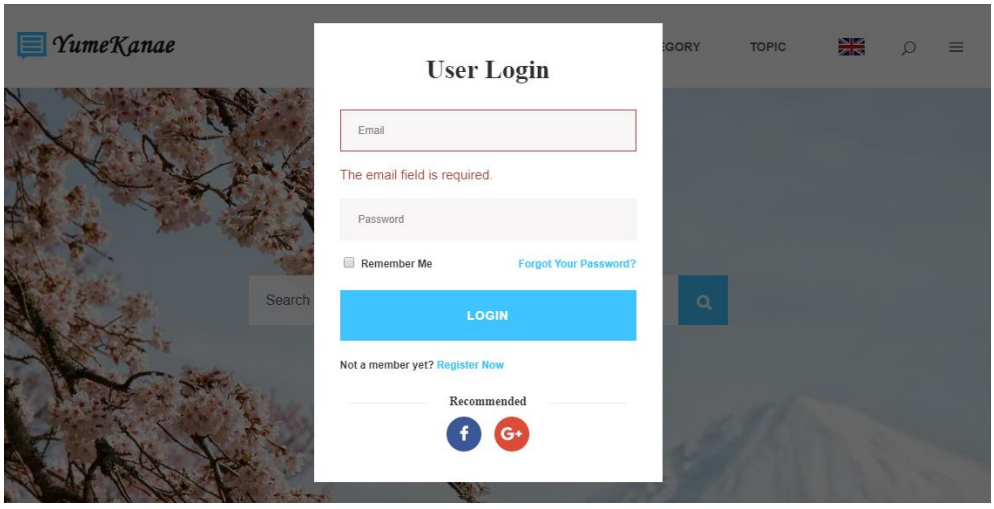
Với giao diện web mobile, thanh header menu sẽ chỉ hiển thị logo của trang Web, khi ấn vào logo sẽ đi đến trang chủ. Các tùy chọn khác sẽ xem ở phần side menu.

Hình 20 Giao diện header menu trên mobile

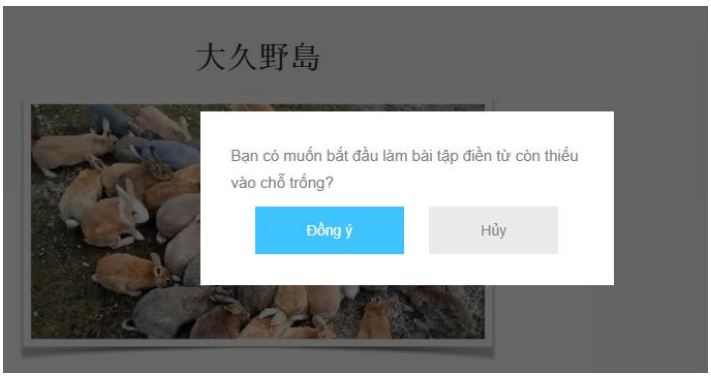


Hình 21 Giao diện side menu trên mobile

Vị trí hiển thị form nhập liệu, các thông báo đều ở chính giữa màn hình. Nội dung bên trong được căn lề so với hai bên và phía trên dưới. Khi xảy ra lỗi nhập liệu sẽ có thông báo lỗi bên dưới ô input. Các nút ấn cũng có màu xanh đặc trưng của trang Web.



Hình 22 Giao diện form nhập liệu



Hình 23 Giao diện hiển thị thông báo

Đối với dữ liệu ảnh bị lỗi không tải được thì hệ thống sẽ tự động thay thế bằng ảnh đại diện của trang Web.

Hình 24 Giao diện ảnh đại diện thay thế ảnh lỗi

Ở một số trang có phần menu bên phải hiển thị danh sách các bài học được gợi ý từ hệ thống và danh sách các bài học được xem nhiều nhất để người dùng dễ dàng cho trong việc lựa chọn bài học.

Hình 25 Giao diện right menu

4.2.2 Thiết kế lớp

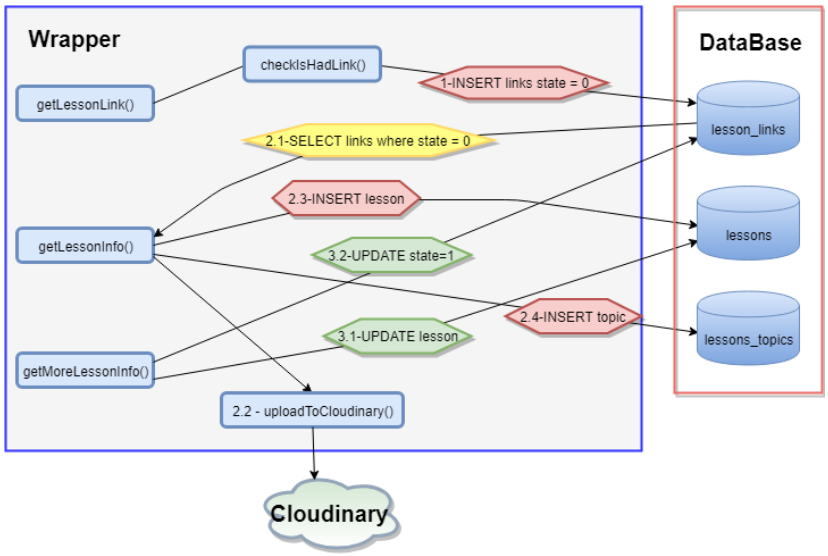
Lớp Crawler của gói Crawler

Các thuộc tính và phương thức của lớp này như sau:

Class Crawler
- sourceURL: string
+ getLessonLink(): void + getLessonInfo(): void + getMoreLessonInfo(): void + checkIsHadLink(link): boolean + uploadToCloudinary(link): string

- **sourceURL**: Địa chỉ URL của trang web cần lấy dữ liệu.
- **getLessonLink()**: Thu thập toàn bộ đường dẫn các bài học vào lưu vào bảng lesson_links.
- **checkIsHadLink()**: Kiểm tra đường dẫn đã có trong bảng lesson_links chưa trước khi thêm dữ liệu vào bảng.
- **getLessonInfo()**: Thu thập toàn bộ thông tin chi tiết của bài học lưu vào bảng lessons.
- **uploadToCloudinary()**: Upload ảnh/audio lên Cloudinary và trả về đường dẫn tương ứng của ảnh/audio trên Cloudinary.
- **getMoreLessonInfo():** Một số dữ liệu không thể lấy trực tiếp thì phải sử dụng Chrome Driver để mở browser giả lập, phân tích cấu trúc HTML từ browser giả lập đó để lấy thêm thông tin bài học và cập vào bảng lessons.

Quy trình hoạt động của Crawler như sau:



Hình 26 Thứ tự thực hiện các bước trong lớp Crawler

Bước 1: Thu thập đường dẫn đến các bài cần lấy dữ liệu lưu vào bảng lesson_links với giá trị trường state bằng 0.

Bước 2.1: Lấy tất cả đường dẫn có giá trị trường state bằng 0 từ bảng lesson_links.

Bước 2.2: Tải ảnh/audio lên Cloudinary và trả về đường dẫn của ảnh/audio ở Cloudinary.

Bước 2.3: Trích xuất dữ liệu chi tiết của bài lưu vào bảng lessons.

Bước 2.4: Với nguồn dữ liệu mà bài học đã được phân loại chủ đề thì lưu thông tin đó vào bảng lessons_topics

Bước 3.1: Mở browser giả lập để trích xuất thêm các thông tin bài học còn thiếu cập nhật vào bảng lessons.

Bước 3.2: Cập nhật trạng thái state bằng 1 ở bảng lesson_links.

Trait ClassifyLesson của gói AdminController

Trait ClassifyLesson
- lesson_data
+ getWordArray(): array + getTopicValue(): array + sortTopicValue(): array + insertTopicValueToDB(): void

Ý nghĩa của các phương thức:

- **getWordArray():** Sử dụng công cụ Mecab để tách nội dung bài học thành từng từ, lưu vào mảng.
- **getTopicValue():** Tính giá trị sử dụng thuật toán Naive Bayes cho 17 chủ đề.
- **sortTopicValue():** Sắp xếp chủ đề theo thứ tự giá trị tính toán thuật toán giảm dần.
- **insertTopicValueToDB():** Lưu id của chủ đề và id của bài học vào bảng lesson_topics.

Trait TaggingLesson của gói AdminController

Trait TaggingLesson
- lesson_data
+ getSpecialNoun(): array + getTFValue(): array + getIDFValue(): array + getTFIDFValue(): array + sortTFIDFValue(): array + insertTagValueToDB(): void

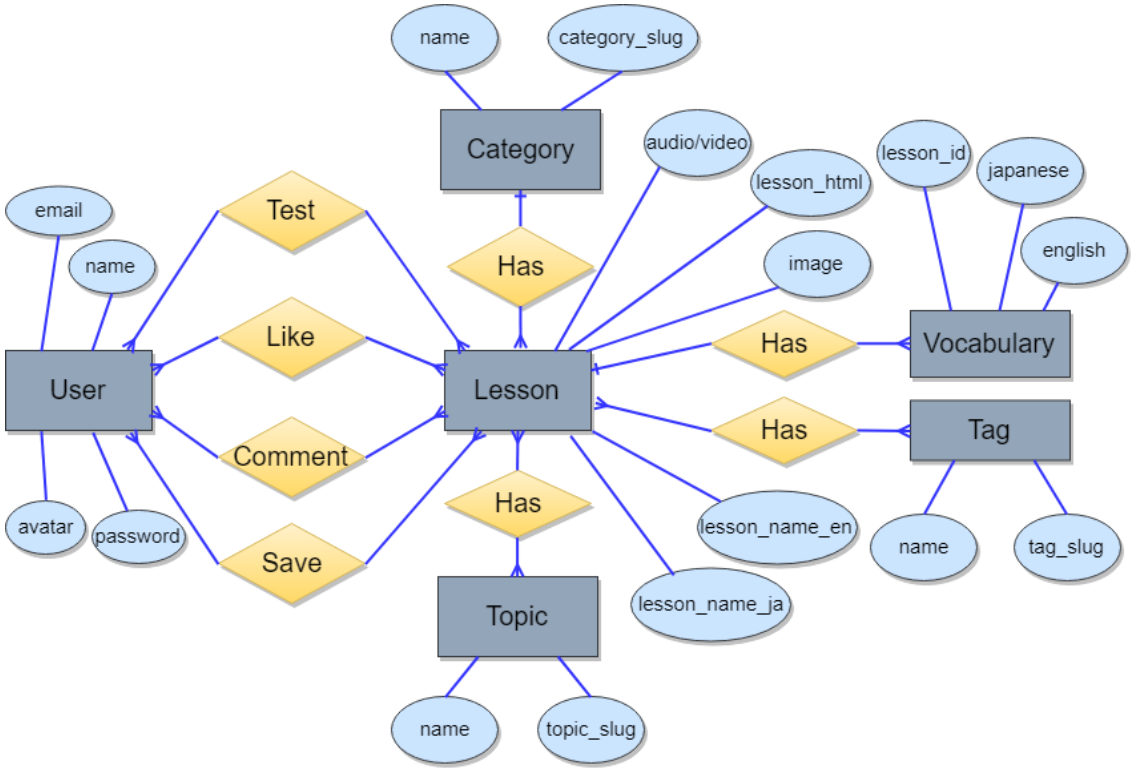
Ý nghĩa của các phương thức:

- **getSpecialNoun():** Sử dụng công cụ Mecab để tách từ và trả về danh từ riêng kèm ý nghĩa danh từ riêng đó là gì (dùng cho chức năng gợi ý tag theo danh từ riêng)
- **getIFValue():** Tính giá trị TF của tất cả danh từ trong bài.
- **getIDFValue():** Lấy giá trị IDF của tất cả danh từ trong bài từ CSDL.
- **getTFIDFValue():** Tính giá trị IF*IDF của tất cả danh từ trong bài.
- **sortTFIDFValue():** Sắp xếp mảng từ theo giá trị IF*IDF từ lớn đến nhỏ.
- **insertTagValueToDB():** Lưu giá trị tag và id của bài học vào bảng lesson_tags.

Một số lớp quan trọng trong gói Model dùng để định nghĩa thuộc tính và phương thức liên kết giữa các bảng trong CSDL với nhau được giải thích chi tiết trong phần phụ lục A.

4.2.3 Thiết kế cơ sở dữ liệu

Trước tiên, em đưa ra biểu đồ ER để thấy được các thực thể, thuộc tính của thực thể và mối quan hệ giữa các thực thể.



Hình 27 Biểu đồ thực thể liên kết ER

Từ biểu đồ thực thể liên kết trên, em thiết kế CSDL sử dụng hệ quản trị cơ CSDL Mysql gồm những bảng sau:

Bảng categories: Lưu thông tin về các thể loại của bài học

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của thể loại
name	varchar	Tên thể loại bằng tiếng anh
category_slug	varchar	Lưu URI để tạo đường dẫn đến category
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 11 Cấu trúc bảng categories

Bảng lesson_links: Lưu thông tin về các links bài học từ các nguồn phục vụ cho việc crawl dữ liệu

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của đường dẫn dữ liệu
lesson_crawl_id	int	Id của bài học đó trong nguồn lấy dữ liệu
topic_id	varchar	Những chủ đề mà bài học đó thuộc vào, lấy từ các nguồn dữ liệu
lesson_link	varchar	Đường dẫn đến bài học cần lấy dữ liệu
lesson_image	varchar	Đường dẫn đến ảnh đại diện của bài học
crawl_link_id	int	Id để phân biệt các nguồn dữ liệu
state	int	Giá trị để nhận biết bài học đó đã được lấy dữ liệu chưa. 0: chưa lấy dữ liệu (mặc định) 1: đã lấy dữ liệu
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 12 Cấu trúc bảng lesson_links

Bảng lessons: Lưu thông tin về các bài học

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id bài học
category_id	int	Id của thể loại bài học, khóa ngoại
lesson_name_en	varchar	Tên bài học bằng tiếng anh
lesson_name_ja	varchar	Tên bài học bằng tiếng nhật
image	varchar	Đường dẫn đến ảnh đại diện cho bài học
audio	varchar	Đường dẫn đến đoạn audio cho bài học
video	varchar	Đường dẫn đến đoạn video cho bài học
like	varchar	Số lượt thích bài học
view	varchar	Số lượt xem bài học
lesson_html	varchar	Nội dung bài học kèm thẻ html để hiển thị trên trình duyệt
lesson_data	varchar	Nội dung bài học
lesson_slug	varchar	Lưu URI để tạo đường dẫn đến lesson
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa
deleted_at	timestamp	Thời gian xóa

Bảng 13 Cấu trúc bảng lessons

Bảng vocabularies: Bảng lưu các từ vựng có trong bài học

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, lưu id của từ vựng
lesson_id	int	Id của bài học, khóa ngoại
japanese	varchar	Từ cần giải thích trong bài học

english	varchar	Giải thích bằng tiếng anh
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 14 Cấu trúc bảng vocabularies

Bảng users: Lưu thông tin người dùng

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id người dùng
social_id	varchar	Id người dùng nếu đăng nhập bằng mạng xã hội
social_type	varchar	Kiểu mạng xã hội
name	varchar	Tên người dùng
email	varchar	Email người dùng
avatar	varchar	Avatar người dùng
password	varchar	Mật khẩu để đăng nhập
is_admin	int	Phân biệt người dùng bình thường và quản trị viên. Mặc định: 0 (người dùng bình thường)
remember_token	varchar	Remember token
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa
deleted_at	timestamp	Thời gian xóa đối

Bảng 15 Cấu trúc bảng users

Bảng user_likes: Bảng lưu thông tin về like của người dùng

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của like
user_id	int	Id của người dùng có hành động like, khóa ngoại

lesson_id	int	Id của bài học ứng với like, khóa ngoại
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 16 Cấu trúc bảng user_likes

Bảng user_saves: Bảng lưu thông tin về save của người dùng

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của save
user_id	int	Id của người dùng có hành động save, khóa ngoại
lesson_id	int	Id của bài học ứng với save, khóa ngoại
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 17 Cấu trúc bảng user_saves

Bảng user_comments: Bảng lưu thông tin về bình luận của người dùng

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của bình luận
user_id	int	Id của người dùng có bình luận, khóa ngoại
lesson_id	int	Id của bài học ứng với bình luận, khóa ngoại
value	text	Nội dung bình luận
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa
deleted_at	timestamp	Thời gian xóa

Bảng 18 Cấu trúc bảng user_comments

Bảng user_logs: Bảng lưu thông tin về lịch sử hành động của người dùng.

