BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

****

ĐỀ TÀI

**XÂY DỰNG HỆ THỐNG**

**QUẢN LÝ SÂN CẦU LÔNG**

**VÀ BÁN SẢN PHẨM ONLINE**

**KẾT HỢP CHAT BOT**

**SINH VIÊN THỰC HIỆN:**

**Lê Minh Chiến**

**MSSV: B1606935**

**Khóa: 42**

**Cần Thơ, Tháng 01/2021**

BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**BỘ MÔN TIN HỌC ỨNG DỤNG**

**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

****

ĐỀ TÀI

**XÂY DỰNG HỆ THỐNG**

**QUẢN LÝ SÂN CẦU LÔNG**

**VÀ BÁN SẢN PHẨM ONLINE**

**KẾT HỢP CHAT BOT**

**GIÁO VIÊN HƯỚNG DẪN: SINH VIÊN THỰC HIỆN:**

**Th.S Nguyễn Cao Hồng Ngọc Lê Minh Chiến**

**MSSV: B1706563**

**Khóa: 43**

**Cần Thơ, Tháng 01/2021**

LỜI CẢM ƠN

Đầu tiên, em xin gửi lời cảm ơn chân thành đến gia đình và bạn bè đã động viên em trong suốt quá trình học tập.

Em xin cảm ơn Bộ môn Công nghệ thông tin, khoa Công nghệ thông tin và Truyền thông, Trường Đại học Cần Thơ đã tạo điều kiện tốt nhất để em thực hiện đề tài này.

Dưới sự hướng dẫn của ThS. Nguyễn Cao Hồng Ngọc, em đã hoàn thành đề tài “Xây dựng web quản lý sân cầu lông kết hợp chatbot”. Em xin bày tỏ lời cảm ơn sâu sắc đến cô vì đã tận tình giúp đỡ, định hướng cho em trong suốt thời gian qua.

Cảm ơn thầy dạy cầu lông của em – thầy Diệp Xuân Tài đã đưa ra đề nghị cũng như ý tưởng để em có ý tưởng thực hiện đề tài luận văn này.

Trong quá trình thực hiện, em đã tích lũy được thêm nhiều kiến thức, giúp bản thân em dần hoàn thiện hơn. Đồng thời, kết hợp với kiến thức nền tảng đã có trong thời gian học tập những năm qua, em đã hoàn thành đề tài này.

Em rất mong nhận được sự góp ý, nhận xét từ quý thầy để đề tài được hoàn thiện nhất.

Cuối lời, em xin chúc quý thầy dồi dào sức khỏe, hạnh phúc và đạt được nhiều thành công trong công việc và cuộc sống.

Trân trọng!

Cần Thơ, ngày 19 tháng 12 năm 2020

Sinh viên thực hiện

Lê Minh Chiến

**NHẬN XÉT CỦA CÁN BỘ HƯỚNG DẪN**

Cần Thơ, ngày … tháng … năm ….  
Cán bộ hướng dẫn

Mục lục

[Mục lục i](#_Toc62855773)

[Danh mục hình ảnh ii](#_Toc62855774)

[Danh mục bảng iii](#_Toc62855775)

[Danh mục các ký hiệu, chữ viết tắt iv](#_Toc62855776)

[Tóm tắt v](#_Toc62855777)

[Abstract vi](#_Toc62855778)

[PHẦN GIỚI THIỆU 1](#_Toc62855779)

[1. Đặt vấn đề 1](#_Toc62855780)

[2. Các nghiên cứu liên quan 2](#_Toc62855781)

[3. Mục tiêu đề tài 3](#_Toc62855782)

[4. Đối tượng và phạm vi nghiên cứu 3](#_Toc62855783)

[5. Phương pháp nghiên cứu 3](#_Toc62855784)

[5.1. Về lý thuyết 3](#_Toc62855785)

[5.2. Về kỹ thuật 4](#_Toc62855786)

[6. Bố cục luận văn 4](#_Toc62855787)

[PHẦN NỘI DUNG 5](#_Toc62855788)

[CHƯƠNG 1: CƠ SỞ LÝ THUYẾT 5](#_Toc62855789)

[1.1. Trí tuệ nhân tạo 5](#_Toc62855790)

[1.2. Chatbot 6](#_Toc62855791)

[1.2.1. Giới thiệu chatbot 6](#_Toc62855792)

[1.2.2. Các khái niệm trong chatbot 7](#_Toc62855793)

[1.2.3. Các phương pháp xây dựng chatbot 7](#_Toc62855794)

[1.2.4. Xử lý ngôn ngữ tự nhiên (NLP) 9](#_Toc62855795)

Danh mục hình ảnh

[Hình 1.1 Trợ lý ảo Siri tích hợp trên Iphone 2](#_Toc62855765)

[Hình 1.2 Chatbot FPT.AI trên nền tảng Facebook 2](#_Toc62855766)

