**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI: NGHIÊN CỨU PHÁT TRIỂN ỨNG DỤNG VIẾT NHẬT KÝ THEO PHONG CÁCH BULLET JOURNAL**

Giảng viên hướng dẫn: NGUYỄN QUANG PHÚC

Sinh viên thực hiện: VĂN THỊ NGÂN HÀ

Lớp: CÔNG NGHỆ THÔNG TIN

Khoá:58

TP. Hồ Chí Minh, tháng 06 năm 2021

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI: NGHIÊN CỨU PHÁT TRIỂN ỨNG DỤNG VIẾT NHẬT KÝ THEO PHONG CÁCH BULLET JOURNAL**

Giảng viên hướng dẫn: NGUYỄN QUANG PHÚC

Sinh viên thực hiện: VĂN THỊ NGÂN HÀ

Lớp: CÔNG NGHỆ THÔNG TIN

Khoá:58

TP. Hồ Chí Minh, tháng 06 năm 20201

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI **CỘNG HÒA XÃ HỘI CHỦ NGHIÃ VIỆT NAM**

**PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH** Độc lập – Tự do – Hạnh phúc

# NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP

BỘ MÔN: **CÔNG NGHỆ THÔNG TIN**

-------\*\*\*-------

**Mã sinh viên:** 5851071021 **Họ tên SV:** Văn Thị Ngân Hà

**Khóa:** 58 **Lớp:** Công Nghệ Thông Tin

1. **Tên đề tài.**

***NGHIÊN CỨU PHÁT TRIỂN ỨNG DỤNG VIẾT NHẬT KÝ THEO PHONG CÁCH BULLET JOURNAL.***

1. **Mục đích, yêu cầu.**
   1. **Mục đích.**

* Xây dựng ứng dụng hỗ trợ người dùng như một người bạn đồng hành trong cuộc sống với các chức năng nổi bật: viết nhật ký, viết ghi chú, ghi lại hành trình và hoạt động hằng ngày.

- Ứng dụng hỗ trợ người dùng viết nhật kí, có thể tùy chỉnh thêm hình ảnh, văn bản, biểu tượng cảm xúc và địa chỉ tạo cảm hứng cho người sử dụng, thỏa sức sáng tạo trong những trang nhật kí của riêng mình.

- Xây dựng chức năng ghi chú, giúp người dùng thuận tiện trong việc ghi lại những việc làm cần thiết. Bên cạnh đó người sử dụng có thể tùy chọn thêm hình ảnh và ghi âm.

- Xây dựng chức năng ghi lại hành trình chuyến đi hỗ trợ thêm hình ảnh giúp hành trình của bạn trở nên sống động hơn.

‘- Xây dựng chức năng ghi lại các hoạt động hàng ngày giúp quản lí công việc một cách tốt hơn.’

**Yêu cầu.**

* Tìm hiểu về ionic framework.
* Tìm hiểu về ngôn ngữ Dart và Flutter để xây dựng ứng dụng.
* Tìm hiểu về cách viết Bullet Journal từ đó xây dựng nên các chức năng.

- Thu thập hình ảnh, thông tin từ các hội nhóm chuyên viết Bullet Journal để hoàn thiện chức năng và phù hợp với nhu cầu thực tế.

* Tìm hiểu về cách thêm hình ảnh, văn bản và âm thanh, sử dụng thư viện lấy dữ liệu từ Google Map API nhận về địa chỉ hiện tại của người dùng.
* Nghiên cứu cách di chuyển các widget trong màn hình và lưu lại vị trí sau khi di chuyển.
* Nghiên cứu cách thêm bớt tọa độ hỗ trợ người dùng tùy chỉnh cuộn lên hoặc cuộn xuống cho phép người dùng có nhiều không gian hơn, thoải mái trong việc sáng tạo trang nhật kí của mình.

1. **Nội dung và phạm vi đề tài.**
   1. **Nội dung đề tài.**

* Giới thiệu về phong cách viết Bullet Journal.
* Nghiên cứu, phân tích phương pháp viết Bullet Journal từ đó phát triển các chức năng cho ứng dụng.
* Nghiên cứu cách xử lí hình ảnh và văn bản cho phép người dùng tùy chỉnh bố cục bài viết theo sở thích của mỗi người.
* Nghiên cứu ........
* Hoàn thiện sản phẩm, đáp ứng các chức năng đã khảo sát.
  1. **Phạm vi đề tài.**

- Nguyên cứu các widget cho phép người dùng di chuyển, tùy chỉnh và xóa.

- Ứng dụng kiến thức đã tìm hiểu để xây dựng ứng dụng viết nhật ký theo phong cách Bullet Journal.

1. **Công nghệ, công cụ và ngôn ngữ lập trình.**
   1. **Công nghệ:** Ionic framework, Google Map API
   2. **Công cụ:** Visual Code, Android Studio
   3. **Ngôn ngữ lập trình:** Dart
2. **Các kết quả chính dự kiến sẽ đạt được và ứng dụng**

* Hoàn chỉnh cuốn báo cáo đề tài.
* Nắm rõ về cấu trúc ngôn ngữ Dart, cách xây dựng một ứng dụng hoàn chỉnh bằng Flutter và thành thạo các thư viện trong Flutter, tiêu biểu là thư viện xử lí về hình ảnh.
* Nắm được qui trình xây dựng một ứng dụng di động từ giai đoạn phân tích yêu cầu đến khi ứng dụng hoàn thành.
* Nắm được các áp dụng thư viện vào dự án, bên cạch đó tùy chỉnh thư viện để phù hợp với các chức năng và giao diện của ứng dụng.
* Sử dụng Flutter để xây dựng được ứng dụng viết nhật ký theo phong cách Bullet Journal.

1. **Giáo viên và cán bộ hướng dẫn**

Họ tên: NGUYỄN QUANG PHÚC

Đơn vị công tác: Trường Đại học Công nghệ Thông tin ĐHQG TP.HCM

Điện thoại: 0906761014 Email: tpnha@utc2.edu.vn

|  |  |
| --- | --- |
| **Ngày tháng 5 năm 2020**  **BM Công Nghệ Thông Tin** | **Đã giao nhiệm vụ TKTN**  **Giáo viên hướng dẫn** |
|  | **Nguyễn Quang Phúc** |

Đã nhận nhiệm vụ TKTN

Sinh viên: Văn Thị Ngân Hà Ký tên:

Điện thoại: 0775881084

Email: nganha2169@gmail.com

**LỜI CẢM ƠN**

Lời nói đầu tiên, em xin gửi lời cảm ơn tới Quý thầy cô trong Bộ môn Công Nghệ Thông Tin và Ban Giám Hiệu trường Đại học Giao thông Vận tải phân hiệu tại Thành phố Hồ Chí Minh, đã tạo điều kiện giúp em được nhận và thực hiện đồ án của mình.

Có rất nhiều yếu tố để em có thể hoàn thành đồ án. Yếu tố đầu tiên chính là sự cố gắng tìm tòi, học hỏi của bản thân. Tiếp theo, em xin gửi lời cảm ơn chân thành nhất đến thầy **Nguyễn Quang Phúc** – người trực tiếp hướng dẫn em, thầy luôn tận tình hướng dẫn, định hướng, giải đáp thắc mắc và giải quyết những khó khăn trong quá trình thực hiện đồ án. Sự hướng dẫn của thầy chính là yếu tố quan trọng nhất giúp em hoàn thành đề tài của mình. Cuối cùng, em xin một lần nữa gửi lời cảm ơn đến Quý thầy cô trong Bộ môn Công Nghệ Thông Tin đã dạy em những kiến thức nền tảng để em có đủ khả năng hoàn thành đồ án.

Mặc dù đã nổ lực hết sức để hoàn thành đề tài, nhưng với năng lực và thời gian có hạn, chắc chắn sẽ khó tránh khỏi những sai sót. Em rất mong nhận được những sự đánh giá, góp ý khách quan nhất của Quý thầy cô, giúp em rút ra cho mình những bài học, kinh nghiệm quý báu sau này.

Cuối lời, em cũng không biết nói gì hơn ngoài kính chúc Quý thầy cô trong Bộ môn Công Nghệ Thông Tin và đặc biệt là thầy **Nguyễn Quang Phúc** thật dồi dào sức khỏe và ngày càng gặt hái được nhiều thành công hơn nữa trong cuộc sống cũng như trong sự nghiệp giảng dạy của mình.

Em xin chân thành cảm ơn !

***TP. Hồ Chí Minh, ngày tháng năm 2020***

**Sinh viên thực hiện**

**Văn Thị Ngân Hà**

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

|  |
| --- |
| ***TP. Hồ Chí Minh, ngày tháng năm 2020***  **Giáo viên hướng dẫn**  **Nguyễn Quang Phúc** |

**MỤC LỤC**

[NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP iii](#_Toc75397038)

[LỜI CẢM ƠN vi](#_Toc75397039)

[NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN vii](#_Toc75397040)

[MỤC LỤC viii](#_Toc75397041)

[DANH MỤC THUẬT NGỮ xi](#_Toc75397042)

[DANH MỤC BẢNG xii](#_Toc75397043)

[DANH MỤC HÌNH ẢNH xii](#_Toc75397044)

[TỔNG QUAN 1](#_Toc75397045)

[Đặt vấn đề 1](#_Toc75397046)

[Tình hình nguyên cứu 3](#_Toc75397047)

[Quá trình nguyên cứu 3](#_Toc75397048)

[Cấu trúc báo cáo đồ án tốt nghiệp 5](#_Toc75397049)

[CHƯƠNG 1: CƠ SỞ LÝ THUYẾT 6](#_Toc75397050)

[1.1 Công nghệ sử dụng. 6](#_Toc75397051)

[1.1.1 Giới thiệu về Machine Learning 6](#_Toc75397052)

[1.1.2 Phân nhóm các thuật toán Machine Learning 9](#_Toc75397053)

[1.1.3 Các bước thực hiện Machine Learning 12](#_Toc75397054)

[1.1.4 Các ứng dụng 14](#_Toc75397055)

[1.2 Ngôn ngữ lập trình. 16](#_Toc75397056)

[1.2.1 Khái niệm 16](#_Toc75397057)

[1.2.2 Những khó khăn trong lĩnh vực xử lý ngôn ngữ tự nhiên 16](#_Toc75397058)

[1.2.3 Một số ứng dụng của xử lý ngôn ngữ tự nhiên 18](#_Toc75397059)

[1.3 Thuật toán áp dụng 19](#_Toc75397060)

[1.3.1 Giới thiệu về Odoo 19](#_Toc75397061)

[1.3.2 Các ưu điểm của Odoo 19](#_Toc75397062)

[1.3.3 Cơ sở dữ dữ liệu PostgreSQL 20](#_Toc75397063)

[CHƯƠNG 2: PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG 20](#_Toc75397064)

[2.1 Biểu đồ usecase của hệ thống 20](#_Toc75397065)

[2.1.1 Tách từ 21](#_Toc75397066)

[2.1.2 Loại bỏ stop word 24](#_Toc75397067)

[2.2 Sơ đồ phân rã chức năng. 24](#_Toc75397068)

[2.2.1 Biểu đồ tổng quát. 24](#_Toc75397069)

[2.2.2 Chức năng quản lí nhật ký. 24](#_Toc75397070)

[2.2.3 Chức năng quản lí ghi chú. 24](#_Toc75397071)

[2.2.4 Chức năng quản lí hành trình. 24](#_Toc75397072)

[2.2.5 Chức năng quản lí công việc. 24](#_Toc75397073)

[2.3 Sơ đồ hoạt động. 24](#_Toc75397074)

[2.2.1 Binary vector 25](#_Toc75397075)

[2.2.2 TF-IDF vector 25](#_Toc75397076)

[CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH 27](#_Toc75397077)

[3.1 Hiện trạng tổ chức 27](#_Toc75397078)

[3.2 Yêu cầu hệ thống 27](#_Toc75397079)

[3.2.1 Yêu cầu chức năng 27](#_Toc75397080)

[3.2.2 Yêu cầu phi chức năng 27](#_Toc75397081)

[3.3 Thu thập dữ liệu 28](#_Toc75397082)

[3.4 Thuật toán áp dụng 30](#_Toc75397083)

[3.4.1 Bài toán phân loại câu hỏi và đề xuất câu hỏi tương tự 30](#_Toc75397084)

[3.3.2 Bài toán dự đoán điểm 35](#_Toc75397085)

[3.7 Xây dựng giao diện chương trình 37](#_Toc75397086)

[3.7.1 Tạo tài khoản và đăng nhập 37](#_Toc75397087)

[3.7.2 Giao diện trang chủ website 38](#_Toc75397088)

[3.7.3 Tương tác bài viết 38](#_Toc75397089)

[3.7.4 Chat trực tuyến 40](#_Toc75397090)

[3.7.5 Dự đoán điểm 41](#_Toc75397091)

[3.7.6 Nhóm chức năng quản lý 42](#_Toc75397092)

[3.7.7 Quản lý người dùng 43](#_Toc75397093)

[3.7.8 Quản lý chat trực tuyến 43](#_Toc75397094)

[3.6.9 Quản lý dự đoán điểm 44](#_Toc75397095)

[KẾT LUẬN 45](#_Toc75397096)

[Kết quả đạt được 45](#_Toc75397097)

[Nhược điểm 45](#_Toc75397098)

[Hướng phát triển 45](#_Toc75397099)

[TÀI LIỆU THAM KHẢO 46](#_Toc75397100)

# DANH MỤC THUẬT NGỮ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **THUẬT NGỮ** | **Ý NGHĨA TIẾNG VIỆT** | **TỪ VIẾT TẮT** | **GHI CHÚ** |
| 1 | Information Technology | Công Nghệ Thông Tin | CNTT |  |
| 2 | Ionic framework |  |  |  |
| 3 | Artificial Intelligence | Trí tuệ nhân tạo | AI |  |
| 4 | Machine Learning | Học máy | ML |  |
| 5 | Combinatory categorial grammar | Văn phạm danh mục kết nối | CCG |  |
| 6 | Knowledge Discovery in Database | Khai phá tri thức | KDD |  |
| 7 | Natural Language Understanding | Hiểu ngôn ngữ tự nhiên | NLU |  |
| 8 | Database | Cơ sở dữ liệu | CSDL |  |
| 9 | Data Mining | Khai phá dữ liệu |  |  |
| 10 | Natural Language Generation | Sinh ngôn ngữ tự nhiên | NLG |  |
| 11 | Association rules | Luật kết hợp |  |  |
| 12 | Classification | Phân lớp |  |  |
| 13 | Clustering | Phân cụm |  |  |
| 14 | Regression | Hồi quy |  |  |
| 15 | Knowledge Discovery from Data | Khai thác tri thức từ cơ sở dữ liệu | KDD |  |
| 16 | Natural Language Processing | Xử lý ngôn ngữ tự nhiên | NLP |  |
| 17 | Support Vector Machine | Máy vectơ hỗ trợ | SVM |  |
| 18 | Linear Least Square Fit |  | LLSF |  |
| 19 | K- Nearest Neighbor |  | KNN |  |
| 20 | Naive Bayes |  | NB |  |
| 21 | Classification and regression trees | Cây quyết định phân lớp | CART |  |
| 22 | Singular Value Decomposition |  | SVD |  |
| 23 | eXtensible Markup Language |  | XML |  |
| 24 | Natural Language Understanding |  | NLU |  |
| 25 | Context-free grammar | Văn phạm phi ngữ cảnh | CFG |  |
| 26 | Dependency grammar | Văn phạm phụ thuộc | DG |  |

**DANH MỤC BẢNG**

[Bảng 1.1 Các phòng ban/trung tâm 4](#_Toc48917793)

[Bảng 2.1 Dữ liệu tên và giới tính 34](#_Toc48917794)

[Bảng 2.2 Kết quả phân loại văn bản với máy học véc-tơ hỗ trợ và cây quyết định 41](#_Toc48917795)

[Bảng 2.3 Kết quả phân loại tài liệu Tiếng Việt bằng thuật toán Naive Bayes 42](#_Toc48917796)

[Bảng 2.4 Kết quả phân loại email spam bằng matlab 44](#_Toc48917797)

[Bảng 3.1 Bảng thống kê số câu train và câu test của các ban ngành/phòng ban 53](#_Toc48917798)

**DANH MỤC HÌNH ẢNH**

[Hình 1.1 Mối quan hệ giữa AI, Machine Learning và Deep Learning. 7](#_Toc48917837)

[Hình 1.2 Mối quan hệ giữa AI, Machine Learning và Deep Learning. 8](#_Toc48917838)

[Hình 1.3 Lựa chọn thuật toán phù hợp trong Machine Learning. 14](#_Toc48917839)

[Hình 2.1 Siêu phẳng với lề cực đại cho một SVM phân tách dữ liệu thuộc hai lớp 31](#_Toc48917840)

[Hình 2.2 Cây quyết định phân lớp mức lương 35](#_Toc48917841)

[Hình 2.3 Quy trình phân loại văn bản với máy học véc-tơ hỗ trợ và cây quyết định 40](#_Toc48917842)

[Hình 2.4 Quy trình phân loại tài liệu Tiếng Việt 41](#_Toc48917843)

[Hình 2.5 Trình tự phân loại tài liệu Tiếng Việt 42](#_Toc48917844)

[Hình 3.1 Dự liệu sau khi lấy trên diễn đàn 48](#_Toc48917845)

[Hình 3.2 Dữ liệu sau khi được chuẩn hóa 48](#_Toc48917846)

[Hình 3.3 Form phân loại câu hỏi sinh viên 49](#_Toc48917847)

[Hình 3.4 Quy trình hóa phân loại văn bản 50](#_Toc48917848)

[Hình 3.5 Các tag phân loại trên diễn đàn 50](#_Toc48917849)

[Hình 3.6 Quá trình tạo mô hình véc- tơ và ma trận câu hỏi 51](#_Toc48917850)

[Hình 3.7 Quá trình chọn câu hỏi-câu trả lời tương tự trên diễn đàn 51](#_Toc48917851)

[Hình 3.8 Quá trình phân tag cho câu hỏi 52](#_Toc48917852)

[Hình 3.9 Tổ chức tập dữ liệu train và test 53](#_Toc48917853)

[Hình 3.10 Kết quả sau khi train tập dữ liệu trên 54](#_Toc48917854)

[Hình 3.11 Công thức dự đoán điểm 55](#_Toc48917855)

[Hình 3.12 Quản lý công thức dự đoán điểm 55](#_Toc48917856)

[Hình 3.13 Sơ đồ usecase tổng quan của hệ thống 56](#_Toc48917857)

[Hình 3.14 Sơ đồ hoạt động đăng nhập 57](#_Toc48917858)

[Hình 3.15 Sơ đồ hoạt động quản lý dự đoán điểm 57](#_Toc48917859)

[Hình 3.16 Sơ đồ hoạt động dự đoán điểm 58](#_Toc48917860)

[Hình 3.17 Sơ đồ hoạt động quản lý người dùng 58](#_Toc48917861)

[Hình 3.18 Sơ đồ hoạt động dự đoán điểm 59](#_Toc48917862)

[Hình 3.19 Sơ đồ hoạt động hỗ trợ trực tuyến 59](#_Toc48917863)

[Hình 3.20 Sơ đồ hoạt động quản lý diễn đàn 60](#_Toc48917864)

[Hình 3.21 Sơ đồ hoạt động quản lý báo cáo thống kê 60](#_Toc48917865)

[Hình 3.22 Giao diện tạo tài khoản đăng nhập 61](#_Toc48917866)

[Hình 3.23 Giao diện đăng nhập 61](#_Toc48917867)

[Hình 3.24 Giao diện trang chủ của website 62](#_Toc48917868)

[Hình 3.25 Giao diện đặt câu hỏi trên diễn đàn 62](#_Toc48917869)

[Hình 3.26 Giao diện trả lời câu hỏi 63](#_Toc48917870)

[Hình 3.27 Các chức năng thao tác với câu hỏi 63](#_Toc48917871)

[Hình 3.28 Giao diện chức năng tìm kiếm câu hỏi 64](#_Toc48917872)

[Hình 3.29 Giao diện chức năng chat trực tuyến 64](#_Toc48917873)

[Hình 3.30 Giao diện người dùng sử dụng chức năng chat trực tuyến 65](#_Toc48917874)

[Hình 3.31 Giao diện chức năng dự đoán điểm 65](#_Toc48917875)

[Hình 3.32 Giao diện bảng điểm sau khi người dùng chọn chức năng dự đoán điểm 66](#_Toc48917876)

[Hình 3.33 Giao diện nhóm chức năng quản lý 66](#_Toc48917877)

[Hình 3.34 Giao diện quản lý người dùng 67](#_Toc48917878)

[Hình 3.35 Giao diện hiện thị người dùng truy cập trang web 67](#_Toc48917879)

[Hình 3.36 Giao diện hiện thị người dùng truy cập trang web 67](#_Toc48917880)

[Hình 3.37 Giao diện theo dõi lịch sử chat 68](#_Toc48917881)

[Hình 3.38 Giao diện quản lý dự đoán điểm 68](#_Toc48917882)

# TỔNG QUAN

## **Đặt vấn đề**

Trong cuộc sống hiện đại và tấp nập, mỗi người trong chúng ta ai cũng có một mối bận tâm cho riêng mình. Phải chăng là nổi lo về cơm áo gạo tiền của các bậc cha mẹ, cũng có thể là những băn khoăn của các em học sinh cuối cấp sắp phải rời xa mái trường thân thuộc và đối mặt với kì thi quan trọng quyết định cuộc đời của các em,...Và đặc biệt nổi lo chung của toàn dân Việt Nam hiện nay chính là sự hoàn hành của Covid-19, dịch bệnh đã làm thay đổi đi cách sống của mỗi người. Bỏ qua những ảnh hưởng tiêu cực từ Covid-19, thời gian này cho phép người ta có thêm thời gian gần gũi với gia đình và suy ngẫm về những việc đã làm cũng như những dự định trong tương lai. Nhưng cho mình hỏi, đã có bao giờ bạn chia sẽ hết những nổi lo trong lòng cho một ai đó? Theo kinh nghiệm sống 20 năm ít ỏi của mình cùng với những kiến thức mình đã đút kết trong sách vở thì sự thật là bạn không thể chia sẽ hết những điều mình suy nghĩ cho người khác. Có thể là vì thời điểm đó chưa phù hợp hoặc người bạn muốn chia sẻ không hiểu hoàn cảnh của bạn hay một điều gì đó rất nghiêm trọng và bạn không muốn cho người thân biết. Từ những lý do trên, mình đã nghĩ ra ý tưởng xây dựng lên một ứng dụng như một người bạn đồng hành có thể chia sẽ những kỉ niệm buồn vui cùng bạn và những bí mật mà bạn chỉ muốn giữ cho riêng mình. *‘Ứng dụng sẽ giúp bạn không còn cảm thấy cô đơn khi phải đối diện nổi lo một mình và bên cạnh đó nó cũng giúp bạn ghi lại những kỉ niệm đẹp bạn từng trải qua.’*

Khi mình đưa ra ý tưởng này, chắc chắn một điều rằng nhiều bạn sẽ thắc mắc liệu ứng dụng này có cần thiết không? Những ai sẽ muốn sử dụng nó? Và có chắc rằng nhiều người sẽ thích viết nhật ký hay không? Đầu tiên, mình xin đưa ra lý do vì sao chúng ta nên tạo thói quen viết nhật ký. Nhật ký không chỉ là nơi để bạn lưu lại ký ức mà còn đem lại rất nhiều lợi ích khác – những cuốn nhật ký giúp nuôi dưỡng sự sáng tạo và tâm hồn bạn. Khi thói quen viết nhật ký được hình thành, kỹ năng viết lách của bạn sẽ ngày càng tốt hơn.