Tên trường	Kiểu dữ liệu	Mô tả

id	int	Khóa chính, id của hành động
user_id	int	Id của người dùng có hành động, khóa ngoại
lesson_id	int	Id của bài học ứng với hành động, khóa ngoại
action_type	int	Kiểu của hành động: 1: Thích bài học 2: Lưu bài học 3: Làm bài tập 4: Xem bài học
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 19 Cấu trúc bảng user_logs

Bảng user_contacts: Bảng lưu các ý kiến đóng góp của người dùng.

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của liên hệ
name	varchar	Tên của người liên hệ
email	int	Email của người gửi liên hệ
user_id	int	Id của người dùng nếu người đó đang đăng nhập
website	varchar	Địa chỉ website của người liên hệ
phone_number	varchar	Số điện thoại của người liên hệ
message	text	Nội dung góp ý
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 20 Cấu trúc bảng user_contacts

Bảng user_exercises: Bảng lưu kết quả làm bài tập của người dùng.

Tên trường	Kiểu dữ liệu	Mô tả

id	int	Khóa chính, id của lần làm bài tập
user_id	int	Id của người làm bài tập
lesson_id	int	Id của bài tập
total_question	int	Tổng số câu hỏi
true_answer	int	Số câu trả lời đúng
wrong_answer	int	Số câu trả lời sai
no_answer	int	Số câu không trả lời
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 21 Cấu trúc bảng user_exercises

Bảng topics: Bảng lưu thông tin về các chủ đề bài học

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của topic
name	varchar	Tên của topic
topic_slug	varchar	Lưu URI để tạo đường dẫn đến trang xem bài học theo chủ đề
lesson_percentage	float	Được tính bằng số bài học của chủ đề đó chia cho tổng số bài học của tất cả các chủ đề. Dùng cho chức năng tự động phân loại bài học
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 22 Cấu trúc bảng topics

Bảng lessons_topics: Bảng trung gian giữa 2 bảng topics và lessons để tạo quan hệ nhiều nhiều giữa 2 bảng đó

Tên trường	Kiểu dữ liệu	Mô tả

id	int	Khóa chính, id của bảng trung gian
topic_id	int	Id của thể loại, khóa ngoại
lesson_id	int	Id của bài học, khóa ngoại
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 23 Cấu trúc bảng lessons_topics

Bảng tags: Bảng lưu thông tin về tag của bài học

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của topic
name	varchar	Tên của topic
tag_slug	varchar	Lưu URI để tạo đường dẫn đến trang xem bài học có cùng tag
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 24 Cấu trúc bảng tags

Bảng lessons_tags: Bảng trung gian giữa 2 bảng tags và lessons để tạo quan hệ nhiều nhiều giữa 2 bảng đó

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính
tag_id	int	Id của tag, khóa ngoại
lesson_id	int	Id của bài học, khóa ngoại
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 25 Cấu trúc bảng lessons_tags

Bảng stop_words: Bảng lưu các từ dừng là những từ không có nhiều ý nghĩa trong bài học như từ nối

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của từ dừng
token	varchar	Từ dừng

Bảng 26 Cấu trúc bảng stop_words

Bảng tag_words: Bảng lưu thông tin của các từ phục vụ cho việc đánh tag tự động

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của từ phục vụ cho việc đánh tag tự động
word	varchar	Giá trị của từ phục vụ cho việc đánh tag tự động
idf	float	Giá trị IDF tương ứng với từ
created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 27 Cấu trúc bảng tag_words

Bảng topic_words: Bảng lưu thông tin của các từ phục vụ cho việc phân loại bài học theo từng topic. Trong đó, tần suất xuất hiện của từ trong một topic = $\text{frequencty} * 10^{-\text{exponent}}$

Tên trường	Kiểu dữ liệu	Mô tả
id	int	Khóa chính, id của từ phục vụ cho việc phân loại bài học
word	varchar	Giá trị của từ phục vụ cho việc phân loại bài học
topic_id	int	Topic Id mà từ đó thuộc vào
frequency	float	Lưu giá trị frequency trong công thức tính tần suất xuất hiện của từ trong một topic
exponent	int	Lưu giá trị exponent trong công thức tính tần suất xuất hiện của từ trong một topic

created_at	timestamp	Thời gian tạo
updated_at	timestamp	Thời gian cuối cùng chỉnh sửa

Bảng 28 Cấu trúc bảng topic_words

4.3 Xây dựng ứng dụng

4.3.1 Thư viện và công cụ sử dụng

Để hoàn thành trang Web, em đã sử dụng các thư viện và công cụ hỗ trợ sau đây:

Mục đích	Công cụ	Địa chỉ URL
Ngôn ngữ lập trình và framework hỗ trợ	PHP 7 và framework Laravel 5.4	https://laravel.com/
IDE lập trình	PhpStorm 64bit	https://www.jetbrains.com/phpstorm/
Trình soạn thảo văn bản	Sublime Text 3 64bit	https://www.sublimetext.com/
Công cụ hỗ trợ	Mecab	https://sourceforge.net/projects/mecab/files/latest/download

Bảng 29 Danh sách thư viện và công cụ sử dụng

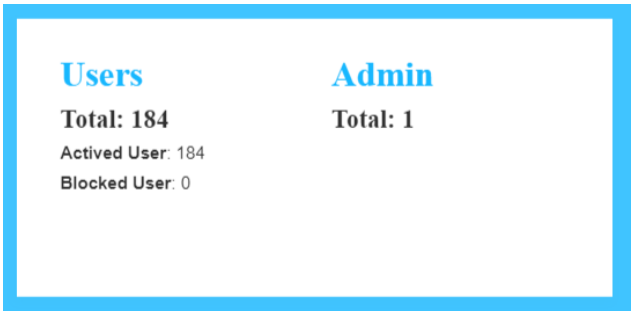
4.3.2 Kết quả đạt được

Hệ thống gồm Module Crawler và trang Web dạy tiếng Nhật. Module Crawler thu thập dữ liệu các bài tiếng Nhật từ các nguồn dữ liệu, được quản trị viên chạy độc lập trên máy tính. Trang Web dạy tiếng Nhật đã được chạy thực tế trên Server với đường dẫn là https://yumekanae.data-io.stream.

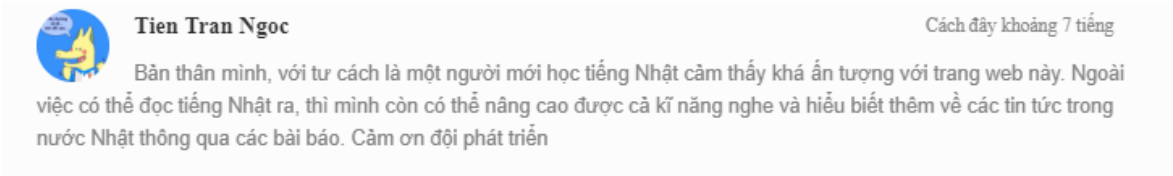
Module	Số dòng code	Số gói	Dung lượng mã nguồn	Sản phẩm đóng gói	Dung lượng sản phẩn đóng gói
Bộ Crawler	1289	1	66.4 kB	Không có	Không có
Trang Web	222603	3	49.1 mB	Không có	Không có

Bảng 30 Thông tin về mã nguồn

Trang Web được triển khai lên Server cho người dùng sử dụng từ ngày 01/05/2018, đến nay đã được gần 200 người sử dụng và cũng nhận được nhiều đánh giá tích cực của người dùng về nội dung và chất lượng.



Hình 28 Thống kê số lượng người dùng trang Web



Hình 29 Đánh giá của người dùng

Hình 30 Đánh giá của người dùng trên mạng xã hội

Sau khi hoàn thành hệ thống và đem so sánh với các trang Web dạy tiếng Nhật thông qua podcast khác, em nhận thấy trang Web của em có đầy đủ tính năng mà các trang Web khác có, ngoài ra còn cung cấp trải nghiệm tốt hơn cho người dùng thông qua:

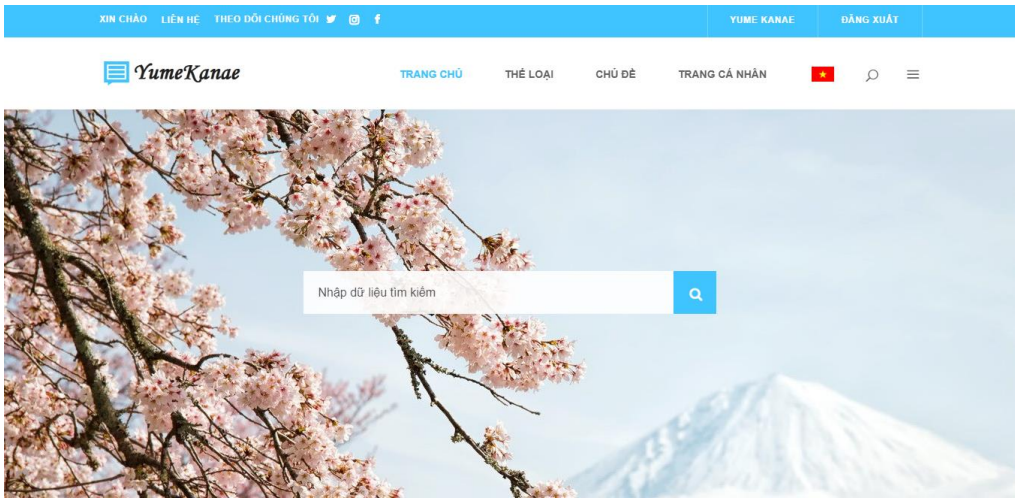
- Chức năng gợi ý bài học mà có thể người dùng quan tâm.
- Có chức năng làm bài tập điền từ còn thiếu vào chỗ trống để việc học tiếng Nhật hiệu quả hơn.
- Mỗi người dùng có trang cá nhân riêng để quản lý thông tin của mình như xem lại các bài đã lưu, xem lịch sử hoạt động và kết quả quá trình học tập…

- Có các chức năng tương tác của người dùng như bình luận bài học, chia sẻ bài học lên mạng xã hội…
- Dữ liệu bài podcast được tổng hợp từ nhiều nguồn nên nội dung phong phú.

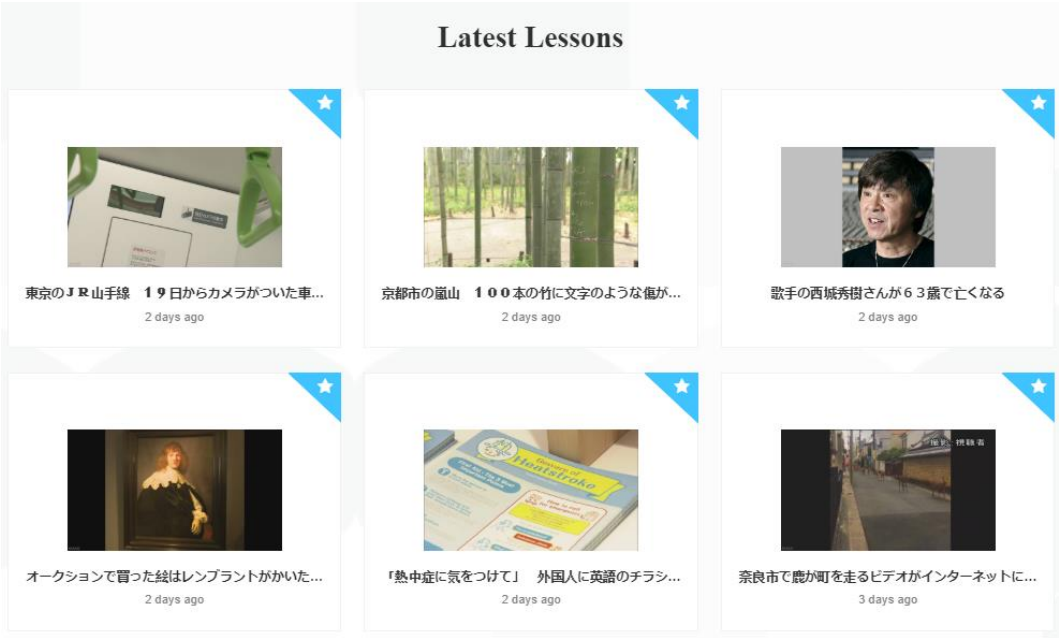
4.3.3 Minh hoạ các chức năng chính

Trang chủ

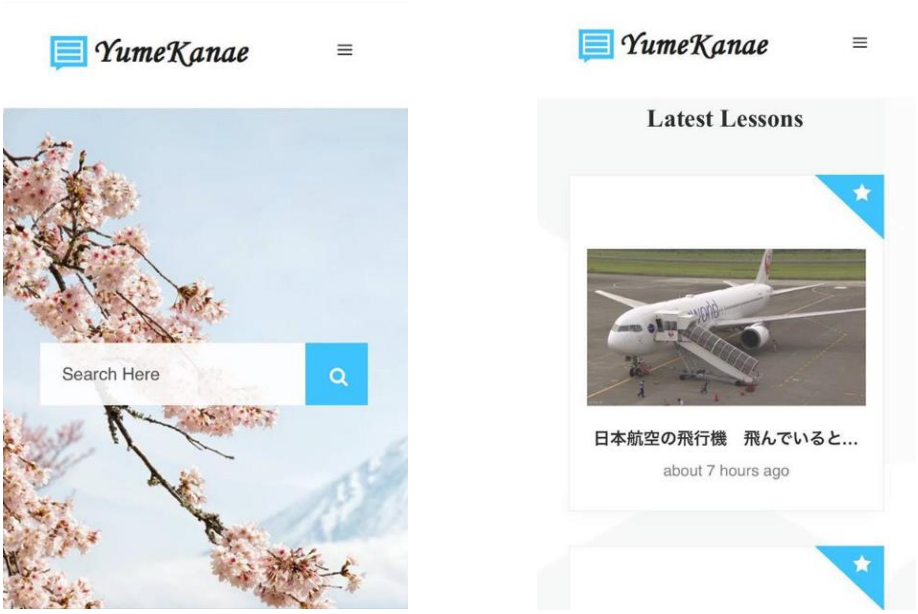
Trang chủ sẽ hiển thị danh sách bài học mới nhất, danh sách bài được xem nhiều nhất và danh sách gợi ý từ hệ thống các bài mà có thể người dùng quan tâm.



