

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG



LUẬN VĂN TỐT NGHIỆP
NGÀNH KHOA HỌC MÁY TÍNH

Đề tài
XÂY DỰNG TRỢ LÝ ẢO
TƯ VẤN NGÀNH HỌC

Sinh viên thực hiện: Võ Ngọc Long

Mã số: B1812282

Khóa: 44

Cần Thơ, 12/2022

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG



LUẬN VĂN TỐT NGHIỆP
NGÀNH KHOA HỌC MÁY TÍNH

Đề tài
XÂY DỰNG TRỢ LÝ ẢO
TƯ VẤN NGÀNH HỌC

Giáo viên hướng dẫn:
Ths. Phạm Xuân Hiền

Sinh viên thực hiện:
Võ Ngọc Long
Mã số: B1812282
Khóa: 44

Cần Thơ, 12/2022

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Cần Thơ, ngày ... tháng ... năm 2022

Giáo viên hướng dẫn

LỜI CẢM ƠN

Lời đầu tiên em xin được phép bày tỏ lòng biết ơn chân thành và sâu sắc đến cô Phạm Xuân Hiền trường CNTT & TT đã tận tình hướng dẫn, chỉ bảo, góp ý trong suốt quá trình thực nghiên cứu và thực hiện đề tài luận văn.

Cùng với đó em xin gửi lời cảm ơn chân thành đến các Thầy Cô Đại học Cần Thơ, đặc biệt là cảm ơn các thầy cô Trường Công nghệ Thông tin và Truyền thông đã hướng dẫn, giảng dạy cho em những kiến thức làm nền tảng thực hiện luận văn. Trong suốt nhiều năm học tập và rèn luyện tại Đại học Cần Thơ, đặc biệt là Trường Công nghệ Thông tin và Truyền thông, đến nay em đã kết thúc khóa học và hoàn thành luận văn tốt nghiệp.

Lời cuối cùng, em xin gửi lời cảm ơn đến cha mẹ, gia đình là chỗ dựa tinh thần vững chắc, là nguồn động lực để em tiếp tục cố gắng trên con đường học tập của mình. Em cảm ơn thầy cô đã luôn lo lắng, dạy dỗ, giúp đỡ em trên con đường học tập. Cảm ơn người sử dụng bè, anh chị cùng trường đã giúp đỡ, chia sẻ khó khăn, cho em những góp ý chân thành.

Do không thể tránh được thiếu sót trong lúc thực hiện luận văn. Nhưng vẫn cố gắng hết sức để hoàn thành luận văn của mình một cách tốt nhất. Mong nhận được sự góp ý quý báu từ thầy cô và các người sử dụng để em có thêm kinh nghiệm thực hiện các nghiên cứu sau này.

Cần Thơ, ngày 02 tháng 12 năm 2022

Người viết

Võ Ngọc Long

MỤC LỤC

MỤC LỤC.....	i
DANH MỤC HÌNH ẢNH	iii
DANH MỤC BẢNG BIỂU	v
DANH MỤC TỪ VIẾT TẮT.....	vi
TÓM TẮT	vii
ABSTRACT	viii
CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN	1
1.1. Đặt vấn đề	1
1.2. Nghiên cứu liên quan.....	2
1.2.1. Website	2
1.2.2. Chatbot	2
1.3. Mục tiêu đề tài	5
1.4. Đối tượng và phạm vi nghiên cứu	5
1.5. Phương pháp nghiên cứu	6
1.6. Nội dung nghiên cứu	6
1.7. Bố cục	7
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT	8
2.1. Mô tả chi tiết bài toán	8
2.1.1. Vấn đề và giải pháp liên quan đến bài toán.....	9
2.1.2. Công cụ xây dựng hệ thống.....	10
2.1.3. Phương pháp học sâu (Deep Learning)	14
2.1.4. Một số giải thuật đánh giá độ chính xác.....	20
2.1.5. Phương pháp đánh giá độ chính xác.....	24
2.2. Mô tả hệ thống trang web	25
2.3. Yêu cầu chức năng.....	26
2.3.1. Người dùng quản trị	26
2.3.2. Người dùng.....	26
CHƯƠNG 3. THIẾT KẾ VÀ CÀI ĐẶT GIẢI THUẬT.....	27
3.1. Thiết kế hệ thống	27
3.1.1. Tập dữ liệu.....	28
3.1.2. Tiền xử lý dữ liệu	33
3.1.3. Xây dựng dữ liệu huấn luyện	34

3.1.4. Xây dựng mô hình huấn luyện	37
3.1.5. Xử lý câu hỏi người dùng và đưa ra dự đoán.....	39
3.1.6. Thu thập và thêm dữ liệu từ câu hỏi người dùng	41
3.1.7. Cài đặt các mô hình	42
3.2. Xây dựng hệ thống	44
3.2.1. Tạo mô hình dự đoán.....	45
3.2.2. Xử lý câu hỏi người dùng và dự đoán câu trả lời.....	48
3.2.3. Xây dựng hệ thống website	50
3.2.4. Giao diện của ứng dụng website	51
CHƯƠNG 4. ĐÁNH GIÁ KẾT QUẢ HUẤN LUYỆN VÀ GIAO DIỆN HỆ THỐNG WEBSITE	52
4.1. Đánh giá và so sánh kết quả huấn luyện.....	52
4.1.1. Độ chính xác của ba giải thuật máy học.....	52
4.1.2. Độ chính xác của mô hình với giải thuật SGD.....	53
4.2. Giao diện website	55
4.2.1. Cấu trúc của website.....	55
4.2.2. Quá trình thu thập câu hỏi	55
CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....	59
5.1. Kết luận nghiên cứu.....	59
5.2. Hướng phát triển hệ thống.....	59
5.2.1. Dữ liệu huấn luyện	59
5.2.2. Chức năng.....	59
5.2.3. Huấn luyện mô hình	60
TÀI LIỆU THAM KHẢO.....	61

DANH MỤC HÌNH ẢNH

Hình 1: Sơ đồ hệ thống Chatbot cho sinh viên Công nghệ thông tin	3
Hình 2: Mô hình Chatbot hỗ trợ khách hàng [3].....	4
Hình 3: Minh họa việc giao tiếp giữa người dùng và Chatbot.....	8
Hình 4: Xử lý dấu câu	10
Hình 5: Xử lý các từ không quan trọng.....	10
Hình 6: Xử lý thanh âm.....	10
Hình 7: Thư viện Keras.....	11
Hình 8: Mô hình ứng dụng web với Flask Framework.....	12
Hình 9: Mô hình web API.....	13
Hình 10: Hệ quản trị cơ sở dữ liệu SQLite	14
Hình 11: Mã hóa từ trong Python	14
Hình 12: Mô hình mạng nơ-ron nhân tạo.....	15
Hình 13: So sánh hội tụ giữa SGD và GD	17
Hình 14: Đồ thị Hàm Sigmoid	18
Hình 15: Đồ thị hàm Tanh	19
Hình 16: Đồ thị hàm ReLU	20
Hình 17: Các đường phân chia SVM.....	21
Hình 18: kNN với k=1 và k=5	22
Hình 19: Định lý Bayes.....	23
Hình 20: Định lý Bayes mở rộng	23
Hình 21: Phân lớp sử dụng Naïve Bayes	24
Hình 22: Mô hình hoạt động của phương pháp đánh giá Hold-out	25
Hình 23: Giao diện ứng dụng website.....	25
Hình 24: Các chức năng của người quản trị.....	26
Hình 25: Các chức năng của người oke	26
Hình 26: Lưu đồ hoạt động của hệ thống.....	27
Hình 27: Nội dung tập tin info.json	29
Hình 28: Cấu trúc tập tin data.json.....	29
Hình 29: Cấu trúc tập dữ liệu đối với nhãn diemchuan	31
Hình 30: Danh sách các ngành Máy tính và Công nghệ thông tin	32
Hình 31: Loại bỏ các ký tự kéo dài	33
Hình 32: Chuyển đổi dữ liệu.....	34
Hình 33: Xử lý câu hỏi tạo tập hợp từ.....	35
Hình 34: Tạo véc-tơ ứng với độ dài tập hợp mảng từ.....	36
Hình 35: Chuyển đổi câu hỏi thành véc-tơ	36
Hình 36: Tạo véc-tơ ứng với số lượng nhãn	37

Hình 37: Câu hỏi được chuyển hóa thành véc-tơ.....	37
Hình 38: Mô hình mạng nơ-ron của mô hình Chatbot.....	38
Hình 39: Quá trình xử lý câu hỏi đơn ý	40
Hình 40: Quá trình xử lý câu hỏi đa ý.....	41
Hình 41: Dữ liệu câu hỏi có câu trả lời	41
Hình 42: Dữ liệu câu hỏi không có câu trả lời	42
Hình 43: Giải thuật SVM với tham số $C=1$ và $C=100$	43
Hình 44: Công thức khoảng cách K – láng giềng	43
Hình 45: Công thức cho giải thuật	44
Hình 46: Tạo mô hình dự đoán	45
Hình 47: Xử lý câu hỏi.....	45
Hình 48: Tạo tập hợp từ	46
Hình 49: Véc-tơ hóa câu hỏi	46
Hình 50: Mô hình mạng nơ-ron	47
Hình 51: Quá trình huấn luyện mô hình	47
Hình 52: Quá trình dự đoán câu trả lời của người dùng	48
Hình 53: Quá trình xử lý câu hỏi người dùng	49
Hình 54: Quá trình đưa ra dự đoán câu trả lời	49
Hình 55: Chatbot trả lời câu hỏi của người dùng.....	50
Hình 56: Mô hình thiết kế website với các ngôn ngữ và công cụ.....	51
Hình 57: Giao diện ứng dụng website Chatbot.....	51
Hình 64: Biểu đồ độ chính xác của 3 giải thuật với nghi thức Hold-out.....	53
Hình 65: Biểu đồ độ chính xác của các giải thuật với nghi thức Hold-out.....	54
Hình 66: Cấu trúc của website	55
Hình 67: Tập hợp câu hỏi của người dùng.....	55
Hình 58: Thông báo đã tồn tại câu hỏi trong tập dữ liệu	56
Hình 59: Hai cách thêm câu hỏi thông qua hệ thống website.....	56
Hình 60: Thêm câu hỏi với nhãn có sẵn	57
Hình 61: Thông báo thêm thành công câu hỏi vào nhãn có sẵn	57
Hình 62: Thêm câu hỏi với nhãn chưa có trong tập dữ liệu.....	57
Hình 63: Thông báo khi thêm thành công nhãn và câu hỏi vào tập dữ liệu	58

DANH MỤC BẢNG BIỂU

Bảng 1: Danh mục từ viết tắt	vi
Bảng 2: Các phương pháp phân lớp và độ chính xác tương ứng	3
Bảng 3: Kết quả nghiên cứu các mô hình cho Chatbot.....	5
Bảng 4: Chức năng các trường trong tập dữ liệu huấn luyện	31
Bảng 5: Số thứ tự các ngành	31
Bảng 6: Danh mục các từ viết tắt	34
Bảng 7: Độ chính xác của ba giải thuật với nghi thức Hold-out.....	52
Bảng 8: Độ chính xác của hai giải thuật với nghi thức Hold-out	54

DANH MỤC TỪ VIẾT TẮT

STT	Từ nguyên gốc	Từ viết tắt
1	Trợ lý ảo	Chatbot
2	Support Véc-tơ Machines	SVM
3	K Nearest Neighbor	kNN
4	Naïve Bayes	Bayes Thơ Ngây
5	Stochastic Gradient Descent	SGD
6	Hypertext Markup Language	HTML
7	Cascading Style Sheets	CSS
8	Application Programming Interface	API
9	Database Management System	DBMS
10	Stochastic Gradient Descent	SGD
11	Gradient Descent	GD
12	Activation Functions	Hàm kích hoạt
13	Rectified Linear Unit	ReLU

Bảng 1: Danh mục từ viết tắt

TÓM TẮT

Hằng năm, vào mỗi mùa tuyển sinh, phụ huynh và các học sinh sẽ có nhu cầu tham khảo về “Tuyển sinh tại các Trường Đại Học, Cao Đẳng”, với tâm thế là trường Đại học lớn tại khu vực Đồng Bằng Sông Cửu Long Đại học Cần Thơ – một điểm Trường nhận được rất nhiều sự quan tâm của phụ huynh và học sinh, những thắc mắc về liên quan đến ngành học và chương trình đào tạo. Do số lượng ngành học tại Đại học Cần Thơ rất lớn, đội ngũ tư vấn của trường thì không thể đáp ứng đáp ứng với số lượng câu hỏi lớn và không thể trả lời một lúc các câu hỏi từ phía học sinh và phụ huynh, nên nghiên cứu này sẽ tập trung ứng dụng Deep Learning để xây dựng hệ thống Chatbot phục vụ trên website. Nghiên cứu “Xây dựng Trợ lý ảo tư vấn ngành học” sẽ thực hiện xây dựng một hệ thống Chatbot giúp người sử dụng có thể trao đổi để tìm kiếm thông tin cần thiết. Hệ thống Chatbot hoạt động dựa trên việc áp dụng giải thuật học sâu Mạng Nơ-ron nhân tạo để xây dựng mô hình huấn luyện và dự đoán câu trả lời, nghiên cứu sử dụng các giải thuật khác như SVM, kNN, Naïve Bayes để so sánh độ chính xác của các mô hình trên tập dữ liệu có 1043 câu hỏi và 101 câu trả lời được thu thập từ nhiều nguồn. Nghiên cứu còn sử dụng những ngôn ngữ lập trình như Python, Javascript, HTML, CSS và một số framework như Flask để xây dựng hệ thống.

Sau khi thực hiện huấn luyện mô hình, kết quả dự đoán chính xác của 4 giải thuật được sử dụng trong nghiên cứu không quá chênh lệch. Hai giải thuật SVM và SGD có độ chính xác xấp xỉ đạt hơn 90% với SGD độ chính xác xấp xỉ đạt 92% nên nghiên cứu quyết định sẽ xây dựng mô hình Chatbot dựa trên thuật toán SGD.

ABSTRACT

Annually, after the national entrance exam, parents and students have a great demand to search for "Admissions at Universities and Colleges". As a major university in the Mekong Delta region, Can Tho University - a university that receives huge attention from parents and students as well as questions related to the field of study and education program of majors. Due to the large number of subjects at Can Tho University, the school's consulting team cannot respond to a large number of questions and cannot answer immediately all the questions from students and parents all the time. This study focuses on applying Deep Learning to build a Chatbot system serving on the website. The study "Building a virtual assistant for academic advising" will build a Chatbot system to help users find the necessary information. Chatbot system work is based on the application of deep learning algorithms Artificial Neural Network to build training models and predict answers. Research uses other algorithms such as SVM, kNN, Naïve Bayes to compare the accuracy of models on a dataset of 1043 questions and 101 answers collected from multiple sources. The research also uses programming languages such as Python, Javascript, HTML, CSS and several frameworks such as Flask to build the system.

After performing model training, the accurate prediction results of the 4 algorithms used in the study were not too different. The two algorithms SVM and SGD have approximate accuracy of more than 90% with SGD approximate accuracy of 92%, so the study decided to build a Chatbot model based on SGD algorithm.

CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN

Chương 1 sẽ trình bày các phần gồm đặt vấn đề, nghiên cứu liên quan, website và ứng dụng, chatbot, máy học, mục tiêu đề tài, đối tượng và phạm vi nghiên cứu, phương pháp nghiên cứu, nội dung nghiên cứu, bố cục luận văn.

1.1. Đặt vấn đề

Hằng năm vào mỗi mùa tuyển sinh, phụ huynh và học sinh có rất nhiều câu hỏi và thắc mắc về các ngành học cũng như những cơ hội nghề nghiệp sau khi ra trường, cách thức tuyển sinh,... Đa phần các trường đại học đều đăng tải những thông tin trên website cũng như đã có đội ngũ tư vấn để hỗ trợ thắc mắc cho phụ huynh và học sinh. Tuy nhiên, các văn bản, thông tin đôi khi sử dụng các thuật ngữ chuyên môn và ít thân thiện với người sử dụng dẫn đến việc phụ huynh, học sinh khó có thể hiểu hết được các thông tin cần thiết và quan trọng của người tư vấn. Vấn đề đặt ra là liệu người tư vấn có thể hỗ trợ những thắc mắc cho mọi người mọi lúc, điều đó đòi hỏi số lượng nhân viên tư vấn nhất định và tốn kém một chi phí khá lớn.

Ngày nay, trong thời đại mà trí tuệ nhân tạo và máy học đang phát triển ngày một rực rỡ, và cho ra đời rất nhiều thiết bị thông minh, không nằm ngoài xu thế đó ứng dụng Chatbot với tích hợp trí tuệ nhân tạo cũng đang hiện hữu rất nhiều trong cuộc sống hiện đại này. Hiện tại Chatbot được ứng dụng trong nhiều lĩnh vực. Nếu người sử dụng muốn trò chuyện, gặp gỡ và trò chuyện về những thứ như nhà hàng, thời tiết và tin tức hàng ngày thì Chatbot Luka có thể dễ dàng đáp ứng được nhu cầu đó. Nhu cầu trong việc tìm một người theo dõi năng suất thể dục hằng ngày một cách thông minh, có tính tương tác thì Chatbot Lark là một huấn luyện viên thể dục mà người sử dụng có thể tìm đến. Tiếng anh không được tốt và muốn cải thiện tiếng anh nhanh chóng và thuận lợi thì Chatbot Andy chính là sự lựa chọn tốt nhất.

Sự ra đời Chatbot mang lại nhiều lợi ích cho rất nhiều lĩnh vực cũng như đáp ứng nhu cầu của mọi người, chúng không chỉ đáp ứng những gì mọi người cần mà vấn đề thời gian cũng được cải thiện. Đặc biệt khi phụ huynh và học sinh muốn biết những thông tin cần thiết cho việc tuyển sinh ở bất kỳ thời gian nào, và họ cần những câu trả lời chính xác với tốc độ nhanh. Để đáp ứng nhu cầu trên thì việc có một Chatbot trả lời tư vấn tuyển sinh là cần thiết.

1.2. Nghiên cứu liên quan

1.2.1. Website

Hiện nay có nhiều loại hình website cung cấp các thông tin tuyển sinh của các trường đại học như: Đại học Cần Thơ¹, Đại học Bách Trường Đại học Quốc gia Thành phố Hồ Chí Minh², ... Tuy nhiên là việc tư vấn cho phụ huynh và học sinh và tìm kiếm thông tin ngành học trên website này còn hạn chế là khi người sử dụng có những thắc mắc về các ngành nghề tuyển sinh, về chỉ tiêu, về tỉ lệ tốt nghiệp và nhiều yếu tố khác. Do chưa hoàn toàn hiểu rõ về trang web, khiến cho người sử dụng sẽ cần một “trợ lý ảo” để hướng dẫn những thao tác, và “trợ lý ảo” cũng có thể cung cấp thêm thông tin cần thiết. Nắm bắt được nhu cầu này, nghiên cứu “Xây dựng Trợ lý ảo tư vấn ngành học” sẽ thực hiện xây dựng một Chatbot với giao diện sử dụng thân thiện cho người sử dụng, giao diện được lấy cảm hứng từ những cửa sổ chat tương tự như những ứng dụng nổi tiếng: Messenger, Zalo,...

1.2.2. Chatbot

Hiện nay chatbot đã trở nên rất phổ biến và ứng dụng trong khá nhiều lĩnh vực khác nhau. Trong đó có những tập đoàn lớn về công nghệ trên thế giới ra sức tạo những trợ lý ảo để người dùng có thể tương tác với các phần mềm, ứng dụng của họ.

Đỗ Viết Mạnh xây dựng chatbot bán hàng sử dụng mô hình sinh trên nền tảng framework Rasa³ và Django⁴. Kết quả thử nghiệm của chatbot đạt độ chính xác là 80% dựa vào các câu hỏi đã được xây dựng trước đó theo kịch bản áp dụng cho hệ thống bán hàng.

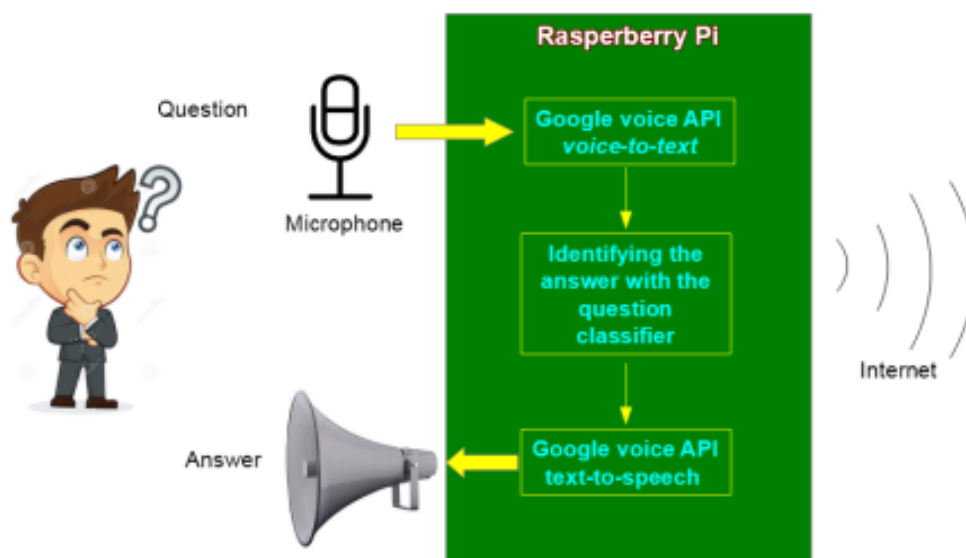
Nghiên cứu của tác giả Đỗ Thanh Nghị đã xây dựng Chatbot trả lời tự động cho Sinh viên Công Nghệ Thông Tin [1]. Chatbot này được máy Raspberry Pi có thể trả lời tự động cho sinh viên ngành Công nghệ thông tin các câu hỏi liên quan đến môi trường học tập và phương pháp học tập bậc đại học, kỹ năng nghề nghiệp và xu hướng công nghệ trong tương lai. Sinh viên thực hiện đặt câu hỏi cho Chatbot bằng cách nói bằng giọng nói Chatbot tiếp nhận xử lý và trả lời sinh viên bằng văn bản và phát ra câu trả lời bằng tiếng nói cho sinh viên nghe. Dữ liệu được thu thập và biên soạn gồm 986 câu hỏi và 213 câu trả lời từ nguồn dữ liệu tài liệu học tập.

¹ <https://tuyensinh.ctu.edu.vn/>

² <http://tuyensinh.hcmut.edu.vn/admission/>

³ <https://rasa.com/>

⁴ <https://www.djangoproject.com/>



Hình 1: Sơ đồ hệ thống Chatbot cho sinh viên Công nghệ thông tin

Nhóm nghiên cứu đã thực hiện huấn luyện mô hình bằng các phương pháp học máy và kết quả được ghi nhận và thống kê trong bảng sau:

Phương pháp	Độ chính xác
Máy học Véc-tơ hỗ trợ	76.77%
Mạng nơ-ron nhân tạo	72.73%
Rừng ngẫu nhiên	71.72%
k láng giềng	65.66%

Bảng 2: Các phương pháp phân lớp và độ chính xác tương ứng

Windiatmoko [2] phát triển bởi các mô hình học sâu, chatbot được tích hợp vào Facebook⁵, được áp dụng bởi một mô hình trí tuệ nhân tạo nhằm tái tạo trí thông minh của con người với một số chương trình đào tạo cụ thể. Loại học sâu này dựa trên RNN⁶ có một số sơ đồ tiết kiệm bộ nhớ cụ thể cho mô hình học sâu, cụ thể là chatbot này sử dụng LSTM⁷ đã được tích hợp bởi khung Rasa. LSTM giúp tiết kiệm hiệu quả một số bộ nhớ cần thiết nhưng sẽ loại bỏ một số bộ nhớ không cần thiết.

Ngoài ra, Mamatha tích hợp chatbot lên website thương mại điện tử giúp tư vấn khách hàng. Chatbot này sẽ hữu ích giúp bộ phận chăm sóc khách hàng phản ứng nhanh, có thể giải quyết nhiều câu hỏi từ khách hàng.