Xuất phát từ tình hình thực tế đó, tôi đã nghĩa ra một ý tưởng đó là xây dựng website hỗ trợ sinh viên cho Phân hiệu trường đại học Giao thông Vận Tải tại tp. Hồ Chí Minh để công tác hỗ trợ sinh viên diễn ra tiện lợi linh hoạt hơn.

Trường Đại học Giao thông vận tải Phân hiệu tại Tp. Hồ Chí Minh, hiện nay có gần 7000 học viên và sinh viên các hệ, mỗi sinh viên đều được nhà trường cung cấp cho một tài khoản gmail cá nhân. Trường hiện tại có 11 phòng ban trung tâm với các vai trò khác nhau. Mỗi phòng ban phụ trách các vấn đề riêng tuy nhiên tất cả đều có một mục đích chung đó là làm cho trường đại học GTVT phát triển tốt hơn, các bạn sinh viên có môi trường học tập, rèn luyện hoàn chỉnh hơn.

Các ban ngành, phong ban trung tâm, giáo viên cố vấn và các giảng viên luôn tích cực cố gắng giảng dạy, hỗ trợ và giúp đỡ các bạn sinh viên. Tuy nhiên, không thể nào như cấp 3 giáo viên luôn quan sát nhắc nhở học sinh. Môi trường đại học các bạn sinh viên cần có tính chủ động và tự giác, khi gặp các vấn đề cần chủ động hỏi để được giúp đỡ, hơn nữa sự thiếu hụt về nguồn nhân lực của nhà trường, công tác hỗ trợ sinh viên còn nhiều khó khăn.

Tôi chọn đề tài *Nguyên cứu các thuật toán trong xử lý ngôn ngữ tự nhiên và xây dưng website hỗ trợ sinh viên UTC2* dựa trên phần mềm odoo - giải pháp quản lí doanh nghiệp dẽ dàng để tạo ra một ứng dụng dễ sử dụng, chi phí thấp,có nhiều chức năng hữu dụng và dễ dàng cập nhập phục vụ với nhà trường và nhu cầu sử dụng của các bạn sinh viên.

**Tình hình nguyên cứu**

Trên thị trường ứng dụng hiện nay, có rất nhiều ứng dụng hỗ trợ người dùng viết nhật kí, viết ghi chú tuy nhiên hầu hết là các ứng dụng riêng lẻ và ít tùy chỉnh. Người dùng

**Quá trình nguyên cứu**

Để hiểu rõ hơn về công tác giải đáp hỗ trợ sinh viên. Tôi đã lựa chọn Fanpage Diễn đàn nghe sinh viên nói, và tìm hiểu vai trò của các phòng ban/trung tâm của trường đại học Giao Thông vận tải phân hiệu tại thành phố Hồ Chí Minh để khảo sát, tìm hiểu và phát triển ứng dụng

Hiện nay tại Phân hiệu Trường Đại học GTVT tại TP. HCM có các ban ngành phòng bạn sau:

*Bảng 1.1 Các phòng ban/trung tâm*

|  |  |
| --- | --- |
| **STT** | **Tên các ban ngành/phòng ban** |
| **1** | Phòng tổ chức hành chính |
| **2** | Phòng đào tạo |
| **3** | Phòng công tác sinh viên |
| **4** | Phòng khảo thí và đảm bảo chất lượng đào tạo |
| **5** | Phòng tài chính kế toán |
| **6** | Phòng thiết bị quản trị |
| **7** | Phòng khoa học công nghệ và đối ngoại |
| **8** | Ban quản lý ký túc xá |
| **9** | Ban thanh tra |
| **10** | Trung tâm thông tin thư viện |
| **11** | Trung tâm đào tạo thực hành |

Việc trả lời các thắc mắc cho các bạn sinh viên cần được diễn ra chính xác và nhanh chóng, tiết kiệm được các quy trình nhân lực trong việc xử lí. Đối tượng sử dụng là các bạn sinh viên, giáo viên; các chức năng chính cần có của hệ thống gồm có:

* **Đối với sinh viên:**
* Đăng nhập và đăng xuất
* Đăng bài trong diễn đàn
* Trả lời câu hỏi trong diễn đàn
* Thao tác với bài viết trên diễn đàn (yêu thích, cắm cờ, báo cáo quản trị, ...)
* Cập nhật thông tin cá nhân
* Tích điểm nâng cấp tài khoản (cấp bậc tài khoản gồm Đồng Bạc Vàng)
* Quản lý trang cá nhân
* Quản lý bài đăng
* Nhận tin tức, thông báo từ nhà trường
* Tạo và tùy chỉnh điểm môn học và điểm tích lũy mục tiêu
* Dự đoán điểm môn học và điểm tích lũy
* **Đối với giảng viên:**
* Giải đáp thắc mắc
* Trả lời chat online
* Cập nhật thông tin cá nhân
* Duyệt bài đăng
* Thao tác với bài đăng (yêu thích, xóa bài đăng vi phạm, xóa bình luận vi phạm, …)
* Tạo công thức dự đoán điểm cho môn học
* **Đối với người quản lý:**
* Quản lý giảng viên
* Quản lý sinh viên
* Quản lý diễn đàn
* Quản lý trang cá nhân của thành viên
* Quản lý các tin tức, thông báo
* Quản lý dự đoán điểm sinh viên
* Thống kê, báo cáo

## **Cấu trúc báo cáo đồ án tốt nghiệp**

Cấu trúc đồ án được chia thành các chương như sau:

Tổng quan – Giới thiệu tổng quan về đề tài đồ án tốt nghiệp.

Chương 1: Cơ sở lý thuyết

* Giới thiệu các công nghệ sử dụng.
* Giới thiệu về ngôn ngữ lập trình.
* Giới thiệu các thuật toán đã áp dụng.

Chương 2: Phân tích và thiết kế hướng đối tượng

* Biểu đồ usecase của hệ thống.
* Sơ đồ phân rã chức năng.
* Sơ đồ hoạt động.

Chương 3: Thiết kế và xây dựng ứng dụng

* Hiện trạng tổ chức và đưa ra giải pháp công nghệ.
* Phân tích yêu cầu hệ thống
* Thu thập dữ liệu
* Thuật toán áp dụng
* Xây dựng giao diện ứng dụng

Kết luận

* Đưa ra kết quả đạt được, những thứ còn tồn tại và hướng phát triển các chức năng thực tế, hữu ích cho ứng dụng trong tương lai.

Tài liệu tham khảo.

# CƠ SỞ LÝ THUYẾT

## **1.1 Công nghệ sử dụng.**

***1.1.1 Giới thiệu về Ionic Framework.***

Trong những năm gần đây, số lượng người sử dụng điện thoại thông minh trên toàn thế giới ngày càng tăng cao và điện thoại dần trở thành thứ không thể thiếu trong cuộc sống của mỗi chúng ta. Vì vậy để cạnh tranh, các doanh nghiệp ngoài việc xây dựng website còn phải phát triển ứng dụng di dộng để có thể tiếp cận nhanh hơn tới khách hàng và nắm bắt được thị hiếu của người tiêu dùng. Việc xây dựng một ứng dụng di động đòi hỏi phải chạy được trên nhiều nền tảng tiêu biểu là Android và IOS, do đó nhiều doanh nghiệp đã lựa chọn công nghệ Ionic framework để tiết kiệm thời gian cũng như chi phí.

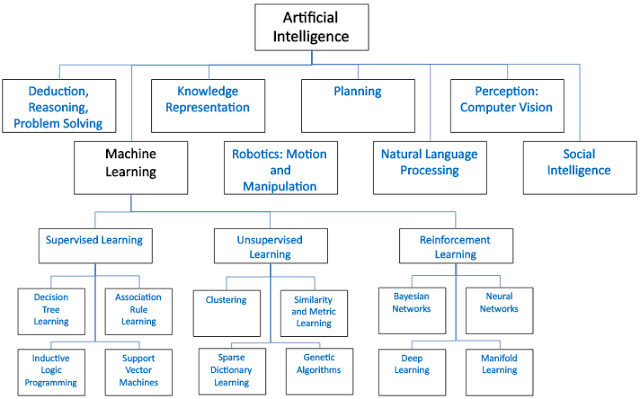
Ionic là một framework dùng để phát triển ứng dụng hybrid dựa trên HTML5. Một ứng dụng hybrid là một ứng dụng di động được xây dựng bằng các công nghệ phát triển web như HTML5, CSS, JavaScript và sau đó được đóng gói trong một môi trường hiển thị nội dung có thể hoạt động trên hầu hết các thiết bị di động giống như trình duyệt web (Native container). Các ứng dụng này sử dụng các thành phần hiển thị nội dung website (các trình duyệt ẩn như UIWebView trên iOS, hay Webview trên Android và các thiết bị khác) để hiển thị các đoạn mã HTML.

Ứng dụng hybrid đem lại nhiều lợi thế như khả năng hiển thị nội dung trên tất cả các thiết bị di động, cũng như tận dụng tối đa các tính năng khác của thiết bị di động như GPS, camera, danh sách liên lạc,… Bên cạnh đó, thời gian và chi phí dùng để tạo nên một ứng dụng hybrid cũng thấp hơn so với các ứng dụng di động thông thường.

Bạn có thể coi Ionic như là một bộ khung front-end giúp bạn kiểm soát hình ảnh và trải nghiệm trên ứng dụng của bạn. Giống như “Bootstrap for Native”, nhưng với sự hỗ trợ của một lượng lớn các thành phần di động, hiệu ứng chuyển động mượt mà và thiết kế đẹp.

Không như các framework khác, Ionic mang đến cho chúng ta những thành phần giao diện và cách bố trí mang phong cách rất giống với như phong cách thiết kế mặc định trên thiết bị di động.

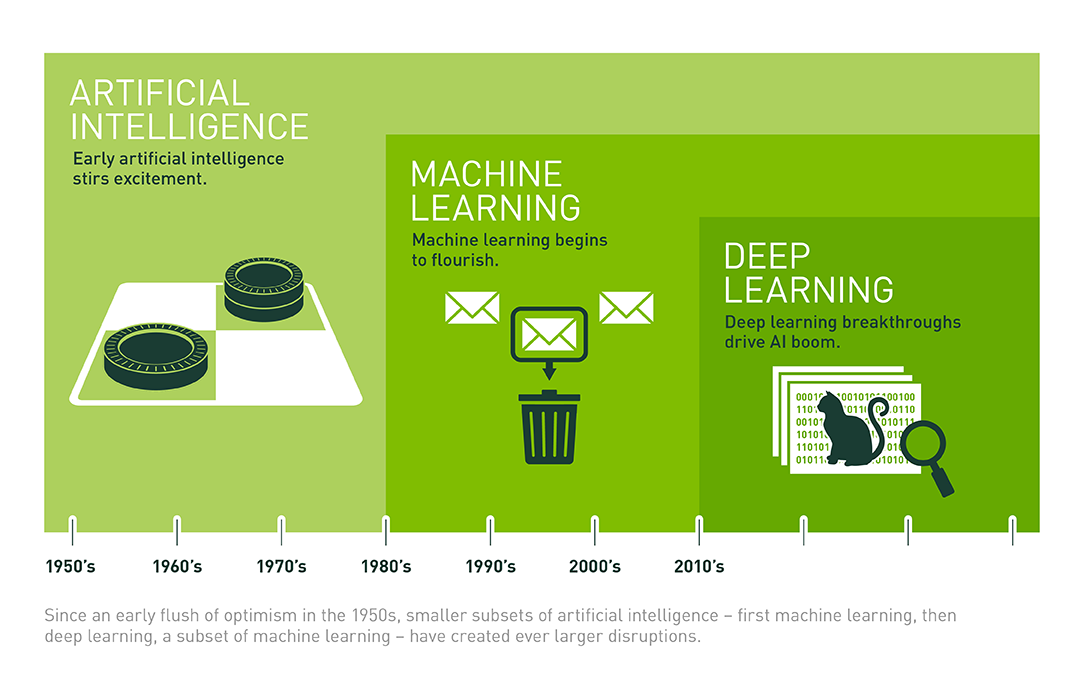
Vì Ionic là một framework HTML5 nên nó cần bộ đóng gói (wrapper) như Cordova hay PhoneGap để có thể chạy được như một ứng dụng di động. Cordova là bộ đóng gói mặc định trong Ionic framework.



*Hình 1.1 Mối quan hệ giữa AI, Machine Learning và Deep Learning.[[1]](#footnote-1)*

Thực tế gần đây, khi mà khả năng tính toán của các máy tính ngày càng được nâng lên một tầm cao mới, song song với đó là sự bùng nổ của BigData, ML đã tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là Deep Learning (DL).

DL đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 10 năm trước như: 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,...



*Hình 1.2 Mối quan hệ giữa AI, Machine Learning và Deep Learning.[[2]](#footnote-2)*

**a) Định nghĩa Machine Learning**

Machine Learning là một thuật toán có khả năng học tập từ dữ liệu, có nghĩa là chương trình máy tính sẽ học hỏi từ kinh nghiệm E (Experience) từ các tác vụ T (Task), với kết quả được đo bằng hiệu suất P. Nếu hiệu suất của nó áp dụng trên tác vụ T khi được đánh giá bởi P (Performance) và cải thiện theo kinh nghiệm E.

Ví dụ 1: Giả sử như bạn muốn máy tính xác định một tin nhắn có phải là SPAM hay không?

* Tác vụ T: Xác định 1 tin nhắn có phải SPAM hay không?
* Kinh nghiệm E: Xem lại những tin nhắn đánh dấu là SPAM xem có những đặc tính gì để có thể xác định nó là SPAM.
* Độ đo P: Là phần trăm số tin nhắn SPAM được phân loại đúng.
* Ví dụ 2: Chương trình nhận dạng số (số từ 0 -> 9)
* T: Là nhận dạng được ảnh chứa ký tự số.
* E: Đặc trưng để phân loại ký tự số từ tập dữ liệu số cho trước.
* P: Độ chính xác của quá trình nhận dạng.

**b) Sự hữu ích của Machine Learning**

Từ lâu đã có nhiều thuật toán ML nổi tiếng nhưng khả năng tự động áp dụng các phép tính phức tạp vào Big Data, lặp đi lặp lại với tốc độ nhanh hơn, chỉ mới phát triển gần đây.

Các ứng dụng của ML đã trở nên quá quen thuộc như:

* Xe tự lái, giảm thiểu tai nạn của Google? Chính là bản chất của ML.
* Các ưu đãi Recommendation Online như của Amazon & Netflix? Ứng dụng của Machine Learning trong cuộc sống hằng ngày.
* Muốn biết người dùng nói gì về bạn trên Twitter? ML kết hợp với sự sáng tạo của quy tắc ngôn ngữ.
* Nhận diện lừa đảo? Một trong những nhu cầu sử dụng hiển nhiên ngày nay.

**c) Đối tượng sử dụng**

Hầu hết mọi ngành công nghiệp đang làm việc với hàm lượng lớn dữ liệu đều nhận ra tầm quan trọng của công nghệ ML. Những cái nhìn sâu sắc từ nguồn dữ liệu này, sẽ giúp các tổ chức vận hành hiệu quả hơn hoặc tạo được lợi thế cạnh tranh so với các đối thủ.

* Các dịch vụ tài chính

Ngân hàng và những doanh nghiệp hoạt động trong lĩnh vực tài chính sử dụng công nghệ ML với 2 mục đích chính: xác định insights trong dữ liệu và ngăn chặn lừa đảo. Insights sẽ biết được các cơ hội đầu tư hoặc thông báo đến nhà đầu tư thời điểm giao dịch hợp lý. Khai phá dữ liệu cũng có thể tìm được những khách hàng đang có hồ sơ rủi ro cao hoặc sử dụng giám sát mạng để chỉ rõ những tín hiệu lừa đảo.

* Chính phủ

Các tổ chức chính phủ hoạt động về an ninh cộng đồng hoặc tiện ích xã hội sở hữu rất nhiều nguồn dữ liệu có thể khai thác insights. Ví dụ, khi phân tích dữ liệu cảm biến, chính phủ sẽ tăng mức độ hiệu quả của dịch vụ và tiết kiệm chi phí. ML còn hỗ trợ phát hiện gian lận và giảm thiểu khả năng trộm cắp danh tính.

* Chăm sóc sức khỏe

ML là 1 xu hướng phát triển nhanh chóng trong ngành chăm sóc sức khỏe, nhờ vào sự ra đời của các thiết bị và máy cảm ứng đeo được sử dụng dữ liệu để đánh giá tình hình sức khỏe của bệnh nhân trong thời gian thực. Công nghệ ML còn giúp các chuyên gia y tế xác định những xu hướng hoặc tín hiệu để cải thiện khả năng điều trị, chẩn đoán bệnh.

* Marketing và sales

Dựa trên hành vi mua hàng trước đây, các trang website sử dụng ML phân tích lịch sử mua hàng, từ đó giới thiệu những vật dụng mà bạn có thể sẽ quan tâm và yêu thích. Khả năng tiếp nhận dữ liệu, phân tích và sử dụng những dữ liệu đó để cá nhân hóa trải nghiệm mua sắm hoặc thực hiện chiến dịch Marketing chính là tương tai của ngành bán lẻ.

* Dầu khí

Tìm kiếm những nguồn nguyên liệu mới. Phân tích các mỏ dầu dưới đất. Dự đoán tình trạng thất bại của bộ cảm biến lọc dầu. Sắp xếp các kênh phân phối để đạt hiệu quả và tiết kiệm chi phí. Có thể nói, số lượng các trường hợp sử dụng ML trong ngành công nghiệp này cực kì lớn và vẫn ngày càng mở rộng.

* Vận tải

Phân tích dữ liệu để xác định mô hình và các xu hướng là trọng tâm trong ngành vận tải vì đây là ngành phụ thuộc vào khả năng tận dụng hiệu quả trên mỗi tuyến đường và dự đoán các vấn đề tiềm tàng để gia tăng lợi nhuận. Các chức năng phân tích dữ liệu và modeling của ML đóng vai trò quan trọng với các doanh nghiệp vận chuyển, vận tải *công cộng và các tổ chức vận chuyển khác.]*

### ***1.1.2 Phân nhóm các thuật toán Machine Learning***

Các thuật toán ML thường được chia làm 4 nhóm:

- Supervise learning (Học có giảm sát)

- Unsupervised learning (Học không giám sát)

- Semi-supervised lerning (Học bán giám sát)

- Reinforcement learning (Học củng cố)

**a) Supervised Learning (Học có giám sát)**

Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới dựa trên các cặp (*input, outcome*) đã biết từ trước. Cặp dữ liệu này còn được gọi là (*data, label*), tức (*dữ liệu, nhãn*). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Một cách toán học, Supervised learning là khi chúng ra có một tập hợp biến đầu vào X={x1,x2,…,xN} và một tập hợp nhãn tương ứng Y={y1,y2,…,yN} trong đó xi,yi là các vector. Các cặp dữ liệu biết trước (xi,yi)∈X×Y được gọi là tập *training data* (dữ liệu huấn luyện). Từ tập traing data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập X sang một phần tử (xấp xỉ) tương ứng của tập Y: yi≈f(xi),  ∀i=1,2,…,N Mục đích là xấp xỉ hàm số f thật tốt để khi có một dữ liệu xx mới, chúng ta có thể tính được nhãn tương ứng của nó y=f(x).

**Ví dụ:** trong nhận dạng chữ viết tay, ta có ảnh của hàng nghìn ví dụ của mỗi chữ số được viết bởi nhiều người khác nhau. Chúng ta đưa các bức ảnh này vào trong một thuật toán và chỉ cho nó biết mỗi bức ảnh tương ứng với chữ số nào. Sau khi thuật toán tạo ra một mô hình, tức một hàm số mà đầu vào là một bức ảnh và đầu ra là một chữ số, khi nhận được một bức ảnh mới mà mô hình chưa nhìn thấy bao giờ, nó sẽ dự đoán bức ảnh đó chứa chữ số nào.

