

**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: NGUYÊN CỨU CÁC THUẬT TOÁN TRONG XỬ LÝ
NGÔN NGỮ TỰ NHIÊN VÀ XÂY DỰNG WEBSITE HỖ TRỢ
SINH VIÊN UTC2**

Giảng viên hướng dẫn: TRẦN PHONG NHÃ

Sinh viên thực hiện: NGUYỄN MINH MÃN

Lớp: CÔNG NGHỆ THÔNG TIN

Khoá: 57

TP. Hồ Chí Minh, tháng 08 năm 2020

**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: NGUYÊN CỨU CÁC THUẬT TOÁN TRONG XỬ LÝ
NGÔN NGỮ TỰ NHIÊN VÀ XÂY DỰNG WEBSITE HỖ TRỢ
SINH VIÊN UTC2**

Giảng viên hướng dẫn: TRẦN PHONG NHÃ

Sinh viên thực hiện: NGUYỄN MINH MÃN

Lớp: CÔNG NGHỆ THÔNG TIN

Khoá: 57

TP. Hồ Chí Minh, tháng 08 năm 2020

NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP
BỘ MÔN: CÔNG NGHỆ THÔNG TIN

-----***-----

Mã sinh viên: 5751071024

Họ tên SV: Nguyễn Minh Mẫn

Khóa: 57

Lớp: Công Nghệ Thông Tin

1. Tên đề tài.

***NGUYÊN CỨU CÁC THUẬT TOÁN TRONG XỬ LÝ NGÔN NGỮ TỰ NHIÊN
VÀ XÂY DỰNG WEBSITE HỖ TRỢ SINH VIÊN UTC2.***

2. Mục đích, yêu cầu.

a. Mục đích.

- Xây dựng trang website hỗ trợ sinh viên UTC2.
- Xây dựng hệ thống diễn đàn sinh viên đề xuất các câu trả lời tương tự trên hệ thống.
- Xây dựng hệ thống dự đoán điểm và tùy chỉnh dự đoán để đặt ra mục tiêu học tập.
- Xây dựng hệ thống hỗ trợ trực tuyến để sinh viên được tư vấn trực tuyến từ các phòng ban trên hệ thống.
- Xây dựng hệ thống giúp giảng viên có thể giải đáp thắc mắc của sinh viên một cách nhanh chóng.
- Xây dựng hệ thống quản lý trên website hỗ trợ sinh viên UTC2.

b. Yêu cầu.

- Tìm hiểu về Machine Learning.
- Nghiên cứu về xử lý ngôn ngữ tự nhiên.
- Nghiên cứu quy trình trong xử lý ngôn ngữ tự nhiên.
- Nghiên cứu các thuật toán trong xử lý ngôn ngữ tự nhiên
- Thu thập dữ liệu câu hỏi, câu trả lời trên diễn đàn nghe nói.

- Thu thập điểm của sinh viên UTC2.
- Ứng dụng thuật toán Linear Regression để dự đoán điểm sinh viên cho website hỗ trợ sinh viên.
- Ứng dụng Support Vector Machine phân loại câu hỏi cho các phòng ban trên website hỗ trợ sinh viên.
- Ứng dụng đo độ tương tự hai vector để đề xuất câu hỏi/ bài viết tương tự được trả lời trên hệ thống.

3. Nội dung và phạm vi đề tài.

a. Nội dung đề tài.

- Giới thiệu tổng quan về trí tuệ nhân tạo và Machine Learning
- Giới thiệu các thuật toán trong xử lý ngôn ngữ tự nhiên.
- Nghiên cứu, phân tích, đánh giá thuật toán.
- Sản phẩm hoàn chỉnh website hỗ trợ sinh viên.
- Ứng dụng thuật toán vào bài toán thực tế, cụ thể là đề xuất câu hỏi tương tự trên diễn đàn.

b. Phạm vi đề tài.

- Nguyên cứu các thuật toán và quy trình trong xử lý ngôn ngữ tự nhiên.
- Ứng dụng các thuật toán và phần mềm Odoo để xây dựng website hỗ trợ sinh viên cho trường Đại học Giao thông Vận tải phân hiệu tại thành phố Hồ Chí Minh.

4. Công nghệ, công cụ và ngôn ngữ lập trình.

a. Công nghệ: Python, JavaScript, Odoo.

b. Công cụ: Một số thư viện mã nguồn mở của Python: Scikit-learn, pandas, numpy, matplotlib, keras

c. Ngôn ngữ lập trình: Python, JavaScript, XML

5. Các kết quả chính dự kiến sẽ đạt được và ứng dụng

- Hoàn chỉnh cuốn báo cáo đề tài.
- Khái quát được tổng quan về Machine Learning.
- Nắm được thuật toán SVM, Linear Regression và có thể áp dụng được thuật toán cho bất kỳ bài toán nào liên quan.
- Nắm được các quy trình trong xử lý ngôn ngữ tự nhiên.
- Nắm được các ưu, nhược điểm của thuật toán, phương pháp tối ưu cho thuật toán.
- Sử dụng Odoo để xây dựng được website hỗ trợ sinh viên UTC2.

6. Giáo viên và cán bộ hướng dẫn

Họ tên: TRẦN PHONG NHÃ

Đơn vị công tác: Bộ môn Công Nghệ Thông Tin – Trường Đại học Giao thông Vận
tải phân hiệu tại 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

Trần Phong Nhã

Đã nhận nhiệm vụ TKTN

Sinh viên: Nguyễn Minh Mẫn

Ký tên:

Điện thoại: 0349183111

Email: minhmanit98@gmail.com

LỜI CẢM ƠN

Lời nói đầu tiên, em xin kính gửi lời cảm ơn chân thành nhất tới Quý thầy cô trong Bộ môn Công Nghệ Thông Tin, cũng như Ban Giám Hiệu Trường Đại học Giao thông Vận tải phân hiệu tại Thành phố Hồ Chí Minh, đã cho phép em thực hiện đề tài tốt nghiệp: ***NGUYÊN CỨU CÁC THUẬT TOÁN TRONG XỬ LÝ NGÔN NGỮ TỰ NHIÊN VÀ XÂY DỰNG WEBSITE HỖ TRỢ SINH VIÊN UTC2.***

Để hoàn thành nhiệm vụ được giao này, ngoài sự nỗ lực học hỏi không ngừng của bản thân còn có sự hướng dẫn tận tình của thầy giáo **Trần Phong Nhã**, người đã hướng dẫn cho em những hướng đi, truyền đạt cho em những kiến thức, kỹ năng để em có thể hoàn thành đề tài tốt nghiệp này.