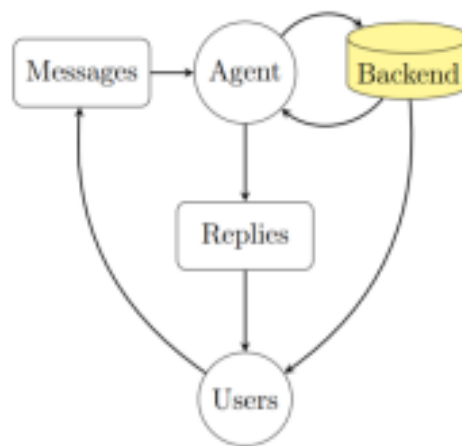
⁵ <https://www.facebook.com/>

⁶ https://en.wikipedia.org/wiki/Recurrent_neural_network

⁷ https://en.wikipedia.org/wiki/Long_short-term_memory

Trong lĩnh vực dịch vụ, nghiên cứu của giáo sư Louis Wehenkel có liên quan đến việc ứng dụng Chatbot trong việc hỗ trợ khách hàng [3]. Nghiên cứu đặt trong những trường hợp thực tế về giải pháp sử dụng Chatbot cho công ty Gaming1. Mục tiêu nghiên cứu là thiết kế một giải thuật Chatbot đa ngôn ngữ có thể mở rộng và dễ dàng giao tiếp với các phần mềm hỗ trợ khách hàng hiện tại của công ty.

Quá trình xây dựng gồm việc xử lý dữ liệu được trích xuất thành một định dạng phù hợp cho các giải pháp máy học và triển khai trên nhiều nền tảng, có thể trả lời người dùng trong thời gian thực. Và yêu cầu người dùng cung cấp thông tin nếu ý định họ không rõ ràng, đặc biệt hỗ trợ nhiều ngôn ngữ.



Hình 2: Mô hình Chatbot hỗ trợ khách hàng [3]

Dữ liệu được sử dụng trong quá trình nghiên cứu được trích xuất từ tập dữ liệu tin nhắn từ công ty có hệ thống hỗ trợ khách hàng. Không gian dữ liệu có khoảng 1.400.000 thông tin được công ty thu thập từ 2013. Khi dữ liệu được trích xuất, sẽ được phân tách theo ngôn ngữ bằng cách sử dụng các thẻ gắn liền với dữ liệu. Hiện có 9 ngôn ngữ được hỗ trợ trong hệ thống Gaming1: tiếng Pháp, tiếng Hà Lan, tiếng Anh, tiếng Bồ Đào Nha, tiếng Tây Ban Nha, tiếng Romania, tiếng Serbia, tiếng Đức và tiếng Thổ Nhĩ Kỳ.

Sau khi dữ liệu được xử lý sẽ được huấn luyện thông qua 4 mô hình gồm: LSTM đơn, GRU đơn, đầu vào GRU đảo ngược, GRU 2 lớp. Kết quả thu lại như bảng sau:

Network type	Best validation loss (lower is better)	Best validation accuracy (higher is better)	Total training time [min]
Single LSTM	0.45	87.6%	314
Single GRU	0.46	87.5%	196
Inverted GRU input	0.49	88.2%	198
Two GRU layers	0.52	86.4%	352

Bảng 3: Kết quả nghiên cứu các mô hình cho Chatbot

Kết quả của nghiên cứu thể hiện rằng giải pháp đáp ứng yêu cầu và cho phép dễ dàng khai và phát triển nếu cần thực hiện các thay đổi. Nó có thể được hoạt động trên mọi hệ thống miễn là nó có thể lưu trữ môi trường Python.

1.3. Mục tiêu đề tài

Trong quá trình xây dựng, nghiên cứu “Xây dựng Trợ lý ảo tư vấn ngành học” đặt ra những mục tiêu phải đạt được về lý thuyết và thực hành. Về mặt lý thuyết hiểu được cách sử dụng các ngôn ngữ lập trình, những công cụ hỗ trợ để thiết kế kế hoạch từng phần trong việc xây dựng một hệ thống trang web có tích hợp Chatbot. Về mặt thực hành, tạo lập biến môi trường, cài đặt các giải thuật phù hợp, sử dụng ngôn ngữ lập trình cho từng mục đích, chức năng riêng biệt. Kết hợp các phương pháp để xây dựng một hệ thống hoàn chỉnh.

Xây dựng trang web tích hợp Chatbot tư vấn tuyển sinh với mục tiêu để đáp ứng được các chức năng sau:

- Thiết kế giao diện dễ sử dụng, thân thiện.
- Hỗ trợ người quản trị trong việc quản lý cơ sở dữ liệu.
- Hỗ trợ chatbot tư vấn và hỗ trợ, giải đáp thắc mắc về các ngành, nhóm ngành Công nghệ Thông Tin.

1.4. Đối tượng và phạm vi nghiên cứu

Chatbot được xây dựng nhằm mục đích hỗ trợ giải đáp các thắc mắc về thông tin tuyển sinh về các ngành của Đại học Cần Thơ. Đối tượng Chatbot hướng tới là những phụ huynh và các em học sinh có ý định thi hoặc xét tuyển vào các nhóm ngành này.

Ngoài ra đối tượng đề tài còn là ngôn ngữ lập trình Python⁸, thư viện Keras⁹, NLTK¹⁰ được sử dụng để lập trình hệ thống Chatbot, sử dụng HTML¹¹, CSS¹² và Javascript¹³ để thiết kế giao diện cho hệ thống.

1.5. Phương pháp nghiên cứu

Để tạo lập tập dữ liệu cho quá trình huấn luyện cần phải thực hiện tìm hiểu thông tin tuyển sinh và thu thập câu hỏi của người dùng. Việc tìm hiểu thông tin tuyển sinh được thực hiện bằng cách theo dõi trang thông tin tuyển sinh của Đại học Cần Thơ về các ngành thuộc Trường Công nghệ Thông tin và Truyền thông, tham khảo về điểm chuẩn, số lượng chỉ tiêu, phương thức xét tuyển,.. để cập nhật nhanh chóng những thông tin mới nhất cho tập dữ liệu. Câu hỏi và câu trả lời sẽ được thu thập bằng việc sử dụng công cụ Google Form để tạo form và thực hiện khảo sát trực tuyến, sử dụng khảo sát thực tế đối với các bạn học sinh tại một số điểm trường cấp 3, thu thập các bộ câu hỏi về “Tuyển sinh các ngành Công nghệ Thông tin”.

Nghiên cứu “Xây dựng Trợ lý ảo tư vấn tuyển sinh” được lập trình chủ yếu bằng ngôn ngữ Python và đặc biệt là có sự hỗ trợ của một số thư viện như Keras, NLTK,... các phương pháp học máy như SVM, mạng nơ ron nhân tạo. Công cụ hỗ trợ cho việc lập trình là phần mềm Pycharm.

Trong quá trình xây dựng chương trình cần tiền xử lý dữ liệu, phân chia các từ trong câu thành từ nhóm và xây dựng một tập dữ liệu. Chia tập dữ liệu thành các phần training và testing dựa trên dữ liệu có sẵn.

Xây dựng mạng nơ-ron bằng cách dùng Sequential API của Keras.

Trang web được xây dựng dựa trên ý tưởng về cửa sổ chat trong các ứng dụng như Facebook, Zalo,... được lập trình trên các ngôn ngữ HTML, CSS và Javascript. Công cụ hỗ trợ xây dựng trang web là phần mềm Visual Studio Code, với Extension hỗ trợ trong việc kiểm tra chạy thử trang web là: Live Server.

1.6. Nội dung nghiên cứu

Các nội dung nghiên cứu của đề tài:

- Nghiên cứu xây dựng front-end webiste bằng các ngôn ngữ: HTML, Bootstrap(CSS), JQuery, JavaScript.
- Nghiên cứu xây dựng back-end website sử dụng.

⁸ <https://vi.wikipedia.org/wiki/Python>

⁹ <https://en.wikipedia.org/wiki/Keras>

¹⁰ https://en.wikipedia.org/wiki/Natural_Language_Toolkit

¹¹ <https://vi.wikipedia.org/wiki/HTML>

¹² <https://vi.wikipedia.org/wiki/CSS>

¹³ <https://vi.wikipedia.org/wiki/JavaScript>

- Nghiên cứu hệ thống phân loại đánh giá sử dụng framework Python Flask.

1.7. Bố cục

Phần giới thiệu

Giới thiệu tổng quát về đề tài.

Phần nội dung

Chương 1: Giới thiệu tổng quan

Chương 2: Cơ sở lý thuyết

Chương 3: Thiết kế và cài đặt giải thuật

Chương 4: Đánh giá kết quả huấn luyện và Giao diện hệ thống website

Phần kết luận

Chương 5: Kết luận và Hướng phát triển

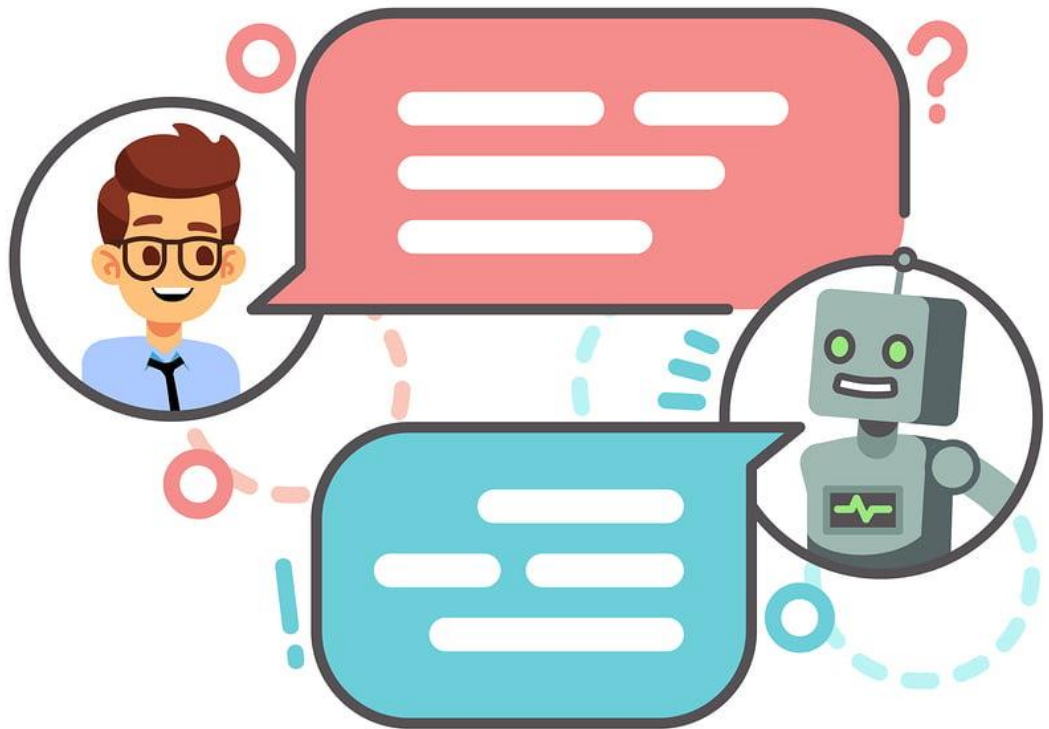
Trình bày kết quả đạt được và hướng phát triển hệ thống.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Chương 2 sẽ trình bày các mô tả chi tiết bài toán, những kỹ thuật dùng trong quá trình xây dựng hệ thống bao gồm những ngôn ngữ lập trình, công cụ, cơ sở dữ liệu, framework, tool,...

2.1. Mô tả chi tiết bài toán

Nhằm đáp ứng nhu cầu giải đáp các thắc mắc liên quan đến việc tuyển sinh trường Đại học Cần Thơ của quý phụ huynh và học sinh. Vì thế, chương trình phải được xây dựng với một cấu trúc đơn giản, dễ sử dụng và thân thiện với người dùng về mặt giao diện và tính năng của chương trình. Câu trả lời phải đáp ứng được các tiêu chí chính xác cao và phải có tốc độ xử lý cao, khi một yêu cầu được đặt ra vào bất kể thời gian nào thì hệ thống phải có câu trả lời cho mọi người.



Hình 3: Minh họa việc giao tiếp giữa người dùng và Chatbot

Để thuận tiện cho việc cung cấp thông tin cho quý phụ huynh và các em học sinh trong quá trình tuyển sinh. Chương trình sẽ được xây dựng cho phép người sử dụng hỏi và nhận được câu trả lời từ hệ thống thông qua văn bản, và hệ thống chatbot sẽ được tích hợp trên một trang web có liên kết đường dẫn đến trang web tuyển sinh chính của Đại học Cần Thơ. Khi quý phụ huynh và học sinh đưa ra các

câu hỏi liên quan đến những thông tin về ngành học, hệ thống sẽ tiếp nhận và đưa ra câu trả lời trực tiếp được hiển thị trong cửa sổ chat trên website.

Hệ thống được xây dựng trên hai thành phần chính là: Chatbot và Trang web. Đối với Chatbot thì nghiên cứu thực hiện xây dựng Chatbot bằng ngôn ngữ lập trình Python, kết hợp với sử dụng những thư viện cần thiết để xây dựng các mô hình giải thuật, sử dụng mô hình giải thuật để huấn luyện Chatbot có thể dự đoán câu trả lời dựa trên tập dữ liệu được thu thập từ thông tin tuyển sinh của Đại học Cần Thơ. Đối với trang website, nghiên cứu thực hiện xây dựng website bằng cách sử dụng ngôn ngữ lập trình HTML và CSS để thiết kế giao diện, sử dụng ngôn ngữ lập trình Javascript để lập trình cho các chức năng của website, truyền dữ liệu từ máy chủ lên trang web.

2.1.1. Vấn đề và giải pháp liên quan đến bài toán

Xử lý văn bản tiếng việt

Hệ thống trợ lý ảo tư vấn tuyển sinh dựa trên mô hình câu hỏi từ người dùng và câu trả lời từ Chatbot đều diễn ra dưới hình thức văn bản.

Khi người dùng đặt câu hỏi cho hệ thống thông qua văn bản, ngôn ngữ được sử dụng sẽ là Tiếng Việt. Tiếng Việt là một ngôn ngữ có dấu câu, đặc biệt trong quá trình đặt câu hỏi, người dùng có thể sẽ nhập vào những từ ngữ viết tắt theo thói quen của người dùng. Ví dụ như:

- “Cnnt diem chuan bn?”
- “Ngành máy tính đào tạo nh j?”
- “Học ngành này có khó ko?”,...

Đa phần mọi người việc hiểu những câu viết tắt, những câu không dấu,... là hoàn toàn dễ dàng nhưng đối với máy tính thì sẽ không thể hiểu được vì máy tính sẽ không có dữ liệu hay bất kỳ căn cứ nào để hiểu được những câu từ không nằm trong bất kỳ văn bản chính thức nào. Nên trước khi vào xây dựng chương trình thì cần xử lý dữ liệu thô là các câu hỏi từ người dùng.

Không chỉ việc xử lý dữ liệu thô là các câu hỏi từ người dùng đặt ra trong hệ thống, việc xử lý dữ liệu cho mục đích huấn luyện mô hình cũng cực kỳ quan trọng. Tất cả câu hỏi sẽ được loại bỏ những thứ không cần thiết như:

- Dấu câu [“.”, “?”, “!”, “:”, “...”]

Phương pháp: xây dựng một hàm có tên là Punctuation(string), thực hiện tạo một mảng bao gồm các dấu câu, cho lặp chuỗi cần xử lý trong mảng, nếu tồn tại dấu câu thì sẽ được thay thế bằng rỗng thông qua hàm “replace()”

và bản đồ tương tác, các chức năng của JavaScript có thể cải thiện trải nghiệm người dùng của trang web. Là ngôn ngữ kịch bản phía máy khách, JavaScript là một trong những công nghệ cốt lõi của World Wide Web.

Ví dụ: Khi sử dụng trong hệ thống Chatbot tư vấn tuyển sinh, những thao tác trao đổi với Chatbot giúp những đoạn văn bản được hiện lên chính là được xử lý bằng Javascript.

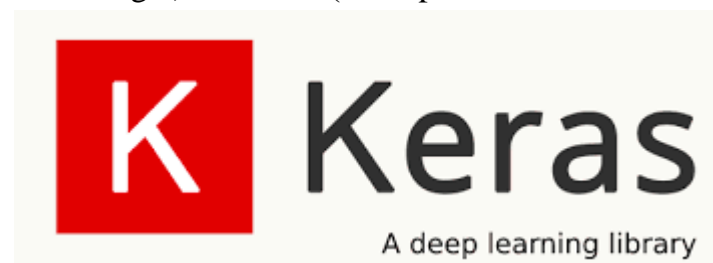
HTML và CSS

HTML là một ngôn ngữ đánh dấu được thiết kế ra để tạo nên các trang web, nghĩa là các mẫu thông tin được trình bày trên World Wide Web.

CSS là định nghĩa về cách hiển thị của một tài liệu HTML. CSS đặc biệt hữu ích trong việc thiết kế web, giúp người thiết kế có thể dễ dàng triển khai các phong cách thiết kế lên bất kỳ trang nào của một website nhanh chóng và đồng bộ.

Thư viện Keras

Keras¹⁴ là một thư viện được phát triển vào năm 2015 bởi Francois Chollet, một kỹ sư nghiên cứu Deep Learning. Keras là một Open Source cho Neural Network được viết bởi ngôn ngữ Python. Và nó được xem là một API bậc cao có thể sử dụng chung với các thư viện Deep Learning nổi tiếng như Tensorflow¹⁵ (được phát triển bởi Google), CNTK¹⁶ (được phát triển bởi Microsoft)...



Hình 7: Thư viện Keras

Keras là API cấp cao của TensorFlow 2: một giao diện dễ tiếp cận, hiệu quả cao để giải quyết các vấn đề về máy học, tập trung vào học sâu hiện đại. Nó cung cấp các khối xây dựng và trừu tượng cần thiết để phát triển và vận chuyển các giải pháp máy học với tốc độ lặp lại cao.

Keras cấp quyền cho các kỹ sư và nhà nghiên cứu để tận dụng tối đa khả năng mở rộng và khả năng đa nền tảng của TensorFlow 2: có thể chạy Keras trên TPU

¹⁴ <https://keras.io/about/>

¹⁵ <https://www.tensorflow.org/federated>

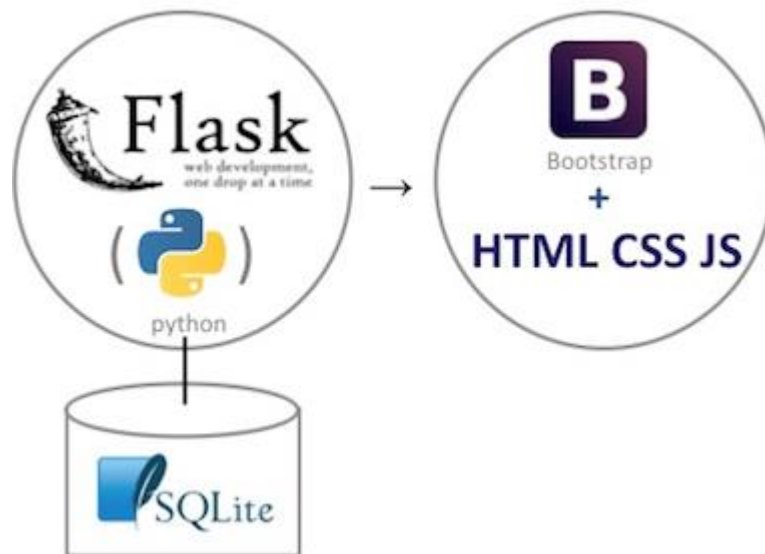
¹⁶ <https://docs.microsoft.com/en-us/cognitive-toolkit/>

hoặc trên các cụm GPU lớn và người sử dụng có thể xuất các mô hình Keras của mình để chạy trong trình duyệt hoặc trên thiết bị di động thiết bị.

Framework Flask

Flask¹⁷ là một web Framework được phát triển bằng ngôn ngữ lập trình Python, không yêu cầu tool hay thư viện cụ thể nào. Flask có hỗ trợ các thành phần tiện ích mở rộng cho ứng dụng như tích hợp: cơ sở dữ liệu, xác thực biểu mẫu, xử lý upload, các công nghệ xác thực, Template, Email, RESTful,.....phụ thuộc vào nhu cầu sử dụng mà người dùng có thể chọn sử dụng tiện ích nào. Người dùng có thể tập trung xây dựng một Web Application ngay từ khi bắt đầu dự án, trong một khoảng thời gian ngắn, theo tiến độ phát triển của dự án mà có thể phát triển ứng dụng theo bất kỳ yêu cầu nào đặt ra.

Để thuận tiện cho người dùng, Flask cung cấp rất nhiều tài liệu hướng dẫn, từ việc cài đặt, thực thi và triển khai, trong đó có cả hướng dẫn nhanh và hướng dẫn chi tiết. Người dùng có thể dễ dàng tìm kiếm những tài liệu này để tham khảo, học tập về ứng dụng web và thực hành cài đặt Flask Framework miễn phí trên internet.



Hình 8: Mô hình ứng dụng web với Flask Framework

API

API¹⁸ là các phương thức, giao thức kết nối với các thư viện và ứng dụng khác. API là viết tắt của Application Programming Interface – giao diện lập trình ứng dụng. API cung cấp khả năng truy xuất đến một tập các hàm hay dùng, từ đó có thể trao đổi dữ liệu giữa các ứng dụng.

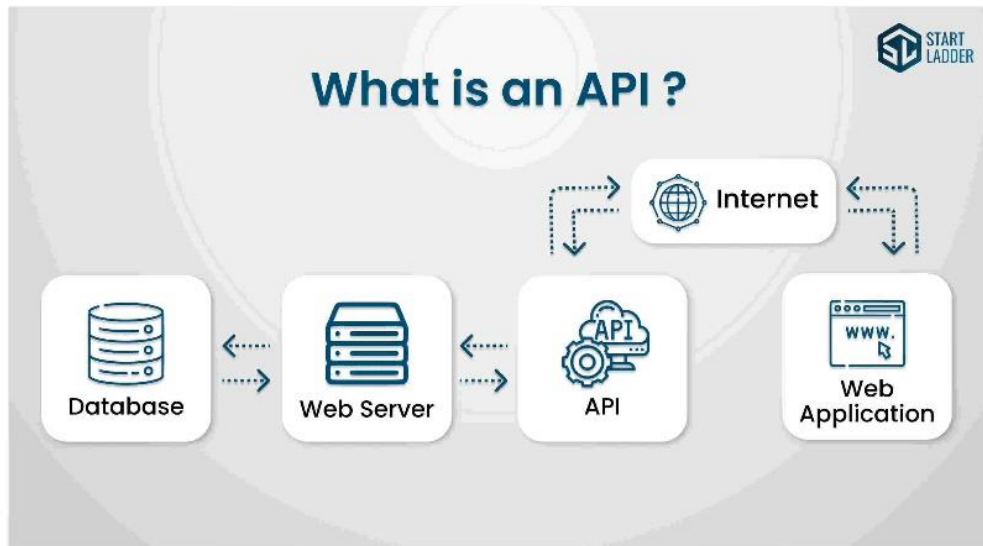
¹⁷ <https://flask.palletsprojects.com/en/2.2.x/>

¹⁸ <https://aws.amazon.com/vi/what-is/api/>

Web API là hệ thống API được sử dụng trong hệ thống website. Đa số các website đều được ứng dụng đến web API cho phép người dùng kết nối, lấy dữ liệu hoặc cập nhật cơ sở dữ liệu.

Ví dụ: Người thiết kế chức năng login vào Google, Facebook, Twitter, Github,... điều này có nghĩa là đang gọi đến API của máy chủ các dịch vụ tương ứng, hoặc như các ứng dụng di động đều lấy dữ liệu thông qua API. Nhóm nghiên cứu đã có một đề tài sử dụng API để lấy dữ liệu từ một ứng dụng du lịch là Booking¹⁹, việc lấy dữ liệu thông qua API sẽ thuận lợi hơn việc cào dữ liệu trực tiếp từ trang web, dữ liệu được lấy về sẽ được lưu trữ dưới dạng tập tin .json.

Phương thức Web API cho phép các ứng dụng khác nhau có thể giao tiếp dữ liệu qua lại. Dữ liệu được web API trả lại thường ở dạng JSON hoặc XML thông qua các giao thức HTTP hoặc HTTPS.



Hình 9: Mô hình web API

Hệ quản trị cơ sở dữ liệu

Trong nghiên cứu này, tập dữ liệu được lưu trữ dưới dạng tập tin .json và những câu hỏi được thu thập sẽ được lưu trữ và sử dụng bằng hệ cơ sở dữ liệu SQLite²⁰.

SQLite là một hệ quản trị cơ sở dữ liệu (DBMS) quan hệ, tương tự như MySQL, Oracle, PostgreSQL,...Đặc điểm nổi bật nhất của SQLite so với các DBMS khác là gọn, nhẹ, đơn giản, đặc biệt là không cần mô hình client – server,

¹⁹ https://github.com/h3x4n1um/CTU-Data_Mining

²⁰ <https://vi.wikipedia.org/wiki/SQLite>

không cần cài đặt, cấu hình hay khởi động nên sẽ không cần user, password hay quyền bên trong SQLite Database, dữ liệu sẽ được lưu ở một file duy nhất.