[Hình 1.1 Mối liên hệ giữa AI, Machine learning và Deep learning 5](#_Toc62855767)

[Hình 1.2 Cơ chế hoạt động chung của một chatbot 6](#_Toc62855768)

[Hình 1.3 Minh họa một Artificial Neural Network 8](#_Toc62855769)

[Hình 1.4 Các ứng dụng của NLP 10](#_Toc62855770)

Danh mục bảng

Danh mục các ký hiệu, chữ viết tắt

|  |  |
| --- | --- |
| **Ký hiệu, chữ viết tắt** | **Diễn giải** |
| AI | Artificial Intelligence |
| API | Application Program Interface |
| CSDL | Cơ sở dữ liệu |
| DL | Deep Learning |
| HTTP | HTTP HyperText Transfer Protocol |
| JSON | JavaScript Object Notation |
| ML | Machine Learning |
| MVC | Model - View - Controller |
| NLP | Natural Language Processing |
| REST | REpresentational State Transfer |
| SVM | Support Vector Machine |
| URL | Uniform Resource Locator |
| XML | eXtensible Markup Language |

Tóm tắt

Nhận thấy nhu cầu của chủ sân cầu lông là cần một ứng dụng để giúp họ quản lý dữ liệu sân tốt hơn khi phải cạnh tranh với nhau trong thời buổi công nghệ. Việc xây dựng một hệ thống website cho phép họ có thể dễ dàng quản lý dữ liệu kinh doanh của mình cũng như việc kết hợp bán sản phẩm online trên hệ thống là lý do giúp em thực hiện đề tài “Xây dựng hệ thống quản lý sân cầu lông và buôn bán sản phẩm kết hợp với chatbot”.

Giao diện website được xây dựng bằng thư viện ReactJS được phát triển và cung cấp bởi Facebook dựa trên nền tảng nodeJS, một thư viện rất nổi tiếng với việc sử dụng component, render dữ liệu dựa vào state và tạo ứng dụng dưới dạng single page. Back-end của hệ thống được xây dựng bằng NodeJS kết nối đến cơ sử dữ liệu NoSQL MongoDB.

Về phần AI (Artificial intelligence) – trí tuệ nhân tạo đã được áp dụng vào nhiều lĩnh vực trong cuộc sống và nâng cao chất lượng con người. Trong hệ thống này em sẽ áp dụng công nghệ này để tạo chat bot giúp trả lời câu hỏi tự động liên quan đến hệ thống.

Kết quả đề tài luậ văn này sẽ là hệ thống website dành cho hai nhóm người dùng chính, chủ sân cầu lông và người dùng mua các sản phẩm trên hệ thống. Các chủ sân có thể dùng các chức năng chính của hệ thống như quản lý sân, quản lý hoá đơn, xem doanh thu, đăng bán sản phẩm trên hệ thống,… Đối với người dùng thông thường họ có thể dùng chatbot để giúp họ giải đáp thắc mắc, mua sắm trên website và thanh toán online. Về phần quản trị viên có thể huấn luyện mô hình chatbot nhằm mở rộng kiến thức của chatbot.

Abstract

PHẦN GIỚI THIỆU

* 1. Đặt vấn đề

Ngày nay, việc quản lý dữ liệu trong kinh doanh một cách số hoá không chỉ được các chủ doanh nghiệp lớn quan tâm mà còn có những người kinh doanh nhỏ lẻ rất quan tâm và trong đó có các chủ sân cầu lông. Họ cần một ứng dụng giúp mình quản lý các dữ liệu trong lúc kinh doanh như doanh thu, hoá đơn, hàng hoá,… Và bên cạnh đó họ cũng bán các sản phẩm liên quan tới cầu lông như quần, áo, giày, vợt,… để tăng doanh thu. Những đối tượng này quan tâm việc khách hàng sử dụng dịch vụ thuê sân hơn và họ chỉ xem việc kinh doanh các sản phẩm còn lại như một cách tăng doanh thu nên họ rất ít đầu tư tiền bạc để nhờ các công ty phát triển cho mình một phần mềm quản lý hay một trang web để buôn bán các sản phẩm mà thường đăng các sản phẩm của mình trên facebook. Vì vậy một hệ thống giúp họ có thể vừa quản lý dữ liệu của sân cầu lông của mình và cho họ một route riêng trên hệ thống để đăng tải các sản phẩm tiện cho việc kinh doanh là một nhu cầu thiết yếu.

Khái niệm chatbot được nhắc đến đầu tiên vào năm 1950, bởi Alan Turing với ý tưởng “xây dựng một chương trình máy tính giả dạng con người để thực hiện cuộc trò chuyện giữa chương trình và con người thật”. Và những năm gần đây, độ tiện lợi của chatbot ngày càng cao khi nó đã cho thấy được tiện ích của mình và xuất hiện ngày càng nhiều hơn trên các mạng xã hội, website, cho đến các thiết bị di động. Chatot có thể giúp cho con người thực hiện một hành động hoặc cung cấp tri thức cho người dùng. Chatbot có thể giap tiếp nhận mệnh lệnh từ người dùng thông qua tin nhắn hoặc giọng nói. Để xác định ý định (intent) của người dùng, chatbot sẽ dùng hệ thống xử lý ngôn ngữ tự nhiên (NLP) để phân tích và dùng các thuật toán máy học (ML) để phân loại câu (text classification) và đưa ra hành động hoặc câu trả lời phù hợp nhất.