Mặc dù đã cố gắng hết sức để hoàn thành đề tài, nhưng chắc chắn rằng sẽ khó tránh khỏi những thiếu sót. Em rất mong nhận được những sự đánh giá, góp ý của Quý thầy cô để em có thể rút ra cho mình những bài học, kinh nghiệm quý báu.

Sau cùng, 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 giáo **Trần Phong Nhã** 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

Nguyễn Minh Mẫn

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày tháng năm 2020

Giáo viên hướng dẫn

Trần Phong Nhã

MỤC LỤC

NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP	iii
LỜI CẢM ƠN.....	vi
NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN	vii
MỤC LỤC	viii
DANH MỤC THUẬT NGỮ	xi
DANH MỤC BẢNG	xii
DANH MỤC HÌNH ẢNH.....	xii
TỔNG QUAN.....	1
Tổng quan về khai phá dữ liệu	1
Khai phá dữ liệu	1
Khai phá dữ liệu văn bản.....	2
Đặt vấn đề.....	2
Tình hình nguyên cứu.....	4
Quá trình nguyên cứu	4
CHƯƠNG 1: CƠ SỞ LÝ THUYẾT.....	6
1.1 Machine Learning.....	6
1.1.1 Giới thiệu về Machine Learning.....	6
1.1.2 Phân nhóm các thuật toán Machine Learning	9
1.1.3 Các bước thực hiện Machine Learning.	12
1.1.4 Các ứng dụng	14
1.2 Xử lý ngôn ngữ tự nhiên.....	16
1.2.1 Khái niệm	16
1.2.2 Những khó khăn trong lĩnh vực xử lý ngôn ngữ tự nhiên.....	16
1.2.3 Một số ứng dụng của xử lý ngôn ngữ tự nhiên	18
1.3 Odoo	19
1.3.1 Giới thiệu về Odoo	19
1.3.2 Các ưu điểm của Odoo	20
1.3.3 Cơ sở dữ liệu PostgreSQL.....	20
1.4 Python.....	20
1.4.1 Giới thiệu về Python.....	20

1.4.2 Đặc điểm của Python.....	21
1.4.3 Một số thư viện liên quan.....	21
1.5 Javascript	22
1.5.1 Giới thiệu về JavaScript	22
1.5.2 Ưu điểm của JavaScript.....	22
1.5.3 Nhược điểm của JavaScript.....	23
1.6 XML	23
1.6.1 Giới thiệu về XML	23
1.6.2 Đặc điểm của XML	23
CHƯƠNG 2: PHÂN LOẠI VĂN BẢN TRONG XỬ LÝ NGÔN NGỮ TỰ NHIÊN .	24
2.1 Tiền xử lý dữ liệu	24
2.1.1 Tách từ.....	24
2.1.2 Loại bỏ stop word.....	27
2.2 Chuyển đổi văn bản sang mô hình không gian véc-tơ	27
2.2.1 Binary vector	28
2.2.2 TF-IDF vector.....	28
2.3 Các phương pháp phân loại văn bản bằng máy học	30
2.3.1 Phương pháp SVM (Support vector machine)	30
2.3.2 Phương pháp Naive Bayes	32
2.3.3 Phương pháp cây quyết định (Classification and regression trees).....	34
2.3.4 K- Nearest Neighbor(KNN)	35
2.3.5 Linear Least Square Fit(LLSF)	36
2.3.6 Các thông số đánh giá giải thuật	37
2.4 Độ tương đồng giữa các véc-tơ (Cosine Similarity)	38
2.5 Các công trình nguyên cứu liên quan.....	38
2.5.1 Phân loại văn bản với máy học véc-tơ hỗ trợ và cây quyết định	39
2.5.2 Xây dựng hệ thống phân loại tài liệu Tiếng Việt	40
2.5.3 Phân loại email spam bằng matlab áp dụng 6 giải thuật	42
CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH.....	46
3.1 Hiện trạng tổ chức	46
3.2 Yêu cầu hệ thống	46
3.2.1 Yêu cầu chức năng	46

3.2.2 Yêu cầu phi chức năng	46
3.3 Thu thập dữ liệu.....	47
3.4 Thuật toán áp dụng	49
3.4.1 Bài toán phân loại câu hỏi và đề xuất câu hỏi tương tự	49
3.3.2 Bài toán dự đoán điểm.....	54
3.5 Sơ đồ usecase của hệ thống	56
3.6 Sơ đồ hoạt động	57
3.6.2 Sơ đồ hoạt động đăng nhập	57
3.6.11 Sơ đồ quản lý dự đoán điểm.....	57
3.5.6 Sơ đồ dự đoán điểm và tạo bảng điểm	58
3.6.8 Sơ đồ quản lý người dùng	58
3.6.7 Sơ đồ hỗ trợ chat trực tuyến	59
3.6.9 Sơ đồ quản lý hỗ trợ trực tuyến	59
3.6.10 Sơ đồ quản lý diễn đàn	60
3.6.1 Sơ đồ báo cáo thống kê.....	60
3.7 Xây dựng giao diện chương trình.....	61
3.7.1 Tạo tài khoản và đăng nhập.....	61
3.7.2 Giao diện trang chủ website	62
3.7.3 Tương tác bài viết.....	62
3.7.4 Chat trực tuyến	64
3.7.5 Dự đoán điểm	65
3.7.6 Nhóm chức năng quản lý.....	66
3.7.7 Quản lý người dùng.....	67
3.7.8 Quản lý chat trực tuyến	67
3.6.9 Quản lý dự đoán điểm	68
KẾT LUẬN	69
Kết quả đạt được.....	69
Nhược điểm	69
Hướng phát triển.....	69
TÀI LIỆU THAM KHẢO	70