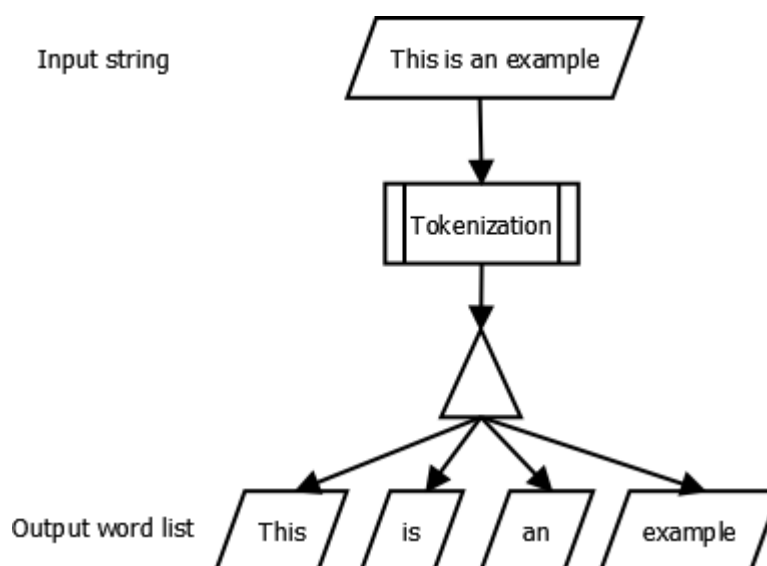


Hình 10: Hệ quản trị cơ sở dữ liệu SQLite

SQLite không yêu cầu một quy trình hoặc hệ thống máy chủ riêng biệt để hoạt động và không cần cấu hình nghĩa là không cần thiết lập hoặc quản trị, SQLite rất nhỏ và trọng lượng nhẹ, dưới 400 KiB được cấu hình đầy đủ. SQLite hỗ trợ hầu hết các tính năng ngôn ngữ truy vấn được tìm thấy trong tiêu chuẩn SQL92.

Word Tokenization

Word Tokenization – Mã hóa từ là quá trình chia nhỏ một mẫu văn bản lớn thành các từ. Đây là một yêu cầu trong các nhiệm vụ xử lý ngôn ngữ tự nhiên, nơi mỗi từ cần được nắm bắt và phân tích sâu hơn như phân loại và đếm chúng cho một cảm xúc cụ thể,... bộ Công cụ Ngôn ngữ Tự nhiên là một thư viện được sử dụng để đạt được điều này. Cài đặt NLTK trong Python trước khi sử dụng để mã hóa từ.

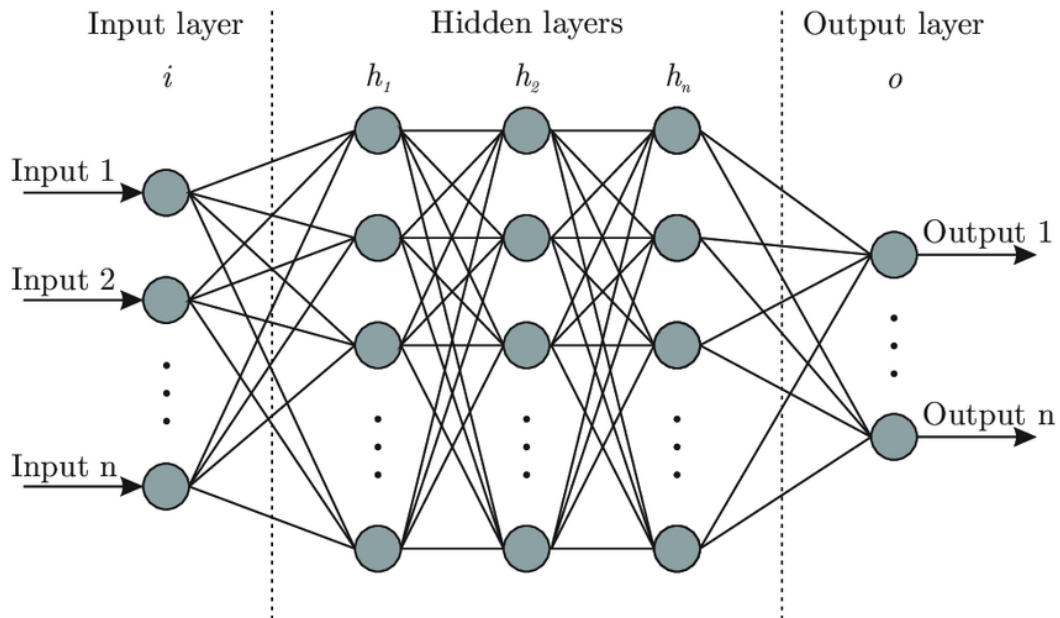


Hình 11: Mã hóa từ trong Python

2.1.3. Phương pháp học sâu (Deep Learning)

Học sâu [4] là một nhánh máy học sử dụng nhiều lớp mạng Nơ-ron nhân tạo (Neural network) để đưa ra một mô hình toán học dựa trên dữ liệu có sẵn. Học sâu

hay Deep Learning²¹ thường được nhắc đến cùng với Dữ liệu lớn (Big Data²²) và Trí tuệ nhân tạo (AI²³) đang được ứng dụng ngày một rộng rãi hơn trong cuộc sống và xuất hiện hầu hết ở những lĩnh vực như: y tế, dịch vụ, trong sản xuất, và được ứng dụng trong cả giảng dạy và rất nhiều lĩnh vực khác.



Hình 12: Mô hình mạng nơ-ron nhân tạo

Các kiến trúc mô hình học sâu khác nhau như: mạng nơ-ron sâu, mã mạng nơ-ron tích chập sâu, mạng niềm tin sâu và mạng nơ-ron tái phát đã được áp dụng cho các lĩnh vực như: thị giác máy tính, tự động nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, nhận dạng âm thanh ngôn ngữ và tin sinh học,... các mô hình học sâu này đã chứng minh là tạo ra được các kết quả rất tốt đối với nhiều nhiệm vụ khác nhau.

Sử dụng phương pháp mạng nơ-ron khi huấn luyện dựa trên tập dữ liệu, mạng nơ-ron sẽ khởi tạo các tập trọng số kí hiệu là weight. Những trọng số này sẽ được tối ưu trong suốt thời gian huấn luyện mô hình, và sẽ tạo ra các trọng số tối ưu hơn. Tổng quát quá trình xử lý một nơ-ron bao gồm các bước như sau:

Bước 1:

Tính tổng các trọng số đầu vào (input) với công thức như sau:

$$Y = \sum (weight \times input) + bias$$

²¹ https://en.wikipedia.org/wiki/Deep_learning

²² <https://topdev.vn/blog/big-data/>

²³ https://en.wikipedia.org/wiki/Artificial_intelligence

Tổng các trọng số sẽ được tính theo công thức sau:

$$x_1w_1 + x_2w_2 + \dots + x_nw_n + bias$$

Bước 2:

Giá trị tính được đưa vào hàm kích hoạt để đưa ra kết quả cho đầu ra (output):

$$activation\ function(x_1w_1 + x_2w_2 + \dots + x_nw_n) + bias$$

Kí hiệu:

- input: giá trị đầu vào
- output: giá trị đầu ra
- x_1 : giá trị đầu vào
- w_1, w_2, w_n : trọng số đầu vào
- bias: độ lệch

Để tính giá trị đầu ra ta cần hai giá trị là trọng số và bias. Bias là độ lệch, là một giá trị không đổi, được thêm vào tổng các tích giữa các giá trị đầu vào và trọng số, chỉ số bias được sử dụng để chuyển đổi kết quả của chức năng kích hoạt về mặt tích cực hoặc tiêu cực.

Ví dụ người dùng muốn mạng nơ-ron trả về 3 khi đầu vào vào là 0. Chúng ta có tổng trọng số đầu vào bằng 0, vậy câu hỏi đặt ra là làm thế nào để đảm bảo nơ-ron của mạng sẽ trả về giá trị bằng 3. Để giải quyết vấn đề này, chỉ cần đặt chỉ số bias là 3.

Nếu chúng ta thêm giá trị bias thì mạng nơ-ron nhân tạo chỉ thực hiện một công việc đơn giản là thực hiện phép nhân ma trận trên các đầu vào trọng số, việc này sẽ tạo nên tập dữ liệu có thể bị overfitting²⁴.

Optimizer SGD (Stochastic Gradient Descent)

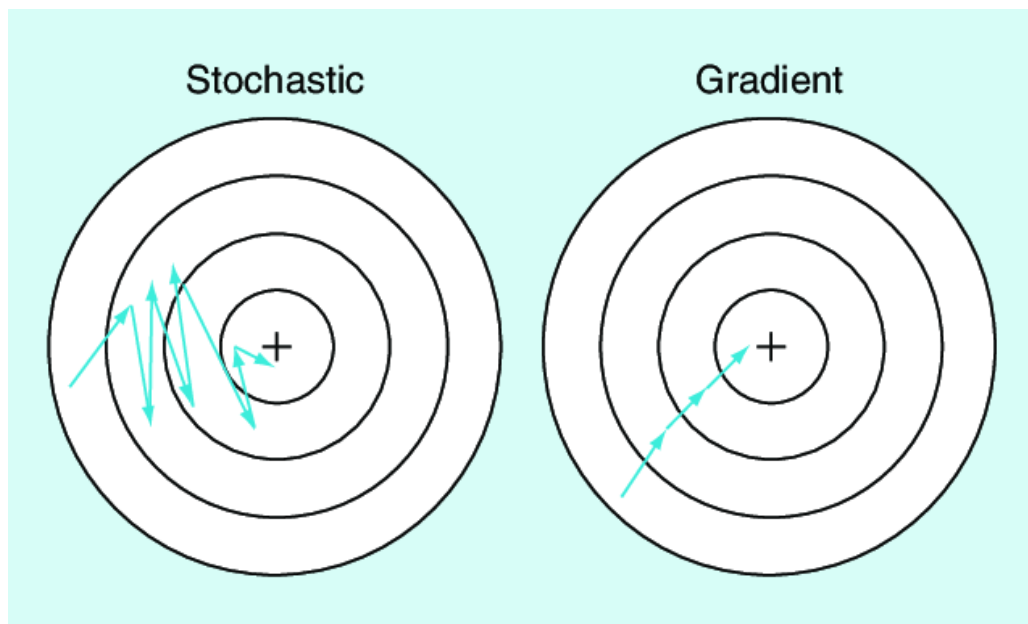
Trước khi nói đến SGD²⁵, việc tìm hiểu về Optimizer cũng rất quan trọng. Đây là thuật toán tối ưu [5]. Về cơ bản, thuật toán tối ưu là cơ sở để xây dựng mô hình Neural Network với mục đích “học” được các “pattern” của dữ liệu đầu vào, từ đó có thể tìm được một cặp trọng số (được gọi là weights) và bias phù hợp để tối ưu hóa mô hình. Phương pháp tìm weights và bias đó là lấy giá trị ngẫu nhiên weights và bias một số lần hữu hạn và hy vọng tại một bước nào đó ta có thể tìm được lời giải cho bài toán. Nếu ta thực hiện theo phương pháp như vậy, là một điều không

²⁴ <https://en.wikipedia.org/wiki/Overfitting>

²⁵ <https://machinelearningcoban.com/2017/01/16/gradientdescent2/>

khả dĩ và lãng phí tài nguyên nghiên cứu cũng như thời gian. Vấn đề đặt ra là cần tìm một giải pháp để cải thiện trọng số weights và bias theo từng bước, đó là lí do các thuật toán optimizer ra đời.

SGD là một biến thể GD (Gradient Descent) và GD còn nhiều hạn chế như phụ thuộc vào giá trị nghiệm khởi tạo ban đầu và tốc độ học của mô hình. Khi tốc độ học quá lớn sẽ khiến cho thuật toán không hội tụ. Thay vì mỗi vòng lặp (được gọi là epoch) chúng ta sẽ cập nhật giá trị trọng số một lần. Đối với GD thì trong mỗi vòng lặp có N điểm dữ liệu, chúng ta sẽ cập nhật trọng số N lần.



Hình 13: So sánh hội tụ giữa SGD và GD

Dựa vào hình ảnh có thể thấy SGD có đường đi đi theo hình zíc-zắc, không mượt như GD. Vậy câu hỏi đặt ra “tại sao phải dùng SGD thay vì dùng GD?”. Bởi vì GD có hạn chế đối với cơ sở dữ liệu lớn, việc tính toán đạo hàm toàn bộ dữ liệu qua mỗi vòng lặp trở nên cồng kềnh. Bên cạnh đó, GD không phù hợp với online learning, khi dữ liệu cập nhật liên tục (ví dụ khi thêm câu hỏi) thì mỗi lần thêm dữ liệu ta phải tính toán lại đạo hàm trên toàn bộ dữ liệu, về mặt này SGD có thể giải quyết dễ dàng vấn đề đó. Ngoài ra, thuật toán SGD có một nhược điểm lớn là về tốc độ học, điểm dữ liệu ban đầu, vì vậy ta phải kết hợp SGD và một số thuật toán khác như Momentum....

Hàm kích hoạt

Hàm kích hoạt là những hàm mô phỏng tỷ lệ truyền xung qua axon của một neuron thần kinh. Trong một mạng nơ-ron nhân tạo, hàm kích hoạt đóng vai trò là thành phần phi tuyến tại output của các nơ-ron. Nếu không có các hàm kích hoạt,

khả năng dự đoán của mạng nơ-ron sẽ bị giới hạn và giảm đi rất nhiều, sự kết hợp của các hàm kích hoạt giữa các tầng ẩn sẽ giúp cho mô hình học được các quan hệ phi tuyến phức tạp tiềm ẩn trong dữ liệu [6].

Một số hàm kích hoạt phổ biến

Hàm sigmoid

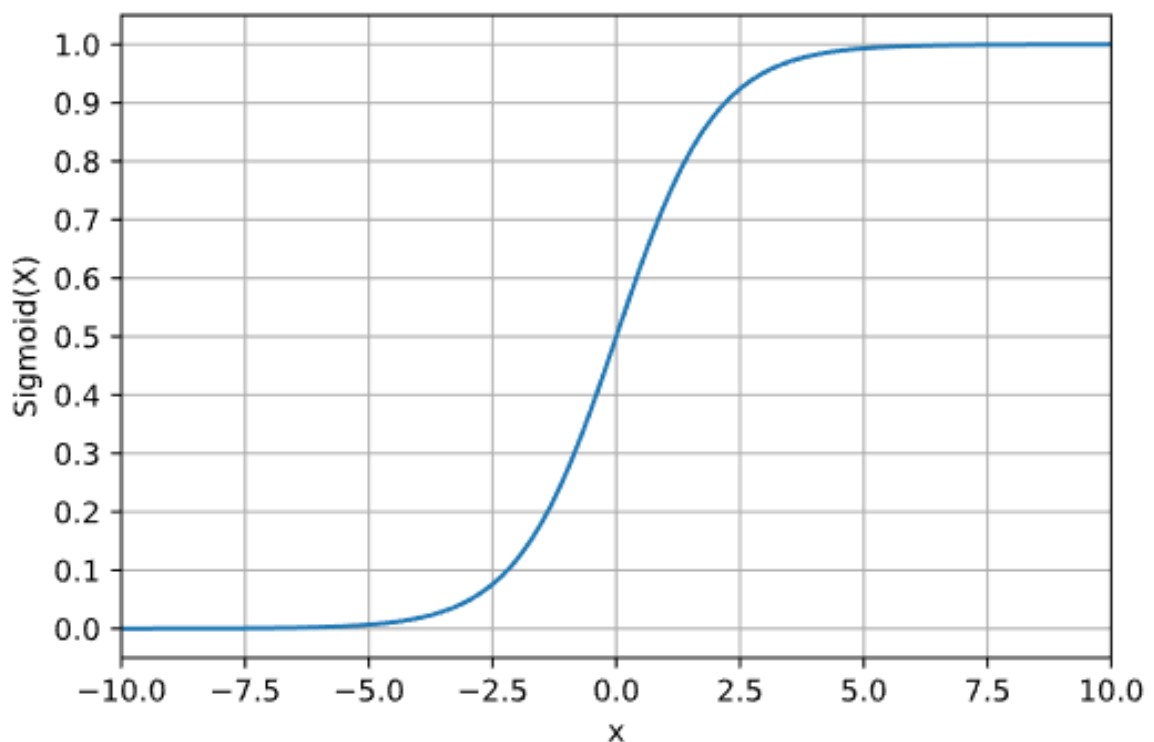
Công thức [6]:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Đạo hàm:

$$f'(x) = \frac{\partial f(x)}{\partial x} = f(x) \times (1 - f(x))$$

Đồ thị:



Hình 14: Đồ thị Hàm Sigmoid

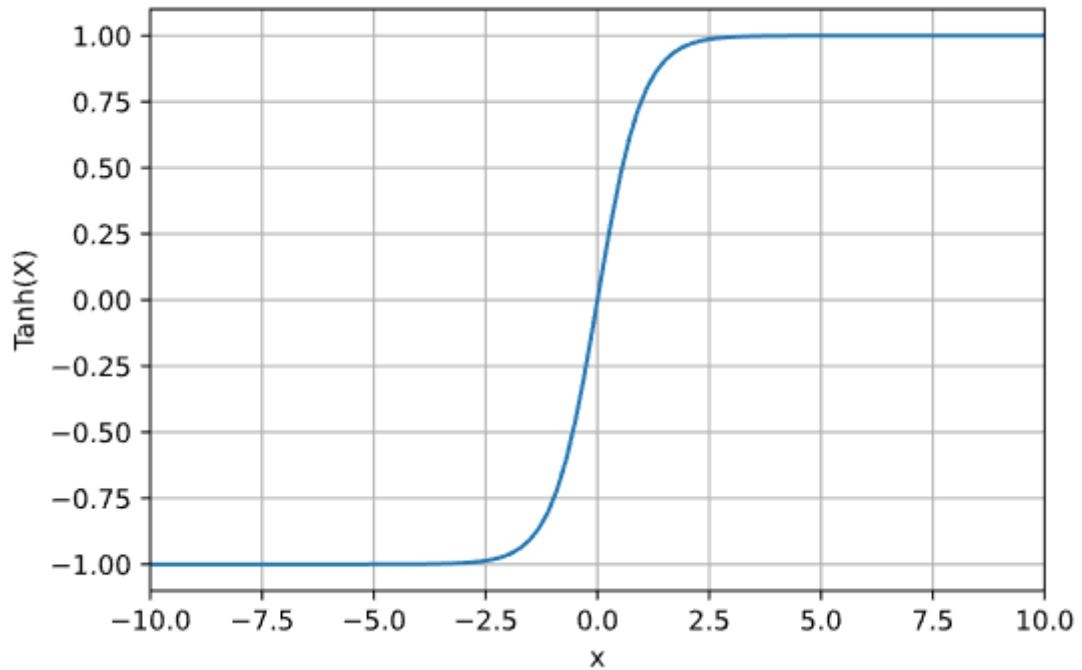
Đầu vào của hàm Sigmoid là một số thực và chuyển thành một giá trị nằm trong khoảng (0; 1). Đầu vào là số thực âm rất nhỏ sẽ cho đầu ra tiệm cận với 0, ngược lại nếu đầu vào là một số thực dương lớn sẽ cho đầu ra tiệm cận với 1.

Hàm Tanh

Công thức [6]:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Đồ thị:



Hình 15: Đồ thị hàm Tanh

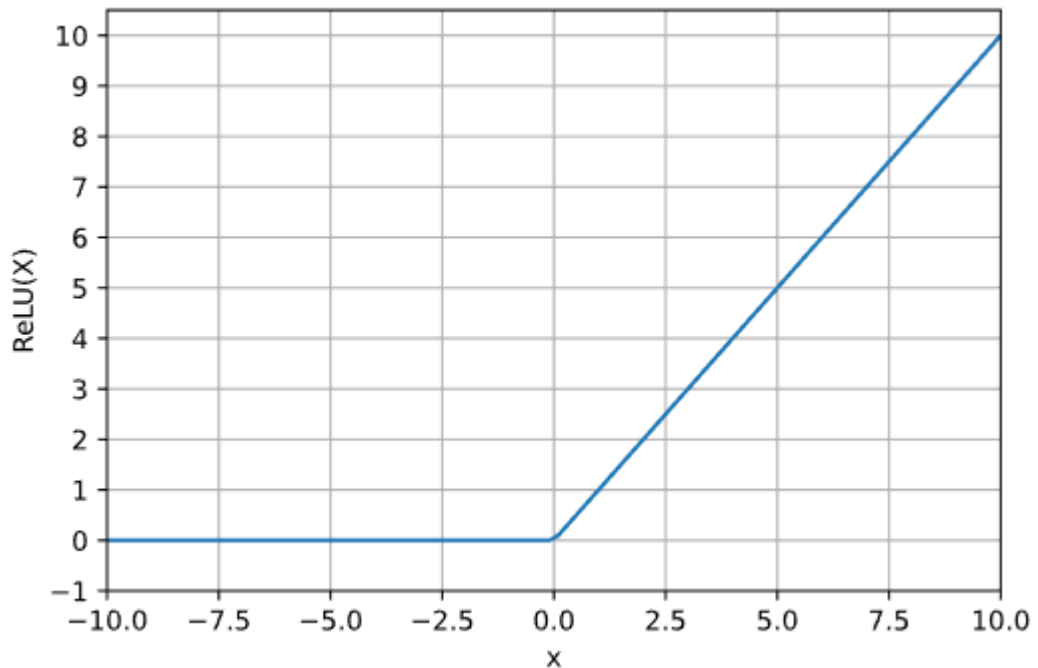
Đầu vào của hàm Tanh là một số thực và chuyển thành một giá trị nằm trong khoảng $(-1; 1)$. Hàm Tanh bị bão hòa của hai đầu (gradient thay đổi rất nhỏ ở hai đầu).

Hàm ReLU (Rectified Linear Unit)

Công thức [6]:

$$f(x) = \max(0, x)$$

Đồ thị:



Hình 16: Đồ thị hàm ReLU

Hàm ReLU đang được sử dụng khá nhiều trong những năm gần đây khi huấn luyện các mạng nơ-ron. ReLU đơn giản là lọc các giá trị nhỏ hơn 0. Một số ưu điểm khá vượt trội của nó so với Sigmoid và Tanh là tốc độ hội tụ nhanh hơn hẳn, điều này có thể do ReLU không bị bão hoà ở 2 đầu như Sigmoid và Tanh. Tính toán của ReLU nhanh hơn do công thức không phức tạp.

Hàm Softmax

Công thức [6]:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

Hàm Softmax xuất ra một véc-tơ có giá trị tổng bằng 1 có thể hiểu là xác suất của các lớp. Softmax thường dùng ở đầu ra cho bài toán phân lớp, giả sử bài toán gồm nhiều lớp khi dùng Softmax kết quả trả về là xác suất của từng lớp.

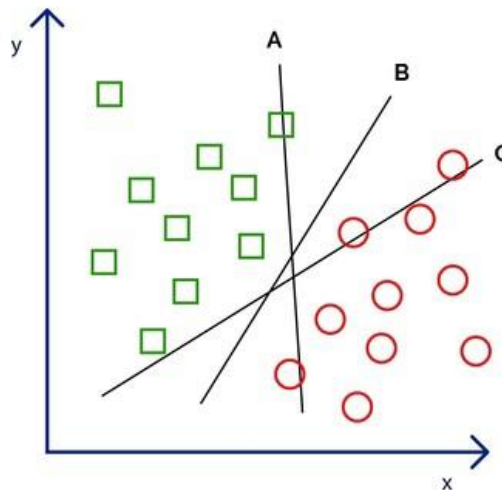
2.1.4. Một số giải thuật đánh giá độ chính xác

Giải thuật Máy học véc-tơ hỗ trợ (Support Véc-tơ Machine)

Máy học véc-tơ hỗ trợ [7] là tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy. SVM dạng chuẩn chấp nhận dữ liệu vào và phân loại dữ liệu và hai lớp khác nhau. Support Véc-tơ được hiểu một

các đơn giản là các đối tượng trên đồ thị tọa độ quan sát, Support Véc-tơ Machine là một biên giới chia cắt các lớp tốt nhất.

Trong SVM điều quan trọng là cần xác định đúng đường Hyperplane. Hyperplane là một đường thẳng có tác dụng phân chia các lớp ra thành các đường riêng biệt.



