TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**KHÓA LUẬN/ĐỒ ÁN TỐT NGHIỆP**

**XÂY DỰNG HỆ THỐNG HỎI ĐÁP**

**DỰA TRÊN CỘNG ĐỒNG**

*Người hướng dẫn*: **PGS.TS LÊ ANH CƯỜNG**

*Người thực hiện*: **NGÔ HÙNG PHÚC-51303129**

**NGUYỄN NHẬT NGUYÊN-51303352**

Lớp **: 13050302**

Khoá  **: 17**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2017**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**XÂY DỰNG HỆ THỐNG HỎI ĐÁP**

**DỰA TRÊN CỘNG ĐỒNG**

*Người hướng dẫn*: **PGS.TS LÊ ANH CƯỜNG**

*Người thực hiện*: **NGÔ HÙNG PHÚC-51303129**

**NGUYỄN NHẬT NGUYÊN-51303352**

Lớp **: 13050302**

Khoá  **: 17**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2017**

LỜI CẢM ƠN

- Lời đầu tiên, em xin chân thành cảm ơn quý thầy cô trong khoa Công nghệ thông tin cũng như các quý thầy cô đang giảng dạy và công tác tại trường Đại học Tôn Đức Thắng đã truyền đạt những kiến thức quý báu cho chúng em trong những năm học vừa qua.

- Đặc biệt, Chúng em xin gửi lời cảm ơn đến thầy Lê Anh Cường đã hỗ trợ tụi em trong suốt thời gian làm luận văn đã đưa ra những ý tưởng để cải tiến cho hệ thống của chúng em giúp chúng em có được một hệ thống hoàn chỉnh hơn, giúp đỡ và động viên em trong suốt thời gian thực hiện đề tài. Và để có được kết quả như ngày hôm nay, em rất biết ơn gia đình đã động viên, khích lệ, tạo mọi điều kiện thuận lợi nhất trong suốt quá trình học tập cũng như quá trình thực hiện đề tài tốt nghiệp này.

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của PGS.TS Lê Anh Cường;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong luận văn còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung luận văn của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Ngô Hùng Phúc*

*Nguyễn Nhật Nguyên*

TÓM TẮT

- Nhận thấy những khó khăn của các bạn sinh viên đặc biệt là các sinh viên mới vào trường khi có những thắc mắc trăn trở trong quá trình học tập sinh hoạt tại trường nhưng không biết trình bày những thắc mắc với ai.

- Thậm chí nếu có thắc mắc và đặt câu hỏi thì đôi khi thông tin không chính xác làm các bạn có thể đánh mất quyền lợi của bản thân.

- Do suy nghĩ đó nên chúng em đã đưa ra ý tưởng xây dựng một hệ thống hỏi đáp giành cho các bạn sinh viên sử dụng để đặt ra những câu hỏi và trò chuyện cùng nhau.

- Hệ thống đã bước đầu đáp ứng được nhu cầu sử dụng của một diễn đàn cho phép hỏi đáp, tìm kiếm những câu hỏi chuẩn xác.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc387692855)

[TÓM TẮT i](#_Toc387692856)

[MỤC LỤC 1](#_Toc387692857)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 2](#_Toc387692858)

[CHƯƠNG 1 – MỞ ĐẦU 1](#_Toc387692860)

[1.1 Giới thiệu hệ thống hỏi đáp tự động 2](#_Toc387692861)

[1.1.1 Hệ thống hỏi đáp là gì 2](#_Toc387692862)

[1.1.2 Ý nghĩa hệ thống hỏi đáp và nhu cầu xã hội 3](#_Toc387692865)

[1.2 Lý do chọn đề tài 3](#_Toc387692866)

[1.3 Mục đích đề tài 3](#_Toc387692866)

[1.4 Đối tượng nghiên cứu 4](#_Toc387692866)

[1.5 Phạm vi nghiên cứu 4](#_Toc387692866)

[1.5.1 Bài toán đưa ra 5](#_Toc387692862)

[1.5.1.1 Đối với hệ thống web 6](#_Toc387692863)

[1.5.1.2 Đối với hệ thống mobile 6](#_Toc387692864)

[1.5.1.3 Cơ sỡ dữ liệu 7](#_Toc387692864)

[1.5.2 Dữ liệu mẫu thu thập 7](#_Toc387692862)

[1.5.3 Các hệ thống đã có và hệ thống trả lời tự động 8](#_Toc387692862)

[1.5.4 Công cụ hiện thực hệ thống web 9](#_Toc387692862)

[1.5.4.1 AngularJS 9](#_Toc387692863)

[1.5.4.2 ExpressJS 9](#_Toc387692863)

[1.5.4.3 NodeJS 10](#_Toc387692863)

[1.5.4.4 MongoDB 11](#_Toc387692863)

[1.5.4.5 Mô Hình MEAN JS 14](#_Toc387692863)

[1.5.5 Công cụ hiện thực hệ thống di động 15](#_Toc387692862)

[1.5.5.1 Mô hình di động 16](#_Toc387692863)

[1.5.5.2 Thư viện sử dụng 16](#_Toc387692863)

[1.6 Server 17](#_Toc387692866)

[1.6.1 Công nghệ sử dụng 17](#_Toc387692862)

[1.6.2 Thư viện sử dụng 17](#_Toc387692862)

[1.7 Thực tiễn của đề tài 18](#_Toc387692866)