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	Stupid Pointless Annoying Messages	Thư rác	SPAM	
3	Artificial Intelligence	Trí tuệ nhân tạo	AI	
4	Machine Learning	Học máy	ML	
5	Deep Learning	Học sâu	DL	
6	Knowledge Discovery in Database	Khai phá tri thức	KDD	
7	Application Programming Interface	Giao diện lập trình ứng dụng	API	
8	Database	Cơ sở dữ liệu	CSDL	
9	Data Mining	Khai phá dữ liệu		
10	Structured Query Language	Ngôn ngữ truy vấn dữ liệu có cấu trúc	SQL	
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	
----	---------------------------------	--	-----	--

DANH MỤC BẢNG

<i>Bảng 1 Các phòng ban/trung tâm</i>	<i>4</i>
<i>Bảng 2.1 Dữ liệu tên và giới tính</i>	<i>33</i>
<i>Bảng 2.2 Kết quả phân loại của [2].....</i>	<i>40</i>
<i>Bảng 2.3 Kết quả phân loại văn bản [4].....</i>	<i>41</i>
<i>Bảng 2.4 Kết quả phân loại theo [5].....</i>	<i>44</i>
<i>Bảng 3.1 Bảng thống kê số câu train và câu test của các ban ngành/phòng ban.....</i>	<i>53</i>

DANH MỤC HÌNH ẢNH

<i>Hình 1.1 Mối quan hệ giữa AI, Machine Learning và Deep Learning.</i>	<i>6</i>
<i>Hình 1.2 Mối quan hệ giữa AI, Machine Learning và Deep Learning.</i>	<i>7</i>
<i>Hình 1.3 Lựa chọn thuật toán phù hợp trong Machine Learning.</i>	<i>13</i>
<i>Hình 2.1 Siêu phẳng với lề cực đại cho một SVM phân tách dữ liệu thuộc hai lớp</i>	<i>31</i>
<i>Hình 2.2 Cây quyết định mức lương.....</i>	<i>34</i>
<i>Hình 2.3 Quy trình phân loại văn bản của [2].....</i>	<i>39</i>
<i>Hình 2.4 Quy trình phân loại văn bản theo [4].....</i>	<i>41</i>
<i>Hình 2.5 Trình tự phân loại văn bản của [4].....</i>	<i>41</i>
<i>Hình 3.1 Dữ liệu sau khi lấy trên diễn đàn</i>	<i>48</i>
<i>Hình 3.2 Dữ liệu sau khi được chuẩn hóa.....</i>	<i>48</i>
<i>Hình 3.3 Form phân loại câu hỏi sinh viên.....</i>	<i>49</i>
<i>Hình 3.4 Quy trình hóa phân loại văn bản.....</i>	<i>50</i>
<i>Hình 3.5 Các tag phân loại trên diễn đàn.....</i>	<i>50</i>
<i>Hình 3.6 Quá trình tạo mô hình véc- tơ và ma trận câu hỏi</i>	<i>51</i>
<i>Hình 3.7 Quá trình chọn câu hỏi-câu trả lời tương tự trên diễn đàn</i>	<i>51</i>
<i>Hình 3.8 Quá trình phân tag cho câu hỏi.....</i>	<i>52</i>
<i>Hình 3.9 Tổ chức tập dữ liệu train và test.....</i>	<i>53</i>
<i>Hình 3.10 Kết quả sau khi train tập dữ liệu trên.....</i>	<i>54</i>

<i>Hình 3.11 Công thức dự đoán điểm</i>	<i>55</i>
<i>Hình 3.12 Quản lý công thức dự đoán điểm</i>	<i>55</i>
<i>Hình 3.13 Sơ đồ usecase tổng quan của hệ thống.....</i>	<i>56</i>
<i>Hình 3.14 Sơ đồ hoạt động đăng nhập.....</i>	<i>57</i>
<i>Hình 3.15 Sơ đồ hoạt động quản lý dự đoán điểm.....</i>	<i>57</i>
<i>Hình 3.16 Sơ đồ hoạt động dự đoán điểm.....</i>	<i>58</i>
<i>Hình 3.17 Sơ đồ hoạt động quản lý người dùng</i>	<i>58</i>
<i>Hình 3.18 Sơ đồ hoạt động dự đoán điểm.....</i>	<i>59</i>
<i>Hình 3.19 Sơ đồ hoạt động hỗ trợ trực tuyến.....</i>	<i>59</i>
<i>Hình 3.20 Sơ đồ hoạt động quản lý diễn đàn</i>	<i>60</i>
<i>Hình 3.21 Sơ đồ hoạt động quản lý báo cáo thống kê</i>	<i>60</i>
<i>Hình 3.22 Giao diện tạo tài khoản đăng nhập</i>	<i>61</i>
<i>Hình 3.23 Giao diện đăng nhập</i>	<i>61</i>
<i>Hình 3.24 Giao diện trang chủ của website</i>	<i>62</i>
<i>Hình 3.25 Giao diện đặt câu hỏi trên diễn đàn.....</i>	<i>62</i>
<i>Hình 3.26 Giao diện trả lời câu hỏi</i>	<i>63</i>
<i>Hình 3.27 Các chức năng thao tác với câu hỏi</i>	<i>63</i>
<i>Hình 3.28 Giao diện chức năng tìm kiếm câu hỏi.....</i>	<i>64</i>
<i>Hình 3.29 Giao diện chức năng chat trực tuyến</i>	<i>64</i>
<i>Hình 3.30 Giao diện người dùng sử dụng chức năng chat trực tuyến.....</i>	<i>65</i>
<i>Hình 3.31 Giao diện chức năng dự đoán điểm</i>	<i>65</i>
<i>Hình 3.32 Giao diện bảng điểm sau khi người dùng chọn chức năng dự đoán điểm...66</i>	
<i>Hình 3.33 Giao diện nhóm chức năng quản lý.....</i>	<i>66</i>
<i>Hình 3.34 Giao diện quản lý người dùng.....</i>	<i>67</i>
<i>Hình 3.35 Giao diện hiện thị người dùng truy cập trang web</i>	<i>67</i>
<i>Hình 3.36 Giao diện hiện thị người dùng truy cập trang web</i>	<i>67</i>
<i>Hình 3.37 Giao diện theo dõi lịch sử chat.....</i>	<i>68</i>
<i>Hình 3.38 Giao diện quản lý dự đoán điểm</i>	<i>68</i>