Hình 17: Các đường phân chia SVM

Theo hình minh họa, tại đây có 3 đường hyperplane (đường A, đường B và đường C). Theo như quy tắc số 1 về chọn lựa đường hyperplane, chọn một hyperplane phân chia hai lớp tốt nhất, trong ví dụ này ta có:

- Đường A: Xuất hiện việc cắt ngang các phân tử tại hai lớp
- Đường C: Xuất hiện việc cắt ngang các phân tử tại hai lớp
- Đường B: Phân chia hai lớp tốt nhất, không cắt qua bất kỳ phân tử nào

SVM là một thuật toán phân lớp hoạt động tốt với những tập dữ liệu có kích thước lớn, thường mang lại kết quả vượt trội so với các giải thuật học có giám sát khác. Một số ưu điểm của thuật toán SVM:

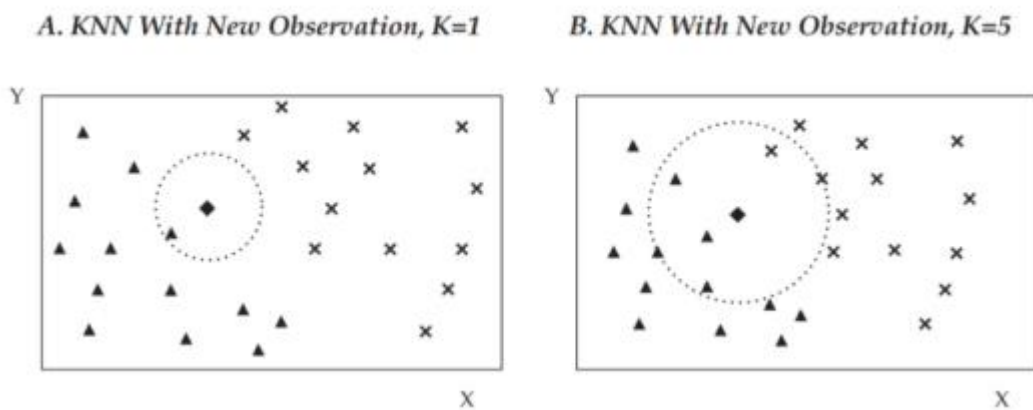
- Ưu điểm 1: Hoạt động hiệu quả với không gian cao chiều (high dimensional spaces): SVM là công cụ tính toán hiệu quả trong không gian cao chiều, trong đó đặc biệt áp dụng cho bài toán phân loại văn bản và phân tích quan điểm nơi chiều có thể cực kỳ lớn.
- Ưu điểm 2: Thuật toán tiêu tốn ít bộ nhớ vì chỉ sử dụng các điểm trong tập hỗ trợ để dự báo trong hàm quyết định.
- Ưu điểm 3: Có thể tạo ra nhiều hàm quyết định từ những hàm kernel khác nhau. Khả năng áp dụng kernel mới cho phép linh động giữa các

phương pháp tuyến tính và phi tuyến tính từ đó cho hiệu suất phân loại tốt hơn.

SVM là một phương pháp hiệu quả cho bài toán phân lớp dữ liệu. Nó là công cụ đặc lực cho các bài toán phân loại văn bản, phân tích quan điểm, SVM cũng từng là mô hình cực kỳ phổ biến trong phân loại ảnh trước khi CNN²⁶ và Deep Learning bùng nổ. Một yếu tố khác làm nên hiệu quả của SVM đó là thuật toán này sử dụng Kernel function khiến cho các phương pháp chuyển không gian trở nên linh hoạt hơn.

Giải thuật K láng giềng (K Nearest Neighbor)

Thuật toán K láng giềng ²⁷ là một kỹ thuật học có giám sát, dùng để phân loại quan sát mới bằng tìm kiếm tương đồng giữa quan sát mới với dữ liệu sẵn có [8].



Hình 18: kNN với k=1 và k=5

Chúng ta cần xác định rõ hình thoi trong cả hai ví dụ đang cần được phân loại và hình dấu nhân hoặc hình tam giác. Hình thoi sẽ được phân loại vào cùng loại với điểm dữ liệu gần nhất:

- Xét $k = 1$: Hình thoi tại thời điểm $k=1$ sẽ được phân loại vào cùng loại với hình tam giác. Vì đây có một điểm dữ liệu hình tam giác nằm gần hình thoi nhất về phía bên trái.
- Xét $k = 5$: Tại thời điểm này, thuật toán kNN sẽ xét 5 điểm dữ liệu gần với hình thoi nhất. Trong đó số lượng hình tam giác gần hình thoi là 3 hình, còn số lượng hình dấu nhân là 2 hình. Dựa theo quy tắc, hình thoi sẽ được phân loại và cùng loại với hình có số điểm dữ liệu lớn hơn trên tổng số 5 điểm dữ liệu gần nhất với hình thoi. Trong trường hợp này, hình thoi được phân loại và hình tam giác.

²⁶ https://en.wikipedia.org/wiki/Convolutional_neural_network

²⁷ https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

KNN là một mô hình đơn giản, trực quan nhưng vẫn có hiệu quả rất cao vì kNN không cần tham số, mô hình không đưa ra bất kỳ giả định nào về việc phân phối dữ liệu, và còn có thể được sử dụng trực tiếp để phân loại đa lớp.

Thuật toán kNN có rất nhiều ứng dụng trong ngành đầu tư, bao gồm: các dự án phá sản, dự đoán đánh giá cổ phiếu, phân bố xếp hạng tín dụng trái phiếu doanh nghiệp.

Khi sử dụng kNN cần phải xác định thước đo để tính khoảng cách giữa các đối tượng cần phân lớp và các đối tượng còn lại trong cơ sở dữ liệu. Vì đây là lựa chọn mang tính chủ quan, nếu chọn một thước đo không phù hợp thì mô hình sẽ không hiệu quả.

Giải thuật Bayes Thơ Ngây (Naïve Bayes)

Naïve Bayes [9] là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê, được ứng dụng rất nhiều trong các lĩnh vực Machine learning dùng để đưa các dự đoán có độ chính xác cao, dựa trên một tập dữ liệu đã được thu thập. Giải thuật thuộc vào nhóm học máy có giám sát.

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là $P(A|B)$, và đọc là “xác suất của A nếu có B”. Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Hình 19: Định lý Bayes

Theo định lý Bayes, $P(A|B)$ sẽ phụ thuộc vào 3 yếu tố:

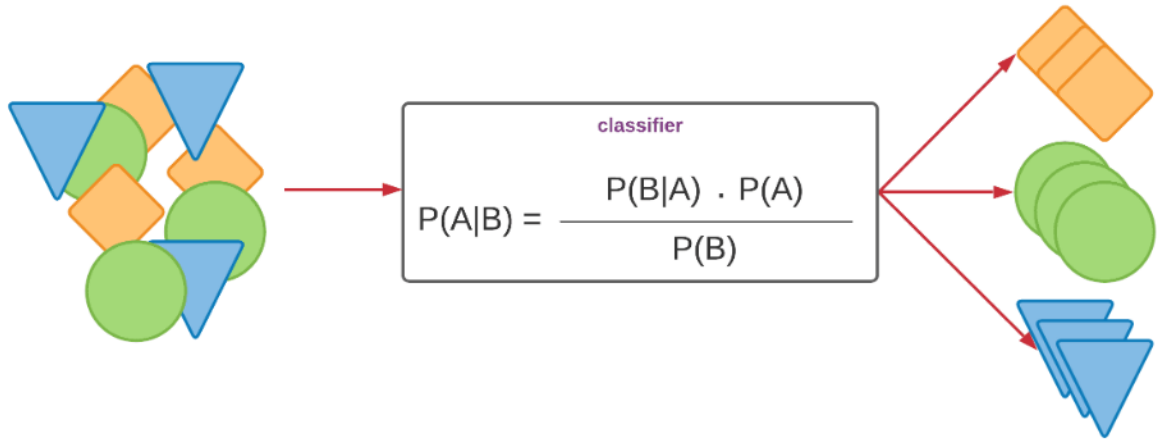
- Xác suất xảy ra A của riêng nó, không quan tâm đến B. Ký hiệu là $P(A)$.
- Xác suất xảy ra B của riêng nó, không quan tâm đến A. Ký hiệu là $P(B)$.
- Xác suất xảy ra B khi biết A xảy ra. Ký hiệu là $P(B|A)$.

Xác suất sự kiện A phụ thuộc vào xác suất sự kiện B, thực tế xác suất A có thể phụ thuộc vào xác suất $B_1, B_2, B_3, \dots B_n$. Nên định luật Bayes có thể được mở rộng bằng công thức:

$$P(A|B) = \frac{(P(B_1|A) \times P(B_2|A) \times P(B_3|A) \times \dots \times P(B_N|A)) \times P(A)}{P(B_1) \times P(B_2) \times P(B_3) \dots \times P(B_n)}$$

Hình 20: Định lý Bayes mở rộng

Thuật toán Naïve Bayes được ứng dụng trong rất nhiều lĩnh vực: hệ thống cảnh báo biến cố, ứng dụng dự đoán nhiều giả thuyết mục tiêu, hệ thống phân loại văn bản hay ngôn ngữ tự nhiên vì tính chính xác của nó hơn các giải thuật khác, hệ thống chống thư rác, phân tích tâm lý thị trường, phân tích thói quen của khách hàng, xây dựng hệ thống gợi ý.



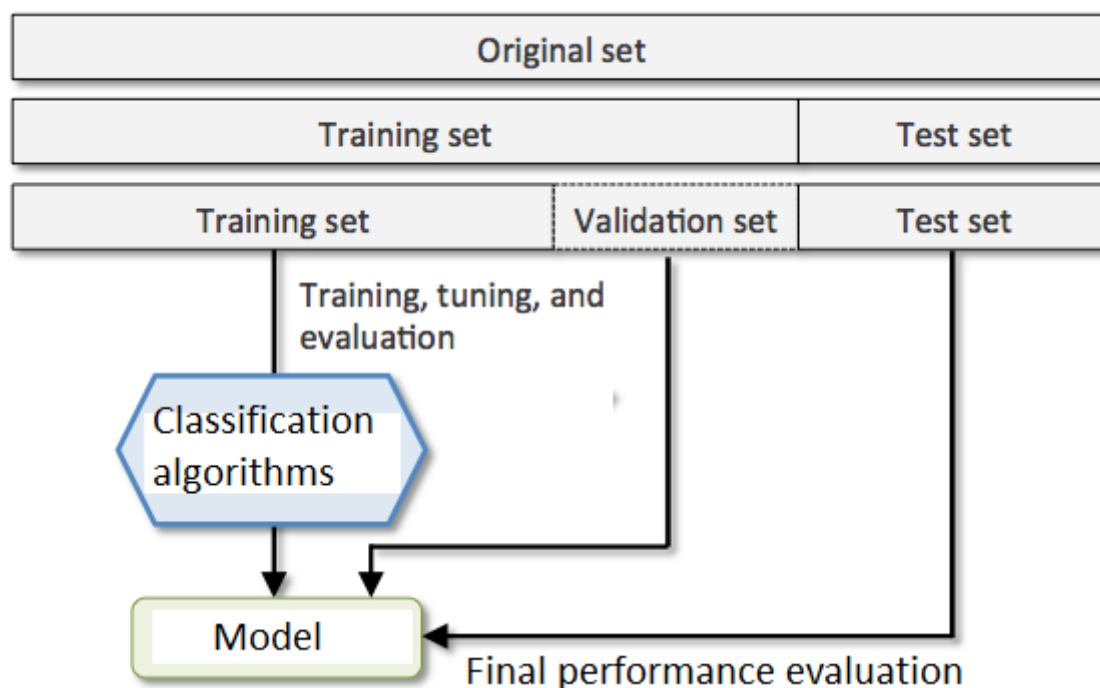
Hình 21: Phân lớp sử dụng Naïve Bayes

2.1.5. Phương pháp đánh giá độ chính xác

Có rất nhiều phương pháp đánh giá độ chính xác của mô hình như [10]: Hold-out, Stratified sampling, Repeated hold-out, Cross validation (K-fold và leave-one-out), Bootstrap sampling. Trong nghiên cứu sử dụng phương pháp đánh giá Hold-out vì đây là phương pháp cài đặt khá đơn giản, phù hợp khi tập dữ liệu sử dụng để huấn luyện có kích thước lớn.

Khi sử dụng phương pháp đánh giá Hold-out. Toàn bộ tập dữ liệu sẽ được chia thành hai tập con là *datatrain* và *datatest*. Tập huấn luyện *datatrain* để huấn luyện hệ thống, tập kiểm thử *datatest* để đánh giá hiệu năng của hệ thống sau khi đã huấn luyện. Dữ liệu *datatrain* không được sử dụng trong quá trình đánh giá hệ thống sau khi huấn luyện và dữ liệu *datatest* không được sử dụng trong quá trình huấn luyện hệ thống.

Ví dụ: Nghiên cứu sử dụng 2/3 cho *datatest* và 1/3 cho *datatrain*.

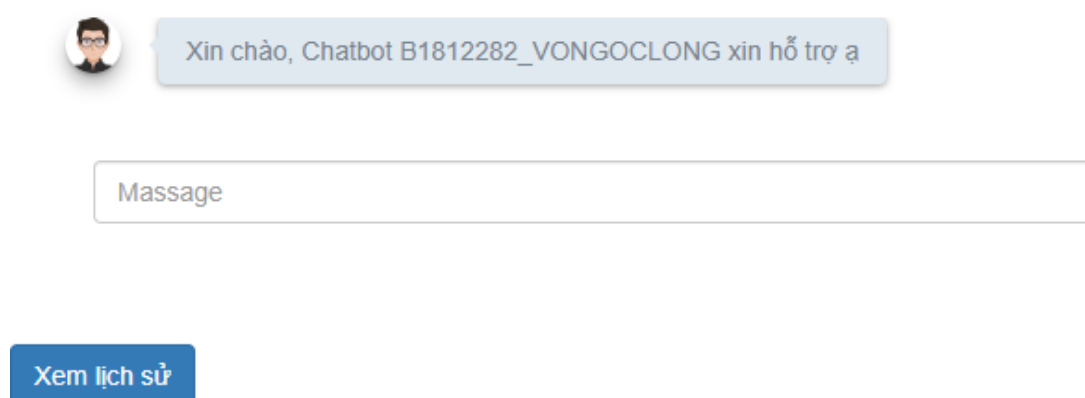


Hình 22: Mô hình hoạt động của phương pháp đánh giá Hold-out

2.2. Mô tả hệ thống trang web

Trang web được thiết kế theo hình thức hỏi – đáp, tương tự với những ứng dụng như: Messenger, Zalo, Whatsapp,... tổng thể của trang web là một ô cửa sổ chat, tập trung vào thể hiện sự trao đổi giữa người dùng và Chatbot. Người dùng có thể đặt câu hỏi về ngành, nhóm ngành Công nghệ Thông tin của Trường Công nghệ Thông tin và Truyền thông, Chatbot sẽ căn cứ vào dữ liệu để trả lời cho câu hỏi của người dùng.

CHATBOT TƯ VẤN TUYỂN SINH - B1812282_VONGOCLONG

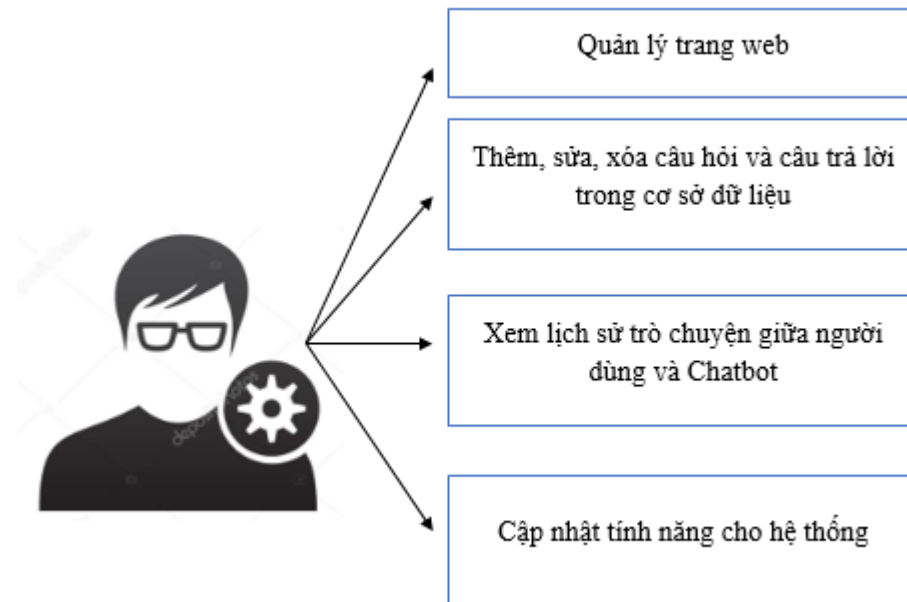


Hình 23: Giao diện ứng dụng website

2.3. Yêu cầu chức năng

2.3.1. Người dùng quản trị

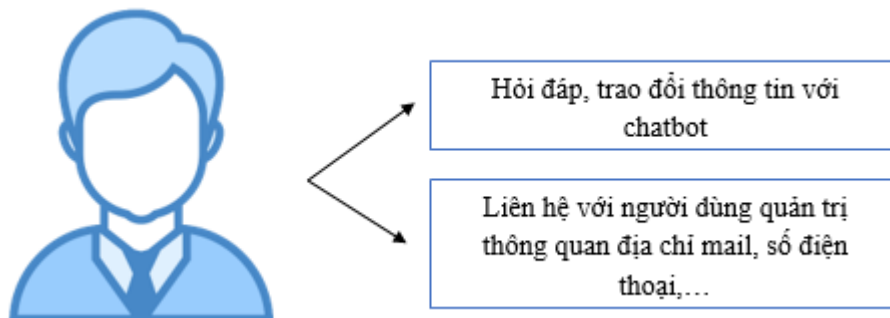
Các chức năng chính của người dùng quản trị:



Hình 24: Các chức năng của người quản trị

2.3.2. Người dùng

Người dùng có những chức năng sau:

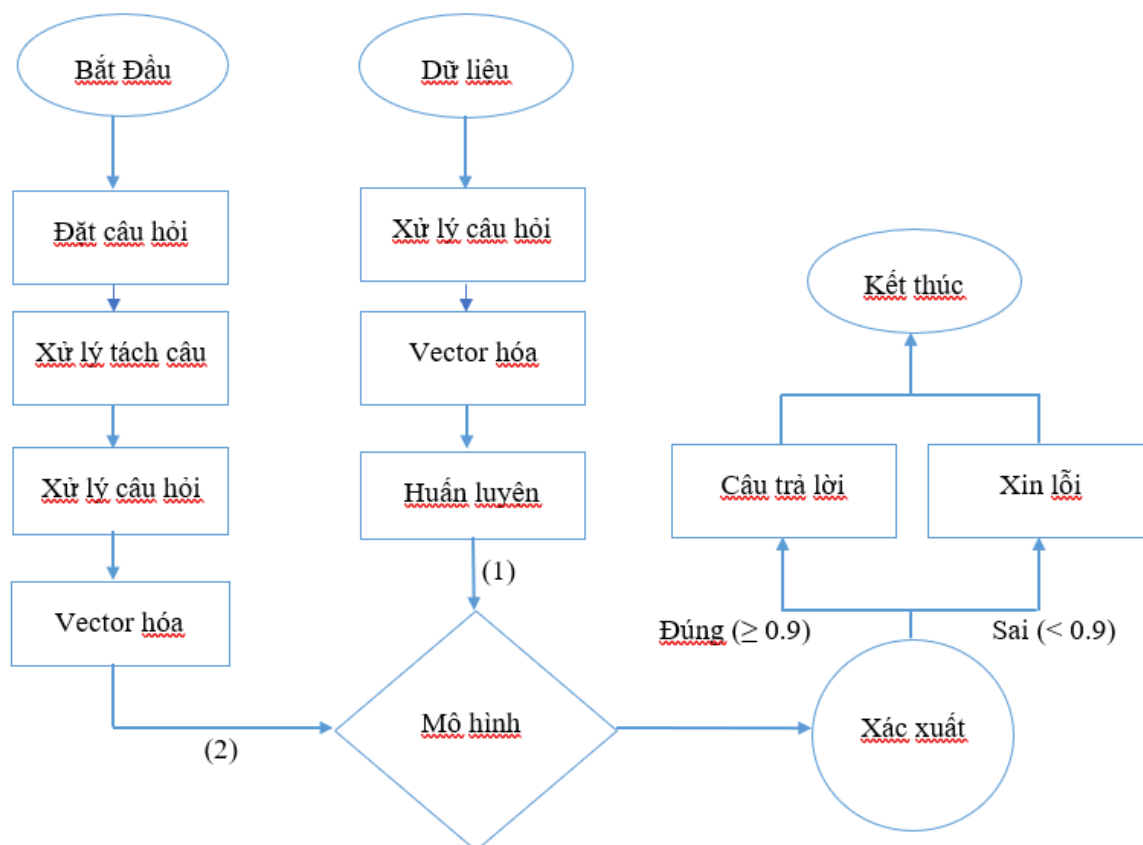


Hình 25: Các chức năng của người 26oke

CHƯƠNG 3. THIẾT KẾ VÀ CÀI ĐẶT GIẢI THUẬT

Chương 3 sẽ giới thiệu về tổng quan về hệ thống chatbot, phương pháp xử lý dữ liệu và huấn luyện mô hình trên ba phương pháp học máy là SVM, K láng giềng, Bayes Thor ngậy, SGD.

3.1. Thiết kế hệ thống



Hình 26: Lưu đồ hoạt động của hệ thống

Quá trình xử lý của hệ thống sẽ chia ra làm 2 giai đoạn:

Giai đoạn (1): Xử lý tập dữ liệu, dùng để tạo mô hình dự đoán. Bao gồm các quá trình:

Xử lý câu hỏi → Véc-tơ hóa → Huấn luyện

Tập dữ liệu sau khi được thu thập, sẽ trải qua quá trình xử lý câu hỏi, sau đó sẽ véc-tơ hóa dữ liệu đã được xử lý, tạo ra mảng véc-tơ. Sau đó đưa mảng véc-tơ vừa được tạo ra vào mô hình mạng nơ-ron 2 lớp để huấn luyện tạo ra Mô hình.

Giai đoạn (2): Xử lý tập câu hỏi người dùng đưa ra để dự đoán. Bao gồm các quá trình:

Tách câu hỏi → xử lý câu hỏi → véc-tơ hóa → Mô hình dự báo

Quá trình bắt đầu từ việc nhận được câu hỏi của người sử dụng, câu hỏi sẽ được tách thành các câu hỏi nhỏ, và được xử lý câu hỏi, sau đó sẽ véc-tơ hóa câu hỏi thành mảng véc-tơ. Mảng véc-tơ sau khi xử lý sẽ đưa vào mô hình dự đoán. Kết quả sẽ là xác suất dự đoán. Tại đây, hệ thống sẽ chia 28okeni trường hợp:

Trường hợp 1: Kết quả dự báo có xác suất lớn hơn hoặc bằng 0.9 ($\geq 90\%$): Kết quả trả về sẽ là câu trả lời cho câu hỏi của người dùng.

Trường hợp 2: Kết quả dự báo có xác suất bé hơn 0.9 ($< 90\%$): Kết quả trả về sẽ là câu “Xin lỗi, hiện tại câu hỏi này bot chưa thể giải đáp, bot sẽ ghi nhận câu hỏi và cải thiện chất lượng”.

3.1.1. Tập dữ liệu

Để thực hiện huấn luyện mô hình chatbot, tôi thực hiện thu thập dữ liệu dùng để huấn luyện mô hình từ trang Fanpage ‘Tư vấn tuyển sinh chính quy²⁸’ của trường Đại Học Cần Thơ, và dữ liệu cũng được thu thập thông qua việc khảo sát trên không gian mạng thông qua sự hỗ trợ của Google Form. Tại các buổi tuyển sinh trực tiếp, trực tuyến của trường Đại Học Cần Thơ để tư vấn cho các người sử dụng học sinh, các bậc phụ huynh về thông tin tuyển sinh của trường. Quý phụ huynh và các người sử dụng học sinh đã có những câu hỏi, thắc mắc của mình về những ngành học và chương trình đào tạo của Trường Đại học Cần Thơ nói chung và trường Công nghệ Thông tin và Truyền thông nói riêng.

Sau khi thu thập dữ liệu về các câu hỏi của các người sử dụng học sinh và các bậc phụ huynh, những câu hỏi sẽ được thực hiện phân loại. Thủ thuật gom nhóm những câu hỏi có cùng câu trả lời sẽ tạo thành một nhóm và những câu hỏi thu thập vẫn chưa được xử lý.