Việc kết hợp một chatbot giúp trả lời khách hàng hiểu rõ hơn về hệ thống sẽ giúp cho họ tiếp cận hệ thống một cách thuận tiện hơn và giúp cho chủ sân đỡ phải trả lời những câu hỏi thường gặp của khách hàng nếu chúng lặp đi lặp lại. Việc lựa chọn website để xây dựng hệ thống cùng nhầm giúp cho khách hàng dễ tiếp cận hơn với hệ thống thay vì phải cài đặt các phầm mềm trên máy tính của họ cũng giúp cho họ dễ tiếp cận với hệ thống hơn. Đây cũng chính là lý do đề tài “Xây dựng hệ thống quản lý sân cầu lông và buôn bán sản phẩm kết hợp chatbot” được thực hiện.

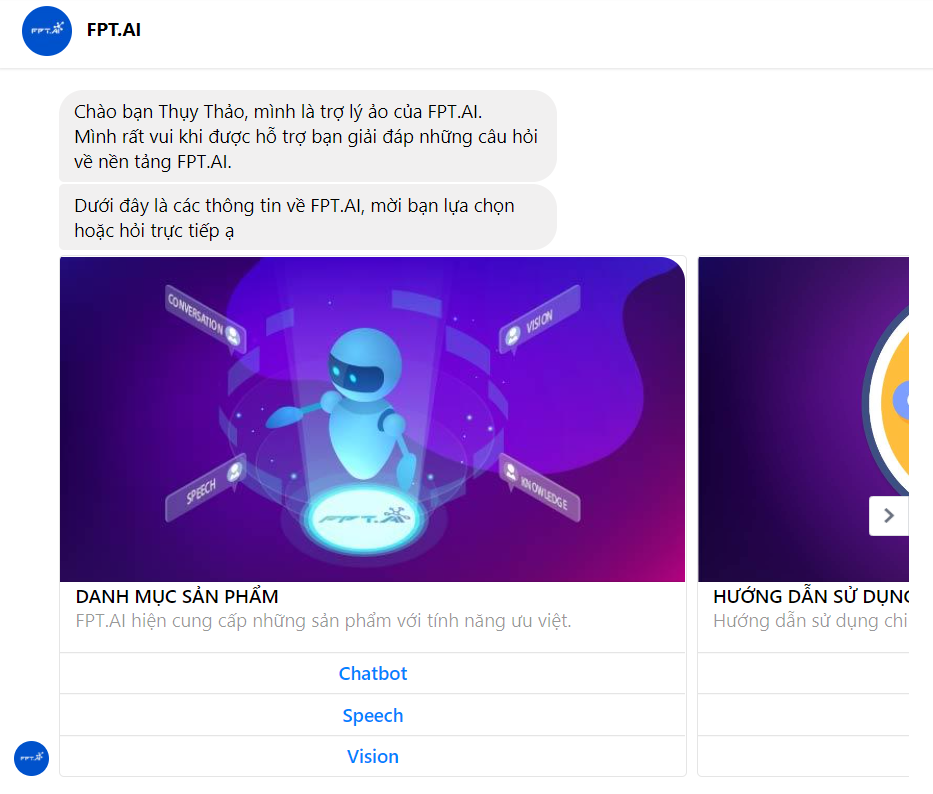
* 1. Các nghiên cứu liên quan

Trợ lý ảo là một dạng chatbot. Siri của Apple, Cortana của Windows hay Google Assistant của Google là những trợ lý ảo rất nổi tiếng được tích hợp vào thiết bị di động hay máy tính để hỗ trợ người dùng. Các trợ lý ảo có thể nhận lệnh từ giọng nói và thay người dùng thao tác như đọc thông báo, đặt lịch hẹn, báo thức, thực hiện cuộc gọi, gửi tin nhắn, …



* + - * 1. Trợ lý ảo Siri tích hợp trên Iphone

Hiện nay, trên nền tảng mạng xã hội Facebook cho phép xây dựng các chatbot. Chatbot dạng này thường đưa ra các lựa chọn cho người dùng chọn để thao tác.