TỔNG QUAN

Tổng quan về khai phá dữ liệu

Khai phá dữ liệu

Trên thế giới cũng như ở Việt Nam, công nghệ thông tin đã trở thành một ngành công nghệ mũi nhọn. Bất kỳ một ngành nghề nào, lĩnh vực nào trong xã hội cũng cần đến sự góp sức của công nghệ thông tin để giải quyết và lượng dữ liệu ngày càng tăng lên cả về số lượng và chất lượng. Tuy nhiên, chỉ một phần nhỏ trong khối dữ liệu khổng lồ đó là có giá trị sử dụng. Nhu cầu tìm kiếm và khai thác tri thức từ khối dữ liệu đó đã mở ra một khía cạnh mới của ngành công nghệ thông tin đó là Khai thác tri thức từ cơ sở dữ liệu (Knowledge Discovery from Data hay KDD).

Khai phá dữ liệu là một bước trong quá trình khai thác tri thức. Bao gồm:

- Xác định vấn đề và không gian dữ liệu để giải quyết vấn đề (problem understanding and data understanding).
- Chuẩn bị dữ liệu (data preparation), bao gồm các quá trình làm sạch dữ liệu (data cleaning), tích hợp dữ liệu (data integration), chọn dữ liệu (data selection), biến đổi dữ liệu (data transformation).
- Khai thác dữ liệu (data mining): xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữ liệu. Kết quả cho ta một nguồn tri thức thô.
- Đánh giá (evaluation): dựa trên một số tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được.
- Triển khai (deployment).

Quá trình khai thác tri thức không thực hiện tuần tự từ bước đầu tiên đến bước cuối cùng mà đó là một quá trình lặp đi lặp lại nhiều lần.

Khai phá dữ liệu có thể hiểu đơn giản là quá trình chắt lọc và khai thác tri thức từ một khối dữ liệu lớn. Việc này cần sử dụng kiến thức từ nhiều ngành và nhiều lĩnh vực khác nhau như thống kê, trí tuệ nhân tạo, cơ sở dữ liệu, tính toán song song,... Đặc biệt, nó rất gần gũi với lĩnh vực thống kê, sử dụng các phương pháp thống kê để mô hình hóa dữ liệu và phát hiện các mẫu. Ứng dụng của khai phá dữ liệu có thể kể đến như: cung cấp tri thức, hỗ trợ ra quyết định, dự báo, khái quát dữ liệu.

Các phương pháp khai phá dữ liệu:

- Bài toán phân lớp (classification): Ánh xạ một mẫu dữ liệu vào một trong các lớp cho trước.
- Bài toán hồi quy (regression): Tìm một ánh xạ hồi quy từ một mẫu dữ liệu vào một biến dự đoán có giá trị thực.
- Bài toán lập nhóm (clustering): Là việc mô tả chung để tìm các tập xác định hữu hạn các nhóm hay các loại để mô tả dữ liệu. Phân loại tin tức tiếng Việt sử dụng các phương pháp học máy 4 Lê Vĩnh Phú – Diệp Minh Hoàng
- Bài toán tổng hợp (summarization): Là việc đi tìm kiếm một mô tả chung tóm tắt từ một tập dữ liệu con.
- Mô hình ràng buộc (dependency modeling): Là việc đi tìm một mô hình mô tả sự phụ thuộc giữa các biến hay giữa các giá trị của các tính năng trong tập dữ liệu.
- Dò tìm biến đổi và độ lệch (change and deviation detection): Là việc tìm những thay đổi lớn nhất trong tập dữ liệu.

Khai phá dữ liệu văn bản

Khai phá dữ liệu văn bản là quá trình khai phá các tri thức đáng quan tâm hay có giá trị từ các tài liệu văn bản phi cấu trúc.

Bài toán Khai phá dữ liệu văn bản là một bài toán đa lĩnh vực bao gồm nhiều kỹ thuật và các hướng nghiên cứu khác nhau : thu thập thông tin (information retrieval), phân tích văn bản (text analysis), chiết xuất thông tin (information extraction), lập đoạn (clustering), phân loại văn bản (categorization)...

Đặt vấn đề

Ở nước ta hiện nay, việc ứng dụng công nghệ thông tin tại các cơ quan, trường học, xí nghiệp, tổ chức đang rất phổ biến và dần trở nên cần thiết. Bởi ngành nghề nào cũng đòi hỏi con người phải xử lý khối lượng công việc khổng lồ, phức tạp với dữ liệu lớn, những kiến thức và đào tạo chuyên sâu. Cụ thể như trong quá trình học tập tại trường nhiều bạn sinh viên, đặc biệt là các bạn sinh viên năm nhất mới bước chân vào giảng đường chắc chắn sẽ có nhiều thắc mắc về quá trình học tập cũng như các quy định rèn luyện thi đua, nhưng đa phần các bạn thường không biết nên hỏi ai và không thể xác định các thông tin chính xác về các vấn đề của mình.