Tổng quan về tập dữ liệu như sau:

- Số lượng câu hỏi được thu thập để tạo thành tập dữ liệu: 1043 câu hỏi.
- Số lượng nhãn hiện tại bao gồm: 25 nhãn.
- Tập dữ liệu được lưu trữ với cấu trúc là một file dạng .json

Trước khi mô tả rõ hơn về tập dữ liệu, cần tìm hiểu trước về JSON [11] và cấu trúc của một file JSON.

JSON là viết tắt của JavaScript Object Notation, là một kiểu định dạng dữ liệu tuân theo một quy luật nhất định mà hầu hết các ngôn ngữ lập trình hiện nay đều có thể đọc được .json là một tiêu chuẩn mở để trao đổi dữ liệu trên web. Cấu trúc của một file .json là theo dạng key – value để dữ liệu sử dụng. Hỗ trợ các cấu trúc dữ

²⁸ <https://tuyensinh.ctu.edu.vn/>

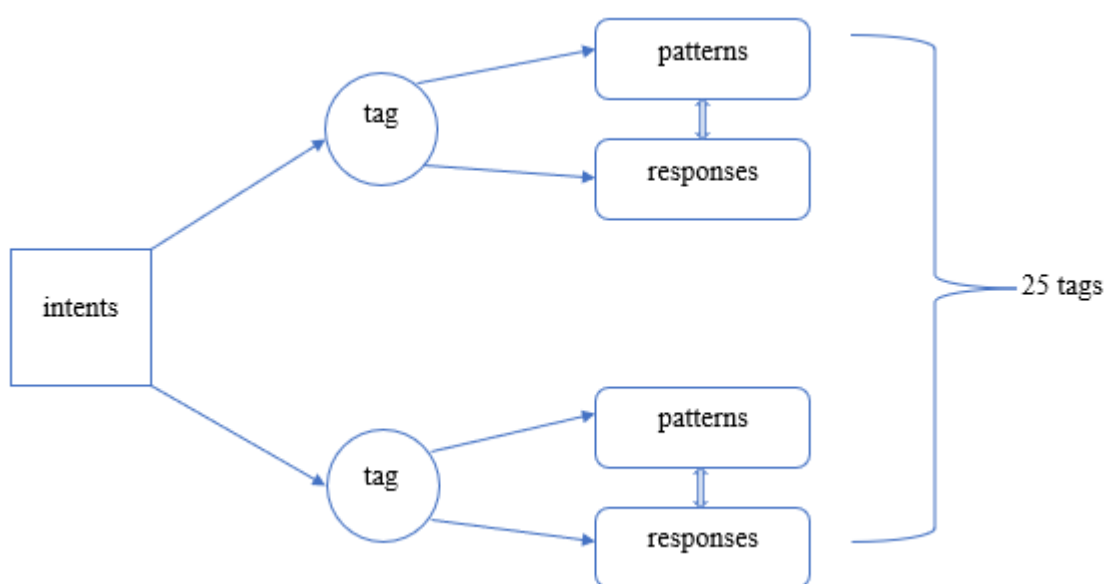
liệu như đối tượng và mảng. Ví dụ một tập tin info.json với nội dung dưới đây sử dụng format kiểu JSON để lưu trữ thông tin:

```
{
  "name" : "B1812282",
  "title" : "Luận Văn Tốt Nghiệp",
  "description" : "Được xây dựng bằng ngôn ngữ lập trình Python và các giải thuật máy học"
}
```

Hình 27: Nội dung tập tin info.json

Tập dữ liệu dùng trong nghiên cứu có tên là data.json, có 1043 câu hỏi được chia vào 25 nhãn. Cấu trúc của tập dữ liệu được mô tả như sau:

- intents: dùng để gọi tập dữ liệu trong quá trình lập trình
- tag: đóng vai trò là nhãn
- patterns: dùng để lưu trữ tất cả câu hỏi thuộc về nhãn hiện hành
- responses: dùng để lưu trữ tất cả câu trả lời ứng với những câu hỏi nằm trong nhãn hiện hành.



Hình 28: Cấu trúc tập tin data.json

Với 25 nhãn trong tập dữ liệu bao gồm các nhãn như: diemchuan, phuongthuc, chitieu, tilevieclam, tiletotnghiep,... Tùy theo nhu cầu của người sử dụng mà số lượng câu hỏi nằm trong trường “patterns” của các nhãn có sự phân hóa rõ ràng. Ví dụ tại nhãn “diemchuan” có số lượng câu hỏi là 176 câu hỏi, nhãn “phuongthuc” có số lượng câu hỏi là 174 câu hỏi, nhưng nhãn “kytucxa” có số lượng câu hỏi khá hạn chế là 11 câu hỏi. Điều này phản ánh rõ mức độ quan tâm của người sử dụng vào những yếu tố như: điểm chuẩn, phương thức xét tuyển, chỉ tiêu,... trong quá trình tuyển sinh.

Đối với mức độ dày đặc của câu hỏi thì phần thiết kế câu trả lời nằm trong trường “patterns” cũng sẽ dựa trên số lượng câu hỏi hiện hành. Ví dụ với nhãn “diemchuan” có 176 câu hỏi – 9 câu trả lời, nhãn “phuongthuc” có 174 câu hỏi – 9 câu trả lời, nhãn “kytucxa” có 11 câu hỏi – 2 câu trả lời.

Trong từng nhãn – trường “tag” bên trong tập dữ liệu, số lượng câu trả lời có nội dung tương tự nhau sẽ được gói vào một nhóm và đưa vào trường “patterns”, số lượng câu hỏi không cố định và không có giới hạn cụ thể, càng nhiều câu hỏi thì dữ liệu dùng cho huấn luyện sẽ càng phong phú hơn. Những câu trả lời tương ứng với các câu hỏi bên trong trường “patterns” sẽ được người quản trị trả lời dựa trên những thông tin thu thập được trong quá trình thu thập dữ liệu, tất cả câu trả lời sẽ được chọn lọc và xử lý những lỗi chính tả, sau đó sắp xếp câu trả lời theo 1 thứ tự nhất định ứng với thứ tự các nhóm câu hỏi hướng tới cùng một câu trả lời để dễ dàng rà soát và xử lý các lỗi phát sinh trong quá trình thử nghiệm, thực hiện đưa tất cả câu hỏi và trường “responses” để thực hiện huấn luyện mô hình.

Minh họa cho tập dữ liệu với nhãn “diemchuan”

```
"intents": [
  {
    "tag": "diemchuan",
    "patterns": [
      "điểm chuẩn là bao nhiêu",
      "điểm chuẩn của từng ngành công nghệ thông tin",
      "điểm chuẩn của năm vừa qua",
      "ngành này tầm bao nhiêu điểm",
      "bao nhiêu điểm để đậu vào ngành này",
      "Cho em hỏi điểm xét học bạ ngành CNTT là bao nhiêu ?",
      "Điểm xét học bạ của CNTT đi ạ ?",
      "Cho em hỏi điểm xét học bạ ngành công nghệ thông tin ạ",
      "Ngành CNTT xét học bạ và thi tốt nghiệp lấy bao nhiêu điểm vậy thầy cô ơi",
      "bao nhiêu điểm để đậu vào ngành này ở trường",
      "điểm trung bình bao nhiêu là đậu",
      "điểm đầu vào năm trước là bao nhiêu",
      "Năm nay điểm học bạ có tăng không ạ ?",
      "Ngành CNTT xét học bạ và thi tốt nghiệp lấy bao nhiêu điểm vậy thầy cô ơi",
      "khoảng bao nhiêu điểm thì được đậu vậy ạ",
      "cntt năm nay xét học bạ tầm bao nhiêu điểm vậy ạ ?",
      "điểm xét học bạ sẽ ra sao ạ",
      "em xin tham khảo điểm xét tuyển CNTT những năm trước ạ",
      "Công nghệ thông tin xét học bạ bao nhiêu điểm thì an toàn v",
      "Xét học bạ công nghệ thông tin thì cần bao nhiêu điểm",
      "em nhớ điểm chuẩn CNTT thi 24.5 đúng k ạ",
      "Điểm chuẩn học bạ năm nay có cao hơn năm ngoái không ạ",
      "ngành cntt bao nhiêu điểm vậy",
    ],
    "responses": [
      "Ngành công nghệ thông tin điểm thi trúng tuyển năm 2021 là 25.75, điểm trúng tuyển học bạ 2021 là 28.50, (",
      "Ngành An toàn thông tin và Ngành Truyền thông đa phương tiện là một ngành mới của trường ĐHCT, nên hiện c",
      "Ngành Khoa học máy tính điểm thi trúng tuyển năm 2021 là 25.00, điểm trúng tuyển học bạ là 27.00",
      "Ngành kĩ thuật máy tính điểm thi trúng tuyển năm 2021 là 23.75 và điểm trúng tuyển học bạ là 24.50",
      "Ngành Mạng máy tính và truyền thông dữ liệu năm 2021 là 24.00, điểm trúng tuyển học bạ năm 2021 là 25.25",
      "Ngành kĩ thuật phần mềm điểm thi trúng tuyển năm 2021 là 25.25 và điểm trúng tuyển học bạ là 27.50",
      "Ngành hệ thống thông tin điểm thi trúng tuyển năm 2021 là 24.25 và điểm trúng tuyển học bạ là 25.75.00",
      "Ngành công nghệ thông tin điểm thi trúng tuyển năm 2021 là 25.75, điểm trúng tuyển học bạ 2021 là 28.50, ("
    ]
  }
]
```

Hình 29: Cấu trúc tập dữ liệu đối với nhãn diemchuan

Tên trường	Chức năng
“tag”	Nhãn của một tập hợp câu hỏi
“patterns”	Chứa tập hợp câu hỏi có chung nhãn
“responses”	Chứa câu trả lời cho một nhãn

Bảng 4: Chức năng các trường trong tập dữ liệu huấn luyện

Trong một nhãn gồm nhiều câu hỏi, mỗi một câu hỏi lại có một câu trả lời riêng biệt, ví dụ khi người sử dụng hỏi về “điểm chuẩn ngành trường học máy tính” và “điểm chuẩn ngành kỹ thuật phần mềm”, hai câu hỏi này hiện đang xếp cùng nhãn đó là nhãn “diemchuan” nhưng người sử dụng lại cần hai câu trả lời khác nhau nên trong trường “responses” được dùng để lưu trữ câu trả lời của các nhóm ngành khác nhau và được sắp xếp theo một thứ tự cố định là:

STT	Tên ngành
0	Công nghệ thông tin
0	Công nghệ thông tin chất lượng cao
1	An toàn thông tin
2	Truyền thông đa phương tiện
3	Trường học máy tính
4	Kỹ thuật máy tính
5	Mạng máy tính và Truyền thông dữ liệu
6	Kỹ thuật phần mềm
7	Hệ thống thông tin

Bảng 5: Số thứ tự các ngành

Thứ tự câu trả lời của các ngành được sắp xếp tương tự theo danh sách các ngành về “Máy tính và công nghệ thông tin” trong trang web tuyển sinh Đại học Cần Thơ, có một sự thay đổi về vị trí của ba ngành: Công nghệ thông tin, Công nghệ thông tin chất lượng cao và Công nghệ thông tin học tại khu Hòa An, ba ngành này sẽ mang số thứ tự là 0, và có cùng 1 câu trả lời chung về ngành Công nghệ Thông tin.

7480202	An toàn thông tin <small>mới</small>	40	A00, A01		
7320104	Truyền thông đa phương tiện <small>mới</small>	100	A00, A01, D01		
7480101	Khoa học máy tính	60	A00, A01	27,00	25,00
7480106	Kỹ thuật máy tính	60	A00, A01	24,50	23,75
7480102	Mạng máy tính và truyền thông dữ liệu	60	A00, A01	25,25	24,00
7480103	Kỹ thuật phần mềm	60	A00, A01	27,50	25,25
7480104	Hệ thống thông tin	60	A00, A01	25,75	24,25
7480201	Công nghệ thông tin	60	A00, A01	28,50	25,75
7480201H	Công nghệ thông tin - học tại khu Hòa An	40	A00, A01	24,25	23,50

Hình 30: Danh sách các ngành Máy tính và Công nghệ thông tin
(chương trình đại trà)

Ví dụ về nhãn “diemchuan” chúng ta có các câu hỏi liên quan đến điểm chuẩn như “điểm chuẩn là bao nhiêu”, “điểm chuẩn ngành trường học máy tính là bao nhiêu”. Với những câu hỏi liên quan đến điểm chuẩn ta sẽ thiết kế câu trả lời liên quan đến điểm chuẩn là: “Ngành Trường học máy tính điểm trúng tuyển 2021 là 25.00, điểm thi trúng tuyển học bạ năm 2021 là 27.00”, thực hiện tương với các câu hỏi và câu trả lời đối với các ngành còn lại trong nhãn “diemchuan”.

Các nhãn khác như “phuongthuc”, “chitieu”, “vitrivieclam”,... cũng thực hiện sắp xếp và thiết kế các câu hỏi trong trường “patterns” và câu trả lời trong trường “responses” tương tự như nhãn “diemchuan”.

3.1.2. Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu được thực hiện với tập câu hỏi sau khi thu thập sẽ được chuyển hóa tất cả thành chữ thường, xử lý thành những câu hỏi không dấu, các từ viết tắt và những chữ đặc biệt sẽ được xử lý về đúng nghĩa của nó, thực hiện lược bỏ những từ ngữ không cần thiết trong quá trình huấn luyện mô hình.

Loại bỏ các từ kéo dài

Các câu hỏi có thể xen lẫn giữa chữ hoa và chữ thường nên cần đưa hết về dạng chữ thường. Vì câu hỏi có thể có những ký tự cố tình viết kéo thì xử lý loại bỏ: Ví dụ như “được khôngggg” ? “cảm ơnnnnn”...

```
#Loại bỏ các ký tự kéo dài VD: onnnnnnnnnn  
s = re.sub(r'([A-Z])\1+', lambda m: m.group(1).upper(), s, flags=re.IGNORECASE)
```

Hình 31: Loại bỏ các ký tự kéo dài

Xử lý dấu câu

Dữ liệu câu hỏi sau khi thu thập và phân loại sẽ được loại bỏ dấu câu trong Tiếng Việt như [“.”, “?”, “!”, “:”, “...”], các thanh âm của câu hỏi cũng cần được loại bỏ là: [“huyền”, “sắc”, “hỏi”, “ngã”, “nặng”]. Việc bỏ dấu câu giúp tiết kiệm thời gian cũng như không gian lưu trữ, và lược bỏ thanh âm là cần thiết vì khi ta không lược bỏ thanh âm dẫn đến từ có dấu và từ không dấu sẽ khác nhau ví dụ như “điểm” và “diem” sẽ khác nhau trong quá trình huấn luyện làm ảnh hưởng rất lớn đến độ chính xác của bài toán.

Loại bỏ từ vô nghĩa và từ viết tắt

Ngoài việc loại bỏ các thanh âm, các từ viết tắt đặc biệt như “cntt, khmt, attt, ktpm, http” được chuyển về đúng ngữ pháp và loại bỏ các từ không quan trọng trong quá trình huấn luyện như “ạ, vậy ạ, sao, em, ời”.

Danh mục các từ viết tắt cần xử lý.

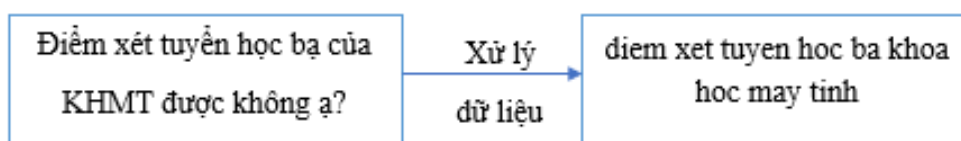
STT	Từ viết tắt	Từ chuyển đổi
1	cntt	cong nghe thong tin
2	attt	an toan thong tin
3	ttdpt	truyen thong da phuong tien
4	khmt	khoa hoc may tinh
5	ktmt	ky thuat may tinh
6	mmt	mang may tinh

7	ktpm	ky thuatphan mem
8	http	he thong thong tin
9	clc	chat luong cao
10	ttdl	truyen thong du lieu
11	bn	bao nhieu
12	ko	khong
13	ntn	nhu the nao
14	ktx	ky tuc xa
15	nhieu	nhieu
16	hc	hoc
17	dc	duoc
18	nganh	nganh

Bảng 6: Danh mục các từ viết tắt

Sau khi tiến xử lý dữ liệu kết quả sẽ đưa về một bộ câu hỏi đáp ứng các tiêu chí đề ra ban đầu:

- Chuyển hóa tất cả thành chữ in thường
- Câu hỏi không mang dấu câu
- Xử lý các từ viết tắt, những ký tự đặc biệt
- Loại bỏ các từ ngữ không cần thiết



Hình 32: Chuyển đổi dữ liệu

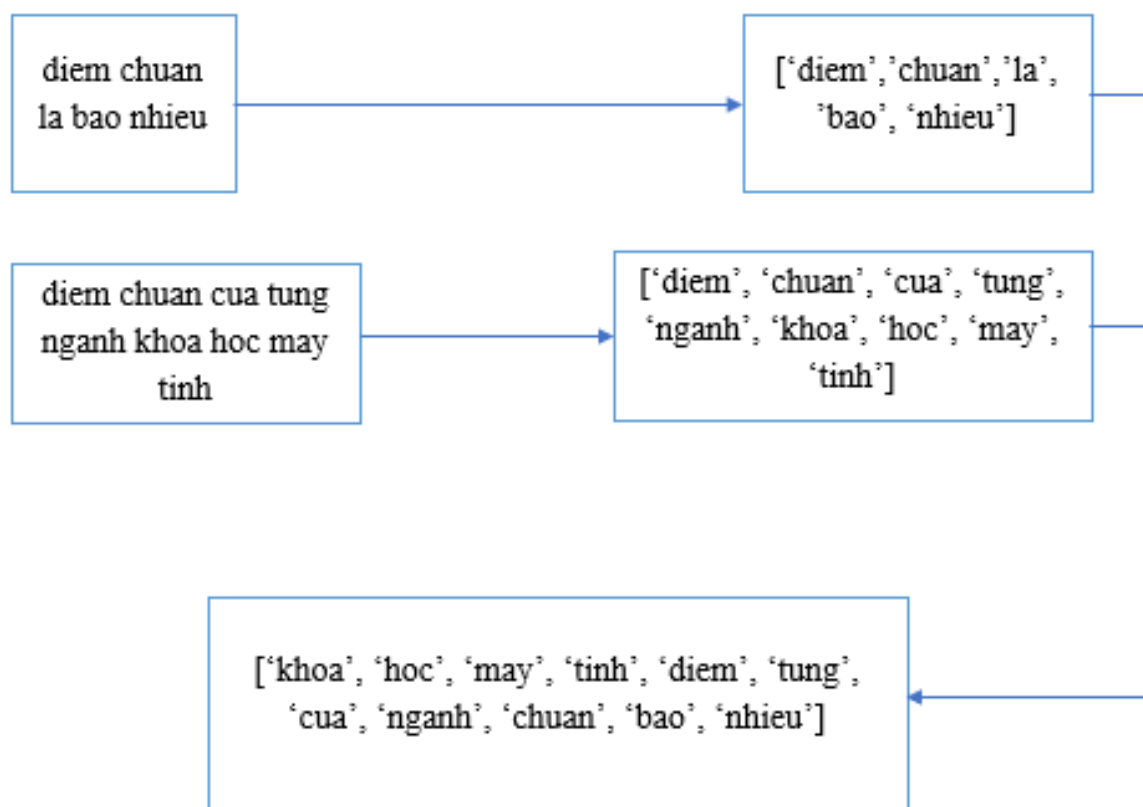
3.1.3. Xây dựng dữ liệu huấn luyện

Sau khi thực hiện tiến xử lý dữ liệu, tất cả các câu hỏi trong tập dữ liệu sẽ được tách thành các từ thêm vào biến “words”. Tất cả các từ sẽ được sắp xếp và đảm bảo các từ trong đó là tồn tại duy nhất để tạo thành một véc-tơ tập hợp tất cả các từ trong câu hỏi.

Xử lý tách câu thành các mảng từ

Quá trình xử lý tách câu thành các mảng từ sử dụng hàm của thư viện NLTK là `word_tokenize`²⁹ có dùng để phân tách các câu thành những mảng từ.

Mô hình xử lý tách những câu hỏi thành mảng các từ:



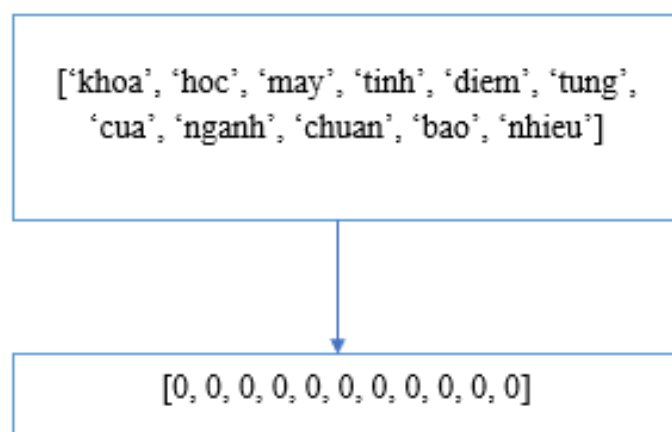
Hình 33: Xử lý câu hỏi tạo tập hợp từ

Tạo véc-tơ

Sau khi tạo được tập hợp những mảng bao gồm tất cả các từ, tiếp tục xử lý khởi tạo một mảng với độ dài bằng với độ dài mảng tập hợp các từ trong tập dữ liệu với giá trị trong mảng khởi tạo là 0.

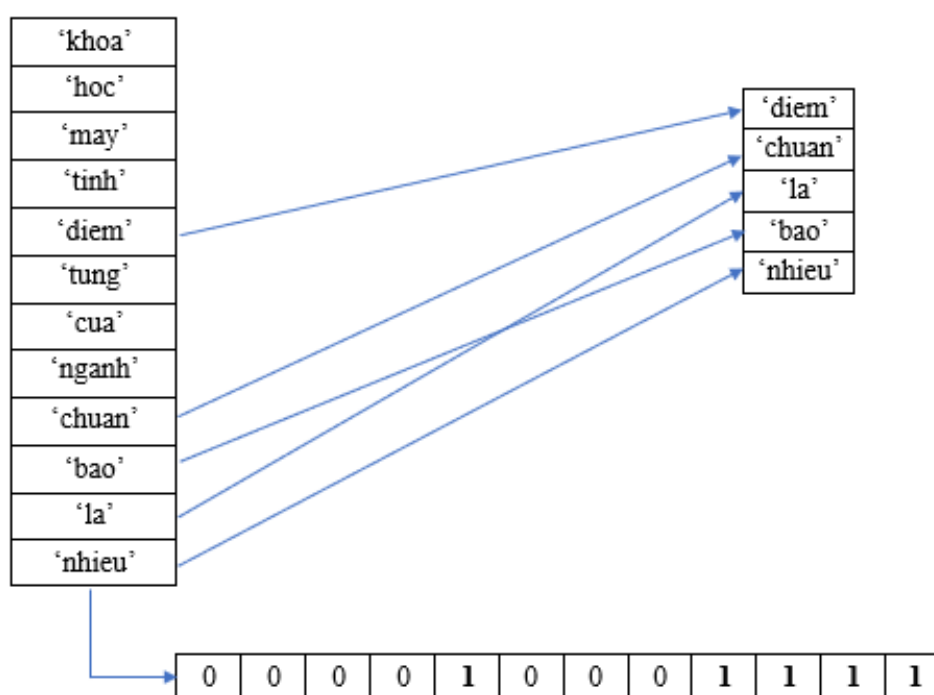
Khởi tạo mảng:

²⁹ <https://product.vinbigdata.org/cac-ky-thuat-tach-tu-trong-xu-ly-ngon-ngu-tu-nhien/>



Hình 34: Tạo véc-tơ ứng với độ dài tập hợp mảng từ

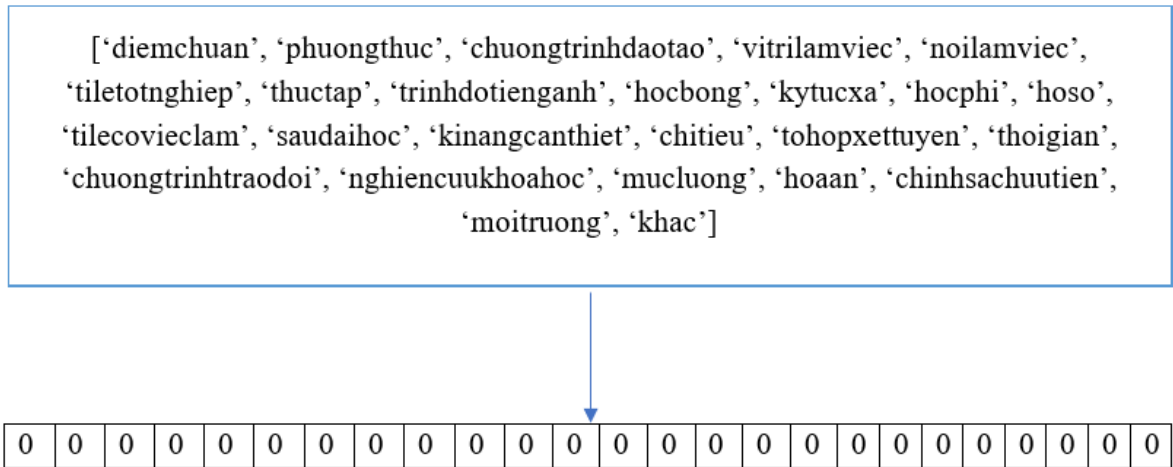
Sau khi tạo thành véc-tơ là mảng tất cả các từ trong tập dữ liệu, xét từng câu hỏi trong câu trong tập dữ liệu, so sánh với các từ trong véc-tơ với các từ trong câu hỏi nếu từ đó có trong câu hỏi sẽ được gán giá trị là 1, ngược lại nếu không tồn tại sẽ được gán giá trị là 0.