* + - * 1. Chatbot FPT.AI trên nền tảng Facebook
  1. Mục tiêu đề tài

Mục tiêu của đề tài là xây dựng được một website hệ thống quản lý sân cầu lông gồm các năng: quản lý sân, quản lý sản phẩm, quản lý hoá đơn, xem các báo cáo doanh thu, mua hàng và thanh toán online cho khách hàng,.. kết hợp với một chatbot có thể trả lời các câu hỏi liên quan tới hệ thống giúp cho khách hàng tiện sử dụng trong quá trình sử dụng hệ thống. Và trang quản lý cho các chủ sân giúp họ dễ dàng quản lý dữ liệu của bản thân. Bên cạnh đó là một trang admin quản lý toàn bộ cơ sở dữ liệu của hệ thống tiện cho việc quản lý dữ liệu của người dùng cũng như huấn luyện cho chatbot.

* 1. Đối tượng và phạm vi nghiên cứu

Đề tài tập trung nghiên cứu các vấn đề:

* Tìm hiểu các chatbot, hệ thống gợi ý hiện có
* Tìm hiểu xử lý ngôn ngữ tự nhiên, thuật toán để huấn luyện mô hình, phân loại ý định người dùng trong chatbot và thiết kế quy trình đưa ra hành động thích hợp.
* Xây dựng thuật toán cho hệ thống gợi ý.
* Xây dựng website dành cho người dùng.
* Xây dựng website dành cho người quản trị.
* Tích hợp chatbot, hệ thống gợi ý vào website.
  1. Phương pháp nghiên cứu
     1. Về lý thuyết
* Tìm hiểu ngôn ngữ Python, các package, thư viện cần thiết cho phần core AI.
* Tìm hiểu các thuật toán sử dụng trong chatbot và hệ thống gợi ý.
* Tìm hiểu framework Flask cho việc xây dựng server cho service AI.
* Tìm hiểu thư viện React và các thư viện hỗ trợ khác cho việc xây dựng frond end của web.
* Tìm hiểu NodeJS và các thư viện hỗ trợ khác xây dựng server backend cho web.
* Tìm hiểu MongoDB và Mongoose cho việc lưu trữ và truy vấn dữ liệu.
* Tìm hiểu kỹ thuật lập trình một website và tích hợp với core AI.
  + 1. Về kỹ thuật
* Sử dụng ngôn ngữ Python cho việc lập trình phần core AI.
* Sử dụng MongoDB và Mongoose để quản lý cơ sở dữ liệu.
* Sử dụng React để xây dựng giao diện và NodeJS để xây dựng server.
  1. Bố cục luận văn

Bố cục quyển báo cáo luận văn gồm 3 phần, mỗi phần có những nội dung như sau:

* Phần giới thiệu: Gồm các nội dung như sau:

+ Đặt vấn đề

+ Các nghiên cứu liên quan

+ Mục tiêu đề tài

+ Đối tượng và phạm vi nghiên cứu

+ Phương pháp nghiên cứu

* Phần nội dung: Gồm các chương 1, 2, 3 với các nội dung như sau:

+ Chương 1: Cơ sở lý thuyết

+ Chương 2: Nội dung, phương pháp nghiên cứu

+ Chương 3: Đánh giá kiểm thử

* Phần kết luận: Gồm các nội dung như sau:

+ Kết luận

+ Hướng phát triển

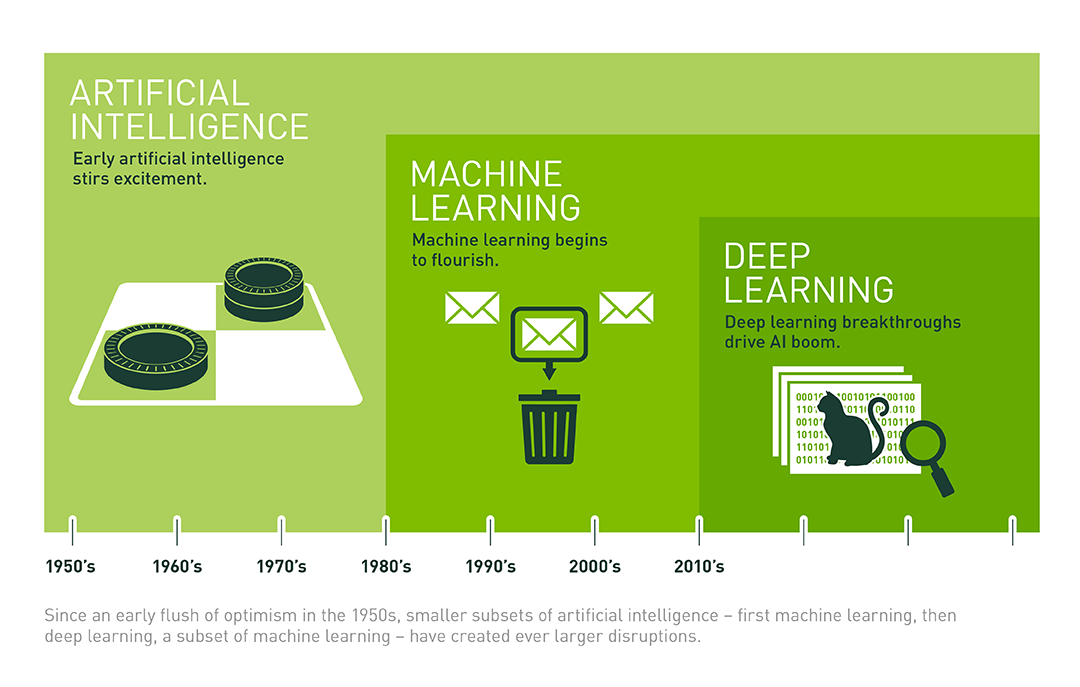
PHẦN NỘI DUNG

1. CƠ SỞ LÝ THUYẾT
   1. Trí tuệ nhân tạo

Những năm gần đây, trí tuệ nhân tạo – AI (Artificial intelligence), nổi lên như cuộc cách mạng công nghiệp lần thứ tư. AI đang len lỏi vào mọi lĩnh vực trong đời sống: xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …, chỉ là một vài trong vô vàn những ứng dụng của AI.

Máy học (Machine learning - ML) là một tập con của AI. Theo định nghĩa của Wikipedia, ***Machine learning is the subfield of computer science that “gives computers the ability to learn without being explicitly programmed”***. Nói đơn giản, ML là một lĩnh vực nhỏ của khoa học máy tính, nó có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể.

Những năm gần đây, khi mà khả năng tính toán của các máy tính được nâng lên một tầm cao mới và lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn, Machine Learning đã tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là Deep Learning - học sâu. Deep Learning đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 10 năm trước: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói và chữ viết của con người, giao tiếp với con người, hay thậm chí cả sáng tác văn hay âm nhạc. [6]



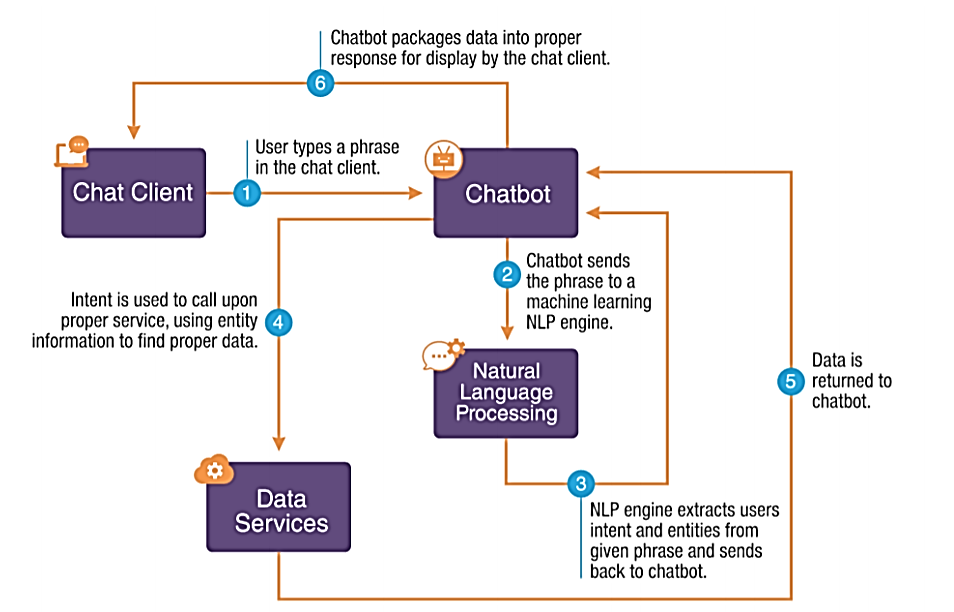
* + - * 1. Mối liên hệ giữa AI, Machine learning và Deep learning
  1. Chatbot
     1. Giới thiệu chatbot

“Bot” là một phần mềm được thiết kế để thực hiện các nhiệm vụ, tác vụ một cách tự động thay con người. Ngoài ra, “Bot” cũng là một chương trình máy tính được thiết kế để giao tiếp với người dùng thông qua Internet. “Chat” nghĩa là giao tiếp và trò chuyện giữa hai người. Vậy “Chatbot” chính là một chương trình máy tính được tạo ra để thực hiện cuộc trò chuyện một cách tự động với người dùng.

Bằng cách huấn luyện tri thức cho chatbot bằng cách sử dụng các thuật toán Machine learning hay Deep learning, chatbot có thể hiểu các câu hỏi, nhu cầu của người dùng và đưa ra các phản hồi sao cho phù hợp. Trong quá trình vận hành, sẽ có những câu hỏi của người dùng nhưng chabot phản hồi sai hoặc không biết. Chatbot sẽ ghi nhận lại và người quản lý hệ thống có thể thu thập các dữ liệu đó và huấn luyện lại cho chatbot.