Ví dụ này khá giống với cách học của con người khi còn nhỏ. Ta đưa bảng chữ cái cho một đứa trẻ và chỉ cho chúng đây là chữ A, đây là chữ B. Sau một vài lần được dạy thì trẻ có thể nhận biết được đâu là chữ A, đâu là chữ B trong một cuốn sách mà chúng chưa nhìn thấy bao giờ.

Thuật toán supervised learning còn được tiếp tục chia nhỏ ra thành hai loại chính:

**Classification (Phân loại)**

Một bài toán được gọi là classification nếu các label của input data được chia thành một số hữu hạn nhóm.

**Ví dụ**: Gmail xác định xem một email có phải là spam hay không; các hãng tín dụng xác định xem một khách hàng có khả năng thanh toán nợ hay không. Ba ví dụ phía trên được chia vào loại này.

**Regression (Hồi quy)**

Nếu label không được chia thành các nhóm mà là một giá trị thực cụ thể.

**Ví dụ:** một căn nhà rộng x m2, có y phòng ngủ và cách trung tâm thành phố z km sẽ có giá là bao nhiêu?

**b) Unsupervised Learning (Học không giám sát)**

Trong thuật toán này, chúng ta không biết được outcome hay nhãn mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm hoặc giảm số chiều của dữ liệu để thuận tiện trong việc lưu trữ và tính toán.

Một cách toán học, Unsupervised learning là khi chúng ta chỉ có dữ liệu vào X mà không biết nhãn Y tương ứng.

Những thuật toán loại này được gọi là Unsupervised learning vì không giống như Supervised learning, chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Giống như khi ta học, không có thầy cô giáo nào chỉ cho ta biết đó là chữ A hay chữ B. Cụm không giám sát được đặt tên theo nghĩa này.

Các bài toán Unsupervised learning được tiếp tục chia nhỏ thành hai loại:

**Clustering (phân nhóm)**

Một bài toán phân nhóm toàn bộ dữ liệu XX thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm.

Ví dụ: phân nhóm khách hàng dựa trên hành vi mua hàng. Điều này cũng giống như việc ta đưa cho một đứa trẻ rất nhiều mảnh ghép với các hình thù và màu sắc khác nhau, ví dụ tam giác, vuông, tròn với màu xanh và đỏ, sau đó yêu cầu trẻ phân chúng thành từng nhóm. Mặc dù không cho trẻ biết mảnh nào tương ứng với hình nào hoặc màu nào, nhiều khả năng chúng vẫn có thể phân loại các mảnh ghép theo màu hoặc hình dạng. [4]

**Association rule (Luật kết hợp)**

Là bài toán khi chúng ta muốn khám phá ra một quy luật dựa trên nhiều dữ liệu cho trước.

Ví dụ: những khách hàng nam mua quần áo thường có xu hướng mua thêm đồng hồ hoặc thắt lưng; những khán giả xem phim Spider Man thường có xu hướng xem thêm phim Bat Man, dựa vào đó tạo ra một hệ thống gợi ý khách hàng, thúc đẩy nhu cầu mua sắm. [4]

**c) Semi-Supervised Learning (Học bán giám sát)**

Các bài toán khi chúng ta có một lượng lớn dữ liệu XX nhưng chỉ một phần trong chúng được gán nhãn được gọi là Semi-Supervised Learning. Những bài toán thuộc nhóm này nằm giữa hai nhóm được nêu bên trên.

Một ví dụ điển hình của nhóm này là chỉ có một phần ảnh hoặc văn bản được gán nhãn (ví dụ bức ảnh về người, động vật hoặc các văn bản khoa học, chính trị) và phần lớn các bức ảnh, văn bản khác chưa được gán nhãn được thu thập từ internet. Thực tế cho thấy rất nhiều các bài toán ML thuộc vào nhóm này vì việc thu thập dữ liệu có nhãn tốn rất nhiều thời gian và có chi phí cao. Rất nhiều loại dữ liệu thậm chí cần phải có chuyên gia mới gán nhãn được. Ngược lại, dữ liệu chưa có nhãn có thể được thu thập với chi phí thấp từ internet.

**d) Reinforcement Learning (Học Củng Cố)**

Reinforcement learning là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao. Hiện tại, Reinforcement Learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi, các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất.

**Ví dụ:** [AlphaGo gần đây nổi tiếng với việc chơi cờ vây thắng cả con người](https://gogameguru.com/tag/deepmind-alphago-lee-sedol/). [Cờ vây được xem là có độ phức tạp cực kỳ cao](https://www.tastehit.com/blog/google-deepmind-alphago-how-it-works/) với tổng số nước đi là xấp xỉ 1076110761, so với cờ vua là 1012010120 và tổng số nguyên tử trong toàn vũ trụ là khoảng 10801080.

Vì vậy, thuật toán phải chọn ra 1 nước đi tối ưu trong số hàng nhiều tỉ tỉ lựa chọn, và tất nhiên, không thể áp dụng thuật toán tương tự như IBM Deep.

Về cơ bản, AlphaGo bao gồm các thuật toán thuộc cả Supervised learning và Reinforcement Learning. Trong phần Supervised Learning, dữ liệu từ các ván cờ do con người chơi với nhau được đưa vào để huấn luyện. Tuy nhiên, mục đích cuối cùng của AlphaGo không phải là chơi như con người mà phải thậm chí thắng cả con người.

Vì vậy, sau khi học xong các ván cờ của con người, AlphaGo tự chơi với chính nó với hàng triệu ván chơi để tìm ra các nước đi mới tối ưu hơn. Thuật toán trong phần tự chơi này được xếp vào loại Reinforcement learning.

***1.1.3 Các bước thực hiện Machine Learning***

Thực hiện Machine Learning bao gồm các bước như sau:

**Thu thập và chuẩn bị dữ liệu**:

Yếu tố ban đầu cần thiết để thực hiện ML là cần có dữ liệu. Dữ liệu có thể được lấy từ nhiều nguồn khác nhau, có thể ít, có thể nhiều, có thể sạch, có thể nhiều dữ liệu lỗi…  
Sau khi thu thập, dữ liệu cần được làm sạch.

**Chọn thành phần**

Với mỗi tập dữ liệu có thể có rất nhiều thành phần, nhưng không phải thành phần nào cũng liên quan tới bài toán mà ta cần giải, việc lựa chọn thành phần (loại bỏ các thành phần không cần thiết) giúp cho việc học của ta trở nên nhanh và hiệu quả hơn. Tuy nhiên, việc lựa chọn đòi hỏi sự thấu hiểu về dữ liệu và bài toán, chủ yếu làm bằng tay và sức người.  
Sau khi chọn được thành phần, nhiều khi ta quay lại bước 1, tiến hành loại bỏ các dữ liệu không liên quan để thu nhỏ tập dữ liệu.

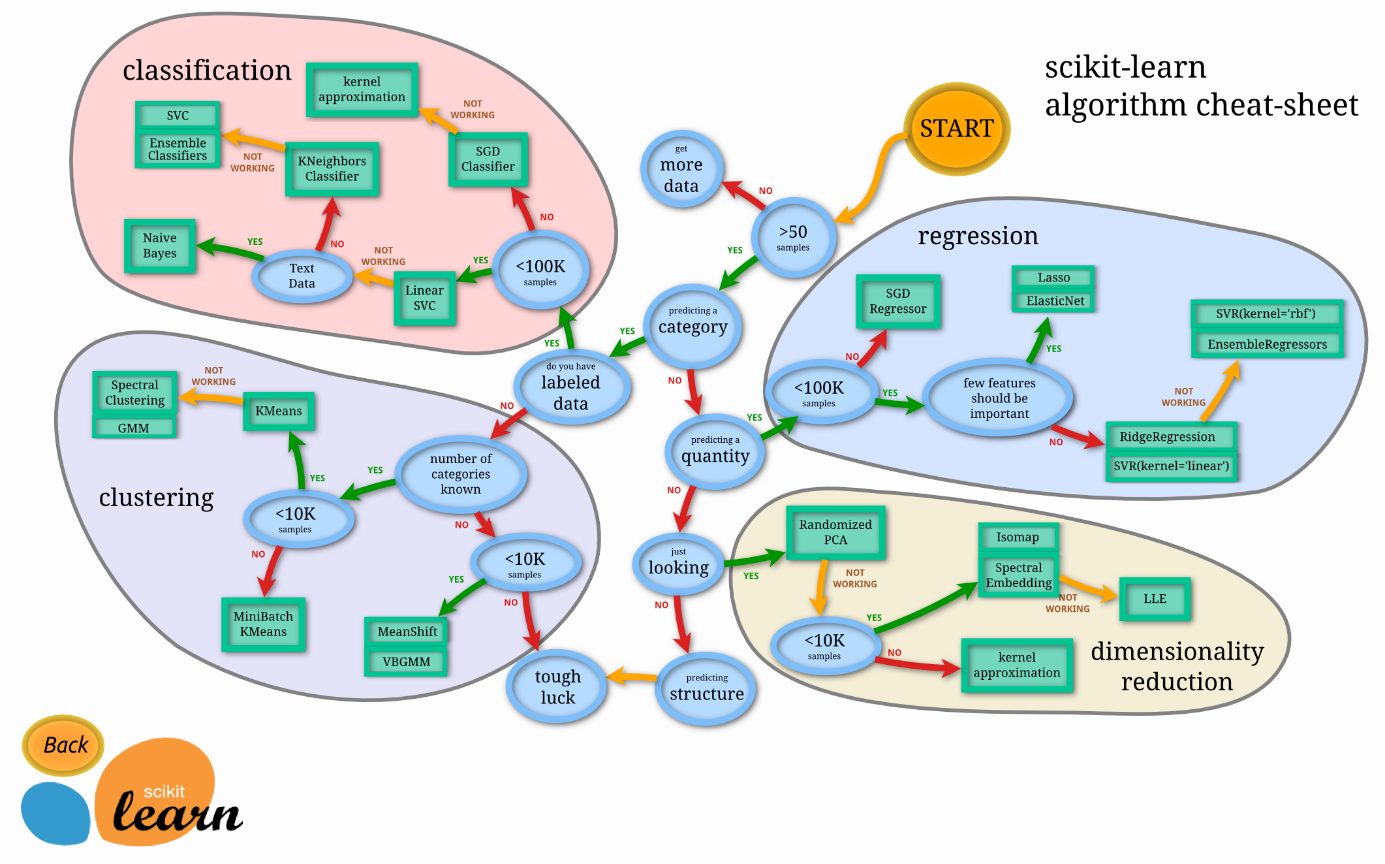
Việc thu nhỏ tập dữ liệu cũng khiến cho việc học của ta tốn ít thời gian hơn, tuy nhiên dữ liệu ít quá cũng khiến việc học đạt độ chính xác không cao, cần cân đối giữa các yếu tố này.

**Chuẩn hoá dữ liệu**

Nhiều khi dữ liệu của từng thành phần có định dạng, kích thước khác biệt lớn, ví dụ thành phần 1 có dữ liệu trong khoảng [0, 1], thành phần 2 có dữ liệu trong khoảng [-1000, 1000], thành phần 3 có dữ liệu là hình ảnh… Để tăng tốc độ và hiệu quả của việc học, ta cần chuẩn hoá dữ liệu, đưa dữ liệu của tất cả các thành phần về cùng một định dạng (số hoá), và cùng khoảng biến thiên.

**Chọn thuật toán**

Tuỳ vào dữ liệu, bài toán mà ta lựa chọn thuật toán ML tương ứng. Đôi khi lựa chọn thuật toán cũng cần dựa trên kinh nghiệm. Một số gợi ý cho việc lựa chọn thuật toán là tham khảo của Scikit-learn.



*Hình 1.3 Lựa chọn thuật toán phù hợp trong Machine Learning.[[3]](#footnote-3)*

**Chọn parameter cho thuật toán**

Tuỳ mỗi thuật toán mà có nhiều các cách cài đặt khác nhau. Hơn thế nữa các tham số tính toán cũng quyết định không nhỏ tới kết quả tính toán. Vì thế khi sau khi chọn thuật toán thì việc chọn parameter phù hợp với dữ liệu cũng khá quan trọng. Việc chọn Parameter chủ yếu dựa trên kinh nghiệm, nhiều khi có thể sử dụng các thư viện support.

**Huấn luyện và Đánh giá**

Sau khi chọn được thuật toán (một hoặc nhiều thuật toán) và parameter tương ứng ta cho dữ liệu vào train. Tiến hành cross-validation để điều chỉnh model.

Sau khi train được model, ta đưa data test vào kiểm tra và đánh giá độ chính xác của model vừa train được.

**Phân chia dữ liệu**

Thông thường với tập data cho trước ta thường chia làm 3 tập để sử dụng với mục đích khác nhau:

* **Tập huấn luyện** (Training set): dùng để huấn luyện model.
* **Tập kiểm chứng** (Validation set): dùng để đánh giá, điều chỉnh model. Ví dụ như ta dùng tập huấn luyện cho nhiều thuật toán khác nhau rồi dùng tập kiểm chứng để chọn thuật toán phù hợp nhất. Hoặc tìm parameter phù hợp nhất cho một thuật toán cụ thể.
* **Tập kiểm tra** (Test set): dùng để đánh giá model sau khi huấn luyện. Mô hình đã được đánh giá bằng tập test phải là mô hình cuối cùng, không được thay đổi nữa.

Việc phân chia tỉ lệ giữa 3 tập này được khuyên là 60% - 20% - 20% [6].

Nếu model cuối cùng thu được sau khi huấn luyện và kiểm chứng cần phải được đánh giá bằng tập kiểm tra. Nếu kết quả không tốt, cần thực hiện huấn luyện lại từ đầu.

***1.1.4 Các ứng dụng***

Có rất nhiều ứng dụng mang tính thực tế cao của máy học mà khó có thể kể hết được. Những ứng dụng dưới đây là những ứng dụng phổ biến và được chọn lọc theo góc nhìn cá nhân.

* Nhận diện và phát hiện khuôn mặt: Nhận diện và phát hiện khuôn mặt là ứng dụng khá thú vị của máy học và được áp dụng khá nhiều vào đời sống. Tiêu biểu là tính năng phát hiện khuôn mặt ở máy chụp ảnh. Ứng dụng được phát triển thêm thành phát hiện chớp mắt, phát hiện cười....
* Xe tự lái: Xe tự lái mặc dù phát triển từ đầu thập niên 90 nhưng cho tới nay vẫn còn là vấn đề được nhiều người quan tâm. Các hãng lớn như Google, NVDIA đang nỗ lực để tạo ra một cỗ máy có thể hoàn toàn tự động lái xe và giảm thiểu tai nạn cho con người.
* Phân lớp ảnh: Tìm kiếm ảnh trên Google hiện rất quen thuộc với nhiều người. Ứng dụng của phân lớp ảnh giúp người dùng sử dụng ảnh làm từ khóa tìm kiếm thay thế cho việc tìm kiếm truyền thống trên Google. Bạn upload một ảnh lên và Google sẽ giúp bạn tìm kiếm những thông tin liên quan đến bức ảnh đó.
* Nhận dạng giọng nói: Các trợ lý ảo như Siri, Conrtana hay Google Now là ví dụ điển hình cho nhận dạng giọng nói. Một ví dụ khác nữa là tính năng dịch thuật trực tuyến của Youtube. Với ứng dụng của DL, khả năng dịch thuật chín xác ngôn ngữ từ các video Youtube đang ngày một phát triển vượt bậc.
* Anti-virus: Có thể nhiều người không nghĩ rằng các phần mềm diệt virus lại áp dụng máy học. Tuy nhiên, áp dụng máy học vào để phân tích và dự đoán xu hướng các loại virus sẽ giúp ích rất nhiều trong việc bảo vệ dữ liệu máy tính.

***1.1.1 Các loại giải thuật***

Các thuật toán học máy được phân loại theo kết quả mong muốn của thuật toán. Các loại thuật toán thường dùng bao gồm:

* Học có giám sát : trong đó, thuật toán tạo ra một hàm ánh xạ dữ liệu vào tới kết quả mong muốn. Một phát biểu chuẩn về một việc học có giám sát là bài toán phân loại: chương trình cần học (cách xấp xỉ biểu hiện của) một hàm ánh xạ một vector [X1, X2, .., Xn] tới một vài lớp bằng cách xem xét một số mẫu dữ\_liệu - kết\_quả của hàm đó.
* Học không giám sát : mô hình hóa một tập dữ liệu, không có sẵn các ví dụ đã được gắn nhãn.
* Học nửa giám sát : kết hợp các ví dụ có gắn nhãn và không gắn nhãn để sinh một hàm hoặc một bộ phân loại thích hợp.
* Học tăng cường : trong đó, thuật toán học một chính sách hành động tùy theo các quan sát về thế giới. Mỗi hành động đều có tác động tới môi trường, và môi trường cung cấp thông tin phản hồi để hướng dẫn cho thuật toán của quá trình học.
* Chuyển đổi : tương tự học có giám sát nhưng không xây dựng hàm một cách rõ ràng. Thay vì thế, cố gắng đoán kết quả mới dựa vào các dữ liệu huấn luyện, kết quả huấn luyện, và dữ liệu thử nghiệm có sẵn trong quá trình huấn luyện.
* Học cách học : trong đó thuật toán học thiên kiến quy nạp của chính mình, dựa theo các kinh nghiệm đã gặp.

Phân tích hiệu quả các thuật toán học máy là một nhánh của ngành thống kê, được biết với tên lý thuyết học điện toán.

**1.2 Ngôn ngữ lập trình.**

***1.2.1 Khái niệm***

Xử lý ngôn ngữ tự nhiên (NLP) là một nhánh của Trí tuệ nhân tạo, tập trung vào việc nghiên cứu sự tương tác giữa máy tính và ngôn ngữ tự nhiên của con người. Mục tiêu của lĩnh vực này là giúp máy tính hiểu và thực hiện hiệu quả những nhiệm vụ liên quan đến ngôn ngữ của con người như: tương tác giữa người và máy, cải thiện hiệu quả giao tiếp giữa con người với con người, hoặc đơn giản là nâng cao hiệu quả xử lý văn bản và lời nói.

Xử lý ngôn ngữ tự nhiên ra đời từ những năm 1940, với rất nhiều công trình nghiên cứu theo hai hướng chính là: 1) ô-tô-mát (automaton) và các mô hình xác suất (probabilistic models) vào những năm 1950; 2) các phương pháp dựa trên ký hiệu (symbolic) và các phương pháp ngẫu nhiên (stochastic) vào những năm 1970. Giai đoạn tiếp theo (1970-1983) chứng kiến sự bùng nổ trong nghiên cứu về xử lý tiếng nói và ngôn ngữ. Ngày nay với sự phát triển nhanh chóng, học máy (machine learning) đã trở thành trung tâm của phần lớn các lĩnh vực thuộc khoa học máy tính, bao gồm xử lý ảnh và thị giác máy tính (computer vision), tin sinh học (bioinformatics), các hệ tư vấn (recommender systems), kỹ nghệ phần mềm, và cả xử lý ngôn ngữ tự nhiên.

***1.2.2 Những khó khăn trong lĩnh vực xử lý ngôn ngữ tự nhiên***

Xử lý ngôn ngữ tự nhiên liên quan tới tương tác giữa máy tính và ngôn ngữ của con người. Ngôn ngữ tự nhiên xuất phát từ cảm xúc, vì thế thường không có quy luật hay tuân thủ theo tính hợp lí logic, kể cả về mặt cú pháp, ngữ nghĩa, và diễn đạt ngôn từ. Nó có tính nhập nhằng cao ở tất cả các mức, bao gồm mức từ vựng, mức cú pháp, mức ngữ nghĩa và mức văn bản. Ta nói rằng ngôn ngữ là nhập nhằng nếu có nhiều cấu trúc ngôn ngữ khác nhau phù hợp với nó. Sự nhập nhằng của ngôn ngữ tự nhiên khiến việc xử lý ngôn ngữ tự nhiên trên máy tính trở nên khó khăn. Hãy cùng xem xét những ví dụ sau đây:

Ví dụ 1:

They book that hotel. (S1)

They read that book. (S2)

Đầu tiên, từ book là nhập nhằng về mặt từ loại. Book có thể là một động từ (trong câu S1) hoặc một danh từ (trong câu S2) tùy thuộc vào ngữ cảnh xuất hiện của nó. Hiện tượng này gây khó khăn cho bài toán gán nhãn từ loại, một bước trong xử lý cú pháp. Không chỉ vậy, book cũng nhập nhằng về mặt ngữ nghĩa. Book có thể là một hành động đặt hàng thứ gì đó (trong câu S1) hoặc có thể là một văn bản viết được xuất bản dưới dạng in ấn hay điện tử (trong câu S2). Hiện tượng này gây khó khăn cho bài toán xác định nghĩa của từ, là một bước trong xử lý ngữ nghĩa.