Hình 35: Chuyển đổi câu hỏi thành véc-tơ

Mỗi câu hỏi sẽ được chuyển đổi để tạo thành một véc-tơ, thực hiện quá trình chuyển đổi toàn bộ câu hỏi trong tập dữ liệu thành véc-tơ để thực hiện huấn luyện mô hình. Cách tạo ra một véc-tơ là thực hiện việc so sánh các từ trong câu hỏi nằm bên trong tập dữ liệu sẽ được gán nhãn. Nhãn của tập dữ liệu huấn luyện là tập hợp

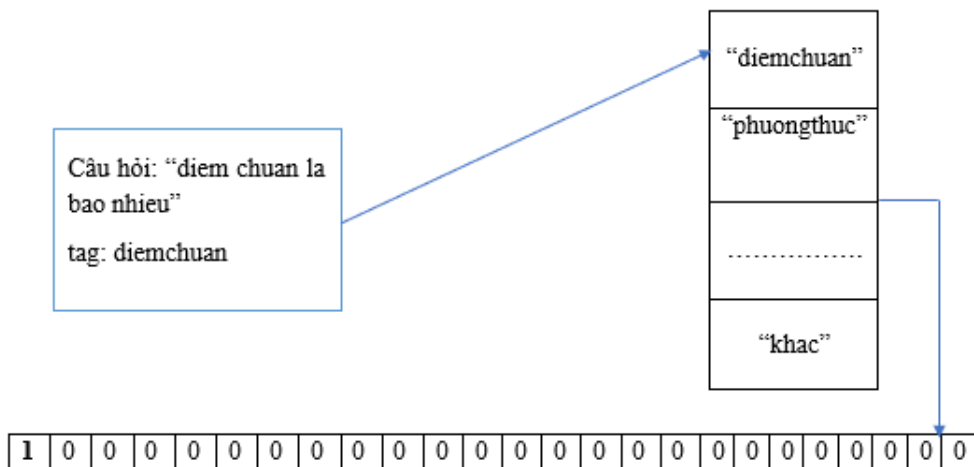
các trường “tag” của tập dữ liệu. Thực hiện khởi tạo một mảng có độ dài bằng với số lượng nhãn trong tập dữ liệu có độ lớn bằng 0.



Hình 36: Tạo véc-tơ ứng với số lượng nhãn

Nếu câu hỏi có nhãn tương ứng với mảng trên vị trí nhãn sẽ là 1, ví dụ câu hỏi “điểm chuẩn là bao nhiêu” có “tag” là “diemchuan” nhãn của câu hỏi đó sẽ là:

[1,0]



Hình 37: Chuyển đổi Véc-tơ nhãn ứng với câu hỏi

Ví dụ: câu hỏi “Điểm chuẩn Khoa học máy tính” qua quá trình xử lý sẽ tạo ra hai véc-tơ [1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, ..., 0] và véc-tơ nhãn sẽ là [1, 0, 0, 0, 0, ..., 0].

3.1.4. Xây dựng mô hình huấn luyện

Sau khi hoàn thành việc xây dựng véc-tơ câu hỏi và mảng các nhãn của câu hỏi tương ứng, và đưa tất cả qua mô hình huấn luyện.

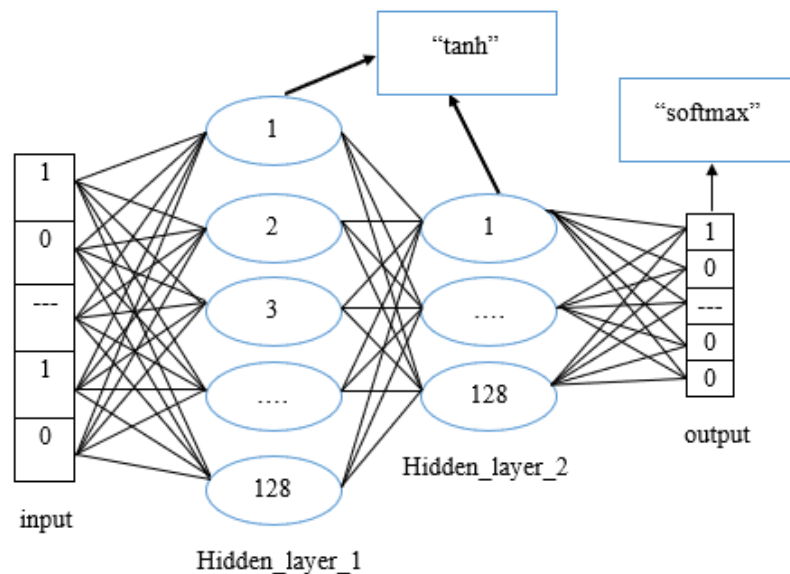
Xây dựng mô hình chatbot

Khi làm việc với dữ liệu văn bản, chúng ta cần thực hiện tiền xử lý trên dữ liệu trước khi chúng ta tạo ra một mô hình Machine-Learning hoặc Deep-Learning, Tokenizing là bước cơ bản nhất có thể làm trên dạng dữ liệu văn bản, đây là quá trình chia toàn bộ văn bản thành các phần nhỏ như các từ. Sau khi thực hiện chia nhỏ các từ, chúng ta sẽ bổ sung từng từ và xóa các từ trùng lặp khỏi danh sách.

Tiếp theo, chúng ta cần tạo một tập training bao gồm các input và output:

- Input: là các mẫu (câu hỏi)
- Output: là các câu trả lời tương ứng

Xây dựng “optimizer³⁰” cho mô hình huấn luyện là SGD với các chỉ số bao gồm tốc độ học là 0.01, với chỉ số “decay³¹” bằng 1e-6 giúp cập nhật lại tốc độ học qua từng vòng lặp (epochs). Và với “Momentum³²” là 0.9 có tác dụng thúc đẩy quá trình vượt qua những điểm ngăn cản quá trình tiến đến mục tiêu nhưng khi tiến đến mục tiêu nó lại mất nhiều thời gian để dừng lại, để khắc phục điều đó với “Nesterov³³” là “true” giúp ta tiến tới quá trình hội tụ được nhanh hơn.



Hình 38: Mô hình mạng nơ-ron của mô hình Chatbot

Mạng nơ-ron xây dựng cho mô hình gồm đầu vào (Input) mảng véc-tơ từ với độ dài là 263 được xây dựng từ trước, tiếp theo 2 lớp ẩn (Hidden_layer_1) bao gồm lớp đầu tiên gồm 128 nút và lớp thứ hai số lượng nút là 128, trong hai lớp ẩn được

³⁰ <https://viblo.asia/p/optimizer-hieu-sau-ve-cac-thuat-toan-toi-uu-gdsgdadam-Qbq5QQ9E5D8>

³¹ <https://viblo.asia/p/demon-momentum-decay-cho-mo-hinh-nn-aWj53jm8l6m>

³² <https://viblo.asia/p/optimizer-hieu-sau-ve-cac-thuat-toan-toi-uu-gdsgdadam-Qbq5QQ9E5D8>

³³ <https://viblo.asia/p/optimizer-hieu-sau-ve-cac-thuat-toan-toi-uu-gdsgdadam-Qbq5QQ9E5D8>

xây dựng thêm hàm kích hoạt “tanh”. Cuối cùng là đầu ra (Output) là mảng giá trị véc-tơ nhãn độ dài là 25 tương ứng số lượng nhãn trong tập dữ liệu và có thêm hàm kích hoạt “softmax”.

Mạng nơ-ron sau khi thêm các lớp và các hàm kích hoạt cần thiết, mô hình sẽ được huấn luyện trên tập dữ liệu gồm 1043 câu hỏi. Với số vòng lặp (epochs) là 100, và số lượng batch_size là 4. Sau khi thực hiện huấn luyện xong, mô hình được lưu với tên là “Chatbot_model.h5”

Đánh giá mô hình tối ưu hóa với giải thuật SGD

Sau khi xây dựng được mô hình, chúng ta lần lượt thực hiện đánh giá lại các hàm kích hoạt đã sử dụng là: relu, sigmoid và tanh. Mô hình vẫn được huấn luyện trên tập dữ liệu gồm 1043 câu hỏi, số lượng vòng lặp (epochs) là 100, số lượng batch_size là 4. Kết quả trả về lần lượt là:

- Đối với hàm “tanh” xấp xỉ đạt: 0.9494
- Đối với hàm “sigmoid” xấp xỉ đạt: 0.8863
- Đối với hàm “relu” xấp xỉ đạt: 0.9452

Trong mô hình, hàm kích hoạt “tanh” cho ra độ chính xác cao nhất, độ chính xác lên đến 94,94%.

3.1.5. Xử lý câu hỏi người dùng và đưa ra dự đoán

Giao diện của website được xây dựng bằng ngôn ngữ HTML, được căn chỉnh bằng CSS và tạo các tính năng bằng ngôn ngữ lập trình Javascript và sự hỗ trợ của Framework Flask. Ứng dụng website được xây dựng đơn giản, thân thiện tối ưu hóa khả năng chat của người sử dụng, ứng dụng được chạy local trên máy với địa chỉ là: <http://127.0.0.1:5000/>.

Ứng dụng website có phương thức trao đổi theo hình thức Người hỏi – Người trả lời. Phương thức này thông qua việc sử dụng văn bản để truyền tải thông điệp theo yêu cầu của người dùng

Người sử dụng sẽ đóng vai trò là người hỏi, sử dụng phương thức nhập câu hỏi thông qua văn bản, có thể xuất hiện các từ viết tắt, viết có dấu hoặc không dấu, viết in hoa trong câu hỏi.

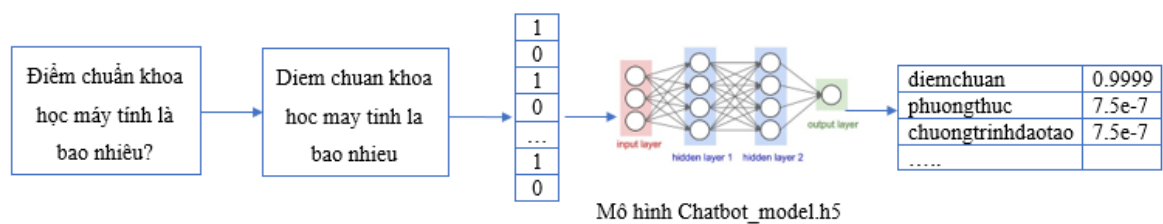
Người trả lời là Chatbot sau khi nhận được câu hỏi của người dùng, sẽ thực hiện các bước bên trong để xử lý câu hỏi của người dùng như các xử lý tập dữ liệu bằng việc chuyển câu hỏi về dạng không dấu, xử lý các từ viết tắt, tách từ,... với các câu hỏi có nhiều ý sẽ được tách thành các ý riêng biệt để trả lời. Sau khi thực

hiện xong quá trình xử lý câu hỏi, ứng dụng sẽ thực hiện tạo mảng véc-tơ đưa vào mạng nơ-ron để đưa ra dự đoán câu trả lời.

Kết quả dự đoán sẽ đưa ra kết quả là xác suất câu hỏi đó thuộc nhãn nào bao nhiêu phần trăm. Nếu nhãn hệ thống dự đoán có kết quả lớn hơn 90% thì hệ thống sẽ tìm kiếm bên trong trường “responses” câu trả lời phù hợp với yêu cầu, ngược lại với xác suất nhỏ hơn 90% thì hệ thống sẽ trả lời người dùng là: “Xin lỗi, hiện tại câu hỏi này bot chưa thể giải đáp, bot sẽ ghi nhận câu hỏi và cải thiện chất lượng”.

Đối với câu hỏi đơn ý

Ví dụ với câu hỏi “Điểm chuẩn khoa học máy tính là bao nhiêu?”

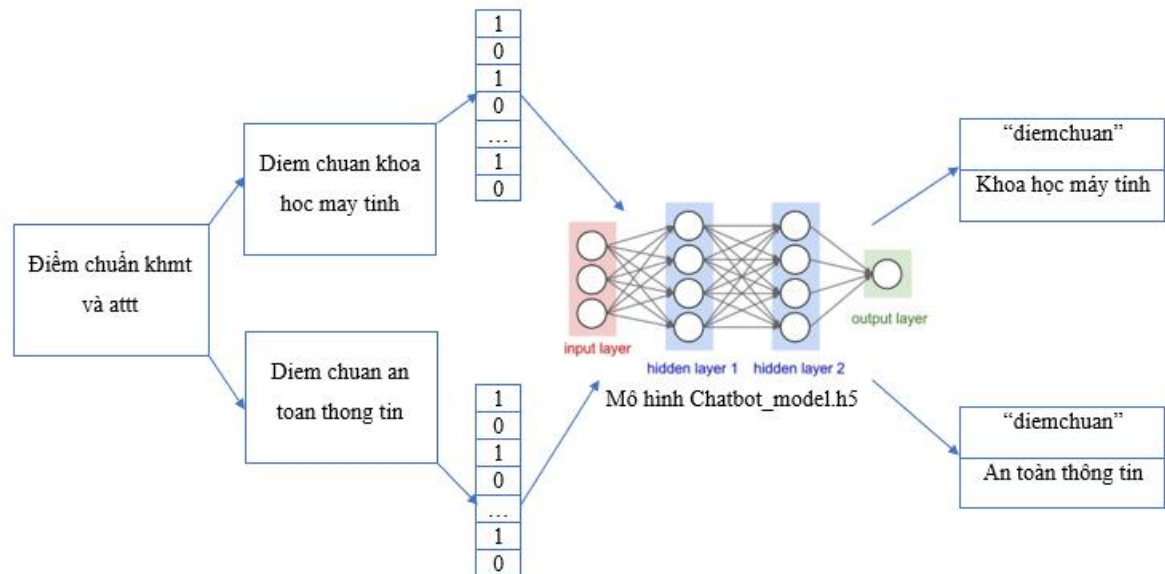


Hình 39: Quá trình xử lý câu hỏi đơn ý

Với câu hỏi đơn ý như: “Điểm chuẩn khoa học máy tính là bao nhiêu?” câu hỏi sẽ được tách ra và lược bỏ những từ hoặc ký tự không quan trọng, sau đó mô hình sẽ đưa ra dự đoán câu hỏi thuộc nhãn “diemchuan” với xác suất là 0.999999 ~ 1.0. Ứng dụng sẽ lấy câu trả lời nằm trong trường “pattern” thuộc nhãn “diemchuan” để trả lời cho người dùng.

Đối với câu hỏi đa ý

Ví dụ câu hỏi: “Điểm chuẩn khmt và attt” quá trình dự đoán sẽ được thực hiện như sau:



Hình 40: Quá trình xử lý câu hỏi đa ý

Khi những câu hỏi có nhiều ý, câu hỏi đó sẽ được xử lý tách thành các ý riêng biệt, và xử lý lọc bỏ những từ không cần thiết. Các câu hỏi nhỏ sau khi được tách ra, véc-tơ hóa các câu hỏi, rồi mô hình sẽ xử lý tương tự như câu hỏi đơn ý.

3.1.6. Thu thập và thêm dữ liệu từ câu hỏi người dùng

Thu thập dữ liệu

Khi người dùng đặt những câu hỏi trên hệ thống, sẽ xuất hiện xác suất có những câu hỏi mà hệ thống Chatbot không thể trả lời do hệ thống chưa từng được huấn luyện. Để tăng thêm độ chính xác của hệ thống và phát triển hệ thống hơn thì việc thu thập câu hỏi người dùng là rất cần thiết. Câu hỏi của người dùng sẽ được thu thập và lưu trữ bên trong cơ sở dữ liệu SQLite để làm dữ liệu huấn luyện sau này.

Cấu trúc của tập tin lưu trữ câu hỏi của người dùng trong quá trình sử dụng hệ thống Chatbot sẽ bao gồm 2 trường dữ liệu là câu hỏi người dùng nhập và nhãn câu trả lời hệ thống.

Ví dụ đối với câu hỏi “Điểm chuẩn Khoa học Máy Tính” được chương trình nhận biết là thuộc nhãn “diemchuan”

```
{'diem chuan khoa hoc may tinh': 3}
[{'intents': 'diemchuan', 'probability': '1.0'}]
Ngành Khoa học máy tính điểm thi trúng tuyển năm 2021 là 25.00, điểm trúng tuyển học bạ là 27.00
['diemchuan3']
```

Hình 41: Dữ liệu câu hỏi có câu trả lời

Nếu trường hợp câu hỏi mà hệ thống không thể trả lời như: “Con gái học có khó không” thì câu hỏi sẽ được lưu vào nhãn “no_answer”

```
Connection to SQLite DB successful
{'con gái học có khó không': -1}
Xin lỗi, hiện tại câu hỏi này bot chưa thể giải đáp, bot sẽ ghi nhận câu hỏi và cải thiện chất lượng
['no_answer']
```

Hình 42: Dữ liệu câu hỏi không có câu trả lời

Việc lưu trữ không chỉ giúp cho hệ thống có thêm dữ liệu để huấn luyện cho mô hình Chatbot, còn giúp nhận biết những câu hỏi mà hệ thống chưa thể trả lời hoặc trả lời sai, đó là tiền đề để điều chỉnh về mặt giải thuật, điều chỉnh mô hình phù hợp nhằm tăng độ chính xác trong quá trình sử dụng hệ thống.

Thêm dữ liệu

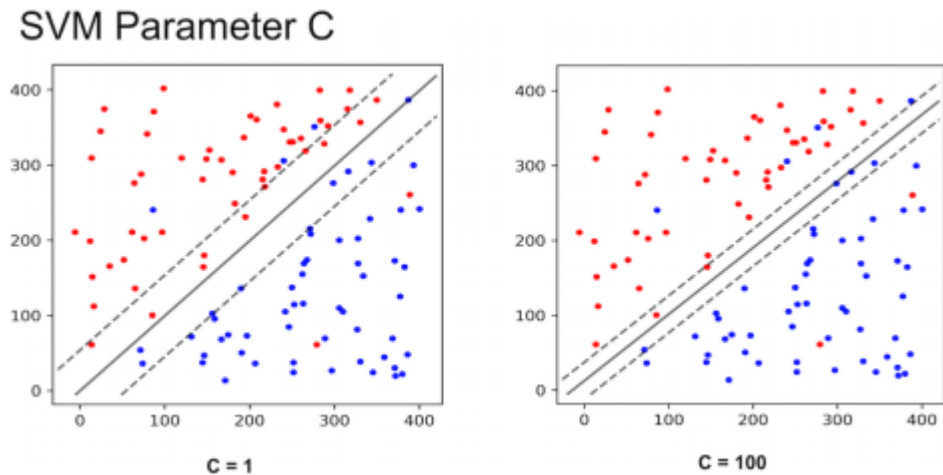
Trong quá trình hệ thống vận hành, người dùng có thể đặt ra những câu hỏi mà Chatbot chưa thể trả lời, để mở rộng tập dữ liệu và làm phong phú hơn đề tài Chatbot có thể cung cấp cho người dùng thì việc thu thập câu hỏi từ chính người dùng đặt ra đóng vai trò rất quan trọng. Người quản trị có thể kiểm tra lịch sử cuộc hội thoại giữa Người dùng và Chatbot từ đó có thể thêm vào những câu hỏi cần thiết. Hệ thống thiết kế việc thu tập câu hỏi trực tiếp thông qua website bằng 2 cách:

- Nếu câu hỏi cần thêm thuộc các nhãn có sẵn trong tập dữ liệu thì người quản trị chỉ cần lựa chọn nhãn thích hợp và thêm câu hỏi đó trong tập dữ liệu.
- Nếu câu hỏi không thuộc trong các nhãn thì người quản trị có thể thêm nhãn mới và tạo câu trả lời cho nhãn mới thêm.

3.1.7. Cài đặt các mô hình

Giải thuật SVM

Dữ liệu huấn luyện sẽ được điều chỉnh nhãn để phù hợp với giải thuật. Đối với SVM [nguyên cứu sử dụng thư viện Sklearn trong Python để tạo mô hình SVM. Mô hình sử dụng Kernel là Linear. Với Linear được sử dụng khi dữ liệu có thể phân tách tuyến tính, tức là có thể được phân tách bằng một dòng duy nhất. Linear được sử dụng khi có một số lượng lớn tính năng trong một tập dữ liệu cụ thể. Một trong những ví dụ về tập dữ liệu có rất nhiều tính năng là phân loại văn bản, vì mỗi bảng chữ cái là một tính năng mới.



Hình 43: Giải thuật SVM với tham số C=1 và C=100

Tham số C cho biết mức độ tối ưu hóa SVM mà người sử dụng muốn để tránh phân loại sai từng ví dụ đào tạo. Đối với các giá trị lớn của C, việc tối ưu hóa sẽ chọn một siêu phẳng có lợi nhuận nhỏ hơn nếu siêu phẳng đó thực hiện tốt hơn việc nhận được tất cả các điểm rèn luyện được phân loại chính xác. Ngược lại, một giá trị rất nhỏ của C sẽ khiến trình tối ưu hóa tìm kiếm siêu phẳng phân tách có lẽ lớn hơn, ngay cả khi siêu phẳng đó phân loại sai nhiều điểm hơn. Đối với các giá trị rất nhỏ của C, người sử dụng sẽ nhận được các ví dụ bị phân loại sai, thường xuyên ngay cả khi dữ liệu đào tạo của người sử dụng có thể phân tách tuyến tính.

Trong nghiên cứu này mô hình SVM được xây dựng với kernel là Linear và C là 0.05.

Giải thuật K láng giềng

Nghiên cứu sử dụng K láng giềng [13] để làm phép so sánh về độ chính xác với mô hình nghiên cứu, giải thuật sử dụng $n_neighbors = 5$ và được xây dựng bằng cách sử dụng thư viện Sklearn trong Python để tạo mô hình.

Trong quá trình xử lý, giải thuật K – láng giềng sẽ dự đoán khoảng cách giữa 2 véc-tơ bằng cách tính toán dựa trên công thức:

$$\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Hình 44: Công thức khoảng cách K – láng giềng

Để đưa ra dự đoán cho câu hỏi, giải thuật sẽ tính khoảng cách câu hỏi đó với toàn bộ câu hỏi trong tập dữ liệu. Nếu khoảng cách câu hỏi nào nhỏ nhất khi xét với

câu hỏi cần dự đoán sẽ lấy câu trả lời của câu hỏi đó ra làm câu trả lời cho câu hỏi cần dự đoán.