Chatbot có hai dạng phổ biến là text (tin nhắn bằng văn bản) và voice (giọng nói). Ví dụ, Alexa của Amazon, Siri của Apple, Cortana của Microsoft là các dạng chatbot voice. Ngoài ra chatbot text cũng xuất hiện trên nhiều trang web và các nền tảng mạng xã hội. [22]

Có hai dạng chatbot chính. Một là mô hình ứng dụng trong miền đóng (closed domain), trả lời theo mô hình truy xuất thông tin (retrieval-based model), trong đó bot đưa ra câu trả lời theo những kịch bản hoặc câu trả lời đã được chuẩn bị sẵn. Hai là mô hình ứng dụng trong miền mở (open domain), người dùng có thể thực hiện cuộc trò chuyện với bot mọi nơi và không giới hạn chủ đề. [3]



* + - * 1. Cơ chế hoạt động chung của một chatbot
    1. Các khái niệm trong chatbot

Các khái niệm chính được sử dụng trong chatbot gồm:

* Intent: Là ý định của người dùng trong câu nói.

Ví dụ: Với câu “Tôi muốn tra cứu thông tin phim Toy Story” từ người dùng, thì ý định của họ ở đây là mong muốn biết các thông tin về bộ phim Toy Story như tên đầy đủ, đạo diễn, thể loại, năm sản xuất, …

* Entity: Là các thực thể bổ nghĩa cho một ý định

Ví dụ: Với câu “Tôi muốn tra cứu thông tin phim Toy Story” từ người dùng, thì thực thể ở đây là “Toy Story”, là tên phim.

* Stop words: Là những từ xuất hiện nhiều trong văn bản nhưng không có ý nghĩa quan trọng trong việc phân lớp
* Negative words: Là các từ mang sắc thái tiêu cực
* Positive words: Là các từ mang sắc thái tích cực
  + 1. Các phương pháp xây dựng chatbot

Chatbot được xây dựng dựa trên các phương pháp sau:

* **Pattern matchers**: Chatbot xây dựng dựa trên phương pháp pattern matchers sẽ tìm kiếm trong câu hỏi của người dùng có khớp với các câu hỏi đã được huấn luyện hay không. Phương pháp pattern matchers dựa theo cấu trúc tiêu chuẩn của ngôn ngữ lập trình “Artificial Intelligence Markup Language” (AIML).

<category>

<pattern>Chatbot là gì</pattern>

<template>Chatbot là là một chương trình máy tính được tạo ra để thực hiện cuộc trò chuyện một cách tự động với người dùng</template>

</category>

<category>

<pattern>Thế nào là chatbot</pattern>

<template>

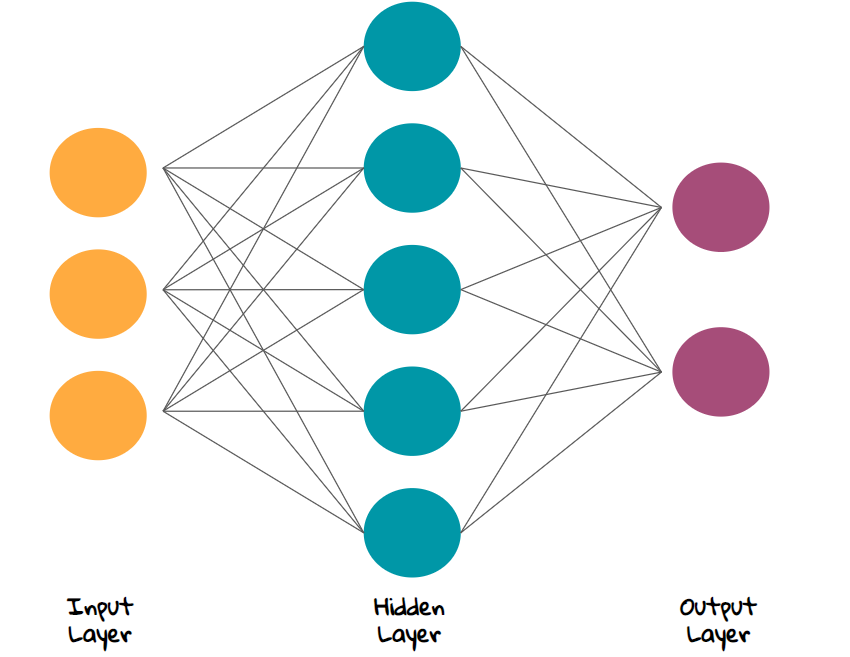
<srai>Chatbot là gì</srai>

</template>

</category>

Nếu người dùng hỏi “Chabot là gì” hoặc “Thế nào là chatbot” thì sẽ nhận được câu trả lời là “Chatbot là là một chương trình máy tính được tạo ra để thực hiện cuộc trò chuyện một cách tự động với người dùng”.