[CHƯƠNG 2 – XÂY DỰNG HỆ THỐNG 1](#_Toc387692867)9

[2.1 Mô Hình Class Diagram 19](#_Toc387692868)

[2.2 Mô Hình Use Case 20](#_Toc387692869)

[2.2.1 Use case chưa đăng nhập 20](#_Toc387692862)

[2.2.2 Use case đăng ký sau đó đăng nhập 21](#_Toc387692862)

[2.2.3 Use case sau khi đăng nhập 21](#_Toc387692862)

[2.2.4 Use case sau khi đăng nhập và đặt câu hỏi 22](#_Toc387692862)

[2.2.5 Use case sau khi đăng nhập và trả lời câu hỏi 22](#_Toc387692862)

[2.3 Mô Hình Sequence Diagram 24](#_Toc387692869)

[2.2.1 Sequence diagram đăng nhập 24](#_Toc387692862)

[2.2.2 Sequence diagram đăng ký 25](#_Toc387692862)

[2.2.3 Sequence diagram đăng câu hỏi 26](#_Toc387692862)

[2.2.4 Sequence diagram trả lời câu hỏi 27](#_Toc387692862)

[2.2.5 Sequence diagram gửi tin nhắn 28](#_Toc387692862)

[2.4 Thiết kế cơ sỡ dữ liệu 29](#_Toc387692868)

[CHƯƠNG 3 – Xây dựng engine Trả Lời Tự Động 30](#_Toc387692870)

[3.1 Tổng quan hệ thống: 30](#_Toc387692871)

[3.2 Phương pháp 31](#_Toc387692872)

[3.2.1 Lọc dữ liệu sử dụng full text search 31](#_Toc387692862)

[3.2.2 Tách từ 32](#_Toc387692862)

[3.2.3 Phân loại câu hỏi sử dụng Maximum entropy classifier 33](#_Toc387692862)

[3.2.4 Maximum entropy là gì? 33](#_Toc387692862)

[3.2.5 Phân loại văn bản sự dụng công cụ Stanford Classifier 34](#_Toc387692862)

[3.2.6 Đo độ tương tự giữa các câu hỏi 34](#_Toc387692862)

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

**CÁC KÝ HIỆU**

**CÁC CHỮ VIẾT TẮT**

* FTS: Full Text Search

CHƯƠNG 1 – MỞ ĐẦU

* 1. Giới thiệu hệ thống hỏi đáp tự động

- Ngày nay hệ thống internet phát triển với một khối lượng dữ liệu khổng lồ dẫn đến việc tìm kiếm sẽ gặp khó khăn cùng với đó là sự nhiễu loạn thông tin cũng có thể dẫn tới việc nắm bắt sai thông tin.

- Thông tin tìm kiếm đôi khỉ chỉ dừng ở mức tài liệu nghiên cứu , còn hệ thống hỏi đáp sẽ cho ta một câu trả lời ngắn gọn nhất có thể và đôi khi là 1 hướng giải quyết vấn đề từ những người đi trước.

- Hiện nay có nhiều hệ thống hỏi đáp phục vụ nhu cầu của cộng đồng từ các mảng lập trình, sửa chữa máy tính, tư vấn học tập, sức khỏe và nhiều hơn nữa nhưng đa phần chỉ được phát triển bằng tiếng Anh.

- Hệ thống hỏi đáp vẫn chưa thật sự phát triển mạnh ở nước ta nên nhận thấy việc xây dựng hệ thống hỏi đáp bằng tiếng việt rất có ý nghĩa và mang tính thực tế.

* + 1. Hệ thống hỏi đáp là gì:

- Hệ thống hỏi đáp là hệ thống cho phép người dùng đặt câu hỏi và nhận được câu trả lời từ cơ sở dữ liệu đã thu thập từ cộng đồng trước đó về những vấn đề mà họ đang gặp khó khăn và chưa tìm ra hướng giải quyết.

- Hệ thống hỏi đáp là nơi tập trung các kiến thức, các kinh nghiệm của những người đi trước chia sẽ lại, nó cũng là thế giới thông tin mở và là kho tàng kiến thức.

- Hệ thống hỏi đáp được chia thành 2 loại chính:

* Hỏi đáp dựa trên cộng đồng: Dữ liệu được thu thập chủ yếu từ những hiểu biết để của người dùng để xây dựng nên hệ thống.
* Hỏi đáp dựa trên bộ sinh: Được xây dựng từ những dữ liệu ban đầu có sẵn và sẽ tự động học được thông tin trong quá trình con người tham gia sử dụng.

- Hệ thống hỏi đáp trong để tài nghiên cứu này sẽ tập trung vào cộng đồng sinh viên trường đại học Tôn Đức Thắng.

1.1.2 Ý nghĩa hệ thống hỏi đáp và nhu cầu xã hội :

- Hệ thống hỏi đáp sẽ là nơi tập trung thông tin các vấn đề, thắc mắc, góp ý v.v….

- Cung cấp một nguồn thông tin đáng tin cậy cho người dùng, tránh được những thông tin trái chiều không chính xác.

- Nhu cầu được giải đáp thắc mắc trong quá trình học tập sinh hoạt luôn có ở bất cứ môi trường học tập nào vì vậy việc ứng dụng hệ thống hỏi đáp vào thực tế cho sinh viên sử dụng sẽ tạo tiền đề và sẽ là cảm hứng phát triển thêm nữa cho các hệ thống khác về sau.