Giải thuật Naïves Bayes

Dữ liệu training sẽ được điều chỉnh nhãn để phù hợp với giải thuật. Đối với Gaussian Naïve Bayes [14] nguyên cứu sử dụng thư viện Sklearn trong Python để tạo mô hình Gaussian Naïve Bayes

GaussianNB triển khai thuật toán Gaussian Naive Bayes để phân loại. Khả năng các tính năng được giả định là Gaussian:

$$P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

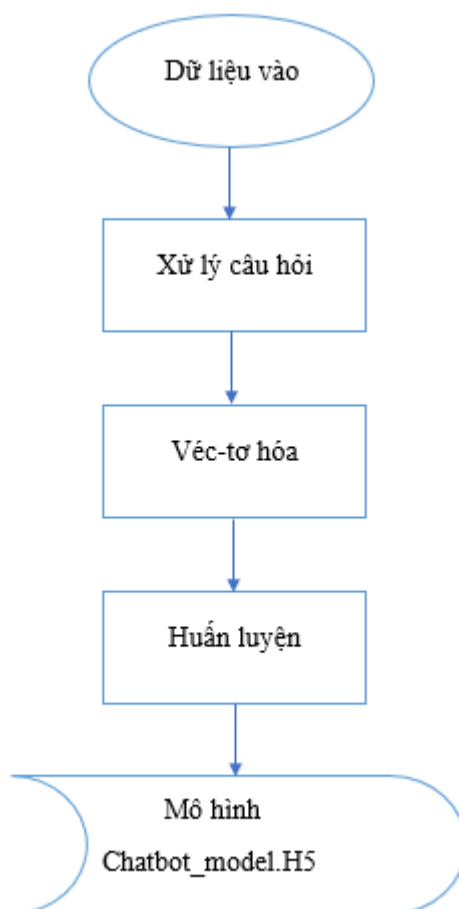
Hình 45: Công thức cho giải thuật

Các tham số σ_y và μ_y được ước tính bằng cách sử dụng khả năng tối đa.

3.2. Xây dựng hệ thống

Cài đặt giải thuật được chia làm hai giai đoạn: Tạo mô hình dự đoán và Dự đoán câu trả lời.

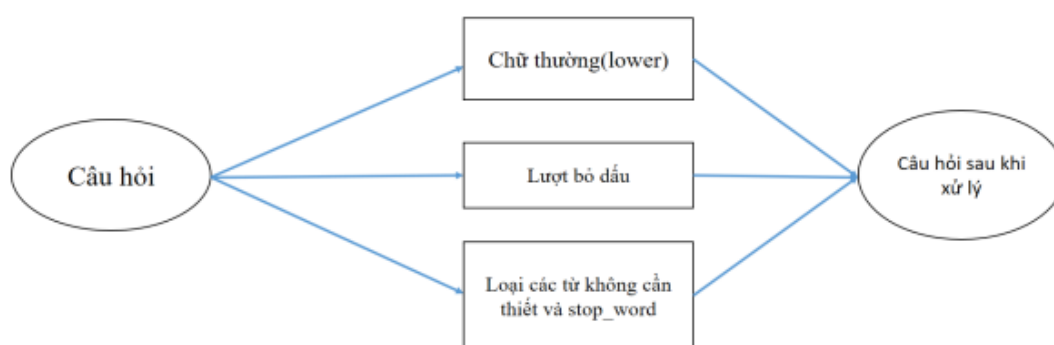
3.2.1. Tạo mô hình dự đoán



Hình 46: Tạo mô hình dự đoán

Xử lý dữ liệu câu hỏi

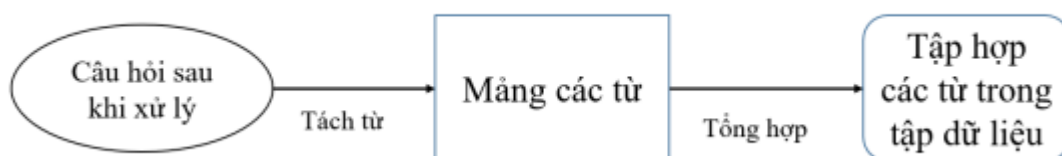
Câu hỏi sẽ được chuyển thành chữ thường, lược bỏ dấu và loại các từ không cần thiết trong câu và stop_word. Quá trình sử dụng các hàm trong python như lower(), replace(), sub().



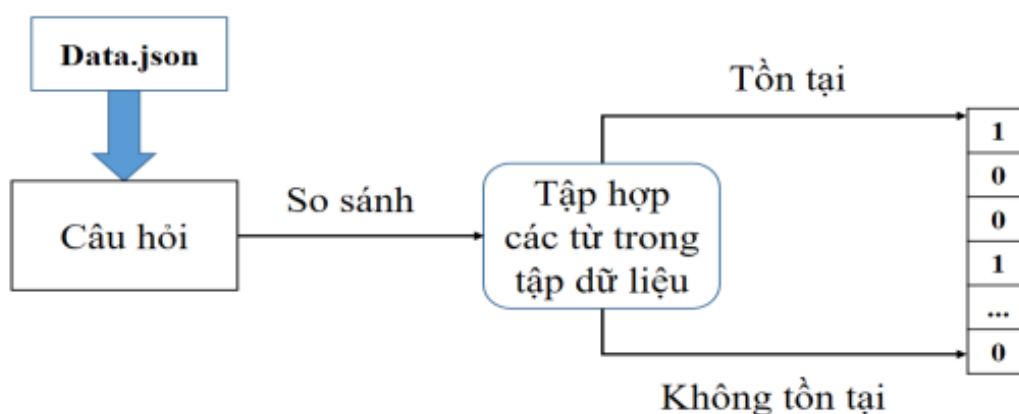
Hình 47: Xử lý câu hỏi

Véc-tơ hóa câu hỏi

Sau khi câu hỏi được xử lý từ bước 1, thực hiện tách các từ trong câu hỏi để tạo thành 1 mảng từ, tổng hợp nhiều mảng từ để tạo thành 1 tập hợp các từ trong tập dữ liệu. Quá trình đó như sau:



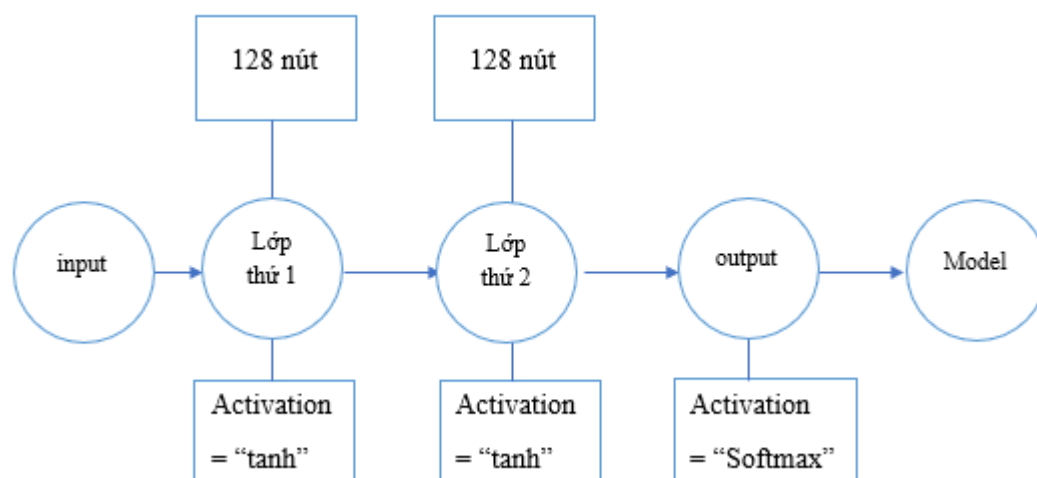
Hình 48: Tạo tập hợp từ



Hình 49: Véc-tơ hóa câu hỏi

Tất cả câu hỏi trong tập dữ liệu sẽ được so sánh với tập hợp các từ trong tập dữ liệu (263 từ). Quá trình véc-tơ sử dụng hàm `word_tokenize()` giúp phân tách các từ trong câu thành mảng các từ. Sau đó chuyển hóa thành véc-tơ mảng số để đưa vào mô hình huấn luyện.

Xây dựng mô hình huấn luyện:

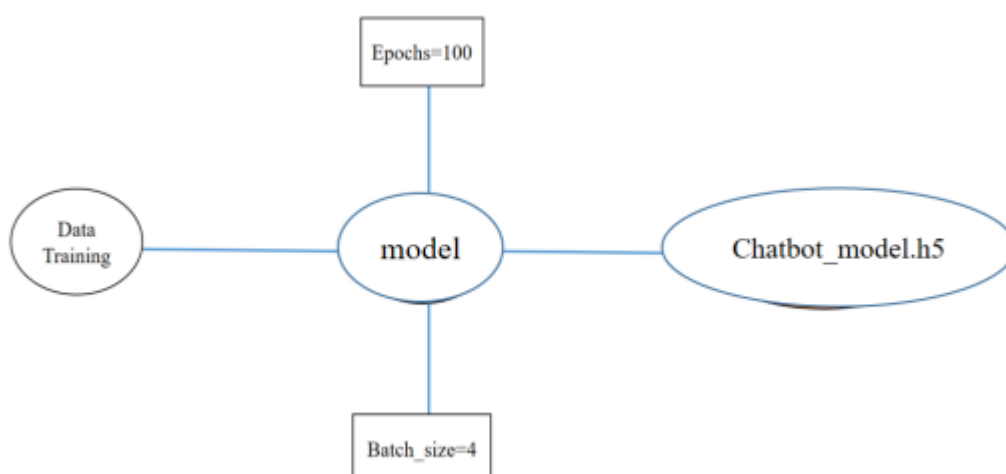


Hình 50: Mô hình mạng nơ-ron

Mô hình được xây dựng thông qua thư viện Keras. Đầu tiên là xây dựng hàm tối ưu hóa SGD với các thông số thích hợp. Tiếp theo xây dựng từng lớp ẩn cho mạng nơ-ron.

Huấn luyện mô hình

Dữ liệu với 1043 câu hỏi sau khi được véc-tơ hóa sẽ được huấn luyện với số vòng lặp là 100 và được chia batch_size là 4. Và kết quả của quá trình là được mô hình “Chatbot_model.h5”.

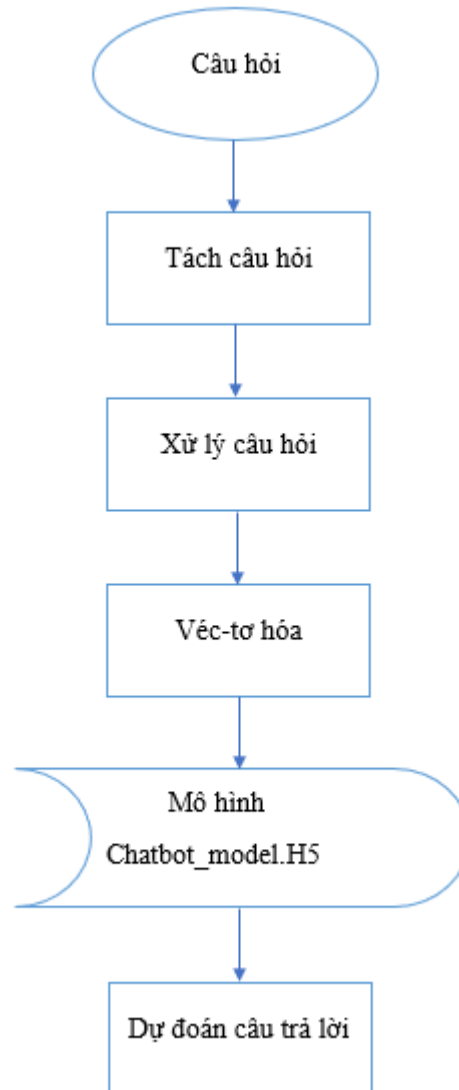


Hình 51: Quá trình huấn luyện mô hình

3.2.2. Xử lý câu hỏi người dùng và dự đoán câu trả lời

Dự đoán câu trả lời

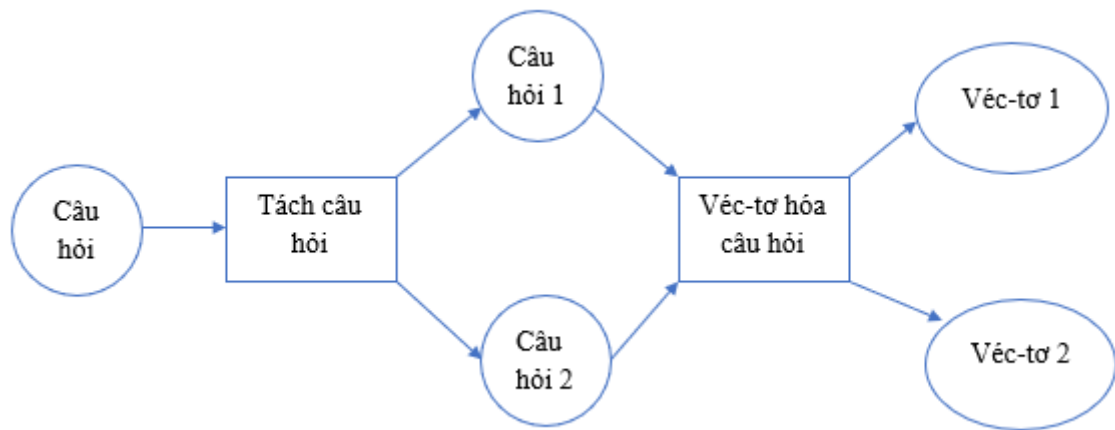
Quá trình dự đoán câu trả lời của người dùng sẽ được minh họa bằng mô hình như sau:



Hình 52: Quá trình dự đoán câu trả lời của người dùng

Xử lý câu hỏi của người dùng:

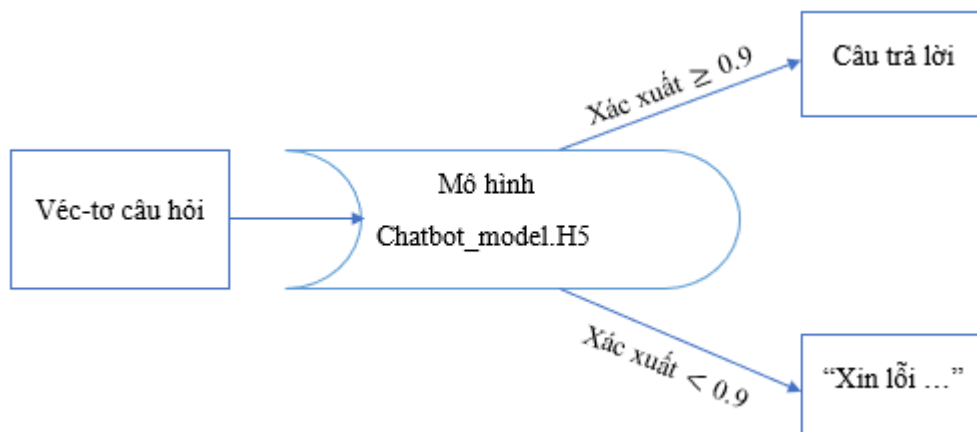
Quá trình xử lý câu hỏi của người dùng được mô hình hóa như sau:



Hình 53: Quá trình xử lý câu hỏi người dùng

Câu hỏi của người dùng khi vào hệ thống sẽ được tách thành từng ý (nếu có), sau đó sẽ trải qua quá trình tiền xử lý dữ liệu và được véc-tơ hóa thành mảng một chiều. Nếu trong câu hỏi có các từ liên quan đến nhóm ngành Công nghệ Thông Tin thuộc Trường Công nghệ Thông tin và Truyền thông bao gồm các ngành: Công nghệ Thông tin, Công nghệ Thông tin Chất lượng cao, Công nghệ Thông tin học tại Khu Hòa An (nhóm ngành Công nghệ Thông tin), An toàn Thông tin, Truyền thông Đa phương tiện, Khoa học Máy tính, Kỹ thuật Máy tính, Mạng máy tính và Truyền Thông dữ liệu, Kỹ thuật Phần mềm và Hệ thống Thông tin sẽ được đánh dấu để trả lời người dùng đúng ngành nghề mà người dùng mong muốn.

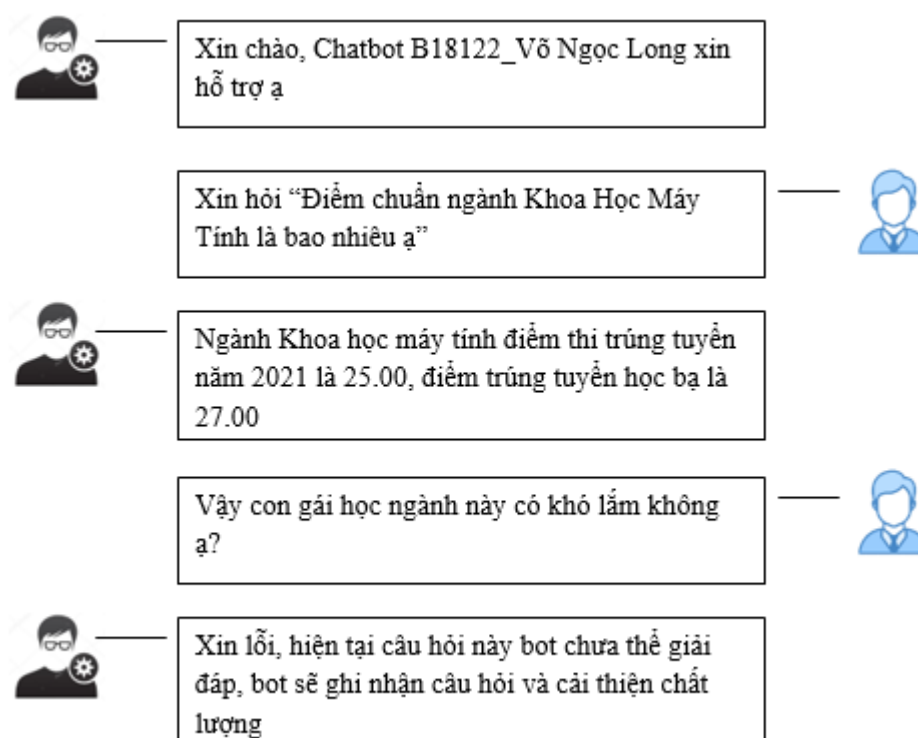
Đưa ra dự đoán câu trả lời:



Hình 54: Quá trình đưa ra dự đoán câu trả lời

Khi véc-tơ hóa câu hỏi người dùng, mô hình “Chatbot_model.h5” sẽ ra dự đoán dựa trên véc-tơ hóa câu hỏi người dùng nếu xác suất nhận dự đoán lớn hoặc bằng 0.9 thì câu trả lời sẽ được đưa ra cho người dùng, ngược lại sẽ được đưa ra

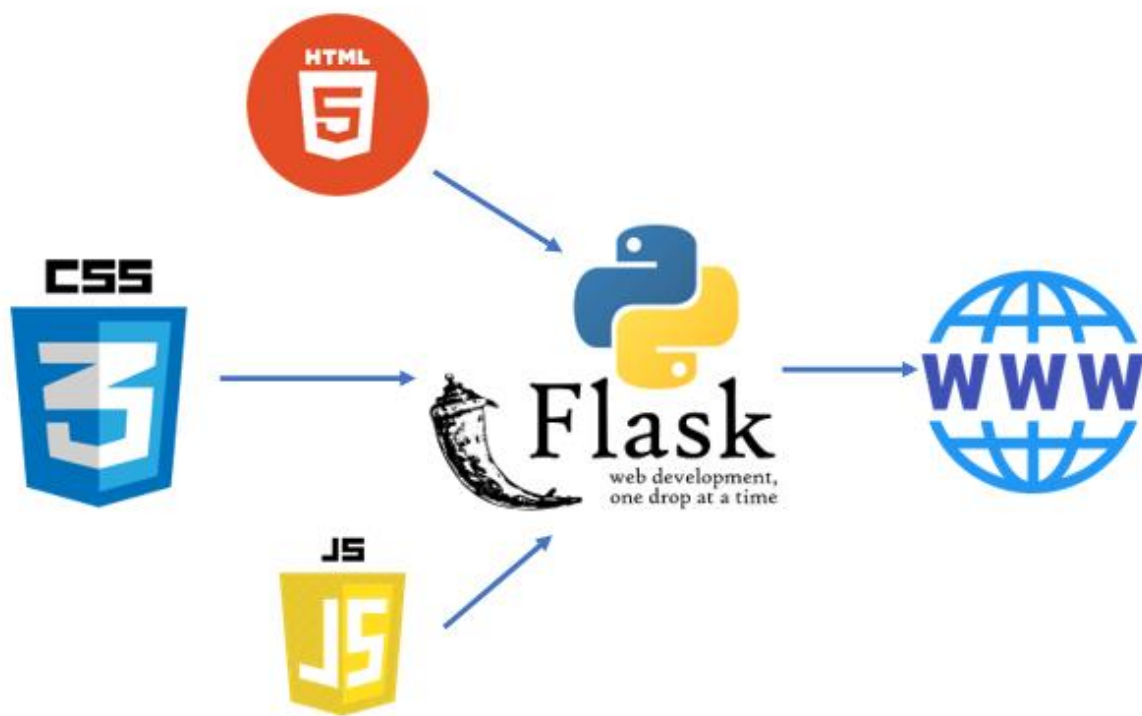
thông báo “Xin lỗi, hiện tại câu hỏi này bot chưa thể giải đáp, bot sẽ ghi nhận câu hỏi và cải thiện chất lượng”.



Hình 55: Chatbot trả lời câu hỏi của người dùng

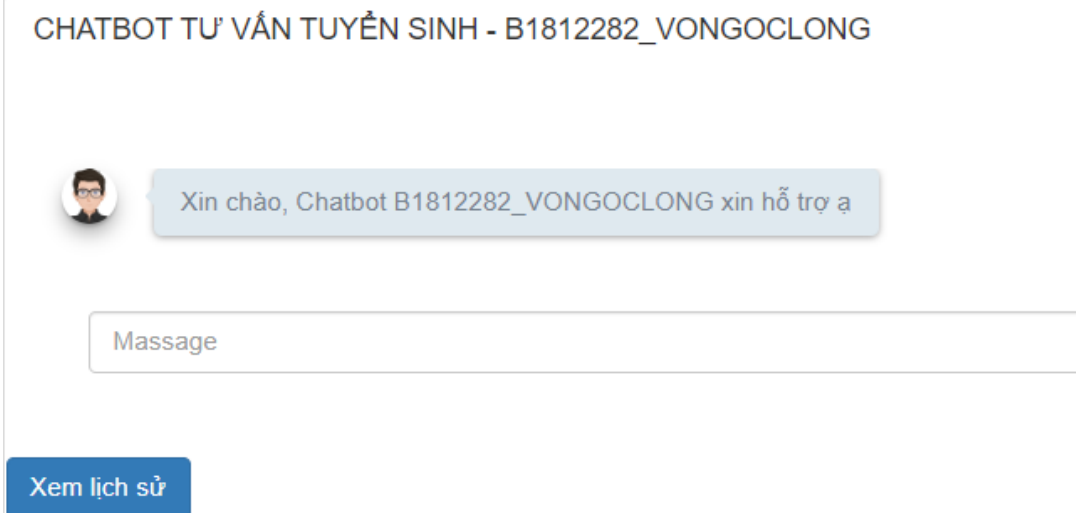
3.2.3. Xây dựng hệ thống website

Hệ thống website được thiết kế bằng Flask Framework của Python. Sử dụng hai ngôn ngữ chính để thiết kế website là HTML và Javascript. Về phần giao diện, sử dụng ngôn ngữ HTML và CSS để thiết kế giao diện, căn chỉnh vị trí của các đối tượng trên website, màu sắc, hình ảnh,... Về phần chức năng, sử dụng Javascript để truyền dữ liệu từ máy chủ lên website, ngoài ra có thể sử dụng Javascript để nâng cấp hệ thống website theo yêu cầu.



Hình 56: Mô hình thiết kế website với các ngôn ngữ và công cụ

3.2.4. Giao diện của ứng dụng website



Hình 57: Giao diện ứng dụng website Chatbot

CHƯƠNG 4. ĐÁNH GIÁ KẾT QUẢ HUẤN LUYỆN VÀ GIAO DIỆN HỆ THỐNG WEBSITE

Chương 4 sẽ giới thiệu về phương pháp đánh giá kết quả huấn luyện độ chính xác của các mô hình, sau đó là phương pháp đánh giá mô hình nào là tối ưu nhất với hệ thống dựa trên so sánh về độ chính xác. Giới thiệu về hệ thống chatbot tích hợp với ứng dụng website, các chức năng cơ bản của hệ thống.