**2.2.3 Những bài toán cơ bản trong NLP**

Xử lý ngôn ngữ tự nhiên bao gồm hiểu ngôn ngữ tự nhiên (Natural Language Understanding – NLU) và sinh ngôn ngữ tự nhiên (Natural Language Generation – NLG). Trong đó, hiểu ngôn ngữ tự nhiên (NLU) bao gồm 4 bước chính sau đây:

* Phân tích hình vị: là sự nhận biết, phân tích, và miêu tả cấu trúc của những hình vị trong một ngôn ngữ cho trước và các đơn vị ngôn ngữ khác, như từ gốc, biên từ, phụ tố, từ loại,… Có hai loại bài toán điển hình trong phần này, bao gồm bài toán tách từ (word segmentation) và gán nhãn từ loại (POS).
* Phân tích cú pháp: là quy trình phân tích một chuỗi các biểu tượng, ở dạng ngôn ngữ tự nhiên hoặc ngôn ngữ máy tính, tuân theo văn phạm hình thức. Văn phạm hình thức thường dùng trong phân tích cú pháp của ngôn ngữ tự nhiên bao gồm Văn phạm phi ngữ cảnh, Văn phạm danh mục kết nối và Văn phạm phụ thuộc. Đầu vào của quá trình phân tích là một câu gồm một chuỗi từ và nhãn từ loại của chúng, và đầu ra là một cây phân tích thể hiện cấu trúc cú pháp của câu đó. Các thuật toán phân tích cú pháp phổ biến bao gồm CKY, Earley, Chart, và GLR.
* Phân tích ngữ nghĩa: là quá trình liên hệ cấu trúc ngữ nghĩa, từ cấp độ cụm từ, mệnh đề, câu và đoạn đến cấp độ toàn bài viết, với ý nghĩa độc lập của chúng. Nói cách khác, việc này nhằm tìm ra ngữ nghĩa của đầu vào ngôn từ. Phân tích ngữ nghĩa bao gồm hai mức độ: Ngữ nghĩa từ vựng biểu hiện các ý nghĩa của những từ thành phần, và phân biệt nghĩa của từ; Ngữ nghĩa thành phần liên quan đến cách thức các từ liên kết để hình thành những nghĩa rộng hơn.
* Phân tích diễn ngôn: Ngữ dụng học là môn nghiên cứu về mối quan hệ giữa ngôn ngữ và ngữ cảnh sử dụng (context-of-use). Ngữ cảnh sử dụng bao gồm danh tính của người hoặc vật, và vì thế ngữ dụng học bao gồm những nghiên cứu về cách ngôn ngữ được dùng để đề cập (hoặc tái đề cập) tới người hoặc vật. Ngữ cảnh sử dụng bao gồm ngữ cảnh diễn ngôn, vì vậy ngữ dụng học cũng bao gồm những nghiên cứu về cách thức cấu tạo nên diễn ngôn, và cách người nghe hiểu người đang đối thoại với mình.

Khía cạnh thứ hai của NLP là sinh ngôn ngữ tự nhiên (NLG). Đây là một nhiệm vụ trong quá trình xử lý ngôn ngữ tự nhiên trong việc sinh ra ngôn ngữ tự nhiên từ một hệ thống máy biểu diễn như một cơ sở tri thức hoặc một dạng biểu diễn logic. NLG đóng vai trò quan trọng trong rất nhiều ứng dụng NLP, bao gồm sinh hội thoại, tương tác người – máy, dịch thuật máy, và tóm tắt văn bản tự động.

***1.2.3 Một số ứng dụng của xử lý ngôn ngữ tự nhiên***

- Truy xuất thông tin (Information Retrieval – IR) có nhiệm vụ tìm các tài liệudưới dạng không có cấu trúc (thường là văn bản) đáp ứng nhu cầu về thông tin từ những nguồn tổng hợp lớn. Những hệ thống truy xuất thông tin phổ biến nhất bao gồm các công cụ tìm kiếm như Google, Yahoo, hoặc Bing search. Những công cụ này cho phép tiếp nhận một câu truy vấn dưới dạng ngôn ngữ tự nhiên làm đầu vào và cho ra một danh sách các tài liệu được sắp xếp theo mức độ phù hợp.

- Trích chọn thông tin (Information Extraction) nhận diện một số loại thực thể được xác định trước, mối quan hệ giữa các thực thể và các sự kiện trong văn bản ngôn ngữ tự nhiên. Khác với truy xuất thông tin trả về một danh sách các văn bản hợp lệ thì trích chọn thông tin trả về chính xác thông tin mà người dùng cần. Những thông tin này có thể là về con người, địa điểm, tổ chức, ngày tháng, hoặc thậm chí tên công ty, mẫu sản phẩm hay giá cả.

- Trả lời câu hỏi (QA) có khả năng tự động trả lời câu hỏi của con người ở dạng ngôn ngữ tự nhiên bằng cách truy xuất thông tin từ một tập hợp tài liệu. Một hệ thống QA đặc trưng thường bao gồm ba mô đun: Mô đun xử lý truy vấn (Query Processing Module) – tiến hành phân loại câu hỏi và mở rộng truy vấn; Mô đun xử lý tài liệu (Document Processing Module) – tiến hành truy xuất thông tin để tìm ra tài liệu thích hợp; và Mô hình xử lý câu trả lời (Answer Processing Module) – trích chọn câu trả lời từ tài liệu đã được truy xuất.

- Tóm tắt văn bản tự động là bài toán thu gọn văn bản đầu vào để cho ra một bản tóm tắt ngắn gọn với những nội dung quan trọng nhất của văn bản gốc. Có hai phương pháp chính trong tóm tắt, là phương pháp trích xuất (extractive) và phương pháp tóm lược ý (abstractive). Những bản tóm tắt trích xuất được hình thành bằng cách ghép một số câu được lấy y nguyên từ văn bản cần thu gọn. Những bản tóm lược ý thường truyền đạt những thông tin chính của đầu vào và có thể sử dụng lại những cụm từ hay mệnh đề trong đó, nhưng nhìn chung được thể hiện ở ngôn ngữ của người tóm tắt.

- Dịch máy (Machine translation – MT) là việc sử dụng máy tính để tự động hóa một phần hoặc toàn bộ quá trình dịch từ ngôn ngữ này sang ngôn ngữ khác. Các phương pháp dịch máy phổ biến bao gồm dịch máy dựa trên ví dụ (example-based machine translation – EBMT), dịch máy dựa trên luật (rule-based machine translation – RBMT), và dịch máy thống kê (statistical machine translation – SMT). Những nghiên cứu gần đây tập trung vào dịch máy thống kê bởi nhiều ưu điểm của nó so với các phương pháp khác. Dịch dựa trên từ (word-based translation), dịch dựa trên cú pháp (syntax-based translation), dịch dựa trên cụm từ (phrase-based translation), và dịch dựa trên cụm từ phân cấp (hierarchical phrase-based translation) là những mô hình dịch máy thống kê thành công nhất.

**1.3** **Thuật toán áp dụng**

***1.3.1 Giới thiệu về Odoo***

* Odoolà một phần mềm quản trị doanh nghiệp mã nguồn mở sử dụng ngôn ngữ lập trình Python 7 (còn có thêm Javascript và XML) [1]. Bao gồm các module bán hàng, chăm sóc khách hàng, quản trị dự án, quản trị kho, quản trị sản xuất, quản lý tài chính và quản trị nguồn nhân lực, …
* Odoo/OpenERP được tích hợp công nghệ điện toán đám mây, cực kỳ phù hợp với các doanh nghiệp vừa và nhỏ trong mọi ngành nghề, lĩnh vực [2].

***1.3.2 Các ưu điểm của Odoo***

* Các ưu điểm của Odoo [1]:
* Là một mã nguồn mở nên nhiều công ty tin học nhỏ có thể tham gia cung cấp triển khai và phát triển bổ sung các module phụ trợ.
* Odoo dễ cài, vận hành thử trên nhiều nền tảng OS
* Công nghệ được cập nhật liên tục
* Kết nối thông minh
* Dễ dàng tùy chỉnh
* Không cần trả phí bản quyền
* Các tiện ích của Odoo:
* Thương mại điện tử.
* Bán hàng.
* Bất động sản.
* Phân tích dữ liệu
* Nắm bắt được những ưu điểm trên, tôi đã sử dụng phần mềm Odoo để xây dựng nên trang web….

***1.3.3 Cơ sở dữ dữ liệu PostgreSQL***

* PostgreSQL là một hệ thống quản trị cơ sở dữ liệu quan hệ-đối tượng (object-relational database management system), hệ thống cơ sở dữ liệu mã nguồn mở tiên tiến nhất hiện nay.
* PostgreSQL được thiết kế để chạy trên các nền tảng tương tự UNIX. Tuy nhiên, sau đó được điều chỉnh linh động để có thể chạy được trên nhiều nền tảng khác nhau như Mac OS X, Solaris và Windows.
* PostgreSQL là một phần mềm mã nguồn mở miễn phí. Vì vậy, bạn sẽ được tự do sử dụng, sửa đổi và phân phối PostgreSQL dưới mọi hình thức.

1. **PHÂN** **TÍCH VÀ THIẾT KẾ HỆ THỐNG**

**2.1 Biểu đồ usecase của hệ thống**

Văn bản là ngôn ngữ phi cấu trúc, để máy có thể hiểu được và tiến hành phân loại tự động, ta cần chuyển chúng về dạng thích hợp, dạng ngôn ngữ có cấu trúc. Giai đoạn tiền xử lý dữ liệu này là bước đệm để việc chuyển đổi văn bản hay véc-tơ hóa văn bản ở bước sau được tiến hành thuận lợi và có hiệu suất cao nhất cho quá trình phân loại sau này. Các việc chính trong giai đoạn này là: Tách từ và loại bỏ stop word.

***2.1.1 Tách từ***

Một số nước châu Á có ngôn ngữ với cấu trúc, hình thái gần tương đồng với ngôn ngữ Tiếng Việt như tiếng Nhật, tiếng Trung, tiếng Hàn đã xây dựng thành công nhiều phương pháp tách từ với kết quả khá tốt. Những phương pháp đó có thể áp dụng vào trong việc tách từ tiếng Việt. Trong luận văn này chúng tôi chỉ trình bày những phương pháp tách từ đã được áp dụng vào tiếng Việt.

**a) Phương pháp Maximum matching**

Phương pháp khớp tối đa (maximum matching) hay còn gọi là Left Right Maximum Matching. Trong phương pháp này, chúng ta sẽ duyệt một câu từ trái qua phải, sau đó chọn từ có nhiều âm tiết nhất trong câu mà có mặt trong từ điển, rồi tiếp tục với các từ còn lại trong câu đến khi hết câu và hết văn bản. Thuật toán này có hai dạng.

Dạng đơn giản: Giả sử chúng ta có một câu S = {l1, l2, l3…, lm} với l1, l2, l3…, lm là các âm tiết đơn được tách nhau bời khoảng trắng trong câu. Chúng ta sẽ bắt đầu duyệt từ đầu chuỗi. Xét xem l1 có phải là từ có trong từ điển không, sau đó tới l1-l2, l1-l2-l3…, l1-l2-l3…-ln với n là số âm tiết lớn nhất của một từ có thể có nghĩa (có trong từ điển tiếng Việt) thông thường sẽ là 4 hoặc 5 đối với tiếng Việt. Sau đó chúng ta chọn từ có nhiều âm tiết nhất mà có trong từ điển và đánh dấu từ đó, rồi tiếp tục quy trình trên với phần còn lại của câu và toàn bộ văn bản. Dạng này khá đơn giản nhưng nó sẽ gặp phải nhiều nhặp nhằng trong tiếng Việt.

Dạng phức tạp: dạng này cũng thực hiện quy trình giống như dạng đơn giản. Tuy nhiên, dạng này có thể tránh được một số nhập nhằng gặp phải trong dạng đơn giản. Giả sử khi duyệt câu và chúng ta có l1 và l1-l2 đều là từ có trong từ điển thì thuật toán sử dụng chiến thuật 3 từ tốt nhất. Tiêu chuẩn 3 từ tốt nhất được Chen & Liu (1992) [11] đưa ra. Nó có nội dung là khi một chuỗi có thể tách thành nhiều cách thì ta chọn cách tách mà sao cho độ dài trung bình của các từ được tách ra từ chuỗi là lớn nhất và sự chênh lệch độ dài các từ được tách ra là nhỏ nhất.

Ví dụ:

Ta có chuỗi có thể tách thành 3 cách:

Thì khi đó cách tách thứ hai sẽ được chọn và từ sẽ được đánh dấu do nó có độ dài trung bình là 2 lớn hơn cách tách đầu và có độ chênh lệch độ dài giữa các từ là 0 nhỏ hơn với cách tách thứ 3.

Ưu điểm của phương pháp này có thể thấy rõ là đơn giản, dễ hiểu, chạy nhanh và chỉ cần dựa vào từ điển để thực hiện. Tuy nhiên nhược điểm của nó cũng chính là từ điển. Nghĩa là độ chính xác khi thực hiện tách từ phụ thuộc hoàn toàn vào tính đủ, tính chính xác của từ điển. Và cũng vì sử dụng từ điển mà thuật toán này gặp phải rất nhiều nhập nhằng cũng như không có chiến lược gì với các từ chưa biết (các từ không có trong từ điển).

**b) Phương pháp Transformation-baseb learning (TBL)**

Phương pháp TBL (Transformation-Based learning) còn gọi là phương pháp học cải tiến, được Eric Brill giới thiệu lần đầu vào năm 1995 [9]. Ý tưởng của phương pháp này là tiếp cận dựa trên tập đã đánh dấu. Nghĩa là chúng ta sẽ huấn luyện cho máy tính biết cách nhận diện ranh giới giữa các từ trong tiếng Việt từ đó có thể tách từ được chính xác. Để thực hiện điều đó chúng ta sẽ cho mấy học các câu mẫu trong tập ngữ liệu đã được đánh dấu, tách từ đúng. Sau khi học xong máy sẽ xác định được các tham số (bộ luật) cần thiết cho mô hình nhận diện từ. Phương pháp TBL có nhược điểm là tốn rất nhiều thời gian để cho máy học và không gian nhớ do trong quá trình học máy sẽ sinh ra các bộ luật trung gian. Ngoài ra việc xây dựng một bộ luật đầy đủ để phân đoạn từ là công việc hết sức khó khăn do bộ luật được máy học tạo nên dựa trên tập ngữ liệu đã được đánh dấu. Cho nên sẽ có khá nhiều nhập nhằng trong việc xảy ra. Tuy nhiên sau khi sinh ra được bộ luật thì TBL tiến hành phân đoạn khá nhanh. Hơn nữa, ý tưởng của phương pháp này là rút ra quy luật ngôn ngữ từ những mẫu sẵn có và “sửa sai” liên tục trong quá trình học là phù hợp với bài toán xử lý ngôn ngữ tự nhiên.

**c) Phương pháp Weighted finite-state transducer (WFST)**

Phương pháp WFST (Weighted Finite-State Transducer) còn gọi là phương pháp chuyển dịch trạng thái hữu hạn có trọng số. Ý tưởng của phương pháp này vào phân đoạn tiếng Việt là các từ sẽ được gán trọng số bằng xác suất xuất hiện của từ đó trong ngữ liệu. Dùng WFST duyệt qua câu cần xét, cách duyệt có trọng số bé nhất sẽ được chọn là cách tách từ. Hoạt động của WFST có thể chia thành 3 bước như sau.

Xây dựng từ điển có trọng số: theo mô hình WFST việc phân đoạn từ được xem như là một sự chuyển dịch trạng thái có xác suất. Chúng ta miêu tả từ điển D là một đồ thị biến đổi trạng thái hữu hạn có trọng số. Giả sử:

+ H là tập các âm tiết của tiếng Việt (các tiếng).

+ P là tập các từ loại của Tiếng Việt

+ Mỗi cung D có thể là: -Từ một phần tử H đến một phần tử của H

- Từ phần tử ε (xâu rỗng) đến một phần tử của P.

Mỗi từ trong D sẽ được biểu diễn bởi một chuỗi các cung, bắt đầu bằng một cung tương ứng với một phần tử của H. Và kết thúc bằng một cung có trọng số tương ứng với một phần tử của ε x P. Trọng số biểu diễn chi phí ước lượng (estimated cost) được cho bằng công thức:

f: tần số xuất hiện của từ

N: kích thước tập mẫu

Xây dựng các khả năng phân đoạn từ: bước này thống kê tất cả khả năng phân đoạn của một câu. Giả sử câu có n tiếng, sẽ có 2n-1 cách phân đoạn khác nhau. Để giảm sự bùng nổ của các cách phân đoạn, thuật toán sẽ loại bỏ ngay những nhánh phân đoạn của những từ không xuất hiện trong từ điển. Lựa chọn khả năng phân đoạn tối ưu: sau khi liệt kê tất cả các khả năng phân đoạn từ, thuật toán sẽ chọn cách phân đoạn tốt nhất, đó là cách phân đoạn có trọng số bé nhất. Ví dụ: Input = “ tốc độ truyền thông tin sẽ tăng cao” (theo [10]). Trong từ điển trọng số chúng ta có trọng số của các từ lần lượt là:

Tốc độ = 8.68

Truyền = 12.31

Truyền thông = 12.31

Thông tin = 7.24

tin = 7.33

sẽ = 6.09

tăng = 7.43

cao = 6.95

Ta sẽ có các cách phân đoạn câu trên như sau:

ID 1 = “ tốc độ # truyền thông # tin # sẽ # tăng # cao “ = 8.68 + 12.31 + 7.33 + 6.09 + 7.43 +6.95 = 48.79

ID 2 = “ tốc độ # truyền # thông tin # sẽ # tăng # cao “ = 8.68 + 12.31 + 7.24 + 6.09 + 7.43 + 6.95 = 48.7

Do ID 2 nhỏ hơn ID 1 nên ID 2 là lựa chọn tốt hơn ID 1.

Ưu điểm của phương pháp này là cho độ chính xác khá cao, ngoài ra mô hình còn cho kết quả tách từ với độ tin cậy kèm theo (trọng số và xác suất). Tuy nhiên cũng như phương pháp TBL, để xây dựng tập ngữ liệu có xác suất là vô cùng công phu và tốn chi phí.

***2.1.2 Loại bỏ stop word***

Stop word hay còn gọi là từ dừng là những từ xuất hiện nhiều trong tất cả các văn bản thuộc mọi thể loại trong tập dữ liệu, hay những từ chỉ xuất hiện trong một và một vài văn bản. Nghĩa là stop word là những từ xuất hiện quá nhiều lần và quá ít lần. Chúng không có ý nghĩa và không chứa thông tin đáng giá để chúng ta sử dụng. Ví dụ như các từ: thì, là, mà, và, hoặc, bởi… Trong việc phân loại văn bản thì sự xuất hiện của những từ đó không những không giúp gì trong việc đánh giá phân loại mà còn nhiễn và giảm độ chính xác của quá trình phân loại. Trong luận văn này chúng tôi tiến hành tách stop word dựa trên tần suất xuất hiện của từ, và kết quả phân loại sau khi loại bỏ stop word hiệu quả hơn nhiều so với không thực hiện. (sẽ được trình bày cụ thể trong chương 5).

**2.2 Sơ đồ phân rã chức năng.**

### ***2.2.1 Biểu đồ tổng quát.***

### ***2.2.2 Chức năng quản lí nhật ký.***

### ***2.2.3 Chức năng quản lí ghi chú.***

### ***2.2.4 Chức năng quản lí hành trình.***

### ***2.2.5 Chức năng quản lí công việc.***

**2.3 Sơ đồ hoạt động.**

Có nhiều cách để chuyển đổi tin tức từ dạng ngôn ngữ tự nhiên (phi cấu trúc) sang dạng ngôn ngữ máy (ngôn ngữ có cấu trúc). Tuy nhiên, trong đề tài luận văn này chúng tôi chỉ tìm hiểu và sử dụng phương pháp biểu diễn văn bản theo mô hình không gian véc-tơ (véc-tơ space model). Đây là cách biểu diễn tương đối đơn giản và hiệu quả. Theo mô hình này, mỗi tin tức sẽ được biểu diễn thành một véc-tơ. Mỗi thành phần của véc-tơ là một từ riêng biệt trong tập tin tức gốc và được gán một giá trị là trọng số của từ đó trong tin tức đó. Do số lượng từ trong tập tin tức là rất nhiều, từ đó khi biểu diễn véc-tơ sẽ dẫn đến một vấn đề đó là tính nhiều chiều của véc-tơ. Để giải quyết vấn đề này, chúng tôi chỉ đưa ra phương pháp loại bỏ stop word để giảm bớt các từ không cần thiết, rút ngắn chiều của véc-tơ và nâng cao hiệu suất phân loại tin tức. Từ tập tin tức gốc ban đầu đã được phân chủ đề, sau khi tiến hành tách từ và loại bỏ stop word chúng tôi tiến hành xây dựng tập từ khóa dựa trên các từ riêng biệt còn lại. Giả sử chúng ta có một tập tin tức gồm m tin tức, \* + và ta có tập từ khóa gồm n từ \* +. Gọi \* + là ma trận trọng số, trong đó là giá trị trọng số của từ ti trong văn bản dj. Sau đây chúng tôi sẽ trình bày hai phương pháp xây dựng véc-tơ từ một tin tức dựa trên tập từ khóa đó.

***2.2.1 Binary vector***

Đây là mô hình biểu diễn véc-tơ với cách tính trọng số của mỗi thành phần véc-tơ cho ra hai giá trị duy nhất là 0 và 1. Nếu trong tin tức đó xuất hiện từ ti thì giá trị trọng số của từ đó trong véc-tơ đại diện sẽ là 1 và ngược lại là 0. Ta có thể biểu diễn nó thành công thức như sau:

***2.2.2 TF-IDF vector***

Tf – Idf (Term Frequency – Inverse Document Frequeny) là một độ đo, cũng có thể xem như là một giải thuật để xác định thứ hạng về một tiêu chí nào đó của từ (cụm từ). Giải thuật này cùng với mô hình không gian véc-tơ được sử dụng phổ biến rộng rãi trong nhiều lĩnh vực: Search engine, text mining…

Nguyên lý cơ bản của giải thuật này là độ quan trọng của một từ sẽ tỉ lệ thuận với số lần xuất hiện của nó trong một tin tức và tỉ lệ nghịch với số lần xuất hiện của nó trong các tin khác trong tập dữ liệu. Tùy mục đích sử dụng mà có nhiều công thức tính Tf-Idf, trong luận văn này, chúng tôi chỉ trình bày và sử dụng công thức phổ biến nhất.