Nhằm khắc phục được vấn đề này, nhiều trường đại học đã thực hiện một số biện pháp để giải quyết như: xây dựng các fanpage, các kênh diễn đàn để sinh viên có thể đặt

các câu hỏi cũng như chia sẻ các khó khăn của mình. Tuy nhiên, cách làm này vẫn tồn tại nhiều phát sinh, các câu hỏi của các bạn sinh viên không thể ngay lập tức đến với những người có thể giải đáp mà phải trải qua quy trình chung gian, là admin của các fanpage lọc câu hỏi và tiếp tục gửi đi tới các bộ phận liên quan có thể giải quyết. Điều này dẫn đến việc các bạn sinh viên phải chờ đợi lâu hơn, quy trình giúp đỡ các bạn sinh viên trở nên rườm rà và còn thiếu tính linh động tương tác.

Xuất phát từ tình hình thực tế đó, tôi đã nghĩ 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, phòng 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 dựng website hỗ trợ sinh viên UTC2* dựa trên phần mềm odoo - giải pháp quản lí doanh nghiệp dễ 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

Hiện nay tại Phân hiệu Trường Đại học Giao Thông Vận Tải tại TP.HCM chưa có ứng dụng phục vụ hỗ trợ cố vấn học tập online trên các trang web để tạo môi trường hỏi đáp, giúp đỡ sinh viên.

Trên thị trường phần mềm Odoo có vai trò hỗ trợ đắc lực của các phần mềm quản lý doanh nghiệp với nhiều chức năng khác nhau như: bán hàng, chăm sóc khách hàng, quản trị dự án, quản lý tài chính và nguồn nhân lực, quản trị sản xuất...phần mềm này được lập trình sẵn 18 ngôn ngữ giúp cho công ty ở nhiều nước có thể dễ dàng tùy chỉnh sử dụng. Bên cạnh đó còn mang đến nhiều lợi ích cho doanh nghiệp như: dễ dàng sử dụng, tiết kiệm chi phí, dễ dàng cài đặt và có nhiều chức năng hữu dụng, công nghệ được cập nhật liên tục.

Quá trình nguyên cứu

Để hiểu rõ hơn về công tác giải đáp hỗ trợ sinh viên. Tôi đã lựa chọn Fanpage Dẫn 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 ban sau:

Bảng 0.1 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 lý. 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, cảm ơn, 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

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

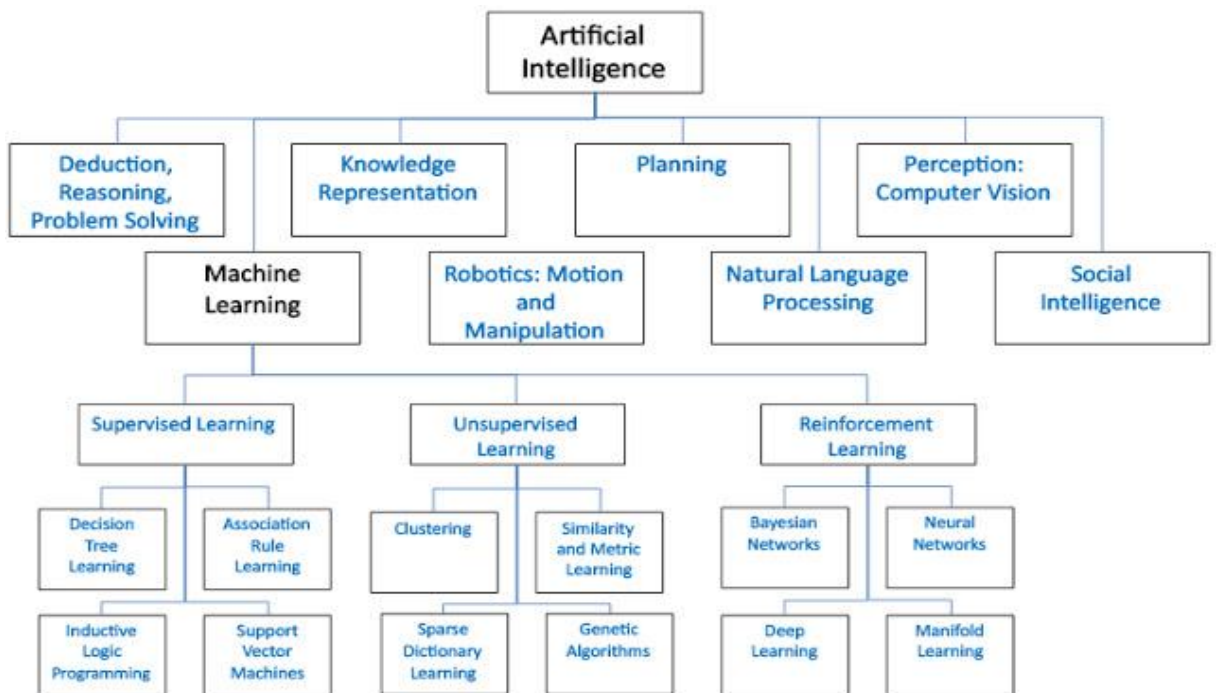
1.1 Machine Learning.

1.1.1 Giới thiệu về Machine Learning

Những năm gần đây, AI – Trí tuệ nhân tạo nổi lên như một bằng chứng của cuộc cách mạng công nghiệp 4.0. Ngày nay, AI càng len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra.

Xe tự lái của Google và Tesla, hệ thống tự nhận diện 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 trí tuệ nhân tạo.

Và Machine Learning chính là một tập con trong trí tuệ nhân tạo. Nó là một lĩnh vực nhỏ trong ngành khoa học máy tính, có khả năng tự học hỏi dựa trên dữ liệu được đưa vào mà không cần phải lập trình cụ thể.