* **Sử dụng thuật toán:** Trong trường hợp có rất nhiều câu hỏi riêng biệt với ý định khác nhau đến từ khách hàng, mà chatbot chỉ có thể phản hồi khi tất cả câu hỏi này đã được thiết lập, huấn luyện cho chatbot với các kịch bản đối thoại đều được lưu trữ trong cơ sở dữ liệu. Do đó khi người dùng hỏi những câu hỏi phức tạp hơn thì cần một phương pháp khác để nhận dạng câu. Vì vậy các nhà phát triển đã đưa các thuật toán vào chatbot để phân loại câu, từ đó tăng độ chính xác cho các phản hồi đưa đến cho người dùng. Các thuật toán có thể sử dụng như SVM, KNN, Naïve Bayes, …
* **Artificial Neural Network (ANN – mạng nơ ron nhân tạo):** Phương pháp này là tính toán tỷ lệ chính xác dữ liệu đầu ra từ dữ liệu đầu vào bằng cách tính toán các trọng số cho mỗi kết nối (connection) từ các lần lặp lại trong khi huấn luyện dữ liệu cho Chatbot. Mỗi bước huấn luyện dữ liệu cho Chatbot sẽ sửa đổi các trọng số dẫn đến dữ liệu đầu ra được xuất ra với độ chính xác cao. Mỗi câu sẽ được phân thành từng từ và mỗi từ chính là dữ liệu đầu vào input đưa vào ANN. Các từ sẽ được kết nối với nhau và khi huấn luyện dữ liệu cho Chatbot, các kết nối này sẽ được đo bằng trọng số cho hàng ngàn lần lặp lại ở các thời điểm khác nhau. Mỗi lần lặp lại trọng số có thể tăng và độ chính xác sẽ cao. [23]



* + - * 1. Minh họa một Artificial Neural Network
    1. Xử lý ngôn ngữ tự nhiên (NLP)

Xử lý ngôn ngữ tự nhiên (NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ-công cụ hoàn hảo nhất của tư duy và giao tiếp.

Trong NLP có hai quan điểm cơ bản:

* Xử lý các từ ngữ bằng máy tính.
* Làm cho máy tính hiểu được các từ ngữ.
  + - 1. Các bước xử lý trong NLP
* **Phân tích hình thái**: Trong bước này từng từ sẽ được phân tích và các ký tự không phải chữ (như các dấu câu) sẽ được tách ra khỏi các từ. Trong tiếng Anh và nhiều ngôn ngữ khác, các từ được phân tách với nhau bằng dấu cách. Tuy nhiên trong tiếng Việt, dấu cách được dùng để phân tách các tiếng (âm tiết) chứ không phải từ. Cùng với các ngôn ngữ như tiếng Trung, tiếng Hàn,tiếng Nhật, phân tách từ trong tiếng Việt là một công việc không hề đơn giản.
* **Phân tích cú pháp:** Dãy các từ sẽ được biến đổi thành các cấu trúc thể hiện sự liên kết giữa các từ này. Sẽ có những dãy từ bị loại do vi phạm các luật văn phạm.
* **Phân tích ngữ nghĩa:** Thêm ngữ nghĩa vào các cấu trúc được tạo ra bởi bộ phân tích cú pháp.
* **Tích hợp văn bản:** Ngữ nghĩa của một câu riêng biệt có thể phụ thuộc vào những câu đứng trước, đồng thời nó cũng có thể ảnh hưởng đến các câu phía sau.
* **Phân tích thực nghĩa:** Cấu trúc thể hiện điều được phát ngôn sẽ được thông dịch lại để xác định nó thật sự có nghĩa là gì.

Tuy nhiên, ranh giới giữa 5 bước xử lý này cũng rất mong manh. Chúng có thể được tiến hành từng bước một, hoặc tiến hành cùng lúc - tùy thuộc vào giải thuật và ngữ cảnh cụ thể.

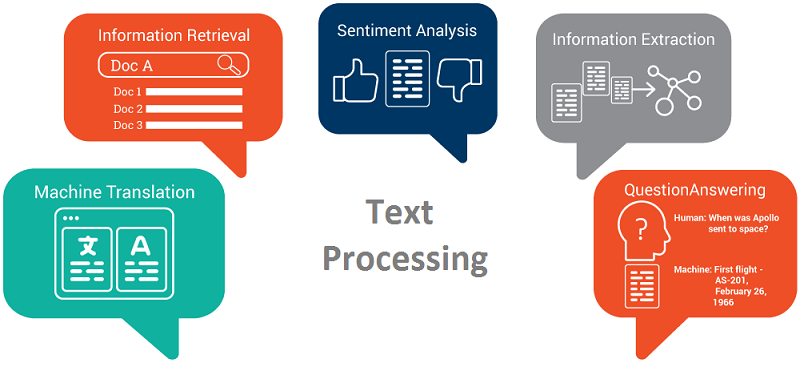
* + - 1. Các ứng dụng cơ bản của NLP và các bài toán trong xử lý tiếng Việt

NLP được ứng dụng vào thực tế như sau:

* Nhận dạng chữ viết
* Nhận dạng tiếng nói
* Tổng hợp tiếng nói
* Dịch tự động
* Tìm kiếm thông tin
* Tóm tắt văn bản
* Khai phá dữ liệu và phát hiện tri thức

Trong tiếng việt, NLP được ứng dụng vào các bài toán như:

* Phân tách câu
* Phân tách từ
* Tự động thêm dấu [24]



* + - * 1. Các ứng dụng của NLP
      1. Phương pháp Bag-of-word

**Bag of Words** (BoW – mô hình túi từ) là một thuật toán hỗ trợ xử lý ngôn ngữ tự nhiên và mục đích của BoW là phân loại câu hay văn bản. Ý tưởng của BoW là phân tích và phân nhóm dựa theo "Bag of Words"(corpus).