**a) TF (TERM FREQUENCY)**

Độ đo tf được dùng để tính tần suất xuất hiện của từ t trong tập tin d. Một từ xuất hiện càng nhiều thì tf của nó càng lớn và ngược lại.

Cách đơn giản nhất để tính tf của từ t trong văn bản d là tính tần suất xuất hiện của t trong d

Trong đó: Ns(t): số lần xuất hiện từ t trong d, W: tổng số từ trong văn bản d. Ngoài công thức trên còn có một công thức đơn giản khác để tính tf đó là công thức tần số tăng cường:

Tử số là tần suất xuất hiện của từ t trong văn bản d. Mẫu số là tần suất xuất hiện của từ xuất hiện nhiều nhất trong văn bản d. Để cho đơn giản, trong luận văn này chúng tôi sử dụng công thức đầu.

Độ đo tf chỉ là tính độ quan trọng của từ ở mức độ cục bộ một tập tin. Chưa thế hiện được mức độ quan trọng của từ đó trong toàn bộ tập tin, do có nhiều stop word xuất hiện rất nhiều lần trong bất kì tập tin nào, vì thế chúng ta tiến hành tính idf để hạn chế mức độ quan trọng của những từ đó.

**b) IDF (INVERSE DOCUMENT FREQUENY)**

Độ đo Idf là tần số nghịch của 1 từ trong tập tin. Nó thể hiện mức độ quan trọng của một từ ở mức độ toàn cục. Tính idf để giảm giá trị của những từ phổ biến.

Trong đó D là số lương tập tin có trong tập dữ liệu và d là số lượng tập tin có trong tập dữ liệu mà nó chứa từ t.

Trong trường hợp nếu t không xuất hiện trong bất kì văn bản d nào của tập D. Thì mẫu số bằng 0, phép chia không hợp lệ, vì thế người ta thường thay mẫu thức bằng 1+ (d D: t d) việc này không làm ảnh hưởng nhiều đến kết quả tính toán.

Chúng ta có thể nhận thấy rằng nếu một từ xuất hiện càng nhiều trong các tập tin của tập dữ liệu thì giá trị idf của nó càng nhỏ và ngược lại. Nghĩa là từ có IDF nhỏ có thể là từ quan trọng, còn từ có IDF lớn chắc chắn là từ phổ biến và cần loại bỏ để tránh gây nhiễu kết quả. Việc một từ có IDF nhỏ có phải là từ quan trọng hay không còn phụ thuộc vào độ đo TF của từ đó, do có những từ hiếm gặp, có thề chỉ xuất hiện trong một vài tập tin của tập dữ liệu nhưng nó không có ích gì trong việc phân loại tin tức đó.

Để xác định từ quan trọng chúng ta tiến hành tính TF-IDF:

Từ có độ đo TF-IDF càng lớn thì nó càng đáng giá và ảnh hưởng càng nhiều đến việc phân loại tin tức. Trong TF-IDF véc-tơ, nếu một từ ti xuất hiện trong văn bản dj thì trọng số của từ đó trong véc-tơ đại diện sẽ là giá trị TF-IDF(ti, dj), ngược lại là 0. Ta có thể biểu diễn nó thành công thức như sau:

# THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

**3.1 Hiện trạng tổ chức**

Các vấn đề mà sinh viên cần được hỗ trợ nhiều hơn tại Trường Đại học Giao thông Vận tải Phân hiệu tại thành phố Hồ Chí Minh :

- Hiện tại nhà trường đã có kênh thông tin giúp sinh viên giải đáp thắc mắc là Diễn đàn Nghe sinh viên nói - UTC2 (<http://facebook.com/nghesinhviennoi.utc2>). Sinh viên sẽ tiến hành đặt câu hỏi qua một đường dẫn cho trước, chờ câu hỏi được xem xét và gửi đến những phòng/ban hoặc cán bộ để giải đáp thắc mắc.

- Sinh viên khi muốn tính điểm tích lũy hay muốn biết mình học môn đó điểm này thì điểm tích lũy sẽ thay đổi như thế nào. Tất cả đều phải tính bằng tay và khó khăn trong quá trình tính toán.

**3.2 Yêu cầu hệ thống**

***3.2.1 Yêu cầu chức năng***

- Diễn đàn giúp sinh viên đăng bài viết, câu hỏi , các thắc mắc cần được giải đáp

- Trang dự đoán điểm giúp sinh viên có thể tạo ra bảng điểm mục tiêu của riêng mình bằng tùy chỉnh hoặc hệ thống dự đoán điểm.

- Kênh chat để sinh viên trao đổi trực tiếp với thầy cô, cán bộ phòng/ban về các thắc mắc cần được giải đáp gấp.

- Trang quản trị quản lý tất cả các chức năng trên.

***3.2.2 Yêu cầu phi chức năng***

Ứng dụng và trang web cần sử dụng các thư viện có tính ổn định cao để đảm bảo chương trình hoạt động tốt trong thời gian dài.

Giao diện đẹp, đơn giản, thân thiện với người dùng. Không bị vỡ giao diện khi thay đổi kích thước ứng dụng cũng như trang web.

Màu sắc không quá bắt, mang đặc trưng của Trường Đại học Giao thông Vận tải Phân hiệu tại thành phố Hồ Chí Minh.

Cả ứng dụng in bằng và trang web tra cứu văn bằng phải dễ sử dụng cho người dùng mới, không quá phức tạp, hoạt động ổn định với lượng dữ liệu lớn.

Tốc độ mở ứng dụng và tốc độ tải dữ liệu từ trang web phải tương đối nhanh, không để người dùng đợi lâu.

Khi có thay đổi trong quá trình xử lý, ứng dụng và trang web phải có thông báo để người dùng biết những gì đang xảy ra.

Trang web có thể tùy chỉnh giao diện dễ dàng.

**3.3 Thu thập dữ liệu**

Thu thập dữ liệu là một trong những bước tiền đề quan trọng để huấn luyện tạo ra mô hình phân lớp dùng để phân loại văn bản. Dữ liệu càng nhiều, càng đa dạng thì mô hình phân lớp huấn luyện được càng bao quát và đầy đủ hơn.

***3.3.1 Xử lý dữ liệu ở dạng html***

Dữ liệu câu hỏi và trả lời được lấy từ các bài viết, bài đăng trên Diễn đàn Nghe sinh viên nói - UTC2 theo cấu trúc:

**<html ..>**

**…**

**<div class="pam \_3-95 \_2pi0 \_2lej uiBoxWhite noborder">**

**<div class="\_3-96 \_2pio \_2lek \_2lel">Diễn đàn Nghe sinh viên nói - UTC2 đã cập nhật trạng thái của mình.</div>**

**<div class="\_3-96 \_2let">**

**<div class="\_2pin">**

**<div>#9087:<br /> Ad cho em hỏi môn đóng học phí bổ sung để mở lớp khi nào nhà trường bố trí lịch cho bọn em học vậy ad. Em thấy lịch học kỳ phụ đợt 2 không có môn đó. Em k56 sắp nhận đồ án tốt nghiệp, hồi giờ cũng chưa học lại bao giờ nên thấy lo lắng lắm. Mong ad phản hồi sớm. Cảm ơn ad nhé!<br /> ---<br /> Ad: bạn liên hệ trực tiếp phòng 10D3 (trường hợp đặc biệt)</div>**

**Câu hỏi**

**Và**

**Trả lời**

**</div>**

**</div>**

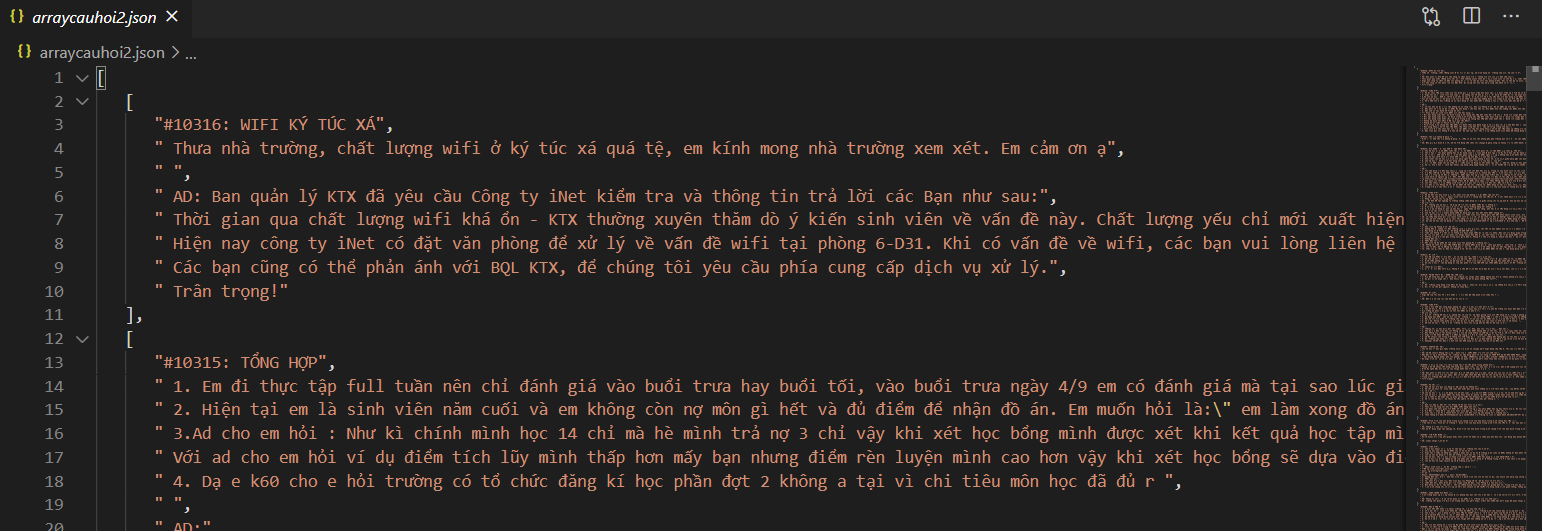
**<div class="\_3-94 \_2lem">11:31, 15 tháng 8, 2019</div>**

**</div>**

**…**

**<html>**

Thực hiện lấy dữ liệu trong những thẻ **<div class=”\_2spin”>**  lưu lại dưới dạng json trong hình 3.1.

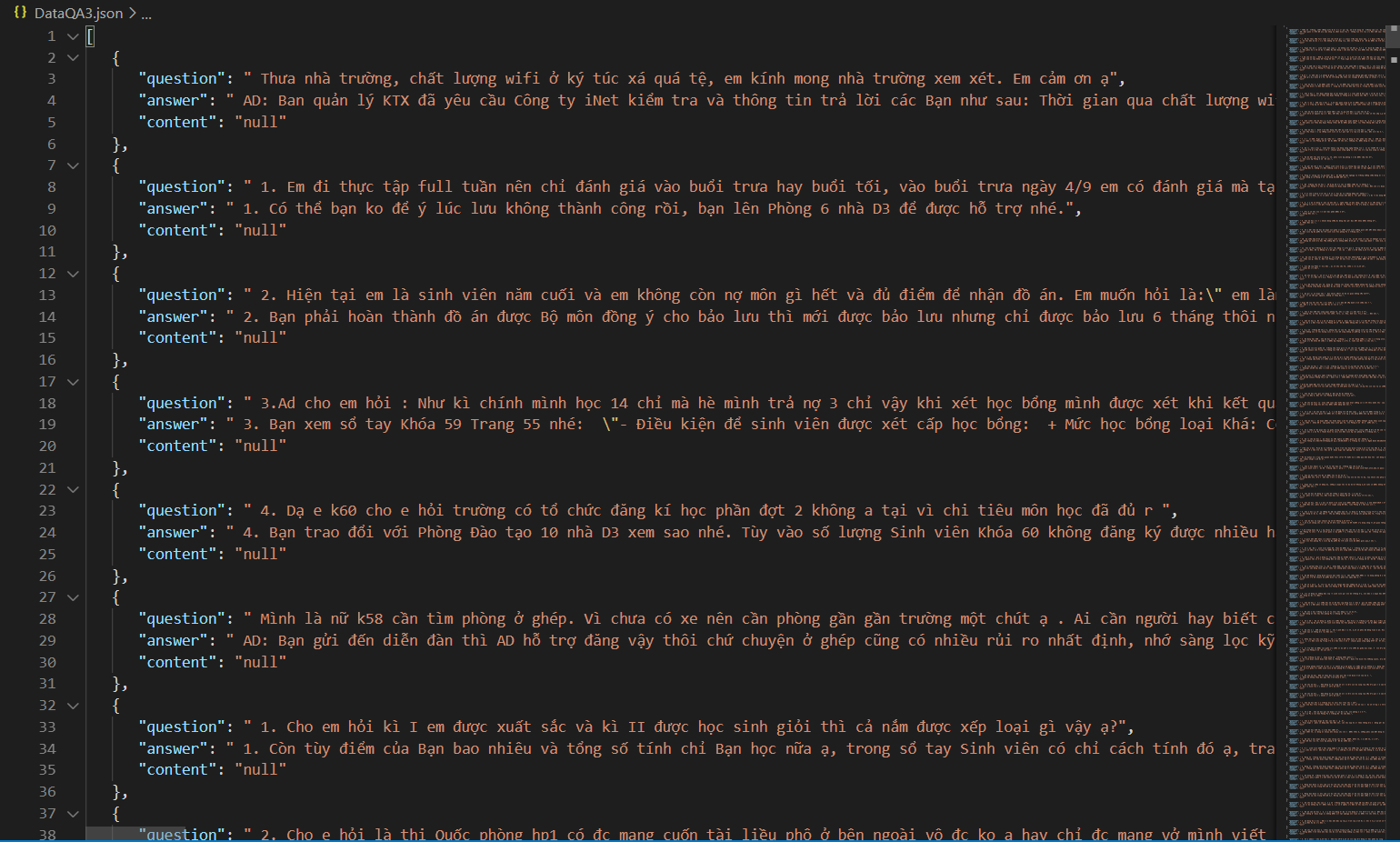


*Hình 3.1 Dự liệu sau khi lấy trên diễn đàn*

***3.3.2 Xử lý dữ liệu câu hỏi và trả lời***

Dữ liệu sau khi lấy ở hình 3.1 chưa được tổ chức dưới đạng câu hỏi và câu trả lời nên cần phải được phân loại và chuẩn hóa.

Nhận thấy các quy tắc và tổ chức của dữ liệu sau khi thu thập dùng Python để chuẩn hóa câu hỏi và câu trả lời.



*Hình 3.2 Dữ liệu sau khi được chuẩn hóa*

***3.3.3 Viết form và phân loại câu hỏi***

Dữ liệu sau khi đã chuẩn hóa cần được phân loại để tiếp tục tiến hành train.



*Hình 3.3 Form phân loại câu hỏi sinh viên*

3.4 Thuật toán áp dụng

***3.4.1 Bài toán phân loại câu hỏi và đề xuất câu hỏi tương tự***

**a) Đặt vấn đề**

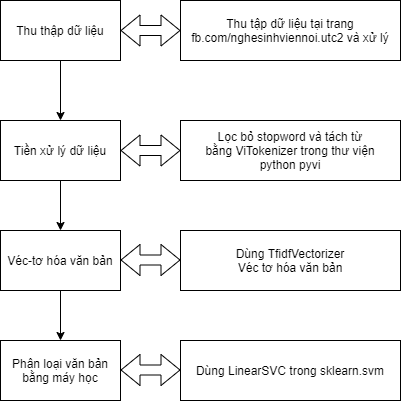
Câu hỏi và câu trả lời trong hệ thống hỏi đáp sinh viên mang những đặc điểm riêng, đó là ở dưới dạng văn bản tự do, không theo một loại câu hỏi nhất định nào, cũng không theo một chủ đề nhất định nào cả. Do đó, một phần hết sức quan trọng trong hệ thống này là phân tích câu hỏi như thế nào lấy được thông tin nhiều nhất khi mà câu hỏi như thế nào để lấy được thông tin nhiều nhất mà câu hỏi không hề có một cấu trúc nhất định nào cả. Hầu hết các hệ thống hỏi-đáp truyền thống đều chỉ trả lời cho các câu hỏi thuộc về một loại câu hỏi nào đó. Do đó, phương pháp mà chúng tôi chọn thử nghiệm cho hệ thống diễn đàn sinh viên là phương pháp dựa trên từ khóa, trích từ khóa và đánh trọng số cho các từ khóa trong văn bản đề tìm kiếm câu hỏi. Ngoài ra, nhằm cải thiện hiệu quả hệ thống, giảm không gian tìm kiếm, trước khi tìm kiếm, các cặp hỏi-đáp được phân thành các cụm gồm các câu hỏi tương tự nhau.

Chúng tôi tiến hành thử nghiệm các phương pháp đề xuất, cải thiện hiệu quả hệ thống trên mỗi bước phù hợp với dữ liệu của diễn đàn sinh viên.

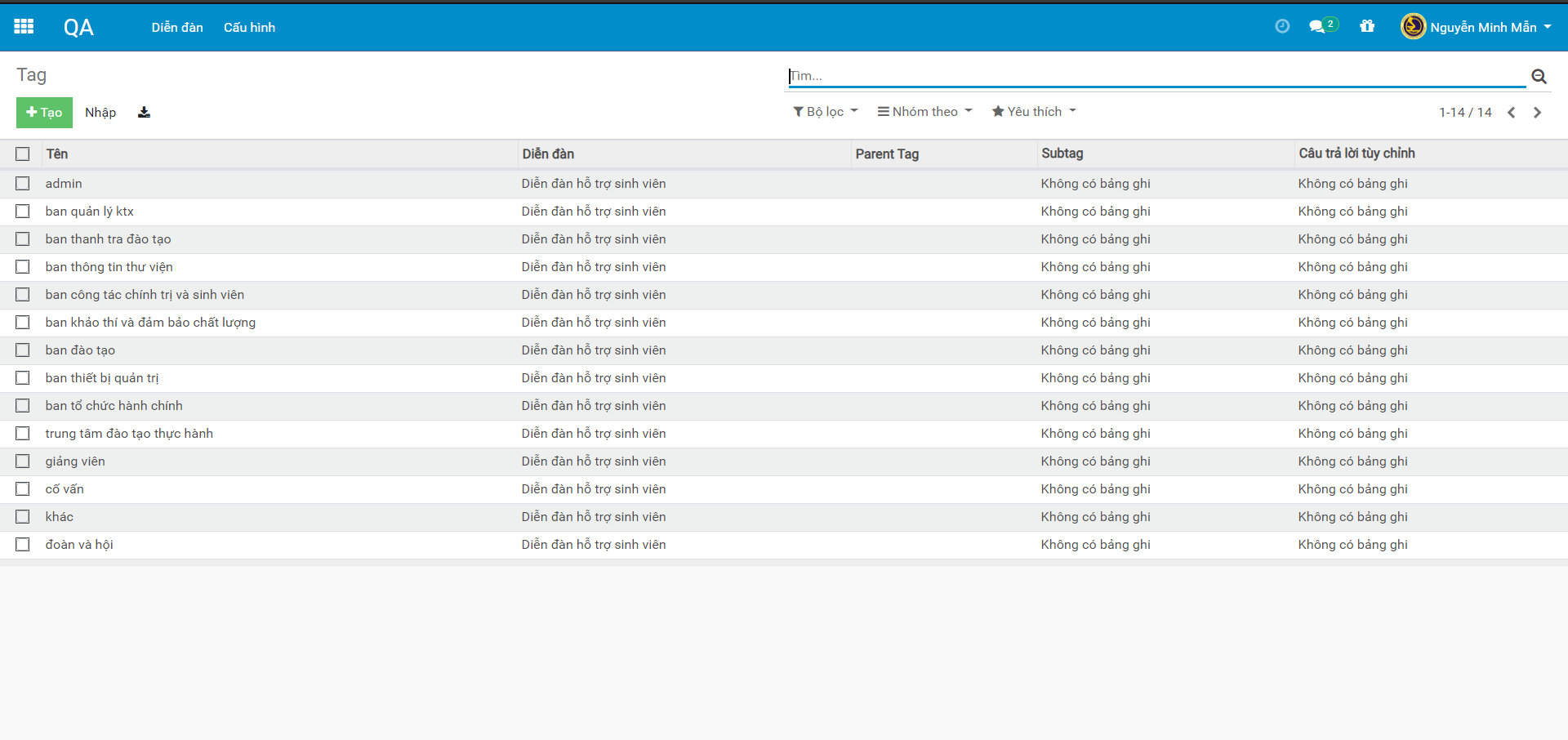
**b) Giải pháp**

Kiến thúc hệ thống diễn đàn sinh viên gồm 2 phần chính yếu nhất là phân loại tag (phân loại câu hỏi), tìm kiếm bài viết/ câu hỏi tương tự trên diễn đàn để người dùng tham gia sau khi đặt câu hỏi.