1.2 Lý do chọn đề tài

- Như đã giới thiệu do nhận thấy nhà trường đã có hệ thống hỏi đáp nhưng vẫn chưa có sự tương tác mạnh giữa sinh viên và nhà trường và số lượng sử dụng còn ít nên chúng em quyết định hiện thực hệ thống cho các bạn sinh viên sử dụng.

- Hệ thống cho phép sinh viên đặt và trả lời các câu hỏi. Cùng với đó là khu vực trò chuyện cho các bạn sinh viên giúp các bạn tiết kiệm thời gian chờ đợi câu hỏi được trả lời.

- Hệ thống sẽ là một địa chỉ đáng tin cậy và tập trung thông tin cho toàn thể sinh viên trường Đại Học Tôn Đức Thắng

1.3 Mục đích đề tài

- Mục đích đề tài là hiện thực các giải thuật tìm kiếm so sánh độ tương tự của câu hỏi từ người dùng nhập vào với các câu hỏi từ cơ sở dữ liệu có sẵn, từ đó tìm ra một câu hỏi tương tự với câu hỏi người dùng nhập vào. Ta phải tìm ra một giải thuật đo độ tương tự và có độ chính xác cao nhất đảm bảo tính đúng đắn của dữ liệu.

- Một hệ thống đáng tin cậy là một hệ thống đảm bảo được sự chính xác có tốc độ nhanh cùng với đó là khả năng bảo mật dữ liệu người dùng.

1.4 Đối tượng nghiên cứu

- Đối tượng nghiên cứu chính của đề tài này là các bạn sinh viên trường đại học Tôn Đức Thắng. Dữ liệu được khai thác và sử dụng chủ yếu từ trang hỏi đáp của trường cùng với đó là các diễn đàn trên Facebook của trường.

- Dữ liệu được khai thác từ chính các bạn sinh viên sẽ đảm bảo độ chân thực và diễn tả chính xác nhất những trăn trở khó khăn mà các bạn gặp phải trong quá trình theo học tại trường và nội quy hoạt động của nhà trường.

- Dữ liệu nếu tự chúng ta suy nghĩ và đưa ra sẽ không đảm bảo được tính khách quan của vai trò người dùng cho hệ thống

1.5 Phạm vi nghiên cứu

1.5.1 Bài toán đưa ra

- Ta cần một hệ thống cho phép người dùng đặt câu hỏi và trả lời, hệ thống cũng cho phép tìm kiếm câu hỏi, tìm kiếm những câu hỏi tương tự,trò chuyện để nhận được những phản hồi nhanh nhất có thể.

- Phạm vi nghiên cứu của đề tài xoay quay 3 đối tượng chính là hệ thống web,ứng dụng di động và máy chủ (Server) để xử lý giải thuật và trao đổi thông tin với Web API

- Vậy để hiện thực được một hệ thống như vậy chúng ta cần những gì :

1. Đối với web: Quản lý theo mô hình MVC

**Model**: có nhiệm vụ thao tác với cơ sở dữ liệu, nghĩa là nó sẽ chứa tất cả các hàm, các phương thức truy vấn trực tiếp với dữ liệu và controller sẽ thông qua các hàm, phương thức đó để lấy dữ liệu rồi gửi qua **View**

**View**: có nhiệm vụ tiếp nhận dữ liệu từ controller trả vể và hiển thị nội dung sang những gì mà người dùng có thể thấy được.

**Controller**: đóng vài trò trung gian giữa Model và View. Controller có nhiệm vụ diều hướng yêu cầu từ người dùng và trả dữ liệu từ tầng Model mà người dùng mong đợi.

* Ý nghĩa của mô hình MVC:
* Dễ dàng quản lý mã nguồn,các tầng có nhiệm vụ riêng
* Dễ dàng nâng cấp mã nguồn
* Có thể dễ dàng tìm lỗi khi gặp sự cố

1. Đối với mobile app: Quản lý theo module.

- Hệ thống sẽ được chia ra theo từng thành phần chức năng

- Kiến trúc mô-đun (module) cho phép chia nhỏ bài toán (hay yêu cầu) của phần mềm thành các phần hầu như không trùng lắp.

- Các module sẽ cung cấp một giao diện (Interface) cho các thành phần khác trong hệ thống có thể gọi và sử dụng.

1. Đối với cơ sỡ dữ liệu: Xây dựng cơ sỡ dữ liệu

* Xây dựng cơ sỡ dữ liệu: dựa trên các bước sau:

1. Xác định vấn đề cần giải quyết: Tự đặt ra vấn đề hệ thống cần thiết kế như thế nào để lữu trữ.Cơ sở dữ liệu được sử dụng cho đối tượng sinh viên và các thông tin của sinh viên cùng với đó là các câu hỏi câu trả lời sinh viên
2. Nghiên cứu hệ thống dữ liệu sẵn có: Các hệ thống hỏi đáp nổi tiếng như stackoverflow,tinhte.
3. Thiết kế cấu trúc dữ liệu:Xác định các đối tượng cần có cho cơ sỡ dữ liệu,các quan hệ giữa các đối tượng,các kiễu dữ liệu.
4. Ràng buộc kiểu dữ liệu: Đảm bảo kiểu dữ liệu phù hợp với nhu cầu sử dụng trách tạo kiểu dữ liệu sai hoặc nhiều hơn thực tế cần để tránh việc dư thừa dữ liệu.
5. Tiến hành xây dựng cơ sỡ dữ liệu.