Hình 31 Giao diện trang chủ



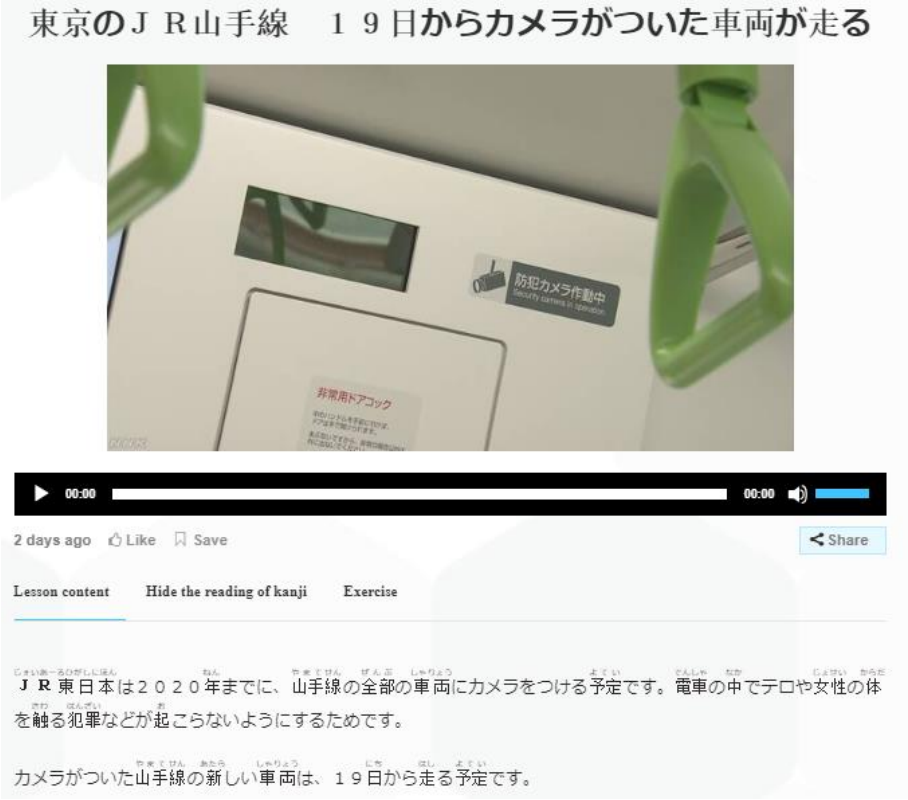
Hình 32 Giao diện trang chủ



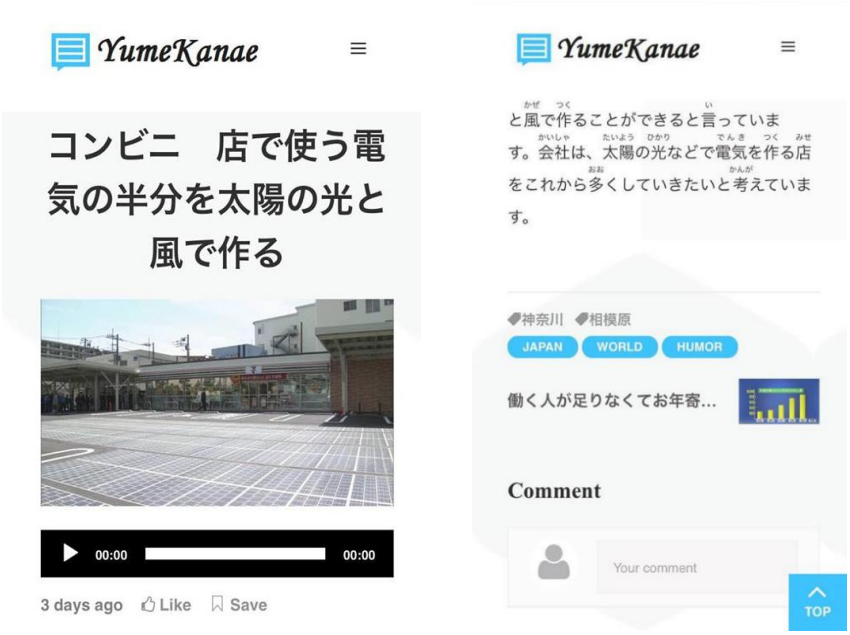
Hình 33 Giao diện trang chủ trên mobile

Chức năng học bài

Người dùng có thể nghe bài học tiếng Nhật dạng audio hoặc video, tùy chỉnh âm lượng, tua đến đoạn muốn nghe, xem nội dung bài học, xem danh sách từ vựng trong bài… Ngoài ra, người dùng có thể thích bài, lưu bài, share bài, bình luận.



Hình 34 Giao diện chức năng học bài

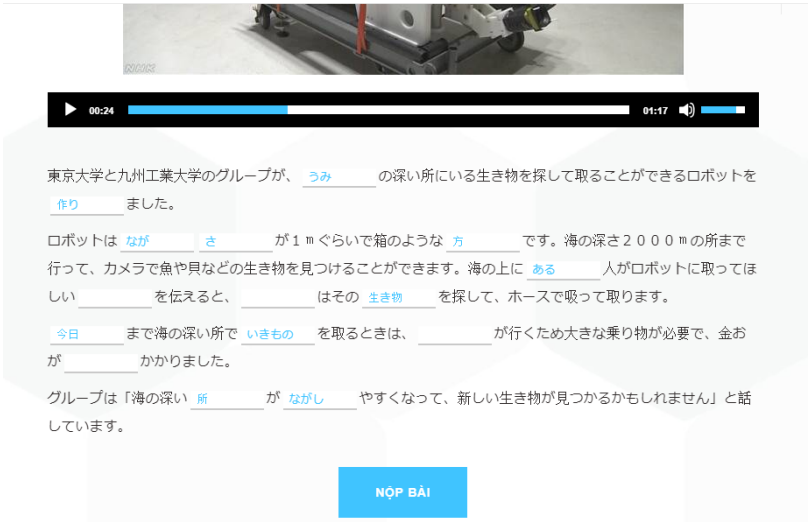


Hình 35 Giao diện chức năng học bài trên mobile

Hình 36 Giao diện chức năng học bài đối với bài học video

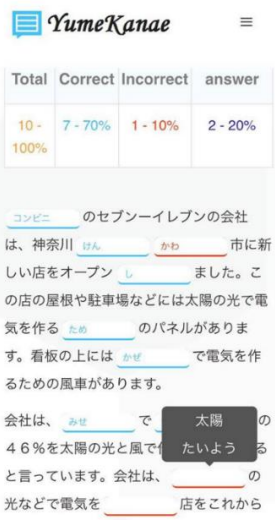
Chức năng làm bài tập điền từ vào chỗ trống

Người dùng vừa nghe bài học và điền từ vào chỗ trống, các từ không nghe được có thể không điền. Khi muốn kết thúc thì ấn vào nút “Nộp bài” hoặc khi hết thời gian làm bài hệ thống sẽ tự hiển thị kết quả. Kết quả sẽ bao gồm thống kê về tổng số câu, số câu trả lời đúng, sai và chưa trả lời. Đồng thời cũng hiển thị đáp án với câu sai.



Hình 37 Giao diện làm bài tập điền từ vào chỗ trống

Hình 38 Giao diện kết quả bài tập điền từ vào chỗ trống



Hình 39 Giao diện kết quả bài tập điền từ vào chỗ trống trên mobile

Hình 40 Giao diện làm bài tập điền từ đối với bài học video

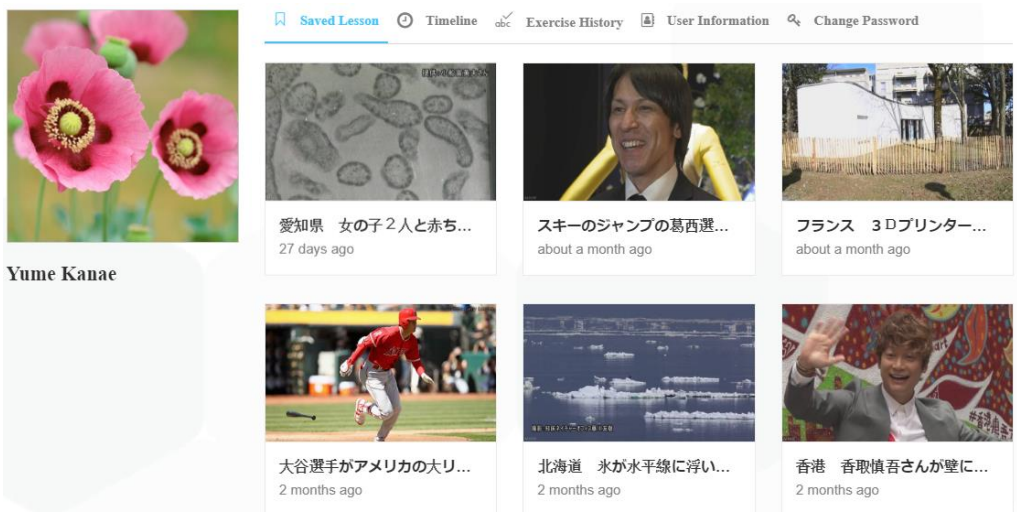
Đối với bài học video, vì nội dung bài học dài, nên kết quả chỉ hiển thị những câu có phần từ còn thiếu và chỉ rõ kết quả của từng câu. Người dùng có thể ấn vào từng câu để tua video đến thời gian câu đó đang nói.

Hình 41 Kết quả bài tập điền từ đối với bài học video

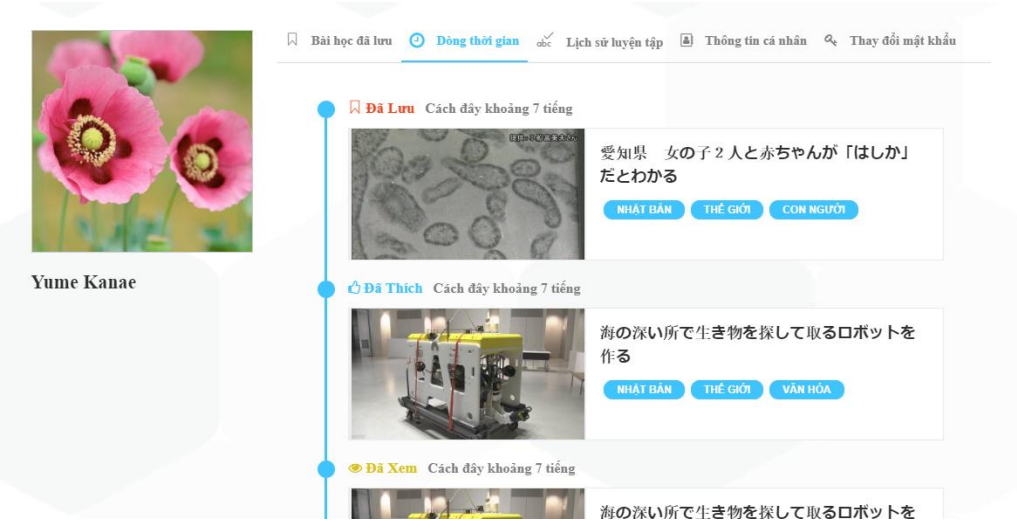
Trang cá nhân quản lý thông tin

Người dùng sẽ thấy được danh sách bài học mình đã lưu, lịch sử hoạt động, lịch sử làm bài tập, thay đổi thông tin cá nhân và đổi mật khẩu.

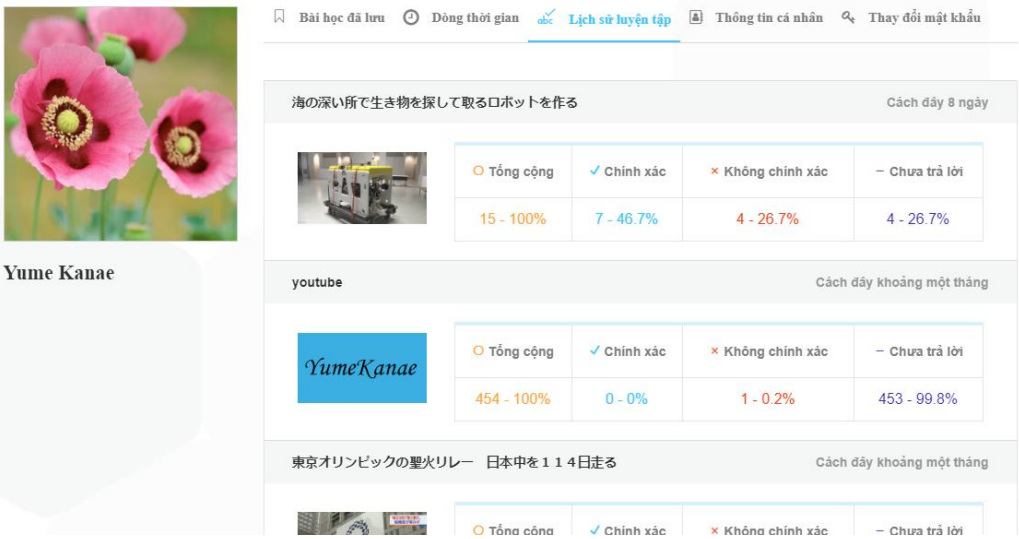
Đối với các màn hình xem bài học đã lưu, lịch sư hoạt động, lịch sử làm bài tập, hệ thống không hiển thị hết thông tin hiện có ngay mà tự động thêm dữ liệu khi có sự kiện scroll của người dùng.



Hình 42 Giao diện xem bài học đã lưu ở trang cá nhân



Hình 43 Giao diện lịch sử hoạt động ở trang cá nhân



Hình 44 Giao diện lịch sử luyện tập ở trang cá nhân

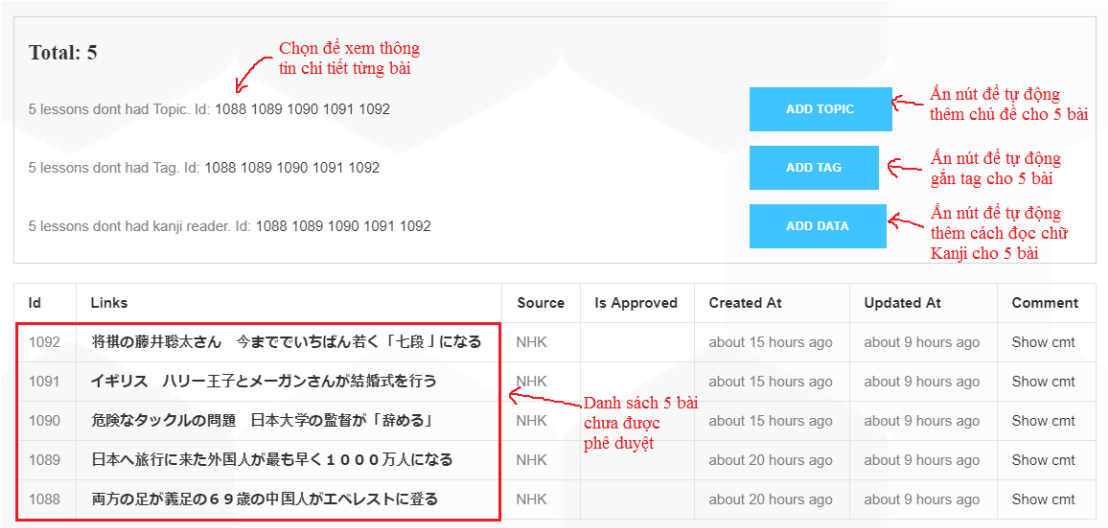
Trang admin quản lý thông tin chung

Ở trang chủ hiển thị ra các thông tin khái quát nhất để quản trị viên nắm bắt được và chọn nhanh đến các trang xem chi tiết. Đó là thông tin bài học như tổng số lượng bài học, số bài học của từng trang lấy dữ liệu, số bài học chưa được phê duyệt của từng trang dữ liệu. Tiếp theo là thông tin về người dùng, số người dùng đang hoạt động và số người đã bị block. Ngoài ra ở trang quản lý chung còn có đường dẫn đến trang huấn luyện dữ liệu và trang quản lý thông tin góp ý của người dùng.