Quá trình mô hình hóa phân loại văn bản được chia làm 4 giai đoạn:

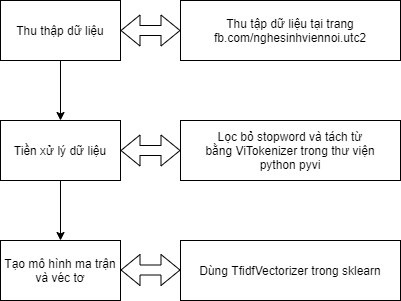


*Hình 3.4 Quy trình hóa phân loại văn bản*



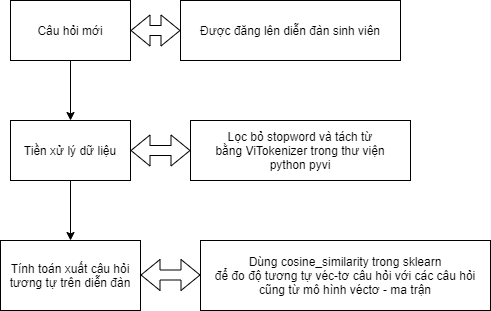
*Hình 3.5 Các tag phân loại trên diễn đàn*

Quá trình tạo mô hình véc-tơ và ma trận câu hỏi:



*Hình 3.6 Quá trình tạo mô hình véc- tơ và ma trận câu hỏi*

**b) Quá trình xử lý khi đưa câu hỏi lên hệ thống**



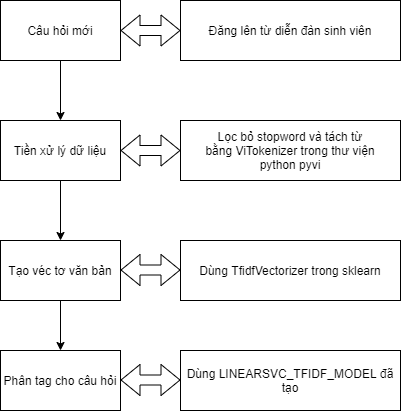
*Hình 3.7 Quá trình chọn câu hỏi-câu trả lời tương tự trên diễn đàn*

Câu hỏi mới được đưa vào hệ thống sẽ được tách câu, tách từ, loại bỏ các hư từ và loại bỏ các cụm từ xuất hiện nhiều nhưng không có ý nghĩa, để còn lại các từ cần thiết (từ khoá) cho việc phân loại và tìm câu tương tự.

Các từ khóa thu được trong giai trước sẽ được sử dụng để xây dụng vector đặc trưng, sau đó xác định các cụm chứa các câu hỏi tương tự nhất. Vector bài viết/câu hỏi sẽ được so khớp với tất cả các vector câ u hỏi trong các cụm đó theo độ đo tương tự cosin. Các giá trị tương tự này được xếp hạng, và hệ thống chọn nQ câu hỏi có giá trị tương tự cao nhất đứa vào giai đoạn tiếp theo.

nQ câu hỏi này được chuyển sang giai đoạn so khớp câu trả để tìm nQ câu trả lời tương ứng. Trong giai đoạn này, vector truy vấn sẽ được so khớp với vector của các câu trả tìm được. Một chiến lược xếp hạng được sử dụng để câu trả lời tốt nhất.

**b) Giai đoạn phân tag cho câu hỏi**



*Hình 3.8 Quá trình phân tag cho câu hỏi*

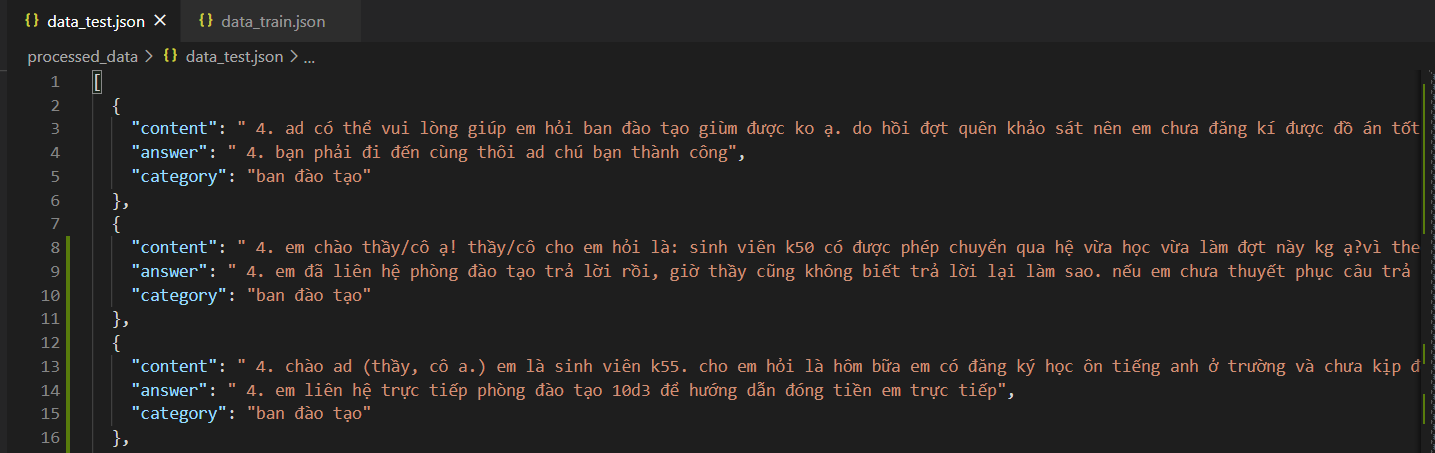
Các từ khóa thu được sau giai đoạn tiền xử lý sẽ được chuyển vào giao đoạn tiếp để tạo véc tơ văn bản sau đó dùng LINEARSVC\_TFIDF\_MODEL đã tạo để dự đoán, tính toán tag cho câu hỏi.

Ứng dụng SVM để phân loại tag cho câu hỏi:

+ Chia thành 2 tập dữ liệu data\_train.json và data\_test.json

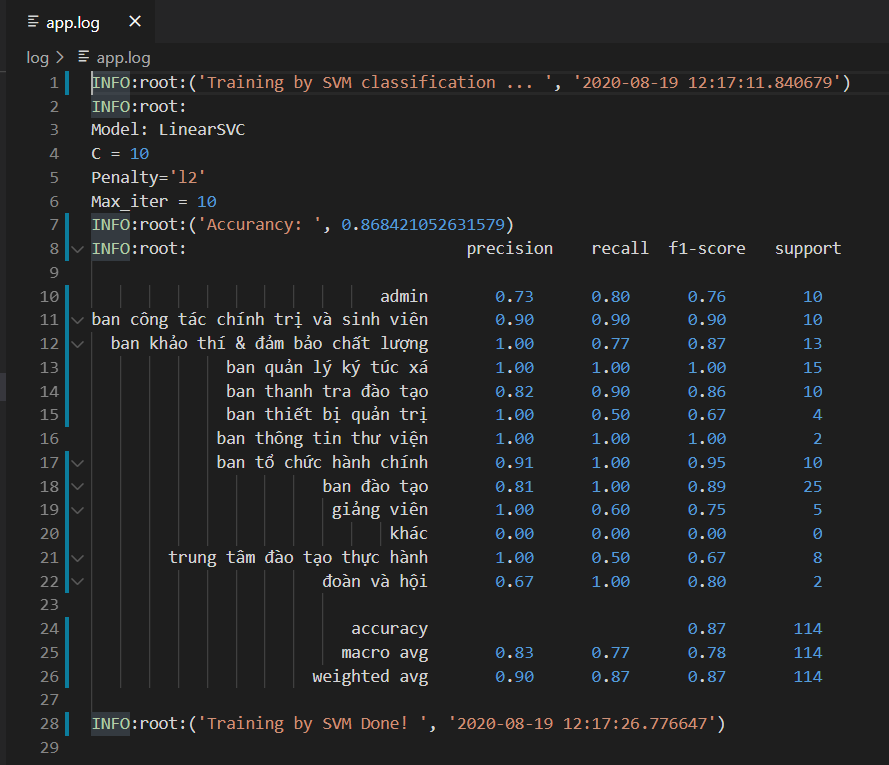
*Bảng 3.1 Bảng thống kê số câu train và câu test của các ban ngành/phòng ban*

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên ban ngành/phòng ban** | **Train** | **Test** |
| **1** | Phòng tổ chức hành chính | 167 | 10 |
| **2** | Phòng đào tạo | 572 | 15 |
| **3** | Phòng công tác sinh viên | 431 | 10 |
| **4** | Phòng khảo thí và đảm bảo chất lượng đào tạo | 82 | 13 |
| **5** | Phòng thiết bị quản trị | 14 | 4 |
| **8** | Ban quản lý ký túc xá | 84 | 15 |
| **9** | Ban thanh tra | 118 | 10 |
| **10** | Trung tâm thông tin thư viện | 10 | 2 |
| **11** | Trung tâm đào tạo thực hành | 15 | 8 |
| **12** | Admin | 191 | 10 |
| **13** | Giảng viên | 43 | 5 |
| **14** | Đoàn và hội | 12 | 2 |



*Hình 3.9 Tổ chức tập dữ liệu train và test*

+ Thực hiện train dữ liệu vào tạo mô hình LinearSVC



*Hình 3.10 Kết quả sau khi train tập dữ liệu trên*

***3.3.2 Bài toán dự đoán điểm***

**a) Đặt vấn đề**

Sinh viên khi học tập tại trường Đại học GTVT phân hiệu tại TP.HCM luôn mong muốn có thành tích cao và ra trường đúng hạn cần phải đặt mục tiêu tính toán điểm, chiến thuật dự đoán điểm hợp lý. Nhưng việc tính toán điểm mục tiêu đặt ra với toàn bộ chương trình sẽ rất khó khăn vì số điểm mục tiêu không giống với thực tế nên việc phải tính toán từ đầu tốn nhiều thời gian và công sức.

Mặc dù có nhiều tài liệu về việc dự đoán kết quả học tập của học sinh, sinh viên nhưng các nghiên cứu đối với các chương trình đại học còn ít vì một số khác biệt so với các chương trình khác. Thứ nhất là đối với chương trình đại học, mỗi sinh viên có nền tảng khác nhau, ngành nghề khác nhau, được chọn những môn học khác nhau dẫn đến khó thống kê hết toàn bộ. Thứ hai là một số môn học không có nhiều thông tin để có thể đưa ra dự đoán chính xác. Ngoài ra còn vì một số lý do nhỏ khác.

Đối với trong nước, qua tìm hiểu thì việc ứng dụng Machine Learning vào việc dự đoán kết quả học tập cho sinh viên còn chưa được chú trọng nhiều. Phần lớn việc ứng dụng Machine Learning tập trung nhiều vào việc phân tích, xử lý hình ảnh, nhận dạng khuôn mặt và khai phá dữ liệu. Đây cũng điều kiện thúc đẩy tìm hiểu đề tài này.

b) Giải pháp

Qua tìm hiểu thuật toán hồi quy tuyến tính thích hợp với bài toán dự đoán điểm sinh viên dự vào các môn liên quan. Ví dụ: Dựa vào điểm của môn tin học đại cương và môn lập trình nâng cao sẽ dự đoán ra điểm môn lập trình hướng đối tượng.

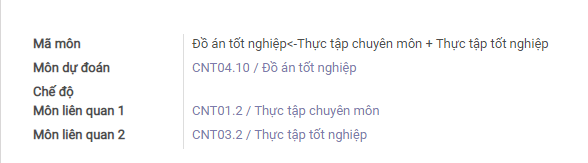
Xây dựng hệ thống giúp sinh viên dự đoán điểm, quản lý mục tiêu của mình. Hệ thống sẽ cho ra 3 loại điểm tích lũy:

+ Điểm tích lũy hiện tại: Hệ thống tính điểm tích lũy hiện tại của sinh viên

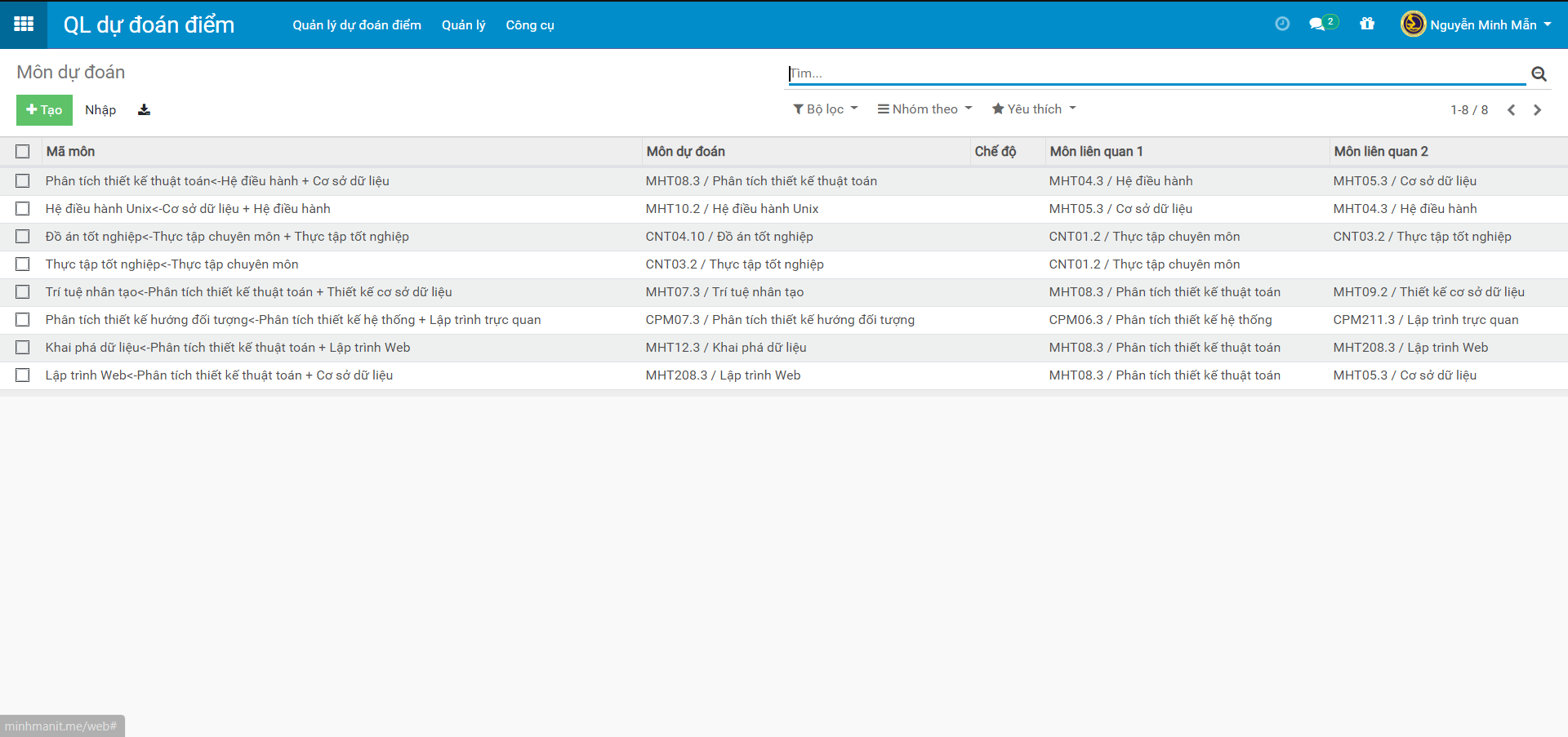
+ Điểm tích lũy mục tiêu: Sinh viên có thể cập nhật điểm mục tiêu của từng môn trong chương trình học từ đó hệ thống sẽ cho ra điểm tích lũy mục tiêu.

+ Điểm tích lũy dự đoán: Sinh viên chọn công thức dự đoán trên những môn để hệ thống dự đoán.

Công thức dự đoán được giảng viên tạo ra bằng các chọn 1 hoặc 2 môn liên quan và một môn đích. Về thời gian sau có thể tích độ tin cậy của công thức bằng nhưng kết quả điểm thực của sinh viên.



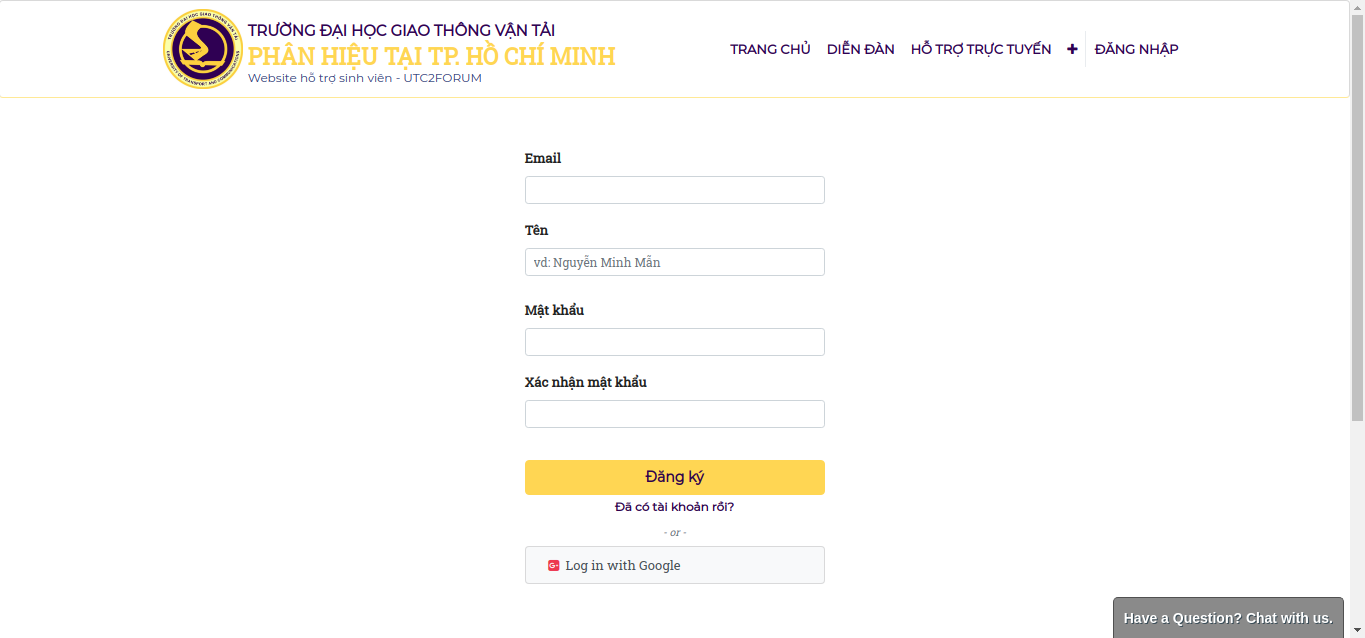
*Hình 3.11 Công thức dự đoán điểm*



*Hình 3.12 Quản lý công thức dự đoán điểm*

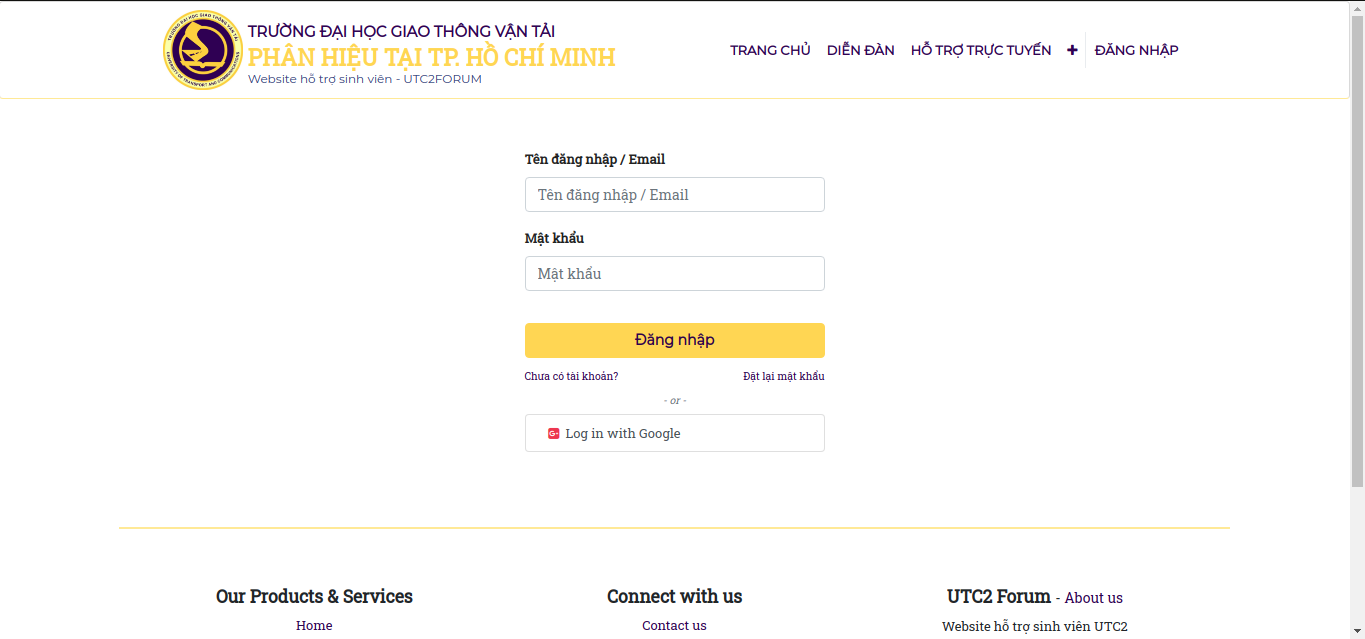
## **3.7 Xây dựng giao diện chương trình**

***3.7.1 Tạo tài khoản và đăng nhập***



*Hình 3.22 Giao diện tạo tài khoản đăng nhập*

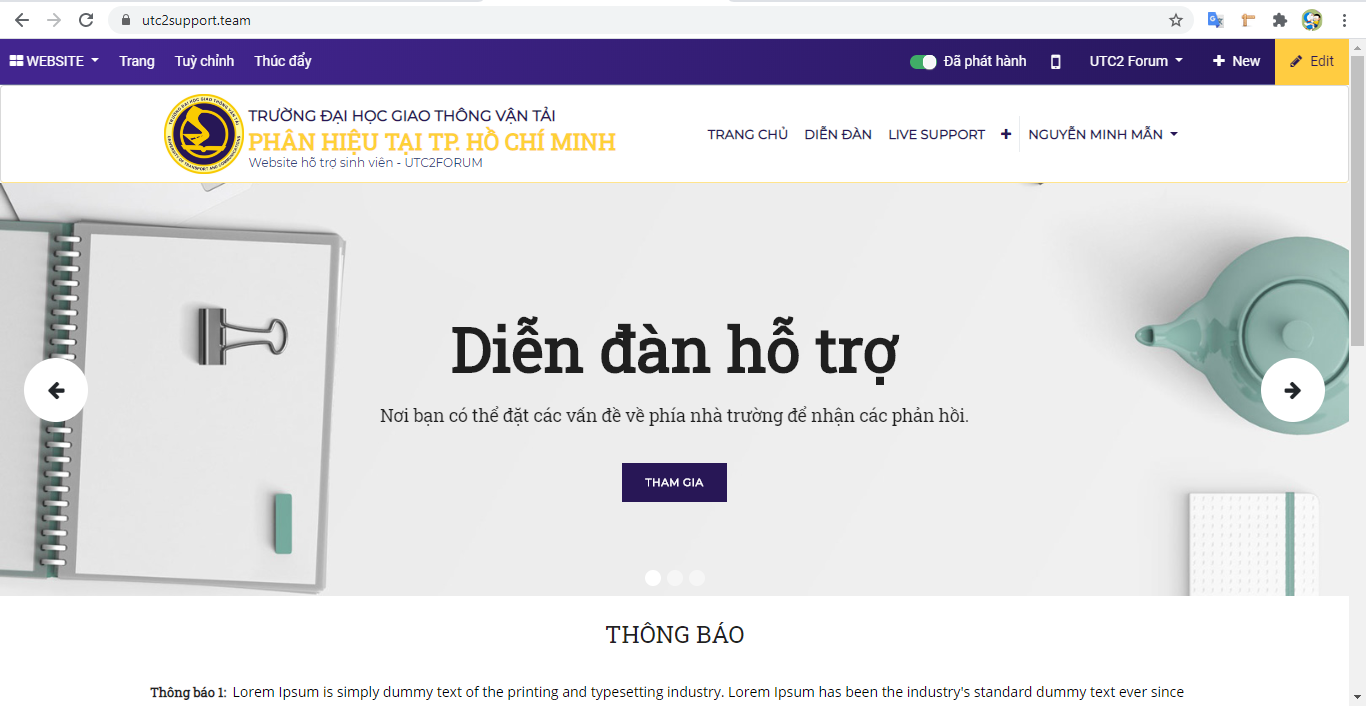
Đối với các sinh viên trường đại học Giao Thông Vận Tải phân hiệu tại TP. Hồ Chí Minh sẽ được cấp một tài khoản gmail. Các bạn sinh viên sử dụng gmail đã được trường cấp để tạo tài khoản đăng nhập.