1.5.2 Về dữ liệu mẫu

- Dữ liệu ban đầu sẽ có 200 câu hỏi và hơn 200 câu trả lời mẫu để tham khảo

- Câu hỏi được khai thác từ các bạn sinh viên trên mạng xã hội và câu trả lời được khai thác từ phía nhà trường nơi có thông tin chính xác và tin cậy nhất.

- Các tài liệu đường link tham khảo chủ yếu được lấy từ nguồn tìm kiếm trên google.

- Dữ liệu khai thác xoay quanh các vấn đề thường gặp như học phí,lịch học,miễn học,quy định,kỷ luật nhà trường v.v…

1.5.3 Nghiên cứu hệ thống hỏi đáp tự động:

a.Các hệ thống đã có tại Việt Nam:

- Diễn đàn hỏi đáp lập trình vb (<http://hoidap.laptrinhvb.net/>): Có các chức năng chính sau:

* Đăng Nhập
* Đăng Ký
* Hot Topic được quan tâm nhiều nhất
* Like Dislike câu hỏi câu trả lời
* Đặt câu hỏi
* Trả lời câu hỏi

- Diễn đàn tinhte (<https://tinhte.vn>):

* Quản lý theo danh mục
* Từ danh mục cha có thêm nhiều danh mục con
* Thống kê số lượng bài viết
* Hiển thị các bài viết mới nhất
* Hot Topic

b.Hệ thống trả lời tự động:

- Giúp đưa ra gợi ý cho người dùng về vấn đề mà họ tìm kiếm.

- Gợi ý được chọn lọc từ những thông tin gần đúng so với dữ liệu người dùng nhập.

- Một hệ thống hỏi đáp sẽ gồm 3 ba bước :

* Phân tích câu hỏi
* Tìm kiếm câu hỏi tương tự
* Đưa ra câu hỏi tương tự
* Có 2 phương pháp phân tích câu hỏi:
* Phương pháp nông: Đo mật độ xuất hiện của câu hỏi người dùng nhập so với các câu hỏi trong hệ thống sau đó loại bỏ các từ dư thừa.Cuối cùng dùng chiến lược xếp hạng,dựa trên các đặc điểm cú pháp hoặc vị trí từ của câu truy vấn.
* Phương pháp sâu: Khi phương pháp nông không thể xử lý được thì các kỹ thuật xử lý cú pháp ngữ nghĩa v.v…. và khả năng xử lý ngôn ngữ tự nhiên để có thể phân tích được chúng.
* Tìm kiếm thông tin cũng là yếu tố quan trọng và có thể áp dụng mô hình xác suất.

Mô hình này coi việc tìm kiếm tài liệu như một công việc suy luận có tính xác suất.Thường thấy trong các mô hình xác suất Bayes,tìm kiếm độc lập nhị phân.

* Quá trình gom nhóm dữ liệu cũng cần thiết cho giải thuật tìm kiếm.Việc chia dữ liệu thành từng cụm sẽ làm giảm quá trình truy vấn và tăng tốc độ xử lý.
* Thuật toán Kmeans là phương pháp được sử dụng nhiều nhất.Ý tưởng chính của thuật toán Kmeans là gán mỗi ứng viên trong hệ thống có giá trị gần nhất và tâm cụm sẽ là giá trị trung bình của các ứng viên.

1.5.4 Công cụ hiện thực (Về hệ thống web)

- Hệ thống web được xây dựng trên nền tảng công nghệ Javascript gồm 4 công nghệ Angular Js, Express , MongoDB, NodeJs.Nền tảng từ 4 công nghệ nêu trên được gọi tắt là MEAN Js

1.5.4.1. Angular Js

- AngularJS là một full-stack Javascript framework, được phát triển bởi Google.

- Ban đầu mục tiêu của Angular là để xây dựng các ứng dụng dựa trên tiêu chuẩn MVC (Model - View - Controller), sau đó Angular dần phát triển và tiến gần hơn về với MVVM và MVP. Sau đó Google đã định nghĩa nó lại là MVW (Model-View-Whatever) để ám chỉ Angular là một framework có tính chất làm mọi thứ mà ta cần.

- Angular Js cho phép ta thao tác dữ liệu phía server trả về và hiển thị cho người dùng thấy ở phần giao diện.

- Angular Js cho phép người dùng lập luận logic if, else, loop, hiển thị, ẩn… các thành phần của giao diện v.v….

- Angular Js cũng cho phép chúng ta quản lý mã nguồn 1 cách linh hoạt bằng việc tổ chức mã nguồn theo tính năng và thành phần của hệ thống web.

1.5.4.2. Express JS

- Express là một web application framework cho NodeJS, cung cấp các tính năng mạnh mẽ cho việc xây dựng một ứng dụng web.

- ExpressJS là framework phổ biến và được sử dụng rộng rãi nhất của NodeJS. Ý tưởng đằng sau ExpressJS là đưa đến một framework nhẹ, dễ dàng tiếp cận để phát triển các ứng dụng web từ nhỏ đến lớn.

- Express cũng có thể sử dụng để xây dựng một API mạnh mẽ và thân thiện với người dùng, vì nó cung cấp rất nhiều tiện ích HTTP và middleware cho việc kết nối.