Hình 45 Trang admin quản lý thông tin chung

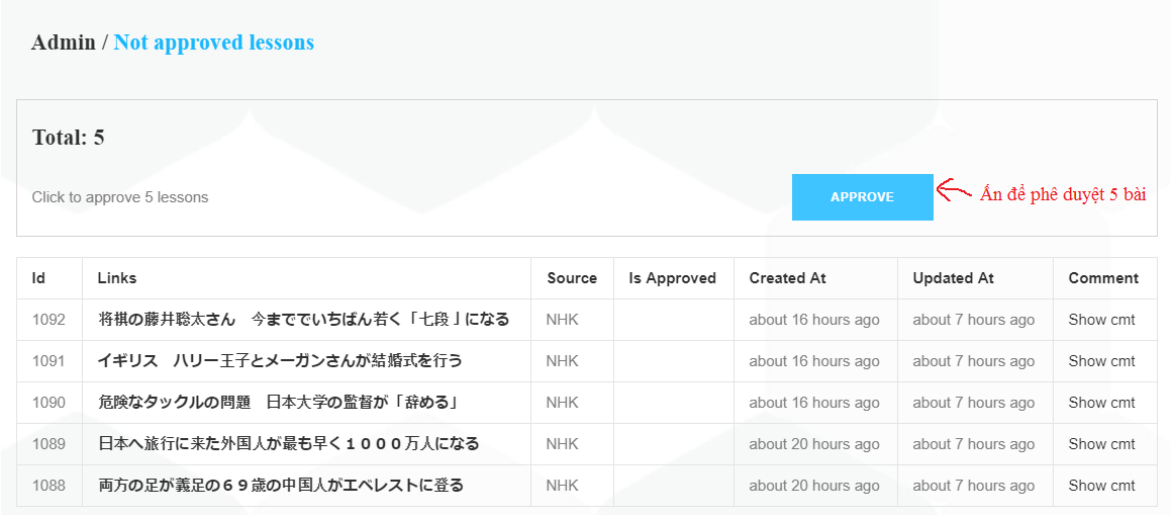
Trang phê duyệt bài học

Hệ thống hiển thị danh sách các bài học vừa được bộ Crawler lấy về, chưa phê duyệt để quản trị viên tiến hành các chức năng thêm dữ liệu như tự động phân loại theo chủ đề, gắn tag và thêm phiên âm cách đọc chữ Kanji cho nội dung bài học.



Hình 46 Giao diện trang phê duyệt bài học

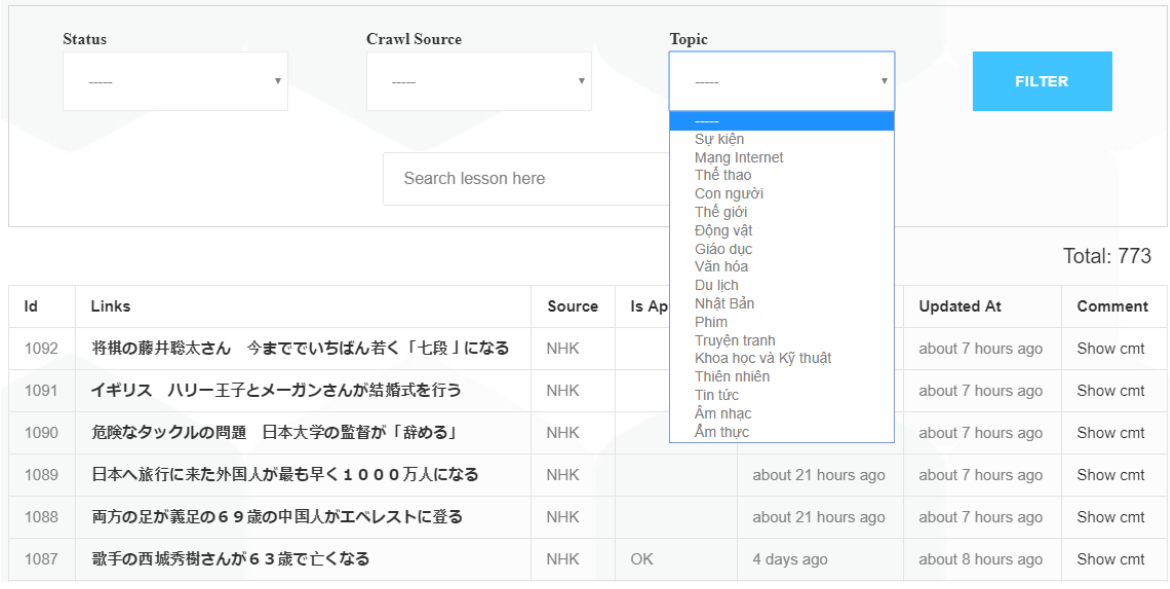
Sau khi các bài học này đã được thêm đầy đủ thông tin, quản trị viên sẽ phê duyệt bài học để hiển thị cho người dùng thấy.



Hình 47 Giao diện trang phê duyệt bài học

Trang admin hiển thị danh sách bài học

Trang này hiển thị danh sách tất cả các bài học có phân theo trang và một số thông tin tiêu biểu của mỗi bài. Quản trị viên cũng có thể chọn chỉ xem bài theo trạng thái (đã phê duyệt, chưa phê duyệt), xem danh sách bài theo nguồn lấy dữ liệu và theo chủ đề của bài.



Hình 48 Trang admin hiển thị danh sách bài học

Quản lý thông tin chi tiết của từng bài

Hệ thống sẽ hiển thị danh sách các chủ đề và các từ có thể dùng làm tag để quản trị viên lựa chọn gắn cho bài học. Bên dưới là thông tin chi tiết về nội dung của bài, và các thông tin liên quan của bài học.

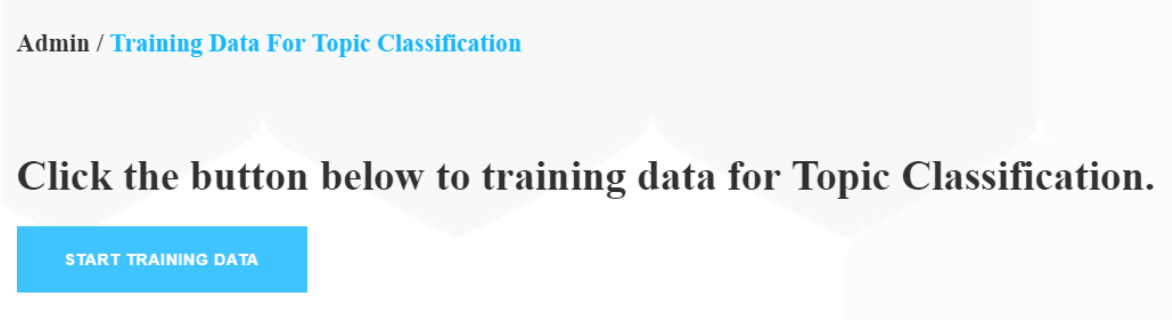
Hình 49 Quản lý thông tin chi tiết của bài học

Hình 50 Quản lý thông tin chi tiết của bài học

Huấn luyện dữ liệu

Để việc tự động phân loại bài học theo chủ đề và gắn tag đạt kết quả chính xác nhất, quản trị viên cần thường xuyên huấn luyện dữ liệu. Quản trị viên chỉ cần ấn nút bắt đầu huấn luyện và dữ liệu phân tích được sẽ tự động cập nhật vào CSDL.

Hình 51 Giao diện huấn luyện dữ liệu cho chức năng gắn tag



Hình 52 Giao diện huấn luyện dữ liệu cho chức năng phân loại bài học

4.4 Kiểm thử

Dưới đây là kịch bản kiểm thử cho một số chức năng quan trọng của trang Web.

STT	Chức năng	Kịch bản kiểm thử	Kết quả mong đợi
1	Đăng nhập	1. Ấn vào nút đăng nhập. 2. Nhập giá trị email và mật khẩu vào ô input. 3. Ấn nút tạo tài khoản.	1. Hiển thị form đăng nhập. 2. Kiểm tra dữ liệu nhập vào, nếu nhập sai hiển thị thông báo ngay bên dưới ô input. 3. Hệ thống kiểm tra dữ liệu nhập vào, nếu đúng hết thì thêm tài khoản mới vào CSDL và cho người tài khoản đó đăng nhập luôn vào hệ thống.

2	Reset mật khẩu	1. Chọn "Quên tài khoản" ở form đăng nhập. 2. Nhập email vào ô input. 3. Kiểm tra email và ấn vào đường dẫn reset mật khẩu, nhập mật khẩu mới.	1. Hiển thị form nhập mật khẩu cần reset. 2. Nếu địa chỉ email đúng thì hiển thị gửi thông tin reset mật khẩu đến email đó, không đúng thì hiển thị thông báo nhập lại email. 3. Cập nhật mật khẩu mới vào CSDL.
3	Xem danh sách bài học theo chủ đề	1. Chọn chủ đề muốn xem. 2. Kéo xuống để hiển thị thêm bài học.	1. Đến trang chủ đề và hiển thị 9 bài gần nhất. 2. Cập nhật thêm dữ liệu và hiển thị.
4	Tìm kiếm bài học	1. Nhập tên bài học cần tìm kiếm bằng tiếng Nhật hoặc tiếng Anh vào ô input tìm kiếm ở trang chủ. 2. Ấn vào icon tìm kiếm bên cạnh ô input. 3. Nhập thông tin tìm kiếm khác vào ô input ở trang tìm kiếm.	1. Hiển thị danh sách 5 bài học phù hợp nhất với kết quả tìm kiếm và có thời gian tạo gần nhất. 2. Đến trang tìm kiếm với nội dung tìm kiếm được nhập ở ô input ở trang chủ và hiển thị hết tất cả các kết quả phù hợp theo thứ tự thời gian tạo. 3. Hiển thị tất cả các kết quả phù hợp với nội dung tìm kiếm.
5	Làm bài tập điền từ vào chỗ trống	1. Ấn nút làm bài tập ở trang chi tiết bài học. 2. Ấn nút xác nhận làm bài tập. 3. Nhập câu trả lời sau đó ấn nút "Submit" hoặc khi hết thời gian làm bài. 4. Ấn nút làm lại bài.	1. Kiểm tra người dùng chưa đăng nhập hiển thị form đăng nhập. Nếu đăng nhập rồi hiển thị thông báo xác nhận bắt đầu làm bài tập. 2. Hiển thị nội dung bài học đã ẩn đi các từ quan trọng. 3. Hiển thị kết quả dưới dạng thống kê số lượng câu đúng, sai và đáp án chi tiết cho từng câu. Lưu kết quả vào CSDL. 4. Hiển thị bài điền từ vào chỗ trống đã ẩn đi các từ khác so với lần trước.
6	Bình luận về bài học	1. Đặt con trỏ chuột vào ô input để nhập nội dung bình luận. 2. Nhập nội dung bình luận và ấn nút "enter". 3. Chọn chỉnh sửa bình luận. 4. Ấn nút "enter"	1. Kiểm tra người dùng chưa đăng nhập hiển thị form đăng nhập. 2. Lưu bình luận vào CSDL. 3. Hiển thị ô input với giá trị là nội dung bình luận hiện tại. 4. Cập nhật nội dung bình luận vào CSDL.

		5. Chọn xóa bình luận	5. Xóa nội dung bình luận trong CSDL.
7	Xem trang cá nhân	1. Chọn xem danh sách các bài học đã lưu và kéo xuống để xem thêm thông tin. 2. Chọn xem lịch sử hoạt động và kéo xuống để xem thêm thông tin. 3. Chọn xem lịch sử luyện tập và kéo xuống để xem thêm thông tin. 4. Chọn mục thông tin cá nhân và thay đổi tên hiển thị trên trang Web, thay đổi avatar. 5. Chọn mục thay đổi mật khẩu rồi nhập mật khẩu hiện tại, mật khẩu mới, mật khẩu xác nhận.	1. Hiển thị danh sách 9 bài học theo thứ tự thời gian lưu gần nhất. Khi có hành động kéo xuống thì tự động tải thêm dữ liệu và hiển thị. 2. Hiển thị danh sách 9 hoạt động gần nhất. Khi có hành động kéo xuống thì tự động tải thêm dữ liệu và hiển thị. 3. Hiển thị kết quả 9 lần làm bài tập gần nhất. Khi có hành động kéo xuống thì tự động tải thêm dữ liệu và hiển thị. 4. Cập nhật các giá trị mới vào CSDL và thông báo đã cập nhật thành công hoặc lỗi nếu có. 5. Kiểm tra thông tin người dùng nhập vào. Cập nhật mật khẩu mới nếu người dùng nhập đúng, thông báo lỗi nếu nhập sai.
8	Phê duyệt tất cả bài học mới được lấy dữ liệu (dành cho quản trị viên)	1. Chọn xem danh sách bài học chưa được phê duyệt. 2. Ấn nút thêm dữ liệu. 3. Ấn nút thêm chủ đề. 4. Ấn nút gắn tag. 5. Ấn nút phê duyệt bài học.	1. Hiển thị danh sách bài học chưa được phê duyệt. 2. Tự động thêm cách đọc chữ Kanji cho tất cả các bài chưa phê duyệt. 3. Tự động thêm chủ đề cho tất cả các bài chưa phê duyệt. 4. Tự động gắn tag cho tất cả các bài chưa phê duyệt. 5. Chuyển trạng thái bài học từ chưa phê duyệt sang đã phê duyệt.
9	Thêm, sửa, xóa thông tin bài học (dành cho quản trị viên)	1. Chọn bài học cần xem. 2. Thêm, sửa, xóa thông tin bài học. 2.1 Gắn tag cho bài, xóa tag 2.2 Thêm chủ đề cho bài, xóa chủ đề 2.3 Thêm, chỉnh sửa thông tin bài học. 3. Phê duyệt bài học.	1. Hiển thị tất cả thông tin bài học và các gợi ý chủ đề, tag từ hệ thống để quản trị viên lựa chọn. 2. Cập nhật các thay đổi vào CSDL. 3. Cập nhật trạng thái bài học.

10	Huấn luyện dữ liệu (dành cho quản trị viên)	1. Chọn huấn luyện dữ liệu phục vụ phân loại bài học theo chủ đề. 2. Chọn huấn luyện dữ liệu phục vụ chức năng gắn tag.	1. Cập nhật dữ liệu mới vào bảng topic_words. 2. Cập nhật dữ liệu mới vào bảng tag_words.

Bảng 31 Bảng phương pháp kiểm thử

Kết quả kiểm thử trên các trình duyệt máy tính:

	Firefox	**Chrome**	**Safari**
Đăng nhập	Pass	Pass	Pass
Reset mật khẩu	Pass	Pass	Pass
Xem danh sách bài học theo chủ đề	Pass	Pass	Pass
Tìm kiếm bài học	Pass	Pass	Pass
Làm bài tập điền từ vào chỗ trống	Pass	Pass	Pass
Bình luận về bài học	Pass	Pass	Pass
Xem trang cá nhân	Pass	Pass	Pass
Phê duyệt tất cả bài học mới được lấy dữ liệu	Pass	Pass	Pass
Thêm, sửa, xóa thông tin bài học	Pass	Pass	Pass
Huấn luyện dữ liệu	Pass	Pass	Pass

Bảng 32 Kết quả kiểm thử trên trình duyệt máy tính

Kết quả kiểm thử trên các trình duyệt điện thoại:

	Firefox Samsung A5 (2015)	**Chrome Iphone6**	**Safari IphoneX**
Đăng nhập	Pass	Pass	Pass
Reset mật khẩu	Pass	Pass	Pass

Xem danh sách bài học theo chủ đề	Pass	Pass	Pass
Tìm kiếm bài học	Pass	Pass	Pass
Làm bài tập điền từ vào chỗ trống	Pass	Pass	Pass
Bình luận về bài học	Pass	Pass	Pass
Xem trang cá nhân	Pass	Pass	Pass
Phê duyệt tất cả bài học mới được lấy dữ liệu	Pass	Pass	Pass
Thêm, sửa, xóa thông tin bài học	Pass	Pass	Pass
Huấn luyện dữ liệu	Pass	Pass	Pass

Bảng 33 Kết quả kiểm thử trên trình duyệt điện thoại

4.5 Triển khai

Trang web đã được triển khai trên server được cung cấp bởi Amazon Web Service, cài hệ điều hành Ubuntu 16.04 có các thông số: ổ đĩa SSD 30GB, dung lượng RAM 1GB, vCPU 1, đặt tại Datacenter Tokyo, Nhật Bản. Địa chỉ truy cập trang Web là https://yumekanae.data-io.stream

Các bước thực hiện để triển khai trang Web hoạt động trên Server:

1. Truy cập vào Server thông qua giao thức SSH.
2. Cài đặt apache.
3. Cài đặt PHP 7.0
4. Cài đặt mysql.
5. Cài đặt git.
6. Tải code từ git về Server lưu ở đường đẫn /var/www/html rồi set quyền truy cập tương ứng.
7. Chạy lệnh tự động tạo key ứng dụng của laravel và chỉnh sửa một số cài đặt của trang Web.