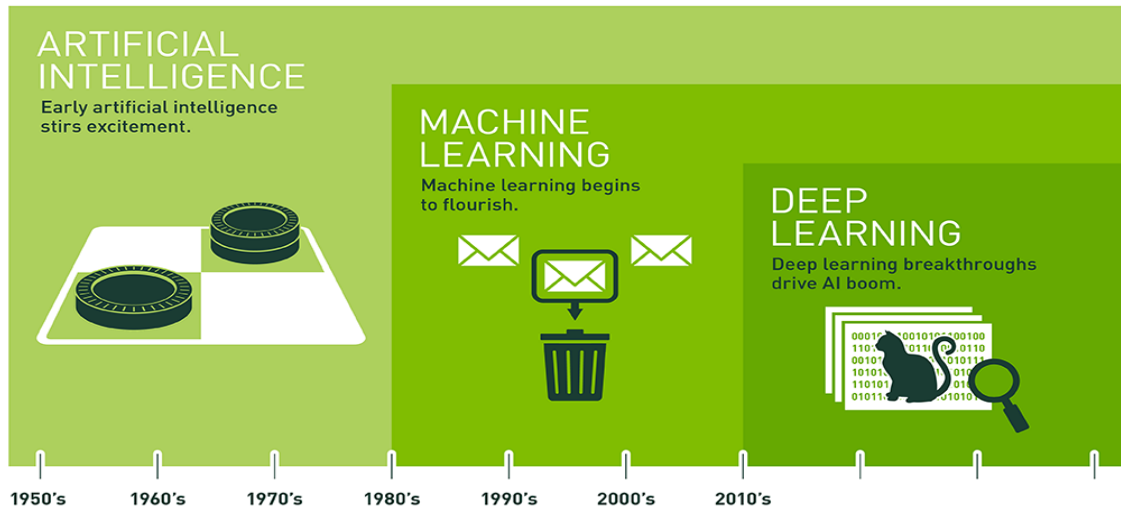


Hình 1.1 Mối quan hệ giữa AI, Machine Learning và Deep Learning.¹

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).

¹ <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>

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

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.

² <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>

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.^[15]

Các ứng dụng của ML đã trở nên quá quen thuộc như: ^[15]

- Xe tự lái, giảm thiểu tai nạn của Google? Chính là bản chất của ML. ^[15]
- Các ưu đãi Recommendation Online như của Amazon & Netflix? Ứng dụng của Machine Learning trong cuộc sống hằng ngày. ^[15]
- Muốn biết người dùng nói gì về bạn trên Twitter? ML kết hợp với sự sáng tạo của quy tắc ngôn ngữ. ^[15]
- Nhận diện lừa đảo? Một trong những nhu cầu sử dụng hiển nhiên ngày nay. ^[15]

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. ^[15]

- 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. ^[15]

- 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. ^[15]

- 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. ^[15]

- 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 lai của ngành bán lẻ. ^[15]

- 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. ^[15]

- 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. ^[15]

1.1.2 Phân nhóm các thuật toán Machine Learning

Các thuật toán ML thường được chia làm 4 nhóm: ^[14]

- Supervise learning (Học có giám sát)
- Unsupervised learning (Học không giám sát)
- Semi-supervised learning (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. ^[14]

Một cách toán học, Supervised learning là khi chúng ta có một tập hợp biến đầu vào $X=\{x_1, x_2, \dots, x_N\}$ và một tập hợp nhãn tương ứng $Y=\{y_1, y_2, \dots, y_N\}$ trong đó x_i, y_i là các vector. Các cặp dữ liệu biết trước $(x_i, y_i) \in X \times Y$ được gọi là tập *training data* (dữ liệu huấn luyện). Từ tập training 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: $y_i \approx f(x_i)$, $\forall i=1, 2, \dots, N$. Mục đích là xấp xỉ hàm số f thật tốt để khi có một dữ liệu x mới, chúng ta có thể tính được nhãn tương ứng của nó $y=f(x)$.^[14]

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.^[14]

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.^[14]

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.^[14]

Regression (Hồi quy)

Nếu *label* không được chia thành các nhóm mà là một giá trị thực cụ thể.^[14]

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?^[14]

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. ^[14]

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. ^[14]

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. ^[14]

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. ^[14]

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. ^[14]

Ví dụ: AlphaGo gần đây nổi tiếng với việc chơi cờ vây thắng cả con người. Cờ vây được xem là có độ phức tạp cực kỳ cao 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. ^[14]

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. ^[14]

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. ^[14]

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.

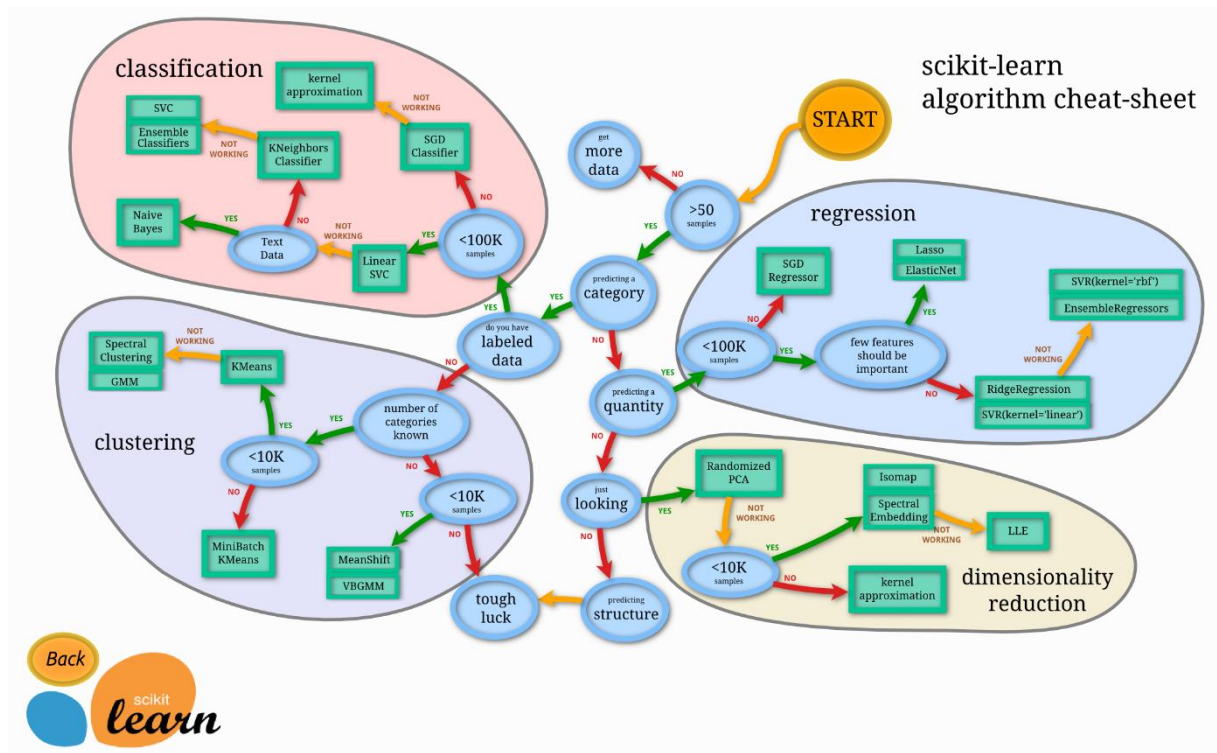
Việc thu nhỏ tập dữ liệu cũng khiến cho việc học của ta tốn ít thời gian hơn, tuy nhiên dữ liệu ít quá cũng khiến việc học đạt độ chính xác không cao, cần cân đối giữa các yếu tố này.