-Express JS sẽ đóng vai trò trung gian cho hệ thống MEAN Js khi tiếp nhận dữ liệu phía người dùng gửi lên hoặc 1 yêu cầu do người dùng yêu cầu phía hệ thống xử lý.Từ đó Express sẽ điều hướng và quyết định nên làm gì tiếp theo cho yêu cầu đó.Express Js có thể chấp nhận hoặc từ chối yêu cầu từ phía người dùng nếu không đủ điều kiện thực hiện.

1.5.4.3. Node JS

- NodeJS là một nền tảng Server side được xây dựng dựa trên Javascript Engine (V8 Engine)

- NodeJs là một mã nguồn mở, đa nền tảng cho phát triển các ứng dụng phía Server và các ứng dụng liên quan đến mạng. Ứng dụng Node.js được viết bằng Javascript và có thể chạy trong môi trường Node.js trên hệ điều hành Window, Linux...

- Node Js chạy phía server side cũng giống như các ngôn ngữ server side như .Net,Php,Java cho phép xử lý logic,thao tác với cơ sở dữ liệu.

-Node Js thực thi các tác vụ bất đồng bộ cho phép hệ thống chạy không theo thứ tự và có thể chạy nhiều tác vụ cùng 1 lúc giúp tiết kiệm thời gian đợi của hệ thống.

- Dưới đây là các lĩnh vực mà Node.js được sử dụng như là một sự lựa chọn hoàn hảo:

* Các ứng dụng về I/O
* Các ứng dựng về luồng dữ liệu
* Các ứng dụng về dữ liệu hướng đến thời gian thực
* Các ứng dụng dựa vào JSON APIs
* Các ứng dụng Single Page Application

1.5.4.4. MongoDB

- MongoDb là cơ sở dữ liệu NoSQL được dùng cho việc lưu trữ dự liệu hệ thống.

- NoSQL được phát triển trên Javascript Framework với kiểu dữ liệu là JSON và dạng dữ liệu theo kiểu key và value (1 đặc trưng về dữ liệu trong JSON)

- Với NoSQL bạn có thể mở rộng dữ liệu mà không lo tới những việc như tạo khóa ngoại, khóa chính, kiểm tra ràng buộc, .v.v…Vì NoSQL không hạn chế việc mở rộng dữ liệu nên tồn tại nhiều nhược điểm như: sự phục thuộc của từng bản ghi, tính nhất quán, toàn vẹn dữ liệu.

- NoSQL cho hiệu suất cao hơn khi giải quyết những bài toán lớn về hệ thống thông tin, phân tán hay lưu trữ dữ liệu.

- Với các CSDL quan hệ, chi phí triển khai cũng như phát triển các ứng dụng sử dụng CSDL quan hệ cũng rất tốn kém và điều quan trọng được bộc lộ rõ khi truy vấn 1 lượng bản ghi lớn trong thời gian rất dài.

- Tiếp đó những thiết bị cầm tay như smartphone thì không phù hợp để triển khai CSDL quan hệ vì dung lượng bộ nhớ thấp và khả năng xử lý kém hơn máy tính.



1.Biểu đồ so sánh tốc độ insert dữ liệu giữa MongoDB và SQL Server

(<https://quantrimang.com/so-sanh-hieu-suat-hoat-dong-cua-mongodb-va-sql-server-2008-84584>)

* Ta co thể thấy việc MongoDB chèn dữ liệu nhanh hơn SQL Server tới hơn 100 lần



2.Biểu đồ so sánh tốc độ truy vấn dữ liệu giữa MongoDB và SQL Server

(<https://quantrimang.com/so-sanh-hieu-suat-hoat-dong-cua-mongodb-va-sql-server-2008-84584>)

* Ta có thể thấy tốc độ truy vấn 50,000 dòng dữ liệ của MongoDB cao gần gấp 30 lần so với SQL Server
* 1.5.4.5. Mô Hình MEAN JS



3.Mô Hình Hoạt Động Của Một Ứng Dụng Mean Js

(Nguồn google)

-Hệ thống sẽ sử dụng 1 file server.js có tác dụng khởi động ứng dụng web

- Từ file server.js hệ thống sẽ trả về giao diện cho người dùng.Từ phía người dùng ta cũng sẽ có mô hình MVC do Angular Js đảm nhận logic từ phía giao diện sau đó chuyển dữ liệu về lại cho phía server xử lý

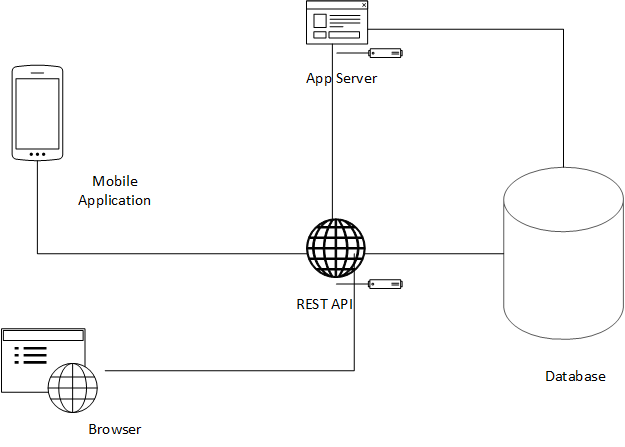
-Phía server cũng áp dụng mô hình MVC controller sẽ xử lý logic và model dùng để thao tác với đối tượng và cơ sở dữ liệu và sẽ trả dữ liệu về giao diện cho người dùng.