Chương 5 Các giải pháp và đóng góp nổi bật

Đây là chương quan trọng nhất giới thiệu về các chức năng chính và đóng góp nổi bật của cả hệ thống. Đó là chức năng tự động sinh bài tập điền từ còn thiếu vào chỗ trống, tự động thêm phiên âm cách đọc cho từ Kanji trong tiếng Nhật, gợi ý bài học có thể người dùng quan tâm, phân loại bài học theo chủ đề có sẵn sử dụng thật toán Naive Bayes, gợi ý đánh tag.

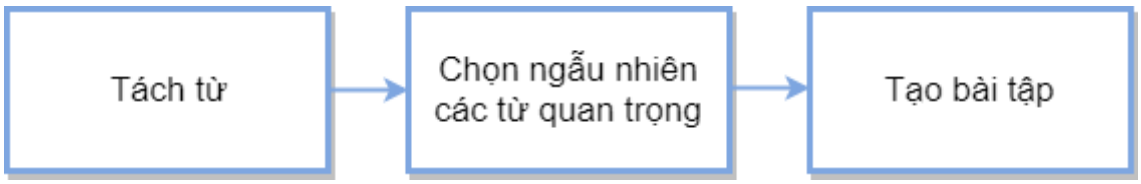
5.1 Tự động sinh bài tập điền từ còn thiếu vào chỗ trống

Hầu như ở các trang web hay ứng dụng khác chỉ cho người dùng nghe bài học và hiển thị nội dung bài học dưới dạng text. Việc chỉ nghe và nhìn nội dung bài học rất dễ gây nhàm chán và không nhớ được từ mới trong bài. Bởi vậy, em đã phát triển thêm tính năng tự động sinh bài tập điền từ vào chỗ trống cho mỗi bài học.

Bài điền từ sẽ được tự động sinh ra mỗi khi người dùng chọn chức năng làm bài tập. Người dùng có thể làm bài tập ứng với một bài học nhiều lần, do đó chức năng làm bài tập điền từ cần đáp ứng các yêu cầu:

(i) Các từ được ẩn đi phải là các từ quan trọng, mang nhiều ý nghĩa trong câu.
(ii) Khi người dùng làm một bài nhiều lần thì các từ ẩn đi đối với mỗi lần làm phải khác nhau.
(iii) Khi hết thời gian làm bài hoặc khi người dùng gửi câu trả lời, hệ thống sẽ hiển thị kết quả thống kê và kết quả chi tiết.

Quá trình tự động sinh bài tập điền từ còn thiếu vào chỗ trống gồm 3 bước:



Bước 1:

+ Sử dụng công cụ Mecab tách nội dung bài học thành từng từ. Trong khi tách, dùng biến đếm để xác định vị trí của từng từ trong câu và vị trí của câu trong bài.

+ Lưu thông tin của tất cả các từ tách được vào mảng words. Mỗi từ trong mảng words đều có các thông tin:

- id: lưu vị trí của từ trong câu
- value: giá trị của từ
- type: thể loại của từ (danh từ, động từ, tính từ, trợ từ, dạng số, ký tự …)
- hiragana: kiểu viết khác của từ dưới dạng chữ hiragana
- sentence: vị trí của câu mà từ đó thuộc vào
- is_question: dùng để xác định xem từ này có là từ ẩn đi không (mặc định ban đầu là false)

Bước 2: Lấy ngẫu nhiên 20% các từ trong mảng words có giá trị "type" là danh từ, động từ hoặc tính từ và lưu vào mảng answers. Mảng answers chứa thông tin của các từ ẩn đi. Cập nhật giá trị "is_question" của các từ được chọn đó trong mảng words thành true.

Bước 3: Ghép giá trị "value" của tất cả từ trong mảng words lại với nhau. Trong khi ghép kiểm tra giá trị "is_question" của từ. Nếu là true thì không ghép giá trị "value" của từ mà thay vào đó là ô input để khi hiển thị lên trình duyệt người dùng sẽ nhập câu trả lời của mình vào ô input này.

Đoạn văn bản sau khi ghép được chính là nội dung bài tập đã ẩn đi các từ quan trọng. Chuyển bài tập đó và mảng answers được tạo ở bước 2 đến Client. Sau khi làm xong bài tập, bên phía Client sẽ dựa vào giá trị mảng answers và nội dung người dùng nhập vào để kiểm tra đáp án và hiển thị kết quả. Đồng thời kết quả đó cũng được gửi về Server để lưu vào CSDL.

Việc chuyển đoạn mã kiểm tra kết quả sang phía Client giúp giảm bớt công việc cần xử lý cho Server, tránh tình trạng quá tải cho Server. Các từ ẩn đi được lấy ngẫu nhiên 20% trong số các từ là danh từ, động từ, tính từ trong bài nên thỏa mãn yêu cầu khác nhau đối với mỗi lần làm bài tập và là những từ mang ý nghĩa quan trọng.

5.2 Tự động thêm phiên âm cách đọc cho từ Kanji trong tiếng Nhật

Dữ liệu các bài học đều được em lấy từ trang thời sự, tin tức của Nhật nên có nội dung chính xác, phong phú nhưng chữ Kanji trong bài chưa có phiên âm cách đọc dưới dạng chữ Hiragana. Chữ Kanji hay còn gọi là Hán tự, được người Trung Quốc phát minh ra và người Nhật đã tiếp nhận, sử dụng nó cùng với chữ Hiragana và chữ Katakana của tiếng Nhật. Mọi chữ Kanji đều có cách đọc và cách viết tương ứng dưới dạng chữ Hiragana.

Ví dụ với đoạn tiếng Nhật “研究会に参加するつもりだったが、思い返して行かないことにした”, dịch sang tiếng Việt là “Tôi dự định tham gia buổi nghiên cứu nhưng nghĩ lại và không tham gia nữa”. Công cụ Mecab hỗ trợ tách văn bản đầu vào thành từng từ có nghĩa và trả về cách đọc dưới chữ Hiragana của tất cả các từ có chứa chữ Kanji trong đoạn trên tương ứng như sau.

Từ tách được	Chữ Hiragana	Từ tách được	Chữ Hiragana
研究 (chữ Kanji)	けんきゅう	、	、
会 (chữ Kanji)	かい	思い返し(chữ Kanji + chữ Hiragana)	おもいかえし
に	に	て	て
参加 (chữ Kanji)	さんか	行か (chữ Kanji + chữ Hiragana)	いか
する	する	ない	ない
つもり	つもり	こと	こと
だっ	だっ	に	に
た	た	し	し
が	が	た	た

Với các từ chỉ gồm chữ Kanji thì chỉ cần thêm cách đọc dạng chữ Hiragana lên trên chữ Kanji "研究(けんきゅう)", "会(かい)", "参加(さんか)". Tuy nhiên với từ gồm cả chữ Kanji và Hiragana là "思い返し" và "行か" thì cần xử lý để chỉ hiển thị phiên âm cách đọc của chữ Kanji thôi, chữ Hiragana giữ nguyên "思(おも)い返(かえ)し", "行(い)か". Do đó, em đã xây dựng thuật toán để tự động thêm cách đọc dưới dạng chữ Hiragana cho tất cả chữ Kanji trong văn bản đầu vào, sử dụng công cụ Mecab hỗ trợ tách từ.

Tư tưởng của thuật toán là tách chữ Kanji và cách đọc dạng Hiragana được trả về từ Mecab thành từng kí tự. Sau đó tìm điểm giống nhau và khác nhau giữa các kí tự vừa tách được, rồi ghép các kí tự đó lại theo đúng thứ tự. Thuật toán chi tiết như sau:

```
word_array = Mecab tách từ (input_kanji_document)
output_kanji_document = ""
for word in word_array
        if word is not kanjiType
                output_kanji_document += word
        else
                Step 1: Tách từ thành từng kí tự lưu vào mảng
                    ▪ Tách word thành từng kí tự lưu vào mảng kanji_array
                    ▪ Tách kiểu chữ Hiragana của word thành từng kí tự lưu vào
                      mảng hiragana_array
                Step 2: Tìm điểm giống nhau và khác nhau
                    ▪ kanji_diff  = diff( kanji_array, hiragana_array)
                    ▪ hiragana_diff  = diff(hiragana_array, kanji_array)
                    ▪ hiragana_same = diff(hiragana_array, kanji_array)
                Step 3: Sử dụng hàm associateWord ghép các phần tử trong mảng
                    ▪ kanji_diff_word = associateWord(kanji_diff)
                    ▪ hiragana_diff _word = associateWord(hiragana_diff )
                    ▪ hiragana_same_word = associateWord(hiragana_same)
                Step 4:
                    ▪ Ghép lần lượt từng kí tự trong 3 mảng kanji_diff_word,
                      hiragana_diff_word, hiragana_same_word với nhau được từ
                      Kanji kèm cách đọc Hiragana
                    ▪ output_kanji_document += từ Kanji kèm cách đọc Hiragana
        end
end
```

Trong thuật toán trên sử dụng hàm associateWord() dùng để ghép các kí tự liên tiếp nhau lại với nhau. Dưới đây là ví dụ của từ "思い返し" và với giá trị mảng kanji_diff,

hiragana_diff , hiragana_same ở Step 2 như sau thì giá trị các mảng kí tự sau khi ghép lại ở Step 3 là:

kanji_diff = [0 => '思', 2 => '返']	kanji_diff_word = associateWord(kanji_diff) = ['思', '返']
hiragana_diff = [0 => 'お', 1 => 'も', 3 => 'か', 4 => 'え']	hiragana_diff_word = associateWord(hiragana_diff) ['おも','かえ']
hiragana_same = [1 => 'い', 3 => 'し']	hiragana_same_word = associateWord(hiragana_same) ['い','し']

Sau đó ghép lần lượt từng kí tự trong 3 mảng kanji_diff_word, hiragana_diff_word, hiragana_same_word, sử dụng thẻ của html là <ruby> và <rt> để hỗ trợ hiển thị chữ Hiragana lên trên chữ Kanji, kết quả hiển thị trên trình duyệt của từ là "思(おも)い返(かえ)し".

Kết quả sau khi sử dụng thuật toán thêm phiên âm cách đọc cho từ Kanji như sau:

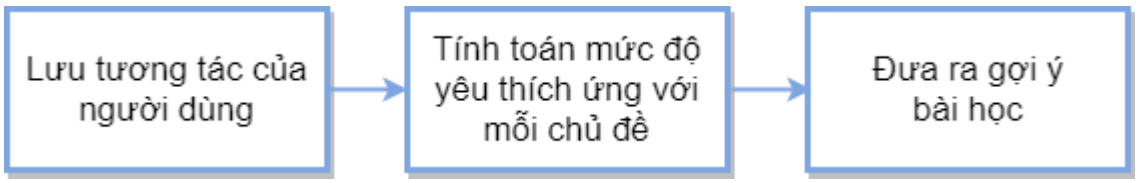
Văn bản đầu vào	**Văn bản đầu ra**
研究会に参加するつもりだったが、思い返して行かないことにした	研究会(けんきゅうかい)に参加(さんか)するつもりだったが、思(おも)い返(かえ)して行(い)かないことにした

Thuật toán sử dụng tuy đơn giản nhưng đây lại là chức năng quan trọng và vô cùng cần thiết nhưng các trang web khác chưa có. Chức năng này giúp hỗ trợ người dùng biết được ngay cách đọc của chữ Kanji mà không phải tra ở từ điển hay ứng dụng khác, từ đó việc học tiếng Nhật cũng đơn giản và hiệu quả hơn.

5.3 Gợi ý các bài học có thể người dùng quan tâm

Nhằm thuận tiện hơn cho người học, hệ thống sẽ tự động gợi ý các bài học mà có thể người dùng quan tâm. Các bài học này là khác nhau đối với mỗi người dùng và được hệ thống đưa ra gợi ý dựa trên tương tác của người dùng.

Các bước xử lý để đưa ra gợi ý cho người dùng



Bước 1: Với mỗi tương tác của người dùng là xem bài, thích, lưu bài học, làm bài tập, hệ thống tự động lưu các thông tin gồm user_id, lesson_id, action_type vào bảng user_logs. Mô tả chi tiết của bảng user_logs đã được trình bày trong phần Thiết kế cơ sở dữ liệu 4.2.3

Bước 2: Sau khi có tương tác của người dùng, hệ thống sẽ tính toán mức độ yêu thích của người dùng ứng với từng chủ đề. Đây cũng là bước quan trọng nhất, quyết định các bài học sẽ gợi ý cho người dùng.

Input: Bảng user_logs lưu tương tác người dùng

Output: Danh sách các chủ đề quan tâm

Thuật toán:

- Tập T lưu danh sách tất cả chủ đề ứng với các bài học mà người dùng có tương tác
- Trong mỗi chủ đề của tập T, đếm số bài học ứng với mỗi kiểu tương tác. Có 4 kiểu tương tác là xem, thích, lưu và làm bài tập (xem: Num_{view}, thích: Num_{like}, lưu: Num_{save}, làm bài tập: $Num_{exercise}$)
- For t in T

$$Weight_t = Weight_{view} + Weight_{like} + Weight_{save} + Weight_{exercise}$$
$$= weight_{view} * Num_{view} + weight_{like} * Num_{like}$$
$$+ weight_{save} * Num_{save} + weight_{exercise} * Num_{exercise}$$

- Sắp xếp T theo thứ tự giá trị Weight giảm dần.

Trong đó trọng số ứng với từng hành động của người dùng: $weight_{view} = 1$, $weight_{like} = 2$, $weight_{save} = 3$, $weight_{exercise} = 4$. Giá trị các trọng số này được đặt dựa theo tính chủ quan của cá nhân em qua quá trình trải nghiệm học tiếng Nhật, dùng thử các trang web hoặc ứng dụng học trực tuyến khác.

- Trước tiên, người dùng chọn vào xem bài, sau đó tùy thuộc vào mỗi người sẽ có thêm các hành động khác để thể hiện rõ sở thích. Vì vậy trọng số ứng với hành động xem có giá trị nhỏ nhất bằng 1.
- Người dùng sau học cảm thấy thích sẽ có thêm tương tác ấn vào nút like bài học. Bởi vậy trọng số của tương tác này bằng 2, lớn hơn của tương tác xem.

- Chức năng lưu bài học để sau đó có thể xem lại dễ dàng hơn mà không phải tìm kiếm. Do đó có thể người dùng đang thấy hứng thú với bài học đó thì mới có muốn xem lại. Em đặt trọng số của tương tác này là 3.
- Với một số lượng bài học rất nhiều, chỉ khi nào thực sự quan tâm đến bài học, chủ đề của bài thì người dùng mới chọn làm bài tập của bài đó. Trọng số của tương tác này bằng 4.

Bước 3: Lấy một số bài học mới nhất mà người dùng chưa xem từ một số chủ đề trong tập T làm gợi ý.

Xác định chủ đề gợi ý dựa trên chênh lệch độ yêu thích. Tập S lưu các chủ đề gợi ý được xác định như sau:

$T_0 - T_1 > \varepsilon_1 : S = \{ T_0 \}$
$T_0 - T_1 <= \varepsilon_1$
- $T_1 - T_2 > \varepsilon_2 : S = \{ T_0 , T_1 \}$
- $T_1 - T_2 <= \varepsilon_2 : S = \{ T_0, T_1, T_2 \}$

Trong đó: $\varepsilon_1 = 5$, $\varepsilon_2 = 3$

Qua thực nghiệm em thấy giá trị chênh lệch độ yêu thích $\varepsilon_1 = 5$, $\varepsilon_2 = 3$ là đủ để lựa chọn gợi ý thêm các bài học thuộc chủ đề $T_1 - T_2$. Với N bài học đưa ra gợi ý là thì số bài học tương ứng của từng chủ đề được xác định:

$Num_S = 1$ => Lấy N bài thuộc T_0
$Num_S = 2$ => Lấy 0.75*N bài thuộc T_0, 0.25*N bài thuộc T_1
$Num_S = 3$ => Lấy 0.5*N bài thuộc T_0, 0.25*N bài thuộc T_1, 0.25*N bài thuộc T_2

Vì một bài học có thể thuộc nhiều chủ đề, do đó để danh sách bài học gợi ý không bị trùng lặp thì khi truy vấn dữ liệu cần kiểm tra thêm điều kiện bài học chưa được gợi ý từ các chủ đề khác trong tập S.