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

³ https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

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 khuyến là 60% - 20% - 20%.

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, NVIDIA 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, Cortana 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 chính 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 $[X_1, X_2, \dots, X_n]$ 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 gán nhãn.
- Học nửa giám sát : kết hợp các ví dụ có gán nhãn và không gán 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 Xử lý ngôn ngữ tự nhiên

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 lý 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 (Context-free grammar – CFG), Văn phạm danh mục kết nối (Combinatory categorial grammar – CCG), và Văn phạm phụ thuộc (Dependency grammar – DG). 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 liệu dưới 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 Odoo

1.3.1 Giới thiệu về Odoo

- *Odoo* là một phần mềm quản trị doanh nghiệp mã nguồn mở sử dụng ngôn ngữ lập trình Python 7 (còn có thêm Javascript và XML) (1995)). 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 (htt).

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 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 (Chaupm, n.d.).

- 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 (Chaupm, n.d.).

- 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.

1.4 Python

1.4.1 Giới thiệu về Python

Python là một ngôn ngữ lập trình bậc cao, thông dịch, hướng đối tượng, đa mục đích và cũng là một ngôn ngữ lập trình động.

Cú pháp của Python là khá dễ dàng để học và ngôn ngữ này cũng mạnh mẽ và linh hoạt không kém các ngôn ngữ khác trong việc phát triển các ứng dụng. Python hỗ trợ mẫu đa lập trình, bao gồm lập trình hướng đối tượng, lập trình hàm và mệnh lệnh hoặc là các phong cách lập trình theo thủ tục.

Python không chỉ làm việc trên lĩnh vực đặc biệt như lập trình web, và đó là tại sao ngôn ngữ này là đa mục đích bởi vì nó có thể được sử dụng với web, enterprise, 3D CAD, ...

Với Python, việc phát triển ứng dụng và debug trở nên nhanh hơn bởi vì không cần đến bước biên dịch và chu trình edit-test-debug của Python là rất nhanh.

1.4.2 Đặc điểm của Python

Dưới đây là một số đặc điểm chính của Python:

- Dễ dàng để sử dụng: Python là một ngôn ngữ bậc cao rất dễ dàng để sử dụng. Python có một số lượng từ khóa ít hơn, cấu trúc của Python đơn giản hơn và cú pháp của Python được định nghĩa khá rõ ràng, ... Tất cả các điều này là Python thực sự trở thành một ngôn ngữ thân thiện với lập trình viên.
- Bạn có thể đọc code của Python khá dễ dàng. Phần code của Python được định nghĩa khá rõ ràng và rành mạch.
- Python có một thư viện chuẩn khá rộng lớn. Thư viện này dễ dàng tương thích và tích hợp với UNIX, Windows, và Macintosh.
- Python là một ngôn ngữ thông dịch. Trình thông dịch thực thi code theo từng dòng, điều này giúp cho quá trình debug trở nên dễ dàng hơn và đây cũng là yếu tố khá quan trọng giúp Python thu hút được nhiều người học và trở nên khá phổ biến.
- Python cũng là một ngôn ngữ lập trình hướng đối tượng. Ngoài ra, Python còn hỗ trợ các phương thức lập trình theo hàm và theo cấu trúc.

Ngoài các đặc điểm trên, Python còn khá nhiều đặc điểm khác như hỗ trợ lập trình GUI, mã nguồn mở, có thể tích hợp với các ngôn ngữ lập trình khác, ...

1.4.3 Một số thư viện liên quan

- Numpy là một thư viện không thể thiếu khi chúng ta xây dựng các ứng dụng Máy học trên Python. Numpy cung cấp các đối tượng và phương thức để làm việc với mảng

nhiều chiều và các phép toán đại số tuyến tính. Trong numpy, chiều của mảng gọi là *axes*, trong khi số chiều gọi là *rank*.

- Pandas là một trong những thư viện được dùng rộng rãi nhất trong Python cùng với Numpy. Pandas cung cấp nhiều đối tượng và phương thức cho các cấu trúc dữ liệu. Pandas là thư viện không thể thiếu cho chúng ta trong suốt quá trình xử lý dữ liệu, từ chuyển đổi hay ánh xạ dữ liệu thô sang dạng dữ liệu mà chúng ta mong muốn, nhằm có thể phân tích dễ dàng hơn.

- Scikit-learn là một thư viện mã nguồn mở dành cho học máy - một ngành trong trí tuệ nhân tạo, rất mạnh mẽ và thông dụng với cộng đồng Python, được thiết kế trên nền NumPy và SciPy. Scikit-learn chứa hầu hết các thuật toán machine learning hiện đại nhất, đi kèm với documentations, luôn được cập nhật.