*Hình 3.23 Giao diện đăng nhập*

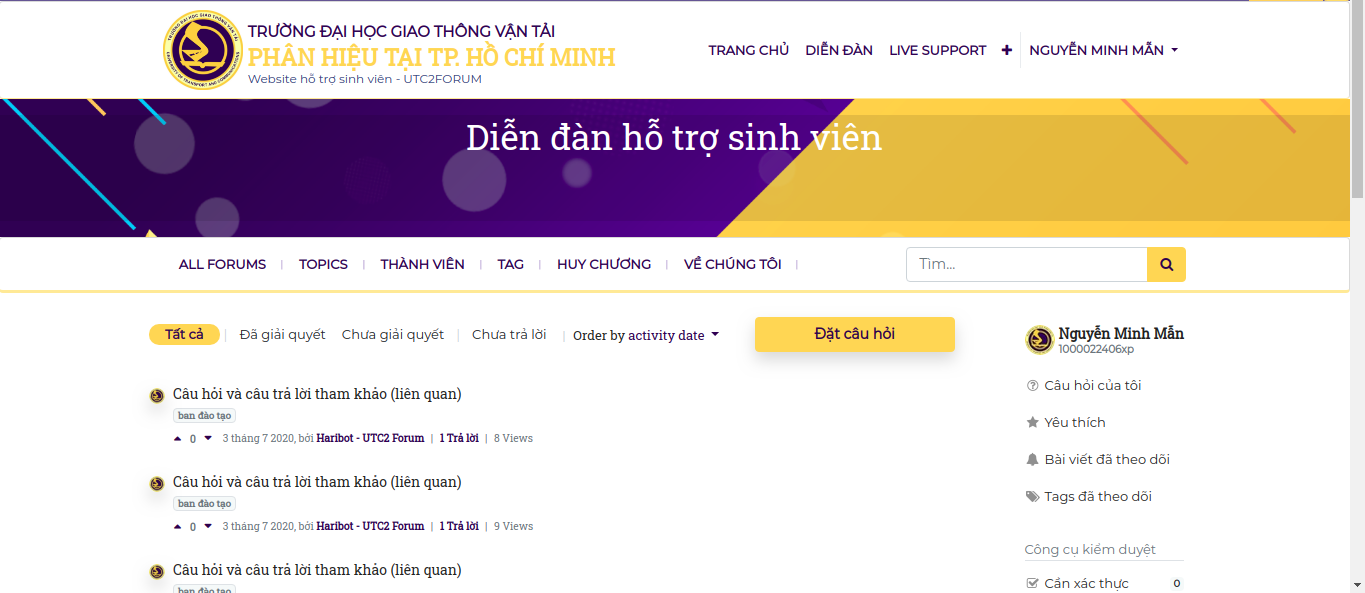
Sau khi đã có tài khoản, người dùng có thể đăng nhập để vào trang web, ngoài ra có thể đăng nhập với tài khoản google

***3.7.2 Giao diện trang chủ website***



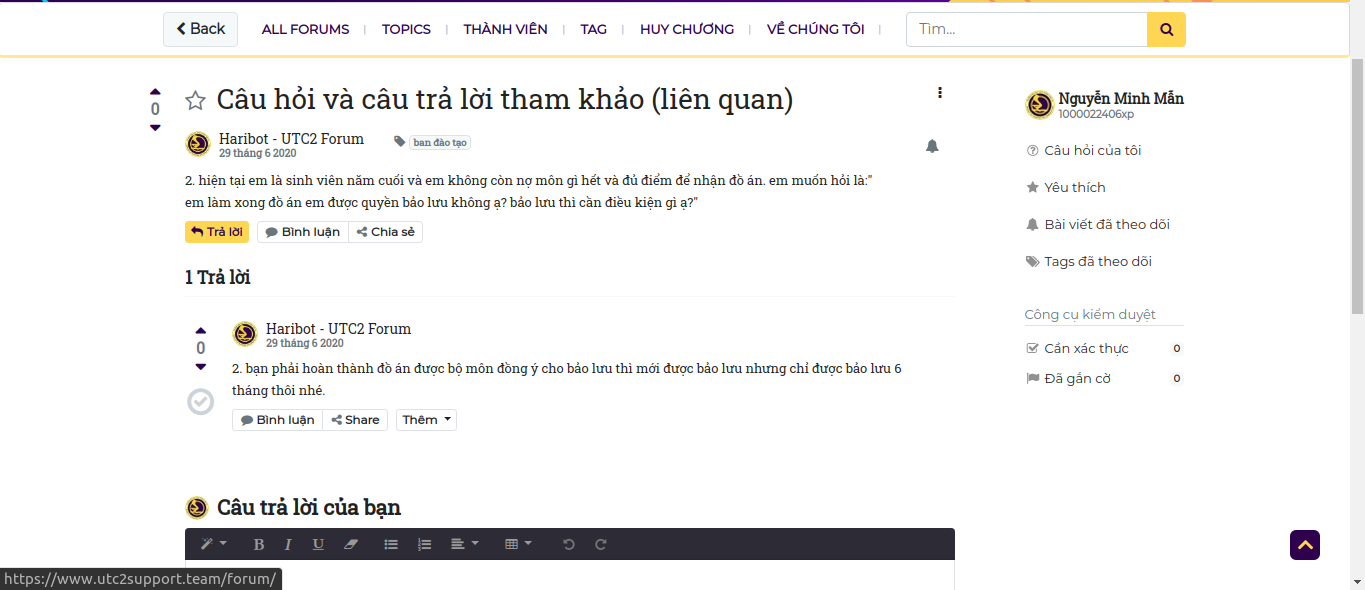
*Hình 3.24 Giao diện trang chủ của website*

***3.7.3 Tương tác bài viết***



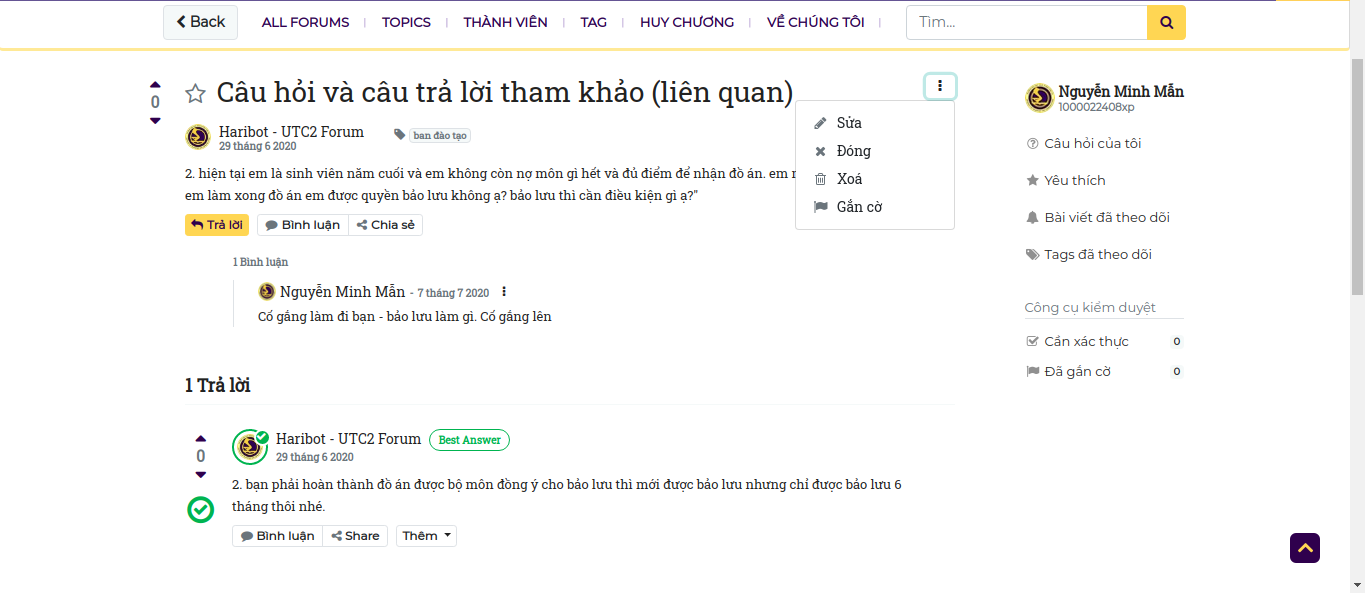
*Hình 3.25 Giao diện đặt câu hỏi trên diễn đàn*

Sau khi đăng nhập vào trang web, người dùng có thể theo dõi các câu hỏi trên diễn đàn, cũng như có thể đăng các câu hỏi.



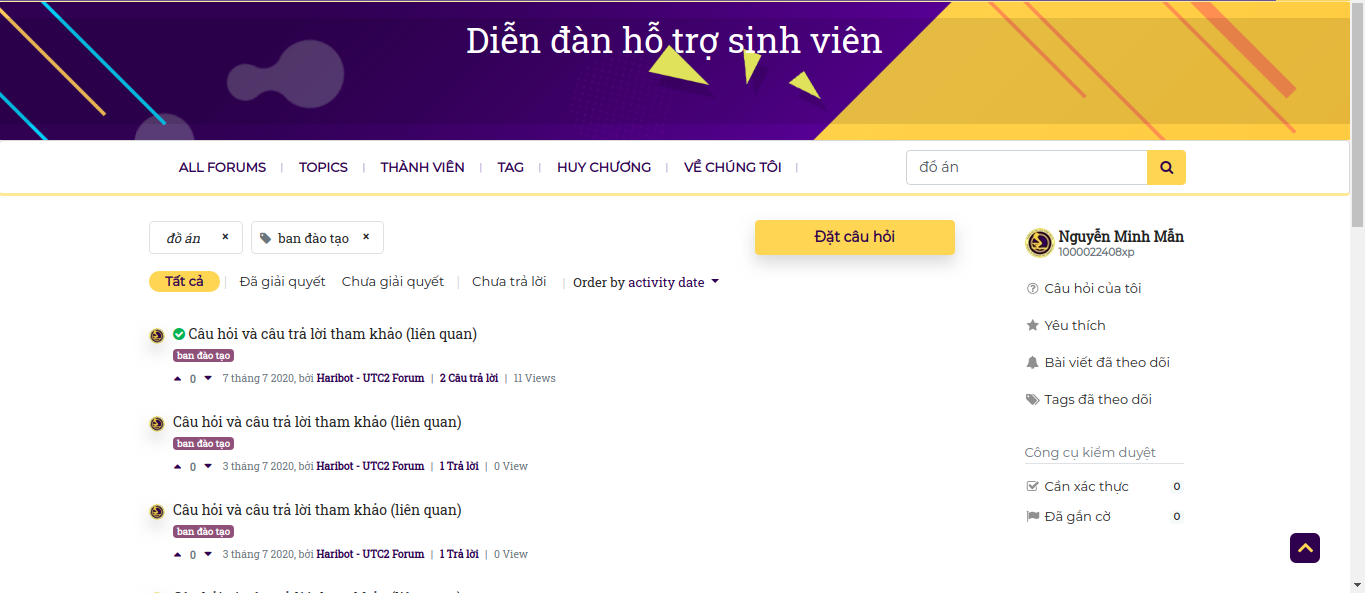
*Hình 3.26 Giao diện trả lời câu hỏi*

Sau khi đăng câu hỏi lên diễn đàn người dùng sẽ nhận được các đáp án phù hợp với câu hỏi của mình, ngoài ra các người dùng có thể nhận thêm được các câu trả lời từ giảng viên, các bạn sinh viên trong trường.



*Hình 3.27 Các chức năng thao tác với câu hỏi*

Người dùng có thể tương tác với bài viết như: Sửa, đóng, xóa, gắn cờ.



*Hình 3.28 Giao diện chức năng tìm kiếm câu hỏi*

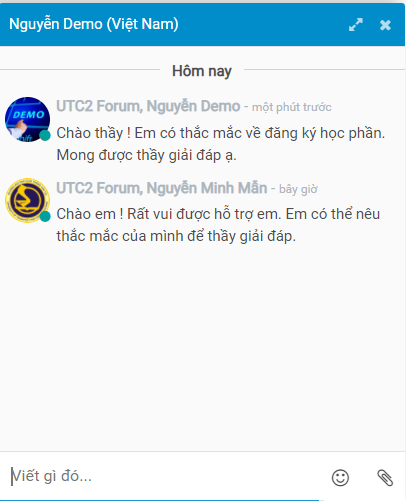
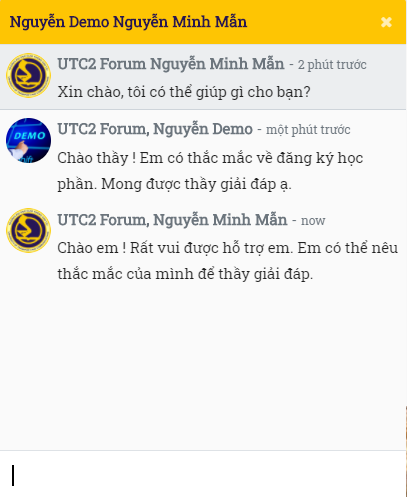
Người dùng chọn tìm kiếm câu hỏi và gõ các từ khóa cần tìm để tìm ra câu hỏi mà mình mong muốn

***3.7.4 Chat trực tuyến***



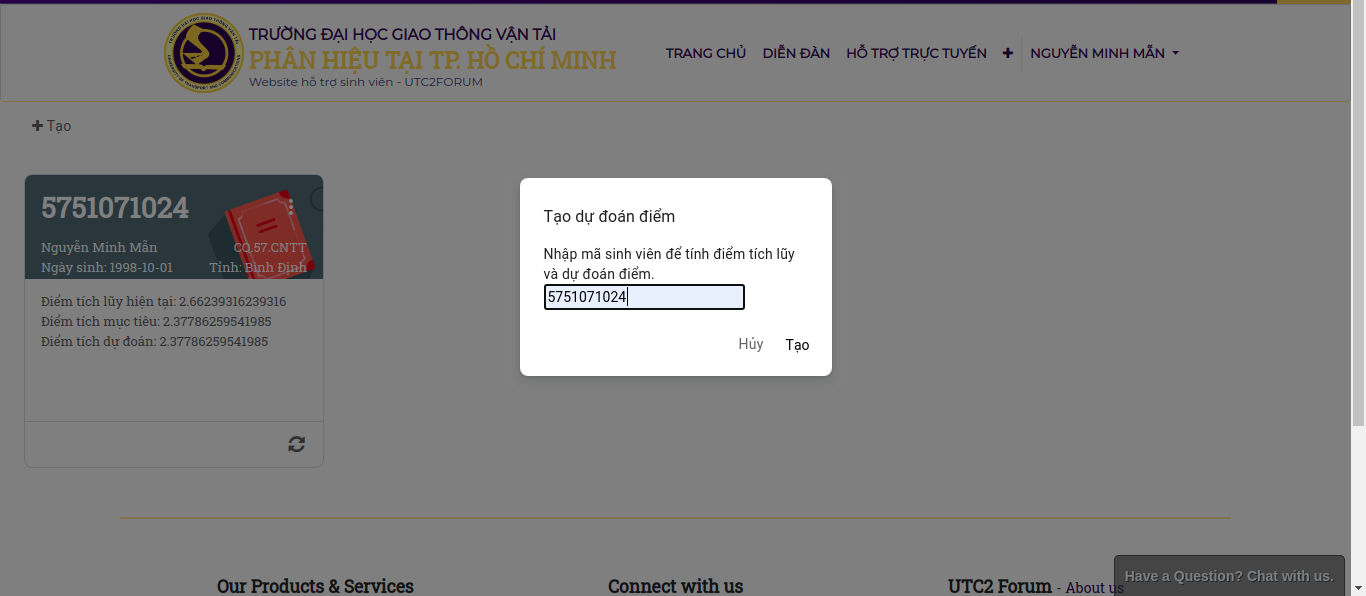
*Hình 3.29 Giao diện chức năng chat trực tuyến*

Khi người dùng chọn chức năng chat trực tuyến sẽ hiện ra kênh chat hỗ trợ trực tuyến, với các phòng ban như: Hỗ trợ- tư vấn, BQL KTX… Người dùng chọn phòng ban muốn hỏi và thực hiện chat.