Chức năng gợi ý này đảm bảo các yêu cầu:

(i) Các bài học được gợi ý là khác nhau đối với mỗi người dùng.

(ii) Các bài học được gợi ý của mỗi người không phải là một số bài cố định mà thay đổi khác nhau dựa trên tương tác của người dùng trong thời gian gần nhất.

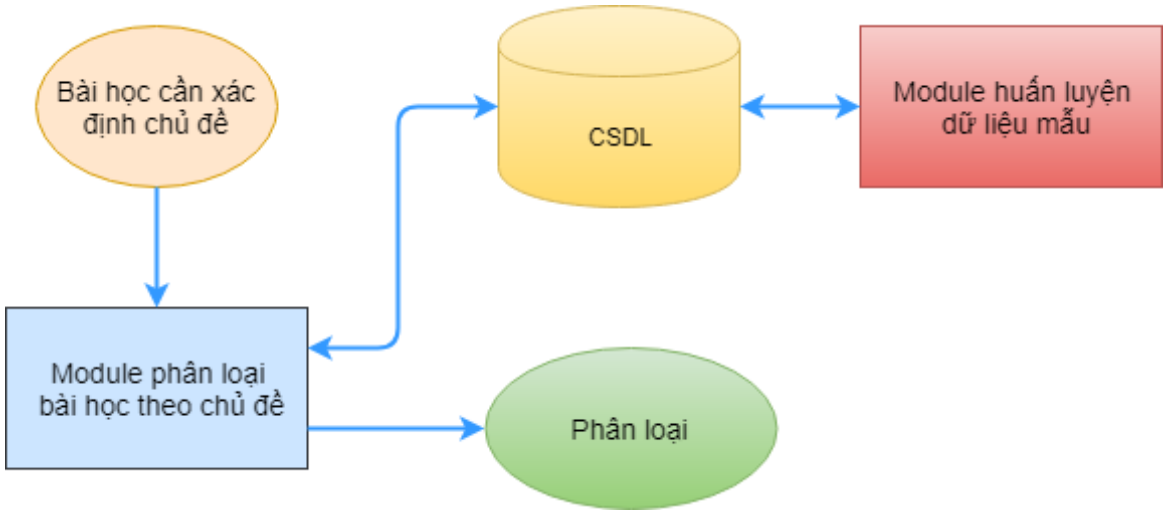
(iii) Các bài mà người học đã xem sẽ không nằm trong danh sách gợi ý.

5.4 Phân loại bài học theo chủ đề có sẵn sử dụng thuật toán Naive Bayes

Để thuận tiện hơn cho người dùng trong việc lựa chọn bài học thì cần thiết phải phân loại bài học theo chủ đề. Nhưng với số lượng bài học hiện tại là hơn 700 bài, con số này sẽ tăng lên theo từng ngày thì việc phân loại thủ công sẽ rất mất thời gian và đòi hỏi người phân loại phải biết về tiếng Nhật.

Do đó, em đã phát triển tính năng tự động phân loại bài học theo chủ đề có sẵn sử dụng thuật toán Naive Bayes. Lý do chọn sử dụng thuật toán và chi tiết về thuật toán đã được trình bày trong phần 3.4 của Chương 3.

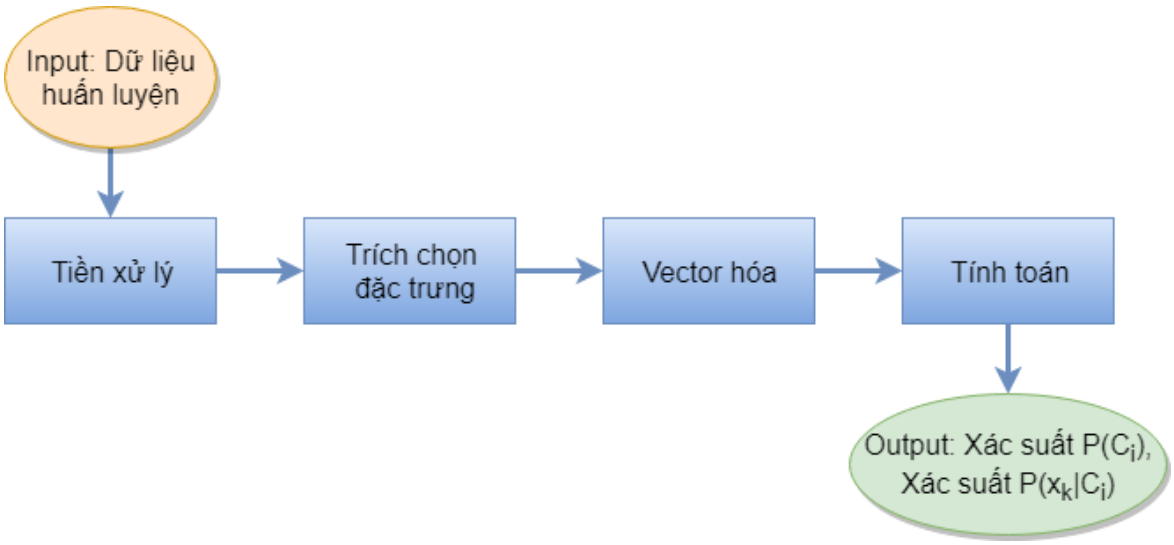
Tính năng phân loại bài học theo chủ đề được xây dựng bởi 2 module chính được mô tả như hình vẽ:



Hình 53 Thiết kế tổng quan chức năng phân loại bài học

5.4.1 Module huấn luyện dữ liệu mẫu

Mục đích của module này là huấn luyện dữ liệu, tính ra các giá trị $P(C_i)$ và $P(x_k|C_i)$ dùng cho thuật toán phân loại Naive Bayes. Các bước huấn luyện dữ liệu được thể hiện như trong hình sau.



Hình 54 Module huấn luyện dữ liệu phân loại bài học

Dữ liệu huấn luyện

Dữ liệu đã được phân loại sẵn theo 17 chủ đề được lấy từ nguồn https://newsinslowjapanese.com. 17 chủ đề gồm:

Sự kiện	Thế giới	Du lịch	Âm nhạc	Ẩm thực
Internet	Động vật	Nhật Bản	Thiên nhiên	Thể thao
Giáo dục	Phim	Tin tức	Con người	Văn hóa
Truyện tranh	Khoa học Kỹ thuật			

Bước 1: Tiền xử lý

Quá trình tiền xử lý rất quan trọng, ảnh hưởng lớn đến kết quả phân loại. Do dữ liệu được lấy từ nhiều nguồn khác nhau nên thiếu tính nhất quán. Vì vậy dữ liệu cần xử lý ở bước này để làm sạch, chuyển đổi trước khi thực hiện các bước tiếp theo. Tiền xử lý sẽ giúp nâng cao hiệu suất phân loại và giảm độ phức tạp của thuật toán.

Công đoạn tiền xử lý sẽ thực hiện: Loại bỏ các ký tự đặc biệt như dấu câu, chữ số, phép toán học.

Dữ liệu trước khi bỏ ký tự đặc biệt	**Dữ liệu sau khi bỏ ký tự đặc biệt**
ロボットは長さが１ｍぐらいで箱のような形です。海の深さ２０００ｍの所まで行って、カメラで魚や貝などの生き物を見つけることができます。	ロボットは長さがぐらいで箱のような形です海の深さの所まで行ってカメラで魚や貝などの生き物を見つけることができます

Bước 2: Trích chọn đặc trưng

Quá trình trích chọn đặc trưng là quá trình loại bỏ các từ stop word được xuất hiện ở hầu hết các văn bản nhưng lại không mang nhiều ý nghĩa quan trọng như các trợ từ và từ nối.

Sau khi trích chọn, dữ liệu còn lại sẽ mang nhiều ý nghĩa, nhiều đặc trưng để việc phân loại đạt kết quả chính xác hơn. Các giá trị stop word được lưu trong bảng stop_words trong CSDL.

Dưới đây là một số từ stop word trong tiếng Nhật

さて	しかし	だから	そのため	そこで
のに	でも	そして	それに	どころか
または	一方	だって	つまり	いわば

Bước 3: Vector hóa

Một văn bản dạng thô (dạng chuỗi) cần được chuyển sang một dạng khác để tạo thuận lợi cho việc biểu diễn và tính toán. Một trong những mô hình đơn giản và thường được sử dụng là mô hình không gian vector.

Trong mô hình này, các văn bản được thể hiện trong một không gian có số chiều lớn, trong đó mỗi chiều của không gian tương ứng với một từ trong văn bản. Có thể biểu diễn một cách hình tượng như sau: mỗi văn bản D được biểu diễn dưới dạng X (vector đặc trưng cho văn bản D). Trong đó, vector X = (X_1, X_2, …, X_n) với X_i là các thuộc tính của vector X, mỗi thuộc tính X_i đều có giá trị trọng số tương ứng là x_i (1 <= i <= n) với n là số chiều của vector X.

Có nhiều cách khác nhau để tính trọng số cho vector đặc trưng của văn bản. Em đã chọn phương pháp word frequency weighting (trọng số tần suất từ), đó là đếm số lần xuất hiện của từ đó trong văn bản D.

Ví dụ sau đã qua các bước tiền xử lý và trích chọn đặc trưng, tập huấn luyện gồm 4 từ: “var”, “bit”, “chip”, “log”. Tập huấn luyện được chia thành 2 phân lớp là Math và Comp. Số lần xuất hiện của các từ như sau:

Docs	**"var"**	**"bit"**	**"chip"**	**"log"**	**Class**
Doc1	42	25	7	56	Comp
Doc2	10	28	45	2	Math
Doc3	11	25	22	4	Math
Doc4	33	40	8	48	Comp
Doc5	8	22	30	1	Math
Total	**104**	**140**	**112**	**111**	

Vector X đặc trưng cho văn bản huấn luyện được biểu diễn là X= ("var", "bit", "chip", "log"). Giá trị ứng với các thuộc tính của vector X là (104, 140, 112, 111)

Bước 4: Tính toán

$$\text{Xác suất } P(C_i) = \frac{\textbf{Số bài học thuộc chủ đề Ci}}{\textbf{Tổng số bài học của tất cả các chủ đề}}$$

C_i là các chủ đề trong 17 chủ đề. Giá trị $P(C_i)$ được lưu vào trường lesson_percentage trong bảng topics.

$$\text{Xác suất } P(\boldsymbol{X_k}|\mathbf{C_i}) = \frac{\textbf{Số lần xuất hiện của } \boldsymbol{X_k} \textbf{ trong chủ đề Ci}}{\textbf{Tổng số từ có trong chủ đề Ci}}$$

Với X_k ứng với mỗi từ trong dữ liệu huấn luyện đầu vào, C_i là các chủ đề trong 17 chủ đề. Các giá trị $P(X_k|C_i)$ thường rất bé, xấp xỉ 0 nên em đổi sang dạng dấu phẩy động trước khi lưu vào CSDL. Ví dụ giá trị $P(X_k|C_i) = 0.015 = 1.5*10^{-2}$ thì lưu 1.5 vào trường frequency, lưu 2 vào trường exponent trong bảng topic_words.

Với dữ liệu huấn luyện ở bước 4, giá trị $P(C_i)$ và $P(X_k|C_i)$ được tính như sau:

- Xác xuất các lớp C_i trong tập huấn luyện:

  $P(C_1 = \text{"Math"}) = 3/5 = 0.6$

  $P(C_2 = \text{"Comp"}) = 2/5 = 0.4$

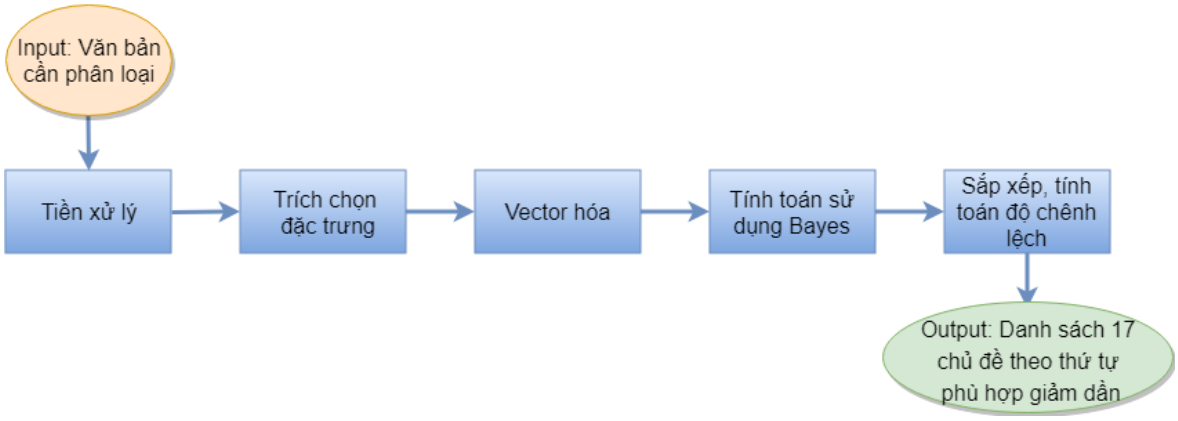
- Xác xuất $P(x_k|C_i)$

Lớp C1 = "Math". Lớp Math có 208 từ	**Lớp C2 = "Comp". Lớp Comp có 259 từ**
P(var | Math) = (10+11+8)/ 208 = 0.139	P(var | Comp) = (42+33)/ 259 = 0.290

P(bit | Math) = (28+25+22)/208 = 0.361 P(chip | Math) = (45+22+30)/208 = 0.466 P(log | Math) = (2+4+1) / 208 = 0.034	P(bit | Comp) = (25+40)/ 259 = 0.251 P(chip | Comp) = (7+8)/ 259 = 0.058 P(log | Comp) = (56+48) / 259 = 0.402

5.4.2 Module phân loại bài học theo các chủ đề

Mục đích của module này từ dữ liệu đã huấn luyện, sử dụng thuật toán Naive Bayes để phân loại bài học theo 17 chủ đề. Các bước huấn luyện dữ liệu được thể hiện theo các bước như trong hình.



Hình 55 Module phân loại bài học

Bước 1: Tiền xử lý

Thực hiện giống quá trình tiền xử lý với dữ liệu huấn luyện ở phần 5.5.1

Bước 2: Trích chọn đặc trưng

Thực hiện giống quá trình trích chọn đặc trưng với dữ liệu huấn luyện ở phần 5.5.1

Bước 3: Vector hóa

Thực hiện giống quá trình vector hóa với dữ liệu huấn luyện ở phần 5.5.1

Văn bản cần phân loại sẽ được biểu diễn thành vector $X^{new} = (X_1, X_2, \ldots, X_n)$.

Ví dụ: Sau đã qua các bước tiền xử lý và trích chọn đặc trưng, văn bản cần phân loại gồm 4 từ: “var”, “bit”, “chip”, “log”, “core”. Vector X^{new} được biểu diễn là X^{new} = (“var”, “bit”, “chip”, “log”, “core”). Tần suất xuất hiện của các từ trong văn bản cần phân loại hay giá trị ứng với các thuộc tính của vector X^{new} lần lượt là (23, 40, 15, 50, 30)

Bước 4: Tính toán sử dụng Naive Bayes

Các bước tính toán được mô tả trong thuật toán dưới đây:

foreach C_i (1 <= i <= 17)

value = 1

Văn bản cần phân loại ứng với vector $X^{new} = (X_1, X_2, ..., X_n)$

for X_k in X^{new} (1<= k <= n, thuộc tính X_k có giá trị trọng số tương ứng là x_k)

Kiểm tra giá trị X_k và C_i đã xuất hiện trong bảng topic_words chưa.