4.1. Đánh giá và so sánh kết quả huấn luyện

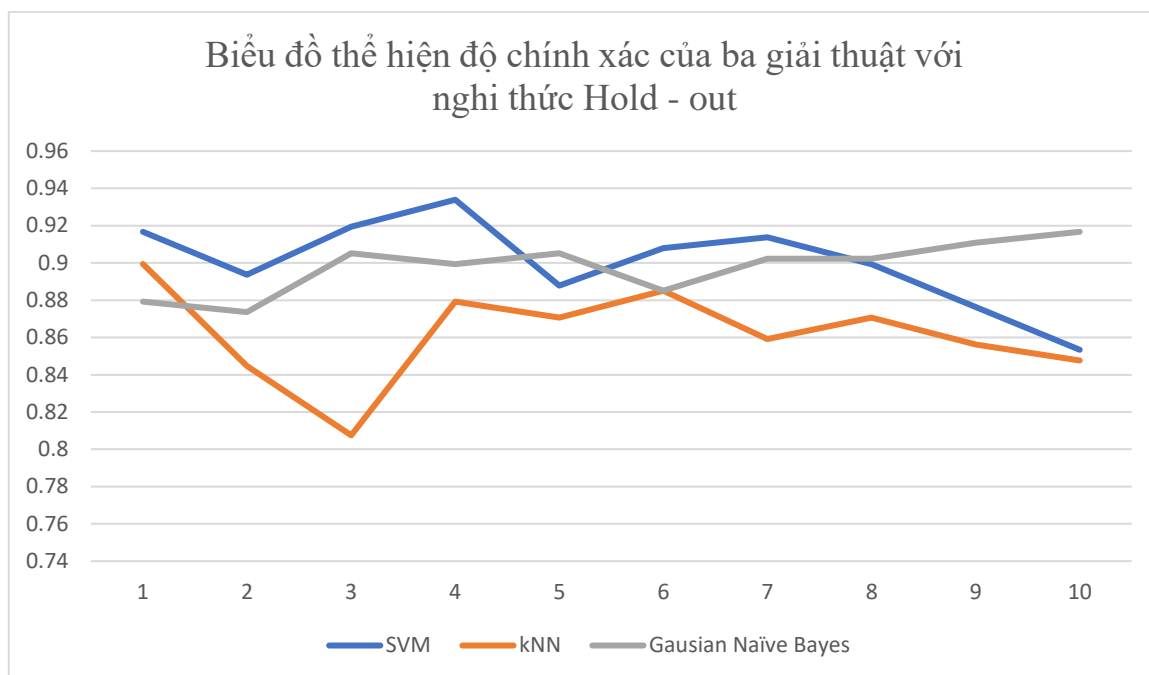
4.1.1. Độ chính xác của ba giải thuật máy học

Với tập dữ liệu với số lượng câu hỏi là 1043, nghiên cứu thực hiện đánh giá độ chính xác bằng cách xử dụng hai nghi thức Hold-out nhằm tính toán chỉ số accuracy_score của bài giải thuật: SVM, kNN, Gaussian Naïve Bayes. Sau đó thực hiện so sánh các giá trị.

Đối với nghi thức Hold-out số lần lặp là 10

	SVM	kNN	Gaussian Naïve Bayes
1	0.9167	0.8994	0.8793
2	0.8937	0.8448	0.8736
3	0.9195	0.8075	0.9052
4	0.9339	0.8793	0.8994
5	0.8879	0.8707	0.9052
6	0.9080	0.8851	0.8851
7	0.9138	0.8592	0.9023
8	0.8994	0.8707	0.9023
9	0.8764	0.8563	0.9109
10	0.8534	0.8477	0.9167
Trung bình	0.8902	0.8779	0.8920

Bảng 7: Độ chính xác của ba giải thuật với nghi thức Hold-out



Hình 58: Biểu đồ độ chính xác của 3 giải thuật với nghi thức Hold-out

Độ chính xác trung bình của ba giải thuật SVM, kNN, Gaussian Naïve Bayes sau 10 lần lặp lần lượt xấp xỉ đạt: 90.02%, 89.79%, 86.20%. Giải thuật SVM là giải thuật có độ chính xác tốt nhất sau 10 lần lặp ở nghi thức Hold-out.

4.1.2. Độ chính xác của mô hình với giải thuật SGD

Độ chính xác của mô hình Chatbot_model.h5 với giải thuật SGD sử dụng tập dữ liệu bao gồm 1043 câu hỏi được tính toán bằng nghi thức Hold-out với số lần lặp là 10 lần, tập train là 2/3 và tập test là 1/3.

So sánh độ chính xác của mô hình với giải thuật SVM, giải thuật có độ chính xác cao nhất khi sử dụng nghi thức Hold-out để tính so với 2 giải thuật còn lại.

	SVM	SGD
1	0.9167	0.9339
2	0.8937	0.9310
3	0.9195	0.9109
4	0.9339	0.9296
5	0.8879	0.9239
6	0.9080	0.9339
7	0.9138	0.9282

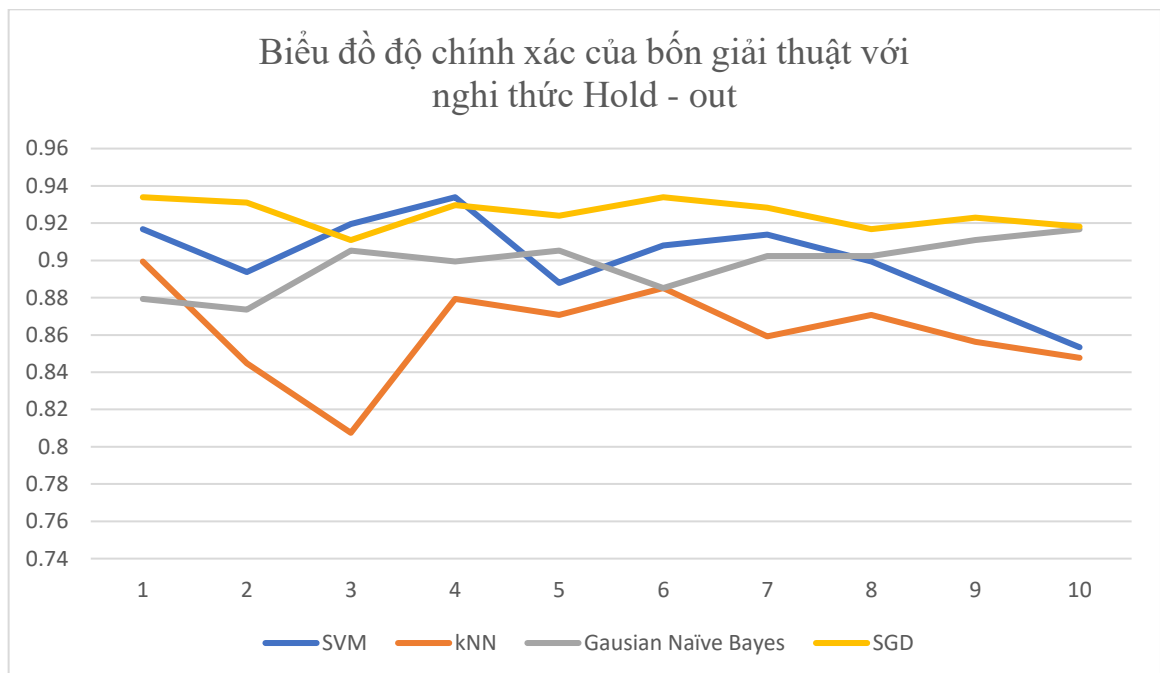
8	0.8994	0.9167
9	0.8764	0.9230
10	0.8534	0.9181
Trung bình	0.9002	0.9247

Bảng 8: Độ chính xác của hai giải thuật với nghi thức Hold-out

Độ chính xác trung bình của mô hình với giải thuật SGD sau 10 lần lặp xấp xỉ: 0.9247.

Cùng với số lần lặp tương tự như thế, độ chính xác của 3 giải thuật còn lại là:

- Giải thuật SVM xấp xỉ đạt: 90.02% xếp thứ 2
- Giải thuật Gaussian Naïve Bayes xấp xỉ đạt: 89.79% xếp thứ 3
- Giải thuật kNN xấp xỉ đạt: 86.20% xếp thứ 4



Hình 59: Biểu đồ độ chính xác của các giải thuật với nghi thức Hold-out

4.2. Giao diện website

4.2.1. Cấu trúc của website



Hình 60: Cấu trúc của website

Người dùng có thể đặt câu hỏi bằng cách nhập câu hỏi vào thanh nhập câu hỏi, Chatbot sẽ căn cứ vào câu hỏi của người dùng để thực hiện phân lớp câu hỏi đó thuộc vào nhãn nào bên trong tập dữ liệu, sau đó Chatbot sẽ đưa ra câu trả lời phù hợp với câu hỏi người dùng.

4.2.2. Quá trình thu thập câu hỏi

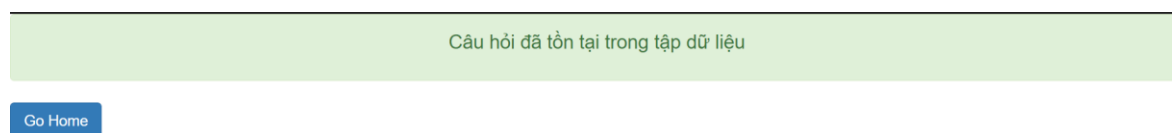
Khi người người dùng đặt câu hỏi cho Chatbot, hệ thống sẽ lưu trữ những câu hỏi đó để dữ liệu huấn luyện sau này, có hai chức năng trong trang thu thập câu hỏi người dùng gồm xóa câu hỏi, thêm câu hỏi vào thập dữ liệu.

92	kha nang tot nghiep nganh khoa hoc may tinh cao khong	tiletotnghiep3	Delete	Add
93	neu hoc xong muon xin hoc bong	hocbong	Delete	Add
94	trường có hỗ trợ kỹ túc xá không	no_answer	Delete	Add
95	kỹ túc xá giá cả ra sao	kytucxa	Delete	Add
96	hồ sơ cần những gì	hoso	Delete	Add
97	Chỉ tiêu của ngành CNTT là bao nhiêu ạ?	chitieu0	Delete	Add
98	thi khối nào sẽ được xét ngành cntt	tohopxettuyen0	Delete	Add
99	ngày nào bắt đầu xét học bạ	no_answer	Delete	Add
100	Cho em hỏi khi nào trường mới nhận xét điểm học bạ ạ	thoigian	Delete	Add

Hình 61: Tập hợp câu hỏi của người dùng

Khi câu hỏi đã tồn tại trong tập dữ liệu hoặc câu hỏi không hợp lý thì admin có thể xóa câu hỏi đó bằng cách bấm “Delete”, khi admin muốn thêm 1 câu hỏi vào tập dữ liệu chỉ cần bấm “Add”.

Khi bấm “Add” nếu câu hỏi tồn tại trong tập dữ liệu thì hệ thống sẽ thông báo ví dụ như câu hỏi “hồ sơ cần những gì” đã tồn tại trong tập dữ liệu hệ thống sẽ báo “Câu hỏi đã tồn tại trong tập dữ liệu”.

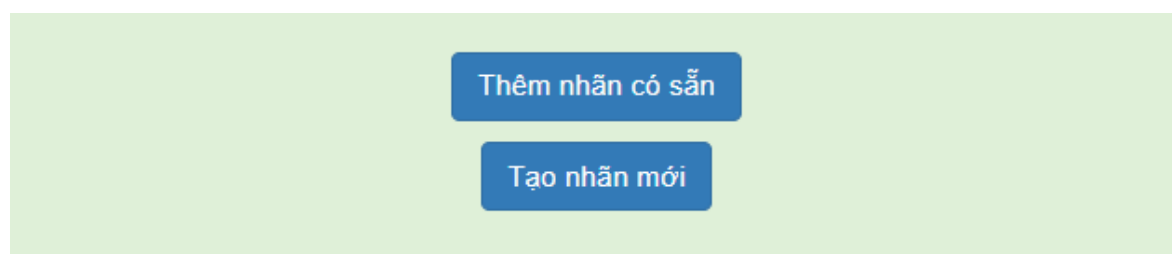


Hình 62: Thông báo đã tồn tại câu hỏi trong tập dữ liệu

Nếu câu hỏi chưa tồn tại thì người quản trị có thể thêm vào tập dữ liệu bằng hai cách:

- Thêm trực tiếp vào tập dữ liệu: Mở tập dữ liệu, thêm câu hỏi vào nhãn, phương pháp này rất mất thời gian chưa thật sự tối ưu
- Thêm vào tập dữ liệu trực tiếp trên website: Sử dụng thêm vào tập dữ liệu bằng hệ thống website.

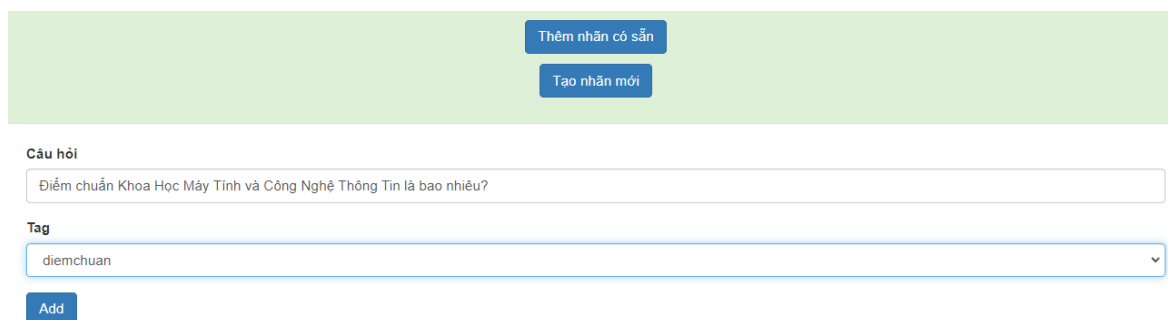
Phương pháp thêm vào tập dữ liệu trực tiếp trên website có thể thêm dữ liệu bằng hai cách.



Hình 63: Hai cách thêm câu hỏi thông qua hệ thống website

Cách 1: Thêm nhãn có sẵn

Nếu câu hỏi cần thêm thuộc vào các nhãn có sẵn trong tập dữ liệu thì người quản trị chỉ cần lựa chọn nhãn thích hợp và thêm câu hỏi đó trong tập dữ liệu.



Thêm nhãn có sẵn

Tạo nhãn mới

Câu hỏi

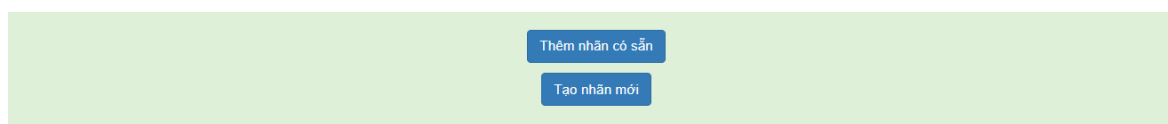
Điểm chuẩn Khoa Học Máy Tính và Công Nghệ Thông Tin là bao nhiêu?

Tag

diemchuan

Add

Hình 64: Thêm câu hỏi với nhãn có sẵn



Thêm nhãn có sẵn

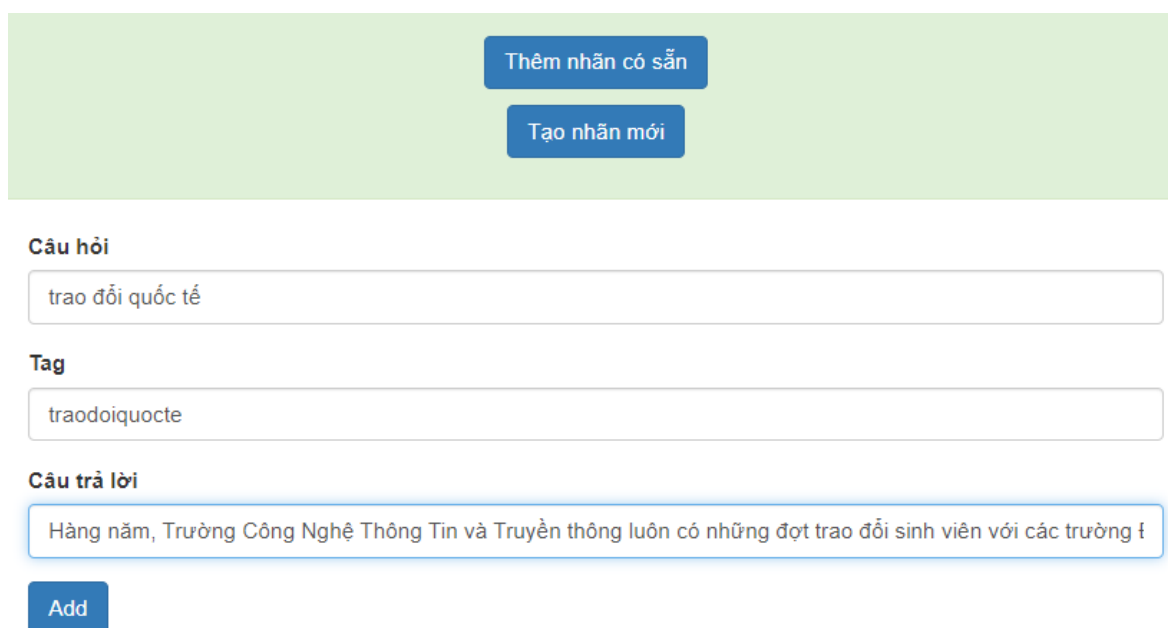
Tạo nhãn mới

Đã thêm câu hỏi Điểm chuẩn Khoa học Máy Tính và Công nghệ Thông Tin là bao nhiêu? vào nhãn diemchuan

Hình 65: Thông báo thêm thành công câu hỏi vào nhãn có sẵn

Cách 2: Tạo nhãn mới – Chưa tồn tại nhãn trong tập dữ liệu

Nếu câu hỏi cần thêm thuộc vào một lĩnh vực khác, không tồn tại bên trong 25 nhãn thuộc tập dữ liệu huấn luyện. Người quản trị cần thực hiện tạo nhãn mới và thêm câu hỏi phù hợp vào nhãn đó.



Thêm nhãn có sẵn

Tạo nhãn mới

Câu hỏi

trao đổi quốc tế

Tag

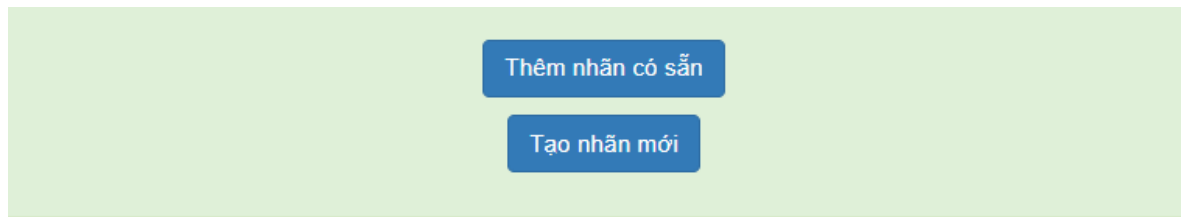
traodoiquocte

Câu trả lời

Hàng năm, Trường Công Nghệ Thông Tin và Truyền thông luôn có những đợt trao đổi sinh viên với các trường t

Add

Hình 66: Thêm câu hỏi với nhãn chưa có trong tập dữ liệu



Đã thêm thành công

Hình 67: Thông báo khi thêm thành công nhãn và câu hỏi vào tập dữ liệu

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận nghiên cứu

Nghiên cứu dựa trên việc huấn luyện Chatbot bằng phương pháp học sâu, với tập dữ liệu là: 1043 câu hỏi. Nghiên cứu xây dựng Mạng nơ-ron 2 lớp tích hợp tối ưu hóa SGD và các hàm kích hoạt “tanh, relu, sigmoid” với hàm kích hoạt “tanh” trả về độ chính xác cao nhất. Chatbot có khả năng trao đổi trực tiếp với người dùng nhằm cung cấp thông tin và có thể lưu trữ câu hỏi để tăng số lượng dữ liệu học.

Chatbot hỗ trợ tư vấn cho các ngành, nhóm ngành Công nghệ Thông Tin thuộc Trường Công nghệ Thông tin và Truyền Thông Đại học Cần Thơ với 8 ngành nghề là: Công nghệ Thông tin, Công nghệ Thông tin Chất lượng cao, Công nghệ Thông tin học tại khu Hòa An (được phân lớp và nhóm ngành Công nghệ Thông Tin), An toàn Thông tin, Truyền thông Đa phương tiện, Khoa học máy tính, Kỹ thuật máy tính, Mạng máy tính và Truyền thông Dữ liệu, Kỹ thuật phần mềm và Hệ thống thông tin.

Các câu hỏi Chatbot có thể trả lời liên quan đến các lĩnh vực như: điểm chuẩn, phương thức xét tuyển, tổ hợp xét tuyển, chương trình đào tạo, chỉ tiêu các ngành,...Ngoài ra còn có các vấn đề khác như: ký túc xá sinh viên, học bổng khuyến khích, môi trường hợp tác, trình độ tiếng anh,...

Một số kết quả đạt được sau khi huấn luyện và vận hành hệ thống Chatbot:

- Chatbot có thể trả lời các câu hỏi với độ chính xác hơn 90%
- Ứng dụng được tích hợp website, có giao diện thân thiện với người dùng.
- Ứng dụng có khả năng lưu trữ và trả lời những câu hỏi về nhiều lĩnh vực khác nhau, ngoài ra có chức năng thu thập dữ liệu người dùng.

5.2. Hướng phát triển hệ thống

5.2.1. Dữ liệu huấn luyện

Tập dữ liệu hiện tại có 1043 câu hỏi và 25 nhãn có nội dung liên quan đến 8 ngành thuộc Trường Công nghệ Thông Tin và Truyền Thông, với hướng phát triển trong tương lai, tập dữ liệu có thể mở rộng về mặt nội dung, số lượng câu hỏi sẽ không giới hạn bởi 8 ngành mà có thể mở rộng lên hơn 1000 ngành thuộc trường Đại Học Cần Thơ.

5.2.2. Chức năng

Các chức năng có thể phát triển:

- Tạo chức năng giúp người dùng trao đổi trực tiếp với người quản trị

- Hệ thống đang thử sử dụng API của Google để tích hợp chức năng nhận dạng giọng nói.
- Thêm các chức năng như tìm kiếm, đăng ký tài khoản,..
- Thêm chức năng thể hiện danh sách các ngành nghề, hệ sau đại học,...

5.2.3. Huấn luyện mô hình

Hiện tại Chatbot thực hiện huấn luyện mô hình bằng các giải thuật học máy là: SVM, kNN, Bayes Thơ Ngây và SGD. Có mô hình giải thuật được lập trình hoàn toàn bằng thuật toán SGD, có các mô hình sử dụng thư viện Sklearn của Python là: SVM, kNN, Bayes Thơ Ngây. Ngoài ra nghiên cứu có thể sử dụng mô hình Chatbot Rasa để thực hiện huấn luyện mô hình. Đối với Rasa có thể sử dụng các giải thuật như SVM, Bayes Thơ Ngây để huấn luyện trực tiếp, một lợi thế khi sử dụng tập dữ liệu được lưu trữ dưới dạng tập tin .json vì có cấu trúc tương tự với dữ liệu huấn luyện Chatbot theo mô hình Rasa.

TÀI LIỆU THAM KHẢO

- [1] Do Thanh Nghi, Hoang Tung. “Chatbot cho sinh viên Công Nghệ Thông Tin,” Can Tho University, 2019.
- [2] Windiatmoko, Yurio, Ridho Rahmadi, and Ahmad Fathan Hidayatullah. “Developing Facebook Chatbot Based on Deep Learning Using RASA Framework for University Enquiries,” IOP Conference Series: Materials Science and Engineering, vol. 1077. No. 1. IOP Publishing, 2021.
- [3] P. L. Wehenkel, Design and implementation of a chatbot in the context of customer support, 2017-2018.
- [4] T. M. Luan. “Deep learning cho Chatbot – Giới thiệu,” VIBLO, 2018.
- [5] T. T. Truc. “Optimizer – Hiểu sâu về các thuật toán tối ưu (GD, SGD, Adam,...),” VIBLO, 2020.
- [6] N. V. Anh. “Các hàm kích hoạt (Activation Function) trong neural network,” Aicurious.io, 2019
- [7] Funda, “Machine Learning cơ bản,” 2017.
- [8] Mai Pham, “Thuật toán K láng giềng gần nhất (K-Nearest Neighbor – kNN),” Vietnambiz, 2019.
- [9] Do Thanh Nghi. “Phương pháp học Bayes Bayesian classification,” Can Tho University, 2019.
- [10] D. N. Ngọc. “Các phương pháp đánh giá độ chính xác của mô hình phân lớp,” 2013.
- [11] “TOPDev”, “Hiểu rõ về JSON là gì? Cách lấy dữ liệu từ JSON”
- [12] “Scikit-Learn,” [Online]. Available: [sklearn.svm.SVC](https://scikit-learn.org/stable/modules/svm.html) — scikit-learn 1.1.3 documentation
- [13] “Scikit-Learn,” [Online]. Available: [sklearn.neighbors.KNeighborsClassifier](https://scikit-learn.org/stable/modules/neighbors.html) — scikit-learn 1.1.3 documentation
- [14] “Scikit-Learn,” [Online]. Available: [sklearn.naive_bayes.GaussianNB](https://scikit-learn.org/stable/modules/naive_bayes.html) — scikit-learn 1.1.3 documentation