*Hình 3.30 Giao diện người dùng sử dụng chức năng chat trực tuyến*

***3.7.5 Dự đoán điểm***



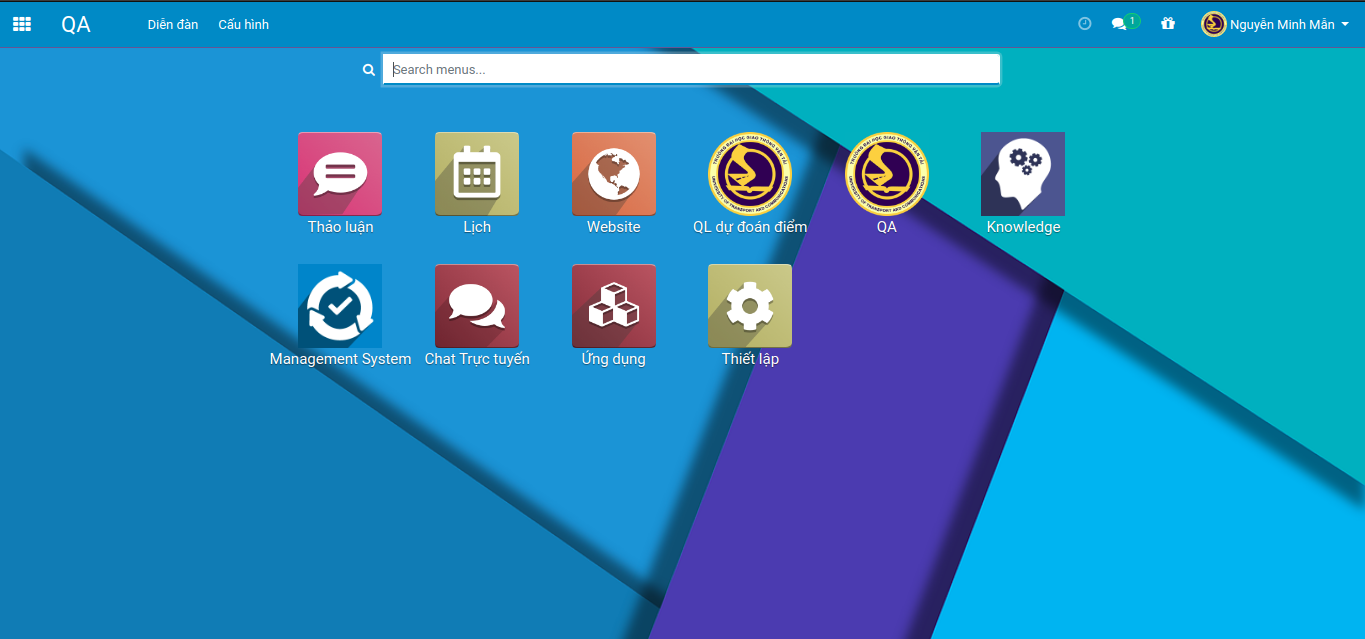
*Hình 3.31 Giao diện chức năng dự đoán điểm*

Khi người dùng chọn chức năng dự đoán điểm sẽ hiện ra thông tin yêu cầu nhập mã sinh viên để theo dõi điểm cũng như dự đoán điểm số



*Hình 3.32 Giao diện bảng điểm sau khi người dùng chọn chức năng dự đoán điểm*

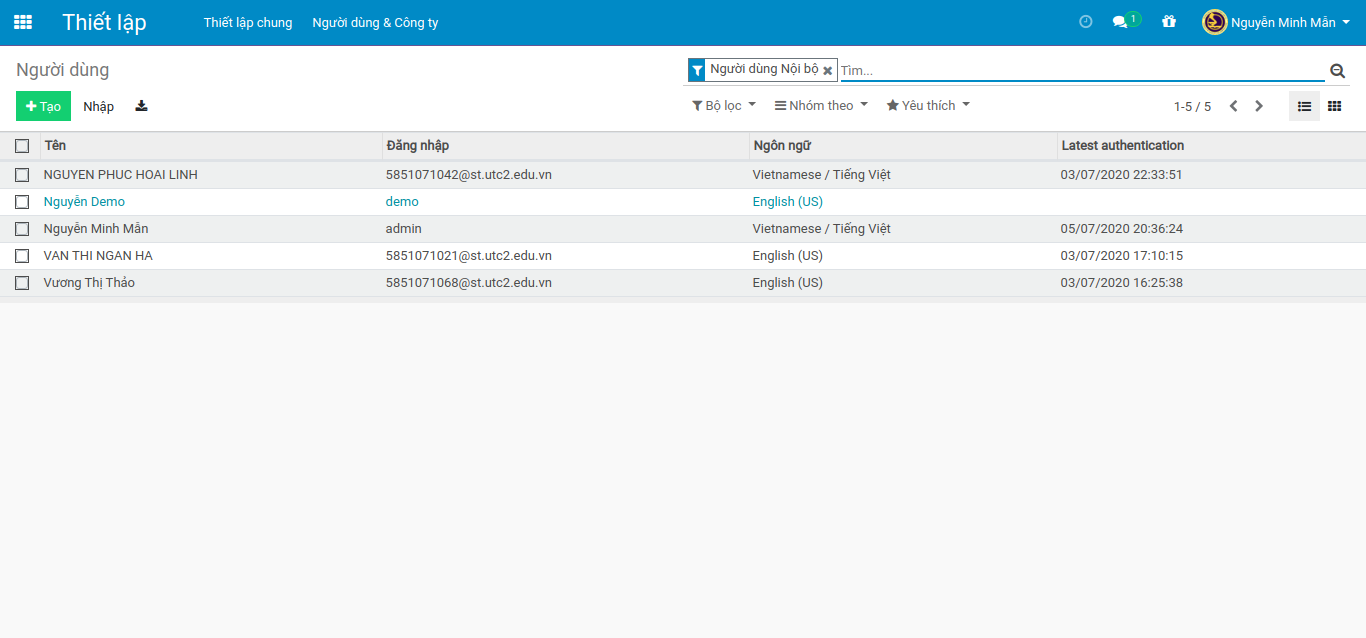
***3.7.6 Nhóm chức năng quản lý***



*Hình 3.33 Giao diện nhóm chức năng quản lý*

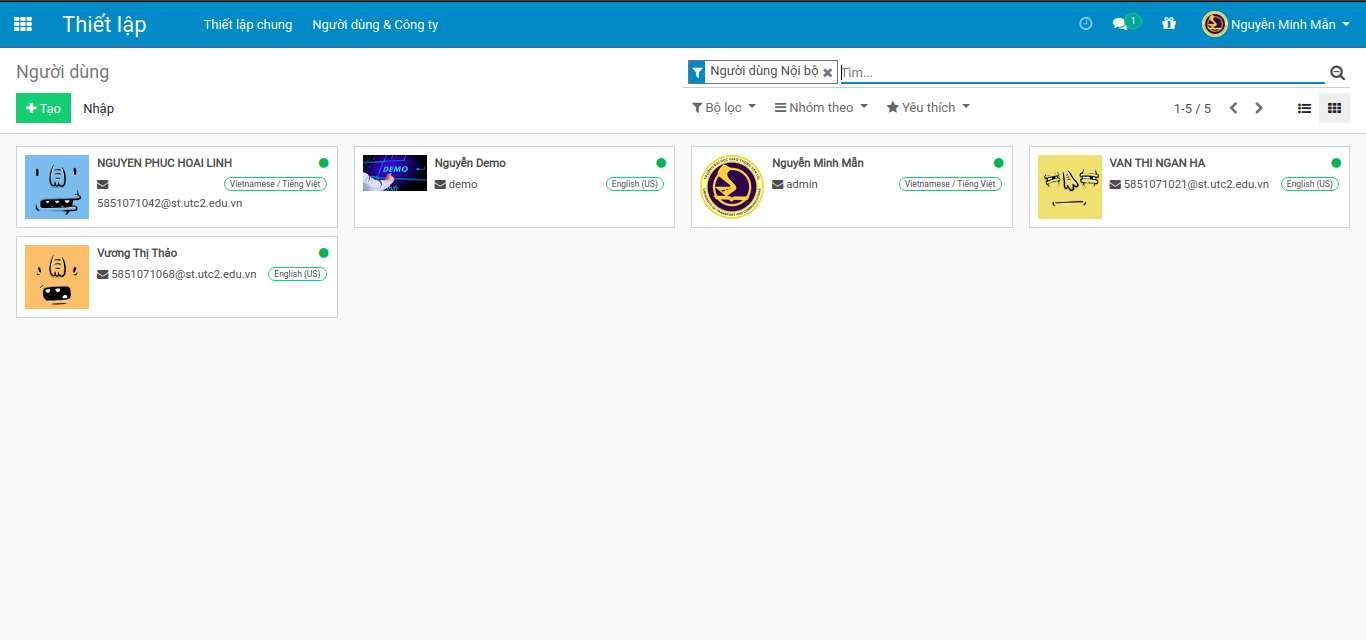
Giảng viên, quản lý(admin) của trang web sẽ có quyền thực hiện các chức năng quản lý website như: Quản lý dự đoán điểm, thiết lập cài đặt cho trang web, quản lý dự đoán điểm, quản lý người dùng, …

***3.7.7 Quản lý người dùng***



*Hình 3.34 Giao diện quản lý người dùng*

Chức năng quản lý người dùng do quản ly(admin) của trang sử dụng, dùng để phân quyền, thêm quyền khi muốn thêm người quản lý trang web.

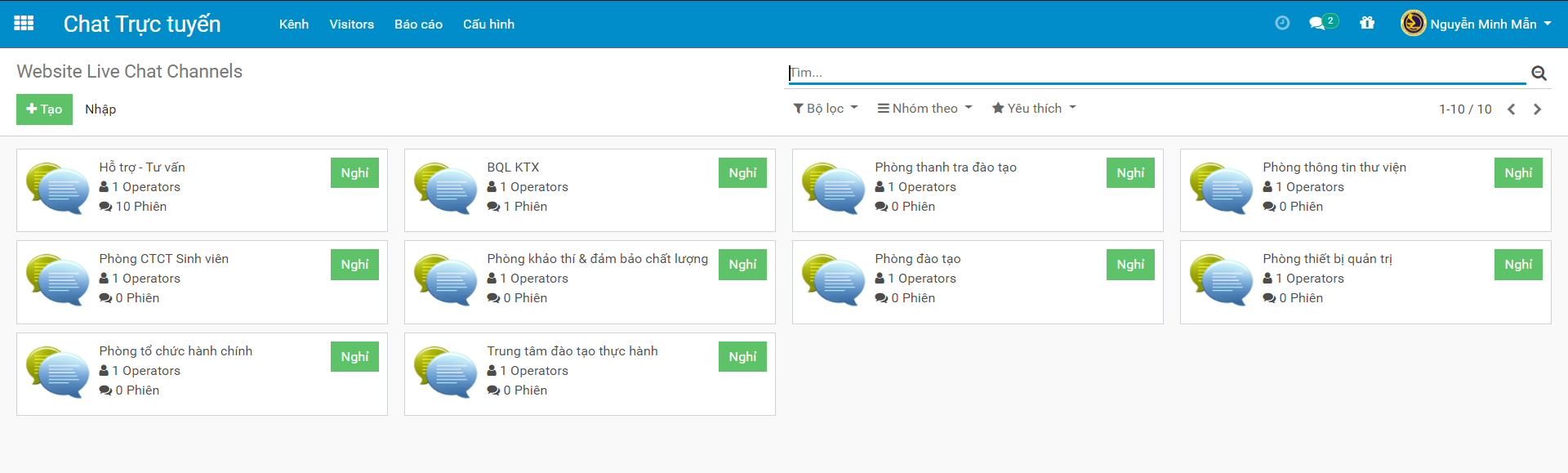
**

*Hình 3.35 Giao diện hiện thị người dùng truy cập trang web*

Quản lý của trang có thể quản lý số người đã truy cập và có tài khoản trong trang web.

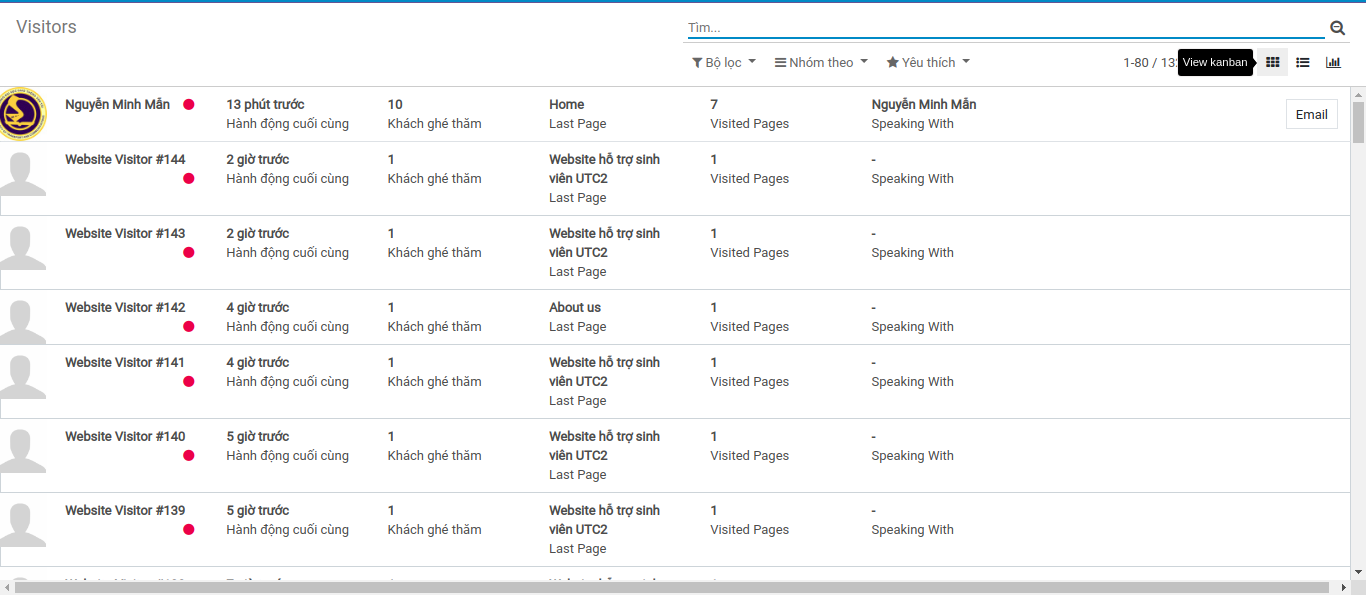
Nếu được cấp quyền admin sẽ hiện thị tài khoản admin, nếu chỉ là thành viên sẽ hiện tên gmail sinh viên truy cập.

***3.7.8 Quản lý chat trực tuyến***



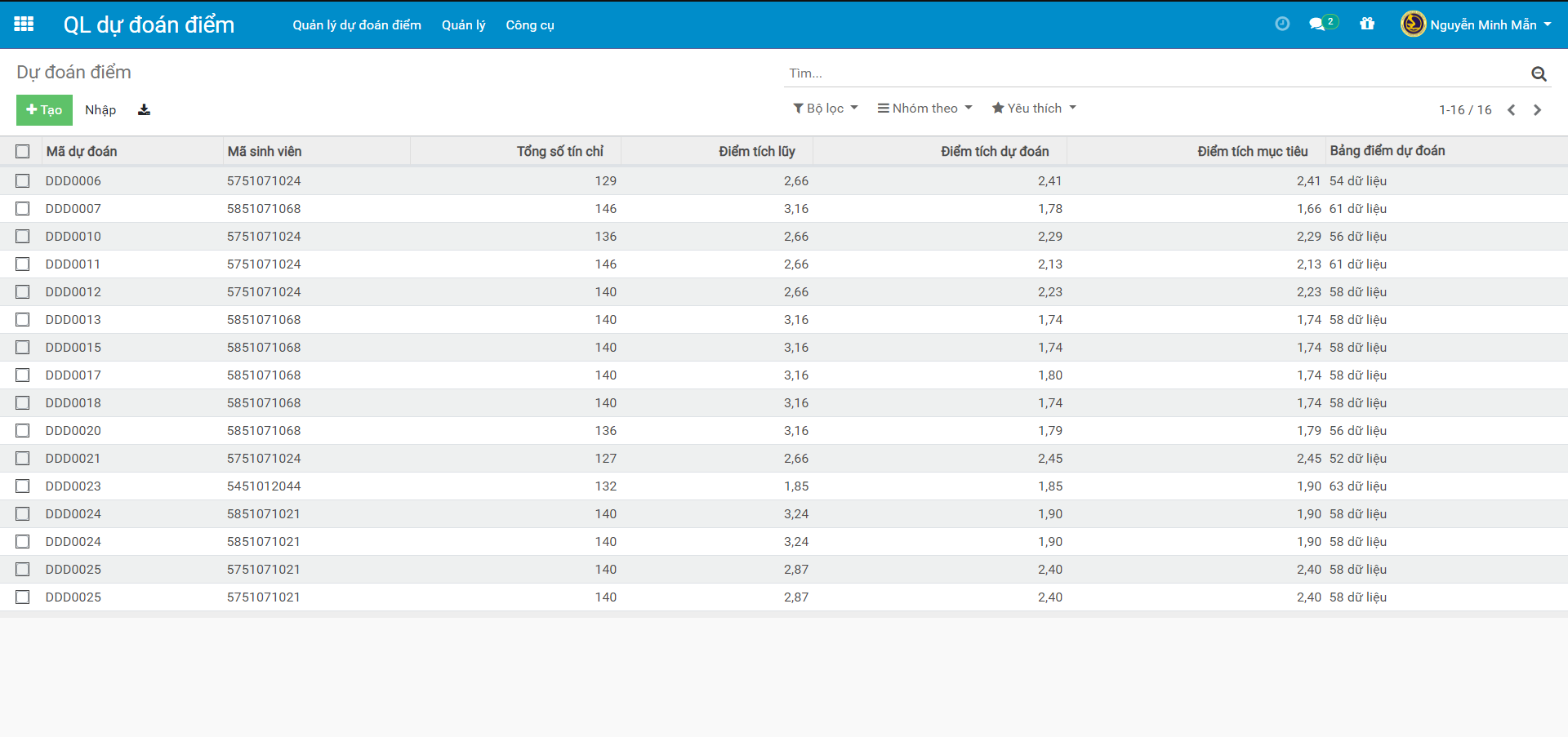
*Hình 3.36 Giao diện hiện thị người dùng truy cập trang web*

Quản lý có thể theo dõi các kênh chat, theo dỗi có người đang hoạt động trong các phòng chat. Số lượng tài khoản nhắn tin, lịch sử chat, …



*Hình 3.37 Giao diện theo dõi lịch sử chat*

***3.6.9 Quản lý dự đoán điểm***



*Hình 3.38 Giao diện quản lý dự đoán điểm*

# KẾT LUẬN

**Kết quả đạt được**

Sau quá trình tìm hiểu, nguyên cứu, phân tích, thực hiện và thử nghiệm trên thực tế, đề tài đã được những yêu cầu đã đặt ra ở mục tiêu đề ra:

- Xây dựng được website hỗ trợ sinh viên UTC2 có đầy đủ các chức năng cần thiết hỗ trợ sinh viên như:

* + Cho phép các bạn sinh viên đăng câu hỏi để được giải đáp các thắc mắc, cũng như có thể tương tác đối với bài đăng. Có thể tìm kiếm các câu hỏi theo từ khóa cần tìm.
  + Sinh viên có thể thực hiện chức năng chat trực tuyến với các bộ phận phòng ban
  + Sinh viên có thể sử dụng chức năng dự đoán điểm

- Hệ thống chạy ổn định với giao diện thân thiện, dễ sử dụng và tương thích với cả máy tính và điện thoại di động, đồng thời website cũng thực hiện tốt trên hầu hết các trình duyệt web như: Chrome, Firefox, Opera,..

**Nhược điểm**

+ Chức năng dự đoán điểm cần phải tạo trước công thức tính điểm, người tạo công thức là thầy cô hoặc sinh viên hiểu biết về các môn này.

+ Chức năng dự đoán điểm chưa tự thêm các môn học theo chương trình khung.

+ Công thức hồi quy dự đoán điểm chưa đưa ra được trọng số. Chỉ tạo công thức dự đoán theo cách khách quan, chưa được khoa học.

+ Chưa kết nối được toàn bộ dữ liệu thật tại trường.

+ Yêu cầu cấu hình máy chủ phải tương đối.

**Hướng phát triển**

- Khắc phục những nhược điểm, những hạn chế mà trang web còn gặp phải để trang web có thể được nhà trường sử dụng vào thực tế, từ đó có thể giúp đỡ được nhiều bạn sinh viên.

- Trang website có thể thêm chức năng tư vấn tuyển sinh cho tân sinh viên.

- Thêm chức năng phân tích ngữ nghĩa câu hỏi và trích nội dung trả lời từ văn bản.

- Công thức dự đoán điểm phải được thử nghiệm thực tiễn và đưa ra trọng số, độ tin tưởng của công thức.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | Daniel Reis, *Odoo 12 Development Essentials 4th Edition*, Packt, 2018. |
| [2] | Trần Cao Đệ và Phạm Nguyên Khang, *Phân loại văn bản với máy học vector hỗ trợ và cây quyết định,* Trường Đại học Cần Thơ, 2012. |
| [3] | Trần Ngọc Phúc, *Phân loại nội dung tài liệu web,* Luận văn thạc sĩ trường Đại học Lạc Hồng, 2012. |
| [4] | Trần Thị Thu Thảo và Vũ Thị Chinh, *Xây dựng hệ thống phân loại tài liệu tiếng Việt*, Khoa CNTT, trường Đại học Lạc Hồng. |
| [5] | Shahar Yifrah & Guy Lev, *Spam Email Filtering*, 2013. |
| [6] | V. M. Sebastian Raschka, *Python Machine Learning - Third Edition*, Packt, 2019. |
| [7] | J. L. E. a. Joseph Labrecque, *The JavaScript Workshop*, Packt, 2019. |
| [8] | J. Robbins, *Learning Web Design: A Beginner's Guide to HTML, CSS, JavaScript, and Web Graphics*, 2020. |
| [9] | Eric Brill, *Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging.* Comput. Linguist. (Cambridge, MA, USA: MIT Press) December 1995. |
| [10] | Dinh Dien, Hoang Kiem, Nguyen Van Toan. *Vietnamese Word Segmentation.The sixth Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, 11/2001. |
| [11] | Chen, K. J., & Liu, S. H *Word identification for Mandarin Chinese sentences,* Proceedings of the Fifteenth International Conference on Computational Linguistics, Nantes: COLING-92, 1992. |
| [12] | Yang and Xin Liu, *A re-examination of text categorization methods*, Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’99), 1999. |
| [13] | J. Han and M. Kamber, *Data mining: concepts and techniques. Morgan Kaufmann*, 2006. |
| [14] | Vũ Hữu Tiệp’s Blog, “*Machine Learning cơ bản*”, [*https://machinelearningcoban.com*](https://machinelearningcoban.com) *,* truy cập ngày 20 tháng 07 năm 2020. |
| [15] | *Tất tần tật về Machine Learning & ứng dụng trong những ngành công nghiệp lớn*, [*https://techtalk.vn/tat-tan-tat-moi-kien-thuc-co-ban-ve-machine-learning.html*](https://techtalk.vn/tat-tan-tat-moi-kien-thuc-co-ban-ve-machine-learning.html) *,* truy cập ngày 20 tháng 07 năm 2020. |
| [16] | W. Foundation, *XML*, <https://vi.wikipedia.org/wiki/XML>*,* truy cập ngày 20 tháng 07 năm 2020. |
| [17] | Wikipedia, “*Linear Regression”,* [*https://en.wikipedia.org/wiki/Linear\_regression*](https://en.wikipedia.org/wiki/Linear_regression)*,* truy cập ngày 21 tháng 07 năm 2020. |
| [18] | Wikipedia, *JavaScript,* <https://vi.wikipedia.org/wiki/JavaScript>, truy cập ngày 21 tháng 07 năm 2020. |

1. https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ [↑](#footnote-ref-1)
2. https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ [↑](#footnote-ref-2)
3. https://scikit-learn.org/stable/tutorial/machine\_learning\_map/index.html [↑](#footnote-ref-3)