Đã xuất hiện	Chưa xuất hiện
value = value * x_k * $P(X_k\|C_i)$ được lấy từ CSDL.	value = value * x_k * 10^{-4} (Coi giá trị $P(X_k\|C_i)$ bằng 10^{-4})

end

end

Với cặp X_k và C_i không được tìm thấy trong CSDL nên chứng tỏ X_k chưa từng xuất hiện trong tập dữ liệu huấn luyện của C_i. Do đó, xác suất phụ thuộc $P(X_k|C_i)$ là rất nhỏ. Xác suất $P(X_k|C_i)$ trong bảng topic_words có giá trị nhỏ nhất là $8*10^{-4}$. Vì vậy, em chọn giá trị 10^{-4} ứng với giá trị $P(X_k|C_i)$ trong trường hợp này.

Theo dữ liệu huấn luyện ở phần 5.5.1 và ví dụ ở bước 3, kết quả tính toán phân loại cho văn bản ứng với vector X^{new} (23, 40, 15, 50, 30) là

Lớp "Math"	**Lớp "Comp"**
P(Math) * P(var|Math) * 23 * P(bit|Math) * 40 * P(chip|Math) * 15 * P(log|Math) * 50 * P(core|Math) * 30 = 0.6 * 0.139 * 23 * 0.361 * 40 * 0.466 * 15 * 0.034 * 50 * 10^{-4} * 30 = 0.987	P(Comp) * P(var|Comp) * 23 * P(bit|Comp) * 40 * P(chip|Math) * 15 * P(log|Comp) * 50 * P(core|Comp) * 30 = 0.4 * 0.29 * 23 * 0.251 * 40 * 0.058 * 15 * 0.402 * 50 * 10^{-4} * 30 = 1.405

Giá trị tính toán theo công thức Naive Bayes ở lớp "Comp" cao hơn, vậy văn bản này thuộc về lớp "Comp".

Bước 5: Sắp xếp, tính toán độ chênh lệch giữa các chủ đề

Bài học được phân loại vào lớp có giá trị tính toán ở bước 4 lớn nhất theo đúng công thức Naive Bayes: $\mathbf{P^{new} = Max (P(C_i)* \prod_{k=1}^{n} P(X_k|C_i))}$

Tuy nhiên, một bài học có thể thuộc nhiều chủ đề, đó đó em đã sắp xếp giá trị tính toán được của từng chủ đề theo thứ tự giảm dần, rồi chuyển về dạng % ứng với độ lớn kết quả tính toán. Tổng của 17 chủ đề là 100%. Do đó, quản trị viên có thể thấy được độ chênh lệch tương đối giữa các chủ đề để dễ dàng lựa chọn chủ đề để gán cho bài học hơn.

5.4.3 Tính năng tự động phân loại bài học theo chủ đề

Quản trị viên có 2 cách để phân loại chủ đề cho bài học:

(i) Vào màn hình chi tiết bài học và dựa vào gợi ý từ hệ thống để lựa chọn chủ đề cho bài học.

(ii) Lựa chọn chức năng tự động phân loại bài học theo chủ đề

Chức năng tự động phân loại bài học theo chủ đề cũng dựa trên kết quả của công thức Naive Bayes. Các chủ đề sau khi được sắp xếp theo giá trị tính toán công thức Naive Bayes giảm dần ở bước 5 mục 5.4.2 sẽ được lưu vào tập T.

Thuật toán sẽ chọn chủ đề của bài học là T_0 và đồng thời kiểm tra thêm giá trị của hai chủ đề là T_1 và T_2. Nếu độ chênh lệch của hai giá trị này so vớ T_0 trong một khoảng định trước thì cũng tương ứng được chọn làm chủ đề của bài.

$T_0 - T_1 > \varepsilon$ chọn chủ đề T_0

$T_0 - T_1 <= \varepsilon$ chọn chủ đề T_0, T_1

$T_0 - T_2 <= \varepsilon$ chọn chủ đề T_0, T_1, T_2

Với $\varepsilon = 1\ \%$

Vì đây là chức năng tự động phân loại nên em chọn độ chênh lệch nhỏ ε = 1 % để kết quả phân loại được chính xác.

Để đánh giá hiệu quả của thuật toán tự động phân loại bài học theo chủ đề có sẵn, sau đây em đưa ra thử nghiệm và kết quả tương ứng.

Kịch bản thử nghiệm

i. Chọn tập mẫu thử nghiệm là ngẫu nhiên 40 bài học thực tế của hệ thống.

ii. Phân loại bài học theo chủ đề bằng con người.

iii. Chạy thuật toán tự động phân loại bài học theo chủ đề và đo thời gian thực hiện.

iv. Đánh giá hiệu quả thông qua công thức.

Công thức đánh giá hiệu quả thuật toán tự động phân loại bài học theo chủ đề

Ứng với mỗi dữ liệu kiểm thử sẽ có 2 thông số để đánh giá là P viết tắt của Precision đo độ chính xác và R viết tắt của Recall là độ bao phủ với:

$$P = \frac{Số\ chủ\ đề\ phân\ loại\ đúng\ bằng\ thuật\ toán}{Tổng\ số\ chủ\ đề\ phân\ loại\ bằng\ thuật\ toán}$$

$$R = \frac{Số\ chủ\ đề\ phân\ loại\ đúng\ bằng\ thuật\ toán}{Tổng\ số\ chủ\ đề\ phân\ loại\ bằng\ con\ người}$$

Dưới đây là kết quả thử nghiệm và các thông số tương ứng.

ID bài học	Chủ đề do con người phân loại	Chủ đề hệ thống tự phân loại	Độ chính xác P	Độ bao phủ R	Thời gian thực hiện (s)
541	thế giới, du lịch	thế giới, du lịch	100%	100%	1.66
542	Nhật Bản, sự kiện	Nhật Bản, sự kiện	100%	100%	1.81
543	sự kiện, Nhật Bản	sự kiện, Nhật Bản	100%	100%	1.74
544	Nhật Bản, khoa học kỹ thuật	Nhật Bản, con người	50%	50%	1.53
545	Nhật Bản, tin tức	Nhật Bản, tin tức	100%	100%	1.23
546	thế giới, tin tức	thế giới, Nhật Bản, tin tức	66.6%	100%	1.41
547	thế giới	Nhật Bản, thế giới	50%	100%	1.32
548	thế giới, sự kiện	thế giới, Nhật Bản	50%	50%	1.11

549	Nhật Bản, sự kiện	Nhật Bản, sự kiện	100%	100%	1.90
550	tin tức, Nhật Bản	thế giới, Nhật Bản	50%	50%	1.82
551	sự kiện, con người	sự kiện, con người	100%	100%	1.06
552	thiên nhiên, Nhật Bản	thiên nhiên, Nhật Bản	100%	100%	1.23
553	ẩm thực, Nhật Bản, văn hóa	ẩm thực, Nhật Bản	100%	66.6%	1.06
554	Nhật Bản, tin tức	thế giới, Nhật Bản, tin tức	66.6%	100%	1.34
555	Nhật Bản, thế giới	Nhật Bản, thế giới	100%	100%	1.35
556	thế giới, con người	thế giới, internet, con người	66.6%	100%	1.21
557	khoa học kỹ thuật, thế giới	con người, thế giới	50%	50%	1.19
558	Nhật Bản, văn hóa	Nhật Bản, văn hóa	100%	100%	1.48
559	Nhật Bản, sự kiện	Nhật Bản, thế giới, sự kiện	66.6%	100%	1.72
560	Nhật Bản, thiên nhiên	Nhật Bản, con người	50%	50%	1.53
561	Nhật Bản, khoa học kỹ thuật	văn hóa, Nhật Bản	50%	50%	1.43
562	Nhật Bản, thế giới	Nhật Bản, thế giới	100%	100%	2.01
563	Nhật Bản, internet, sự kiện	Nhật Bản, internet	100%	66.6%	1.69
564	thế giới, khoa học kỹ thuật	Nhật Bản, thế giới	50%	50%	1.85

565	thế giới, tin tức	thế giới	100%	50%	1.65
566	Nhật Bản, tin tức	Nhật Bản, tin tức	100%	100%	1.73
567	thế giới, tin tức	thế giới, tin tức	100%	100%	1.89
568	Nhật Bản, thế giới	Nhật Bản, thế giới, văn hóa	66.6%	100%	1.53
569	Nhật Bản	Nhật Bản, thế giới	50%	100%	1.78
570	thế giới, con người	thế giới, con người	100%	100%	1.98
571	Nhật Bản, động vật	Nhật Bản	100%	50%	1.34
572	Nhật Bản, du lịch	Nhật Bản	100%	50%	1.56
573	Nhật Bản, tin tức, giáo dục	Nhật Bản, tin tức, văn hóa	66.6%	66.6%	1.34
574	Nhật Bản, động vật	Nhật Bản, văn hóa	50%	50%	1.42
575	thế giới, con người, tin tức	thế giới, con người	100%	66.6%	1.65
576	ẩm thực	ẩm thực	100%	100%	1.43
577	con người, Nhật Bản	thế giới, con người, Nhật Bản	66.6%	100%	1.34
578	Nhật Bản, ẩm thực	Nhật Bản, ẩm thực	100%	100%	1.69
579	Nhật Bản, con người	Nhật Bản	100%	50%	1.32
580	thế giới, con người	thế giới	100%	50%	1.45
Trung bình			**82%**	**80%**	**1.55 (s)**

Thời gian trung bình để phân loại chủ đề cho tập dữ liệu mẫu 40 bài là khá nhanh 1.55s. Kết quả trung bình độ chính xác và độ bao phủ của 40 mẫu kiểm thử là trên 80%. Con số này cho thấy hệ thống tự động phân loại bài học theo chủ đề có độ chính xác khá cao và đã xác định được hầu hết chủ đề để gắn cho bài học.

5.5 Gợi ý đánh tag

Tag được coi như là từ khóa, là đối tượng mà bài học tập trung giới thiệu. Vì tag liên quan đến nội dung của bài học nên để gắn tag phù hợp thì đòi hỏi quản trị viên phải biết về tiếng Nhật.

Vì vậy, để hỗ trợ cho quản trị viên, hệ thống sẽ tự động đưa ra gợi ý tag cho quản trị viên lựa chọn. Có 2 kiểu gợi ý tag: gợi ý tag theo danh từ riêng trong bài và gợi ý tag sử dụng thuật toán TF-IDF.

5.5.1 Gợi ý đánh tag theo danh từ riêng

Các bài học thường xoay quanh nói về chủ đề liên quan đến người, địa điểm, tổ chức nào đó. Vì vậy dùng danh từ riêng như tên người, tên địa điểm, tên tổ chức để đánh tag cho bài sẽ có độ chính xác cao. Công cụ Mecab hỗ trợ tách từ và trả về thông tin của từ, đồng thời cũng nhận biết được danh từ riêng trong văn bản đầu vào.

Ví dụ từ “日本” (Nhật Bản) là danh từ riêng. Mecab sẽ nhận biết và trả về thông tin:

名詞 danh từ	固有名詞 danh từ riêng	地域 địa điểm	国 đất nước	*	*	日本 từ ở thể từ điển	ニッポン cách đọc	ニッポン cách phát âm

Đối với những danh từ là danh từ riêng thì giá trị trả về ở phần thông tin thứ 2 của từ là “固有名詞” (danh từ riêng), phần thông tin thứ 3 sẽ chỉ rõ hơn danh từ riêng đó là gì. Công cụ Mecab nhận biết được các kiểu danh từ riêng sau đây:

地域：Địa điểm	国：Đất nước	組織：Tổ chức	人名：Tên người
名：Tên	姓：Họ	一般：Chung	

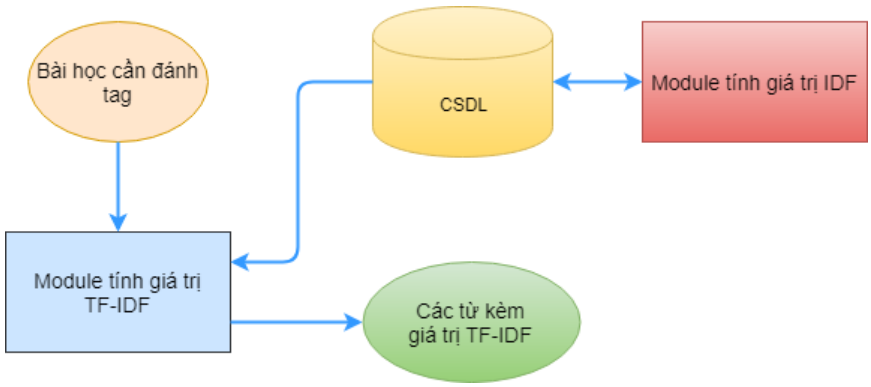
Với chức năng tự động đánh tag, hệ thống sử dụng công cụ Mecab nhận biết các danh từ riêng rồi tự động gắn các danh từ riêng này làm tag của bài học.

5.5.2 Gợi ý đánh tag sử dụng thuật toán TF-IDF

Theo đúng lý thuyết, hệ thống sẽ lần lượt tính tính giá trị TF và IDF của từng từ trong bài học. Tuy nhiên, giá trị IDF còn liên quan đến dữ liệu của các bài học khác nên việc tính toán mất nhiều thời gian. Vì vậy, em đã tính trước giá trị IDF của tất cả các từ xuất hiện trong các bài học hiện có trong CSDL và lưu vào bảng tag_words.

Như vậy, với từ cần tính giá trị IDF thì chỉ cần lấy trong bảng tag_words ra. Giá trị IDF lấy ra từ CSDL có thể sẽ có một chút sai lệch so với tính giá trị IDF thực tế. Nhưng với tập dữ liệu các bài học hiện tại là khá nhiều khoảng hơn 700 bài học thì sai lệch này sẽ rất nhỏ và không gây ảnh hưởng mấy đến kết quả.

Do đó, tính năng gợi ý đánh tag sử dụng thuật toán TF-IDF được xây dựng bởi 2 module chính là Module tính giá trị IDF và Module tính giá trị TF-IDF, được mô tả như hình vẽ dưới đây:



Hình 56 Thiết kế tổng quan chức năng gắn tag sử dụng TF-IDF

5.5.2.1 Module tính giá trị IDF

Giá trị IDF của từng từ trong tất cả các bài học hiện có sẽ được tính toán sẵn và lưu vào bảng tag_words. Thuật toán được mô tả như sau:

```
lesson_datas = ['lesson_data1', 'lesson_data2', ... , 'lesson_data-n'] (có n bài học)
words = array() // Mảng lưu tất cả từ trong tất cả bài học và số bài học có từ đó xuất hiện.
for lesson_data in lesson_datas
        lesson_words = []
        + Loại bỏ kí hiệu, dấu câu và các từ trong bảng stop_words ra khỏi lesson_data.
        + Mecab tách nội dung còn lại của lesson_data thành từng từ.
```

```
        + Lưu giá trị các từ vừa tách được vào mảng lesson_words. Một từ xuất
        hiện nhiều lần trong lesson_data cũng chỉ lưu một lần trong mảng
        lesson_words.
        + for word_value in lesson_words
                if word_value in words
                        array_ push(words, ['word_value' => 1])
                else
                        words['word_value'] += 1
                end
        end
end
for word in words
        IDF = log_e(n / word ["word_value"])
        insert word_value, IDF to tag_words table
end
```

5.5.2.2 Module tính giá trị TF-IDF

Input: Văn bản cần gắn tag. Output: Danh sách các từ có thể làm tag được sắp xếp theo thứ tự giá trị TF-IDF giảm dần. Các bước thực hiện như sau:

Bước 1: Tiền xử lý

Loại bỏ các kí hiệu, dấu câu và các từ không mang nhiều ý nghĩa như từ nối, trợ từ. Sau đó, đếm tổng số từ còn lại trong bài học.