1.5.5 Về hệ thống di động

Ngoài hệ thống web, chúng em còn cung cấp thêm một ứng dụng được xây dựng trên nền tảng Android để thuận tiện trong việc hỏi đáp.

1.5.5.1 Hệ thống di động

- Xây dựng trên nền tảng Android sử dụng ngôn ngữ lập trình Java đồng thời kết hợp với Web API của hệ thống để trao đổi dữ liệu.



4.Mô Hình Hoạt Động Của Web API Service

* Như trên hình chúng ta có thể thấy web api cho phép người dùng có thể truy vấn hoặc thao tác với dữ liệu thông qua web hoặc các hệ điều hành di động khác nhau.
* Với mô hình web api ta có thể mở rộng không giới hạn số lượng thiết bị truy cập vào tài nguyên của hệ thống.
* Web API Services cũng cho phép 1 số lương lớn truy cập có thể lên tới hàng ngàn người.

1.5.5.2 Thư viện sử dụng

- Retrofit (<http://square.github.io/retrofit/>) giúp hệ thống android thao tác trao đổi dữ liệu, thông tin với API của website.

- Thư viện cho phép người dùng thực thi các yêu cầu thêm, xóa, sửa, cập nhật cơ sở dữ liệu thông qua API của website.

- Ngoài ra còn sử dụng các thư viện Design-support của google để thiết kế và hiện thực giao diện người dùng, thư viện facebook-sdk cho phép người dùng đăng nhập, đăng ký thông qua tài khoản Facebook.

***1.6 Server***

Server là nơi tiếp nhận những câu hỏi của người dùng từ Web API sau đó áp dụng, xử lý các giải thuật về phân loại câu hỏi, đo độ tương tự của câu hỏi đó với các câu hỏi trong cơ sở dữ liệu để tìm ra câu hỏi gần với câu người dùng tìm kiếm nhất và trả về lại cho Web API, để Web API xử lý và hiển thị lên cho người dùng cuối.

1.6.1. Công nghệ sử dụng:

- Server được viết trên ngôn ngữ lập trình Java, sử dụng giao thức TCP trong lập trình mạng để trao đổi dữ liệu với Web API.

1.6.2. Thư viện sử dụng

- vnTokenizer: thư viện cho phép chúng ta tách các từ trong trong văn bản Tiếng Việt thành các từ đơn hoặc từ ghép có nghĩa

- Stanford Classifier: là công cụ sử dụng trong học máy, hiện thực dựa trên phương pháp Maximum Entropy Classifier, ứng dụng trong việc phân loại các câu hỏi trong Tiếng việt thành các loại câu về “cái gì (what)”, “tại sao (why)”, “như thế nào (how)”…

- MongoDB JDBC Driver: thư viện cho phép kết nối, lấy, truy vấn cơ sở dữ liệu MongoDB trên Java.

1.7 Thực tiễn của đề tài

- Hệ thống sẽ được áp dụng thực tế cho nhà trường và các bạn sinh viên là người sử dụng hệ thống để đặt câu hỏi và giải đáp thắc mắc cho nhau.

- Là cầu nối giữa nhà trường và sinh viên.

- Nơi để sinh viên bày tỏ nguyện vọng của mình và cũng là nơi để nhà trường tiếp nhận ý kiến đóng góp của sinh viên để từng bước cải thiện quá trình đào tạo.

- Tránh làm mất quyền lợi của các bạn sinh viên khi không được giải đáp kịp thời.

CHƯƠNG 2 – XÂY DỰNG HỆ THỐNG

2.1 Mô Hình Cơ Sỡ Dữ Liệu

2.1.1-Mô Hình Class Diagram



Hình 2.2.1 Class Diagram hệ thống

2.2.2-Mô Hình Usecase

****

Hình 2.2.2.1 Use case chưa đăng nhập

****

Hình 2.2.2.2 Use case đăng ký sau đó đăng nhập



Hình 2.2.2.3 Use case sau khi đăng nhập



Hình 2.2.2.4 Usecase sau khi đăng nhập và đặt câu hỏi



Hình 2.2.2.5 Usecase sau khi đăng nhập và trả lời câu hỏi

2.2.3-Mô Hình Sequence Diagram



Hình 2.2.3.1 Sequence Diagram đăng nhập



Hình 2.2.3.2 Sequence Diagram đăng ký



Hình 2.2.3.3 Sequence Diagram đăng câu hỏi



Hình 2.2.3.4 Sequence Diagram trả lời câu hỏi



Hình 2.2.3.5 Sequence Diagram gửi tin nhắn

2.6 Thiết kế cơ sỡ dữ liệu



Hình 2.6.1 Thiết kế cơ sỡ dữ liệu

CHƯƠNG 3 – XÂY DỰNG ENGINE TRẢ LỜI TỰ ĐỘNG

3.1 Kiến trúc tổng quan



3.2 Phương pháp:

3.2.1 Lọc dữ liệu sử dụng Full Text Search

* Để tăng hiệu năng, thời gian truy xuất dữ liệu trong cơ sở dữ liệu, đồng thời loại bỏ những dữ liệu không liên quan đến câu hỏi từ người dùng nhập vào, chúng ta sẽ sử dụng một kỹ thuật gọi là Full Text Search (FTS)
* FTS là một kỹ thuật tìm kiếm trên cơ sở dữ liệu dạng văn bản. Sở dĩ FTS có hiệu năng và tốc độ truy xuất cao hơn các kỹ thuật tìm kiếm thông thường là nhờ sử dụng phương pháp đánh chỉ mục (Indexing). Mỗi từ trong văn bản mới chèn vào sẽ được lưu vào một mảng gồm địa chỉ các văn bản chứa từ đó

Ví dụ:

D1= “Đây là văn bản thứ nhất”.