Xét hai ví dụ sau:

1. Tôi thích xem phim. Em tôi cũng vậy.
2. Bạn tôi thì thích xem kịch.

Dựa vào hai văn bản trên, xây dựng được danh sách như sau:

1. “tôi”, “thích”, “xem”, “phim”, “em”, “tôi”, “cũng”, “vậy”
2. “bạn”, “tôi”, “thì”, “thích”, “xem”, “kịch”

Dựa vào hai danh sách trên, ta xây dựng được một từ điển gồm 10 từ như sau:

[ “tôi”, “thích”, “xem”, “phim”, “em”, “cũng”, “vậy”, “bạn”, “thì”, “kịch”]

Mỗi văn bản sẽ tạo ra một vector đặc trưng có số chiều bằng 10, mỗi phần tử đại diện cho số từ tương ứng xuất hiện trong văn bản đó. Hai văn bản trên sẽ có hai vector đặc trưng là:

1. [2, 1, 1, 1, 1, 1, 1, 0, 0, 0]
2. [1, 1, 1, 0, 0, 0, 0, 1, 1, 1]

Tuy nhiên, vector đại diện này không bao gồm thứ tự của các từ trong câu gốc. [6]

* + - 1. Phương pháp TF-IDF

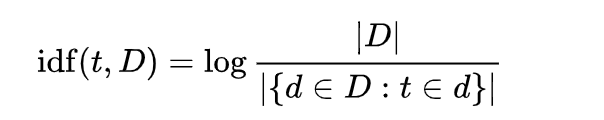
**TF-IDF** (Term Frequency – Inverse Document Frequency) là 1 kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. Một vài biến thể của tf-idf thường được sử dụng trong các hệ thống tìm kiếm như một công cụ chính để đánh giá và sắp xếp văn bản dựa vào truy vấn của người dùng. Tf-idf cũng được sử dụng để lọc những từ stopwords trong các bài toán như tóm tắt văn bản và phân loại văn bản

**TF**: Term Frequency (Tần suất xuất hiện của từ) là số lần từ xuất hiện trong văn bản. Vì các văn bản có thể có độ dài ngắn khác nhau nên một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn là một văn bản ngắn. Như vậy, term frequency thường được chia cho độ dài văn bản (tổng số từ trong một văn bản).

Trong đó:

* tf (t, d): tần suất xuất hiện của từ t trong văn bản
* f (t, d): Số lần xuất hiện của từ t trong văn bản d
* max ({f(w, d) : w ∈ d}): Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d

**IDF**: Inverse Document Frequency (Nghịch đảo tần suất của văn bản), giúp đánh giá tầm quan trọng của một từ. Khi tính toán TF, tất cả các từ được coi như có độ quan trọng bằng nhau. Nhưng một số từ như “là”, “hoặc”, “và” thường xuất hiện rất nhiều lần nhưng độ quan trọng là không cao. Như thế cần giảm độ quan trọng của những từ này xuống.



Trong đó:

* idf (t, D): giá trị idf của từ t trong tập văn bản
* |D|: Tổng số văn bản trong tập D
* |{d ∈ D : t ∈ d}|: thể hiện số văn bản trong tập D có chứa từ t.

Cơ số logarit trong công thức này không thay đổi giá trị idf của từ mà chỉ thu hẹp khoảng giá trị của từ đó. Vì thay đổi cơ số sẽ dẫn đến việc giá trị của các từ thay đổi bởi một số nhất định và tỷ lệ giữa các trọng lượng với nhau sẽ không thay đổi. (nói cách khác, thay đổi cơ số sẽ không ảnh hưởng đến tỷ lệ giữa các giá trị IDF). Việc sử dụng logarit nhằm giúp giá trị tf-idf của một từ nhỏ hơn, do công thức tính tf-idf của một từ trong 1 văn bản là tích của tf và idf của từ đó. [6]

Xét với ví dụ nêu trên, sau khi thực hiện vector hóa bằng BoW, ta có kết quả TF-IDF như sau:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| tôi | thích | xem | phim | em | cũng | vậy | bạn | thì | kịch |
| 0.00 | 0.00 | 0.00 | 0.09 | 0.09 | 0.09 | 0.09 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.12 | 0.12 |