Bước 2: Tính giá trị TF

- Dùng công cụ Mecab để tách dữ liệu bài học trên thành từng từ, đồng thời đếm số lần xuất hiện của từ. Mỗi từ được lưu vào mảng words dưới dạng [“word”=> word_count].
- Duyệt mảng words, giá trị TF của mỗi từ trong mảng được tính bằng giá trị word_count chia tổng số từ trong bài học được đếm ở bước 1.
- Cập nhật giá trị TF của từ vào mảng words. Mỗi phần tử trong mảng words lúc này có dạng [“word”=> TF].

Bước 3: Tính giá trị TF-IDF

- Duyệt mảng words, ứng với mỗi từ, tìm từ đó đã xuất hiện trong bảng tag_words hay chưa.

Đã xuất hiện	Chưa xuất hiện
Nhân giá trị TF của từ với giá trị IDF trong CSDL	Nhân giá trị TF của từ với 7, tức là coi giá trị IDF bằng 7.

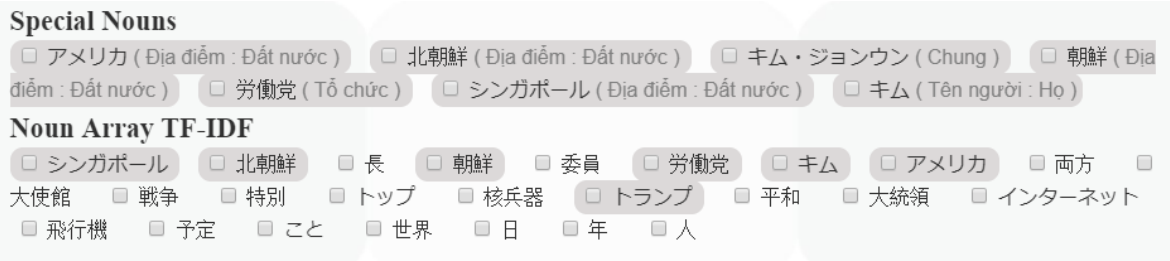
- Các từ không tìm thấy trong bảng tag_words tức là chưa từng xuất hiện trong bất kỳ bài học nào nên mức độ quan trọng của từ là cao. Trong CSDL giá trị IDF lớn nhất hiện tại là 6.32 nên em chọn 7 là giá trị IDF của những từ này.
- Cập nhật giá trị TF*IDF của từ vào mảng words. Mỗi phần tử trong mảng words lúc này có dạng [“word”=> TF*IDF].

Bước 4: Sắp xếp mảng words theo thứ tự giá trị TF*IDF từ lớn đến nhỏ và hiển thị cho quản trị viên lựa chọn để đánh tag cho bài.

Kết quả kiểm thử

Dưới đây là ví dụ về dữ liệu bài học thực tế của trang Web có id = 1061 và kết quả gợi ý tag từ hệ thống theo 2 cách là danh từ riêng và sử dụng thuật toán TF-IDF tính mức độ quan trọng của từ.

Dữ liệu bài học	Dịch sang tiếng Việt
アメリカのトランプ大統領は１０日の夜、インターネットのツイッターに「北朝鮮のキム・ジョンウン朝鮮労働党委員長と６月１２日にシンガポールで会います。世界の平和のために特別な時間にしたいです」と書きました。 シンガポールには、アメリカと北朝鮮の両方の大使館があります。キム委員長の飛行機は途中で止まらないでシンガポールまで行くことができます。 アメリカと北朝鮮のトップが会うのは初めてです。２人は、北朝鮮の核兵器をなくすことや、６５年前から続いている朝鮮戦争を終わらせることなどを話す予定です。	Tối ngày 10, Tổng thống Mỹ Trump đã viết trên trang Twitter của mình như sau "Tôi sẽ gặp ông Kim Jong-un chủ tịch Đảng Lao động Triều Tiên tại Singapore vào ngày 12 tháng 6. Tôi muốn biến đây thành thời điểm đặc biệt cho hòa bình thế giới". Singapore có đại sứ quán của cả và Mỹ và Triều Tiên. Máy bay của chủ tịch Kim có thể bay thẳng đến Singapore mà không dừng lại trên đường. Đây là lần đầu tiên lãnh đạo của Mỹ và Triều Tiên gặp nhau. 2 người dự định trao đổi về việc loại bỏ vũ khí hạt nhân ở Triều Tiên và chấm dứt chiến tranh Triều Tiên đã kéo dài suốt từ 65 năm trước.



Hình 57 Kết quả gợi ý tag đối với dữ liệu bài học kiểm thử

Những từ được bôi đậm là tag hiện tại của bài học. Trong đó:

Các danh từ riêng được tự động gắn làm tag từ hệ thống sử dụng là "**アメリカ**" (nước Mỹ), "**北朝鮮**" (bắc Triều Tiên), "**キム・ジョンウン**" (ông Kim Jong-un), "**朝鮮**" (Triều Tiên), "**労働党**" (đảng lao động), "**シンガポール**" (Singapore), "**キム**" (ông Kim)

Quản trị viên dựa thêm các gợi ý từ hệ thống sử dụng thuật toán TF-IDF đã gắn thêm tag cho bài là "**トランプ**" (ông Trump).

Từ kết quả thử nghiệm trên em nhận thấy việc gắn tag cho hệ thống sử dụng danh từ riêng khá chính xác và sát với nội dung bài học đang muốn nói đến. Đồng thời, gợi ý tag từ thuật toán TF-IDF cũng giúp quản trị viên trong việc lựa chọn gắn thêm các tag phù hợp mà hỗ trợ của công cụ Mecab không nhận biết được.

Trong chương này, em đã trình bày về các chức năng chính và đóng góp nổi bật của cả hệ thống. Đó là chức năng tự động sinh bài tập điền từ còn thiếu vào chỗ trống, tự động thêm phiên âm cách đọc cho từ Kanji trong tiếng Nhật, gợi ý bài học có thể người dùng quan tâm, phân loại bài học theo chủ đề có sẵn sử dụng thật toán Naive Bayes, gợi ý đánh tag. Ở chương tiếp theo cũng là chương cuối cùng của báo cáo, em sẽ tổng kết những kiến thức học được thông qua làm đồ án tốt nghiệp, đồng thời tự đưa ra nhận xét, đánh giá về trang web em đã xây dựng.

Chương 6 Kết luận và hướng phát triển

Đây là chương kết để thông qua đó em nhìn lại kết quả mình đã đạt được trong suốt quá trình làm đồ án tốt nghiệp, tổng kết kiến thức đã học và các kinh nghiệm rút ra. Tiếp đến em đưa ra một vài hướng phát triển để cải thiện, nâng cao chất lượng dạy học cho trang web nhưng vì thời gian còn hạn chế nên em chưa kịp thực hiện.

6.1 Kết luận

Trong suốt quá trình làm đồ án tốt nghiệp, em đã hoàn thành trang web dạy học tiếng Nhật trực tuyến thông qua video, audio tiếng Nhật với giao diện đẹp, dễ sử dụng, đầy đủ tính năng cần thiết. Đây là trang web rất bổ ích dành cho các bạn đang học tiếng Nhật và có tính ứng dụng thực tế rất cao. Trang web có các tính năng nổi bật:

- Tổng hợp các bài tiếng Nhật từ nhiều nguồn uy tín nên nội dung phong phú, chính xác và được cập nhật liên tục.
- Tự động gợi ý các bài học mà có thể người dung thích.
- Tính năng làm bài tập điền từ vào chỗ trống và theo dõi được kết quả quá trình làm bài.
- Dữ liệu sau khi được thu thập về, hệ thống sẽ đưa ra gợi ý chủ đề phù hợp để quản trị viên lựa chọn gắn cho bài học. Ngoài ra, quản trị viên cũng có thể chọn tính năng tự động phân loại bài học theo chủ đề từ hệ thống.
- Hệ thống sẽ đưa ra gợi ý các từ có thể dùng làm tag để quản trị viên lựa chọn gắn cho bài học. Quản trị viên cũng có thể chọn tính năng tự động gắn tag từ hệ thống.

Bên cạnh những tính năng nổi bật, em cũng nhận thấy trang web vẫn còn một vài hạn chế:

- Bộ Crawler không thu thập được bài học podcast dạng video từ nguồn youtube. Bởi vì youtube không cung cấp API để lấy dữ liệu nội dung đoạn video (video subtitle). Người biên tập phải truy cập vào đường dẫn https://downsub.com để tải file video subtitle về.
- Một số bài có giải thích nghĩa của từ vựng sang tiếng Anh, còn lại vẫn chưa có giải thích của từ vựng.

Đồng thời thông qua đồ án, em đã học hỏi thêm được rất nhiều điều và tự rút ra được cho mình những bài học kinh nghiệm:

- Thiết kế giao diện vô cùng quan trọng, thiết kế đẹp và có sự thông nhất toàn cục trong trang web sẽ gây ấn tượng và giúp người dùng trải nghiệm tốt hơn.
- Để trang web có được nhiều người sử dụng, quá trình quảng bá quan trọng không kém với quá trình phát triển. Nếu quảng bá không tốt dù sản phẩm có tốt đến đâu thì cũng rất khó đến tay người dùng.
- Trước khi tích hợp thư viện bên thư 3 nên tìm hiểu các thư viện tương tự, phân tích và so sanh ưu nhược điểm để tìm ra thư viện phù hợp với ứng dụng của mình nhất.
- Cần đặt mục tiêu sản phẩm chạy tốt lên đầu chứ không phải sản phẩm nhiều tính năng. Vì dù sản phẩm có nhiều tính năng hay nhưng khi sử dụng lại lỗi nhiều thì người dùng cũng sẽ gỡ bỏ ngay lập tức.

6.2 Hướng phát triển

Để hoàn thành trang web, em đã tìm hiểu, học hỏi và hoàn thiện các kiến thức, kỹ năng như:

- Nắm vững các công nghệ để phát triển trang web: PHP, Laravel Framework, HTML, CSS, Javascript, AngularJs và một số thư viện hỗ trợ lập trình web khác.
- Sử dụng thành thạo Mysql để truy vấn CSDL.
- Sử dụng thành git để quản lý code.
- Lập trình thành thạo ngôn ngữ Ruby và biết cách trích xuất dữ liệu từ trang web.
- Tìm hiểu cách tự động upload ảnh, audio, video lên server trung gian là Cloudinary để lưu trữ.

- Tìm hiểu cách sử dụng công cụ Mecab trong việc tách từ trong văn bản tiếng Nhật và vận dụng công cụ đó hỗ trợ một số chức năng cho trang web.
- Nghiên cứu và vận dụng thuật toán Naive Bayes cho việc phân loại bài học theo chủ đề.
- Nghiên cứu và vận dụng thuật toán TF-IDF cho việc xác định mức độ quan trọng của từ trong văn bản.
- Cách thức deploy trang Web lên Srver và quản lý trang Web trên Server.

Trong phạm vi đồ án, em đã hoàn thành đầy đủ chức năng cơ bản của trang web dạy tiếng Nhật trực tuyến, đồng thời áp dụng một số thuật toán giúp hỗ trợ việc thêm thông tin cho bài học nhưng vì thời gian còn hạn chế nên em xin đề xuất hướng phát triển vẫn chưa kịp tiếp theo cho trang web như sau:

- Phân tích nội dung bài học để tìm ra cấu trúc ngữ pháp được sử dụng trong bài và hiển thị ý nghĩa của cấu trúc đó là gì. Nếu có thể, hiển thị thêm một số câu ví dụ sử dụng cấu trúc ngữ pháp đó.
- Thêm tính năng tra từ điển cho các từ trong bài để người dùng không phải mở ứng dụng khác ra tra từ.
- Phát triển trang Web thành các ứng dụng mobile để thuận tiện hơn với người dùng.

Thông qua quá trình làm đồ án, dưới sự hướng dẫn tận tình của PGS.TS. Cao Tuấn Dũng, em đã hoàn thành hệ thống hỗ trợ học tiếng Nhật trực tuyến thông qua podcast. Đồng thời em học thêm được rất nhiều kiến thức mới, tăng khả năng phân tích, tìm hiểu, nghiên cứu và giải quyết vấn đề, kỹ năng lập trình cũng được nâng cao. Em mong rằng trang Web của em sẽ giúp ích cho việc học tiếng Nhật của người dùng trở nên hiệu quả và dễ dàng hơn.

Tài liệu tham khảo

[1] Trang chủ lập trình Laravel https://laravel.com

[2] Trang chủ Ruby https://www.ruby-lang.org/en

[3] Hướng dẫn cài đặt và sử dụng công cụ Mecab cho PHP

https://github.com/nihongodera/php-mecab-documentation

[4] Jonathan Robie, Texcel Research, What is the Document Object Model, https://www.w3.org/TR/WD-DOM/introduction.html, last visited May 2018.

[5] G. K. Zipf, Human Behavior and the Principle of Least Effort, an Introduction to Human Ecology, Addison-Wesley, Reading, Mass., 1949.

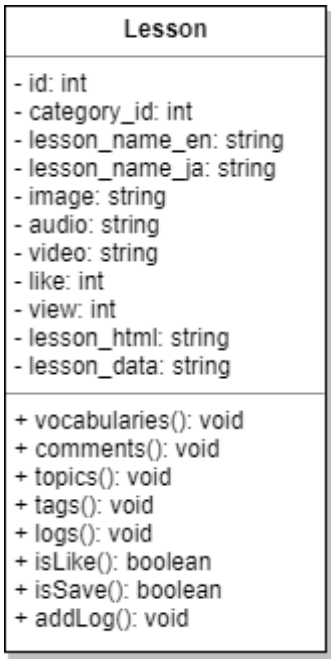
[6] Nigam, “Learning to classify text from label”, Kamal Nigam; Andrew McCallum; Sebastian Thrun; Tom Mitchell, 2000.

[7] Juan Ramos, Using TF-IDF to determine word relevance in document queries, Rutgers University, January 2003.

Phụ lục

A Thiết kế lớp

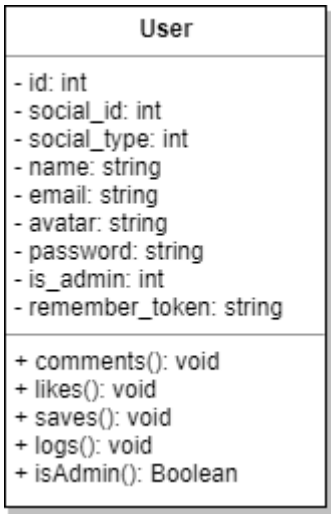
A.1 Thiết kế lớp cho Model Lesson



Ý nghĩa các phương thức:

- vocabularies(): Tạo quan hệ một – nhiều với lớp Vocabulary
- comments(): Tạo quan hệ một – nhiều với lớp Comment.
- categories(): Tạo quan hệ một – nhiều với lớp Category.
- topics(): Tạo quan hệ nhiều – nhiều với lớp Topic.
- tags(): Tạo quan hệ nhiều – nhiều với lớp Tag.
- logs(): Tạo quan hệ một – nhiều với lớp Log.
- isLike(): Kiểm tra người dùng hiện tại đã thích bài học chưa.
- isSave(): Kiểm tra người dùng hiện tại đã lưu bài học chưa.
- addLog(): Thêm dữ liệu log của bài học hiện tại vào bảng user_logs.

A.2 Thiết kế lớp cho Model User



Ý nghĩa các phương thức:

- comments(): Tạo quan hệ một – nhiều với lớp Comment.
- likes(): Tạo quan hệ một – nhiều với lớp UserLike.
- saves(): Tạo quan hệ một – nhiều với lớp UserSave.
- logs(): Tạo quan hệ một – nhiều với lớp UserLog.
- isAdmin():Kiểm tra người dùng hiện tại có phải là quản trị viên hay không.