D2= “Đây là hai văn bản”.

D3=” Một hai”.

* Việc tạo index như vậy sẽ giúp cho việc tìm kiếm nhanh hơn. Thay vì phải tìm kiếm từng văn bản 1, ta chỉ tìm kiếm dựa trên các phép kết.
* Các câu hỏi sau khi tìm kiếm được sẽ là một danh sách mà mỗi câu hỏi trong danh sách này sẽ chứa ít nhất 1 từ trùng với các từ trong câu hỏi người dùng nhập vào.

3.2.2 Tách từ

* Việc đầu tiên trong xử lý câu hỏi từ người dùng nhập vào đó là Tách từ. Tách từ là phương pháp xử lý xác định các từ có nghĩa trong một văn bản Tiếng Việt. Mục đích của việc tách từ nhầm lọc ra những từ có nghĩa trong Tiếng Việt để nâng cao khả năng tìm kiếm các câu hỏi tương tự.
* Ví dụ:

Cho một câu: “Xã hội ngày càng phát triển”

Sau khi tiến hành xử lý tách từ, ta sẽ nhận được một câu: “Xã\_hội ngày càng phát\_triển”.

* Trong luận văn này, chúng em sử dụng công cụ tách từ vnTokenizer của tác giả Lê Hồng Phương, cộng cụ này được sử dụng trong quá trình phân tích câu hỏi người dùng nhập vào và phân tích các câu hỏi trong cơ sở dữ liệu.

3.2.3 Phân loại câu hỏi sử dụng Maximum entropy classifier

- Sau khi câu hỏi từ người dùng nhập vào được chuyển thành chữ in thường và đi qua giai đoạn tách từ. Chúng ta sẽ phân loại, kiểm tra xem câu hỏi đó thuộc loại nào trong các loại “What”, “When”, “Where”, ”Which”, ”Who”, ”How”, ”YesNo”. Để phân loại thanh các câu hỏi như vậy chúng ta sẽ sử dụng phương pháp tính Maximum entropy

3.2.4 Maximum entropy là gì?

- Maximum entropy là một kỹ thuật học máy có mục địch. Nó cung cấp các ước tính ít sai lệch nhất có thể dựa vào những thông tin nhất định cho trước được gọi là tập dữ liệu mẫu.

3.2.5 Maximum entropy classifier

- Maximum entropy classifier là một bộ phân loại thường được sử dụng trong xử lý ngôn ngữ tự nhiên. Nó cho phép xây dựng các đặc trưng từ tập dữ liệu huấn luyện. Tự động phân loại 1 câu theo 1 chủ đề dựa trên các đặc trưng đó.

- Để áp dụng Maximum entropy classifier chúng ta cần chọn ra một tập hợp các đặc trưng từ tập dữ liệu dữ liệu huẩn luyện để cài đặt các truy vấn. Cụ thể, chúng ta sử dụng số lượng từ xuất hiện như là các đặc trưng.

- Với mỗi từ ta đưa ra được một đặc tính như sau:

Trong đó: - là số lần từ *w* xuất hiện trong loại d.

- *N(d)* là số lượng từ trong loại d.

- Như vậy nêu 1 từ w thường xuyên xuất hiện trong loại d nào đó, ta sẽ tính được trọng số của từ này. Nếu 1 từ có trọng số cao thì khả năng nó thuộc loại d cao.

- Ví dụ: Trong lĩnh vực phân loại các loại câu hỏi trong bai toán này, chúng ta có 7 loại câu hỏi là What, Where, When, Which, How, Why, YesNo.

- Các thống kê dữ liệu cho rằng 90% các loại câu hỏi When có chứa từ “khi nào”. Như vậy nếu D có chứa từ “khi nào” thì xác suất nó thuộc loại câu hỏi When là 90%, đây cũng gọi là 1 ràng buộc của mô hình và xác suất vào các loại còn lại là 2,5 %

3.2.6 Phân loại văn bản sự dụng công cụ Stanford Classifier:

- Stanford classifier là một công cụ cho phép chúng ta chia dữ kiệu thành các chủ đề từ một tập các dữ liệu đào tạo cho trước, tạo ra một bộ phân loại dưa trên Maximum entropy classifier

- Trong bài toán này, chúng ta sẽ phân loại các câu hỏi trong Tiếng Việt thành 7 chủ để: “What”, “Where”, “When”, “Which”, “How”, “Why”, “YesNo”

- Trước tiên chúng ta cần tạo ra một tập dữ liệu đặc trưng, Mỗi chủ đề chúng ta sẽ lấy từ 25-30 câu. Tập dữ liệu đặc trưng là một tập tin có phẩn mở rộng là “.train”

- Tạo một tập tin có phần mở rộng là “.test” chứa nội dung câu hỏi và chủ đề chúng ta mong đợi sau khi chạy qua chương trình.

- Tiếp đó là tập tin có phần mở rộng là “.prop” trình bày các thuộc tính, các cấu trúc quy định trong tập tin .train mà chương trình sẽ học.

- Sau khi có cả 3 tập tin trên, chúng ta sẽ tiến hành cho chương trình tính toán.

3.2.7 Đo độ tương tự giữa các câu hỏi.

- Giải thuật dựa trên 3 yếu tố chính sau:

* Sự ánh xạ độ tương tự từ ngữ: 1 chuỗi ít khác biệt có thể xem là tương tự.
* Sự thay đổi trật tự từ: 2 chuỗi bao hàm từ giống nhau nhưng theo 1 trật tự khác có thể được xem là tương tự
* Không phụ thuộc vào ngôn ngữ : Giải thuật không phải chỉ hoạt động được với tiếng anh mà còn nhiều thứ tiếng khác
* Ví dụ ta có 2 câu hỏi như sau:

1. Tín chỉ bao nhiêu tiền
2. Một tín chỉ bao nhiêu

Là 2 câu có độ tường tự bằng nhau vì câu 1 và câu 2 chỉ khác nhau ở vị trí các từ.

- Ta có 1 bài toán sau:

1. Chúng ta sẽ tính độ tương tự bằng cách chia các từ thành các bi gram(2-gram)

- Chúng ta sẽ tách ra 3 cặp mỗi cặp 2 chữ cái

1. S1 = FRANCE: {FR, RA, AN, NC, CE}
2. S2 = FRENCH: {FR, RE, EN, NC, CH}

Độ tương tự(s1,s2) =

* Sự giống nhau giữa hai chuỗi s1 và s2 được tính bằng số lượng các cặp từ chung ở cả hai chuỗi nhân 2 chia cho tổng số các cặp từ trong hai chuỗi. Giải thuật sẽ đo độ tương tự giữa 2 chuỗi bằng 0 nếu như 2 chuỗi đó khác nhau, bằng 1 nếu như 2 chuỗi đó giống hệt nhau. Nêu kết quả cho ra càng gần bằng 1 thì khả năng 2 chuỗi đó tương tự nhau càng cao và ngược lại nếu kết quả càng gần với số 0 thì 2 chuỗi đó càng khác nhau.
* Để so sánh từ FRANCE và FRENCH, các số liệu được tính như sau

= = 0.4

- Độ tương tự đo được luôn nằm trong khoảng từ 0 – 1.Ở giải thuật này ta tính được độ tương tự của 2 từ FRANCE và FRENCH là 40%.

CHƯƠNG 4– CƠ SỞ LÝ THUYẾT / NGHIÊN CỨU THỰC NGHIỆM

4.1 Quá trình xây dựng các câu hỏi thử nghiệm:

4.2 Đánh giá giải thuật:

4.2.1 Mô hình 1-gram

- Để đánh giá tính đúng đắn của giải thuật chúng ta sẽ chạy bộ đánh giá.Việc chạy bộ đánh giá được chia ra 2 trường hợp:

* Lấy 1 câu hỏi có số điểm cao nhất
* Lấy 3 câu hỏi có số điểm cao nhất
* Sau đây là bảng so sánh 2 trường hợp so sánh:

|  |  |  |
| --- | --- | --- |
| Tiêu chí đánh giá | Lấy 1 câu hỏi có số điểm cao nhất | Lấy 3 câu hỏi có số điểm cao nhất |
| Số lượng câu đánh giá | 30 | 30 |
| Số lượng câu nguồn để kiểm tra | 44 | 44 |
| Số lượng câu đúng |  |  |
| Tỉ lệ đúng | 56% | 80% |

* Vậy ta có thể thấy rõ việc lấy 1 câu hỏi có số điểm cao nhất cho kết quả đúng thấp hơn
* Việc lấy ra 3 câu hỏi tương tự được khuyến khích hơn do tỉ lệ xuất hiện câu hỏi tương tự trong cở sở dữ liệu cao hơn.
* Tạo điều kiện cho người sử dụng dễ dàng tìm được câu trả lời cho thắc mắc của mình hơn.

4.2.2 Mô hình 2-gram

|  |  |  |
| --- | --- | --- |
| Tiêu chí đánh giá | Lấy 1 câu hỏi có số điểm cao nhất | Lấy 3 câu hỏi có số điểm cao nhất |
| Số lượng câu đánh giá | 30 | 30 |
| Số lượng câu nguồn để kiểm tra | 44 | 44 |
| Số lượng câu đúng |  |  |
| Tỉ lệ đúng | 60% | 63% |

**TÀI LIỆU THAM KHẢO**

**Tiếng Việt**

**Tiếng Anh**

**PHỤ LỤC**

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung luận văn như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của luận văn

[**http://ktmt.github.io/blog/2013/10/27/full-text-search-engine/**](http://ktmt.github.io/blog/2013/10/27/full-text-search-engine/)

[**http://viet.jnlp.org/kien-thuc-co-ban-ve-xu-ly-ngon-ngu-tu-nhien/machine-learning-trong-nlp/thuat-toan-entropy-cuc-dai**](http://viet.jnlp.org/kien-thuc-co-ban-ve-xu-ly-ngon-ngu-tu-nhien/machine-learning-trong-nlp/thuat-toan-entropy-cuc-dai)

**http://luanvan.co/luan-van/xay-dung-mot-he-thong-hoi-dap-tu-dong-phuc-vu-tu-van-ghi-danh-truc-tuyen-36867/**