1.5 Javascript

1.5.1 Giới thiệu về JavaScript

JavaScript là ngôn ngữ lập trình phổ biến nhất trên thế giới trong suốt 20 năm qua. Nó cũng là một trong số 3 ngôn ngữ chính của lập trình web:

- HTML: Giúp bạn thêm nội dung cho trang web.
- CSS: Định dạng thiết kế, bố cục, phong cách, canh lề của trang web.
- JavaScript: Cải thiện cách hoạt động của trang web.

Chi tiết cụ thể hơn có thể tham khảo trong tài liệu [18].

1.5.2 Ưu điểm của JavaScript

JavaScript có rất nhiều ưu điểm khiến nó vượt trội hơn so với các đối thủ, đặc biệt trong các trường hợp thực tế. Sau đây chỉ là một số lợi ích của JavaScript:

Bạn không cần một compiler vì web browser có thể biên dịch nó bằng HTML.

- Nó dễ học hơn các ngôn ngữ lập trình khác.
- Lỗi dễ phát hiện hơn và vì vậy dễ sửa hơn.
- Nó có thể được gắn trên một số element của trang web hoặc event của trang web như là thông qua click chuột hoặc di chuột tới.
- JS hoạt động trên nhiều trình duyệt, nền tảng...
- Bạn có thể sử dụng JavaScript để kiểm tra input và giảm thiểu việc kiểm tra thủ công khi truy xuất qua cơ sở dữ liệu.
- Nó giúp website tương tác tốt hơn với khách truy cập.

- Nó nhanh hơn và nhẹ hơn các ngôn ngữ lập trình khác.

1.5.3 Nhược điểm của JavaScript

Mọi ngôn ngữ lập trình đều có các khuyết điểm. Một phần là vì ngôn ngữ đó khi phát triển đến một mức độ như JavaScript, nó cũng sẽ thu hút lượng lớn hacker, scammer, và những người có ác tâm luôn tìm kiếm những lỗ hổng và các lỗi bảo mật để lợi dụng nó. Một số khuyết điểm có thể kể đến là:

- Dễ bị khai thác.
- Có thể được dùng để thực thi mã độc trên máy tính của người dùng.
- Nhiều khi không được hỗ trợ trên mọi trình duyệt.
- Có thể bị triển khai khác nhau tùy từng thiết bị dẫn đến việc không đồng nhất.

1.6 XML

1.6.1 Giới thiệu về XML

XML là ngôn ngữ đánh dấu với mục đích chung do W3C đề nghị, để tạo ra các ngôn ngữ đánh dấu khác. Đây là một tập con đơn giản của SGML (một hệ thống tổ chức và gắn thẻ tài liệu.), có khả năng mô tả nhiều loại dữ liệu khác nhau (Foundation, XML, n.d.).

Mục đích chính của XML là đơn giản hóa việc chia sẻ dữ liệu giữa các hệ thống khác nhau, đặc biệt là các hệ thống được kết nối với Internet.

1.6.2 Đặc điểm của XML

XML là dữ liệu độc lập. Đây cũng là ưu điểm lớn nhất của file XML. Nó được sử dụng để mô tả dữ liệu dưới dạng text. Vì vậy, hầu hết các phần mềm hay chương trình bình thường đều có thể đọc được nó.

File XML có thể dễ dàng đọc và phân tích các nguồn dữ liệu. Do đó, nó được sử dụng chính vào mục đích trao đổi dữ liệu giữa các chương trình, hệ thống với nhau.

File XML được tạo một cách dễ dàng chỉ với vài thao tác đơn giản.

File XML được sử dụng dành cho Remote Procedure Calls với mục đích phục vụ các dịch vụ trong thiết kế website.

CHƯƠNG 2: PHÂN LOẠI VĂN BẢN TRONG XỬ LÝ NGÔN NGỮ TỰ NHIÊN

2.1 Tiền xử lý dữ liệu

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 $S = \{l_1, l_2, l_3, \dots, l_m\}$ với $l_1, l_2, l_3, \dots, l_m$ là các âm tiết đơn được tách nhau bởi khoảng trắng trong câu. Chúng ta sẽ bắt đầu duyệt từ đầu chuỗi. Xét xem l_1 có phải là từ có trong từ điển không, sau đó tới $l_1-l_2, l_1-l_2-l_3, \dots, l_1-l_2-l_3-\dots-l_n$ 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 nhập 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 l_1 và l_1-l_2 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 $L_1 - L_2 - L_3 - L_4$ có thể tách thành 3 cách:

$L_1, L_2 - L_3, L_4$

$L_1 - L_2, L_3 - L_4$

$L_1 - L_2 - L_3, L_4$

Thì khi đó cách tách thứ hai sẽ được chọn và từ $L_1 - L_2$ 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-based 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 máy 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 $\epsilon \times P$. Trọng số biểu diễn chi phí ước lượng (estimated cost) được cho bằng công thức:

$$C = -\log \frac{f}{N}$$

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 nhiễu 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 Chuyển đổi văn bản sang mô hình không gian véc-tơ

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 t_i trong văn bản d_j . 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 t_i 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:

$$w_{ij} = \begin{cases} 1, & t_i \in d_j \\ 0, & t_i \notin d_j \end{cases}$$

2.2.2 TF-IDF vector

Tf – Idf (Term Frequency – Inverse Document Frequency) 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

$$Tf(t, d) = f(t, d) = \frac{Ns(t)}{W}$$

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:

$$Tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}$$

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 thể 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 FREQUENCY)

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.

$$Idf(t, D) = \log\left(\frac{D}{d \in D : t \in d}\right)$$

Trong đó D là số lượng 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 $1 + (d \in D : t \in d)$ 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 thể 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:

$$Tf - Idf(i, j) = Tf(i, j) * Idf(i)$$

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 t_i xuất hiện trong văn bản d_j thì trọng số của từ đó trong véc-tơ đại diện sẽ là giá trị TF-IDF(t_i, d_j), ngược lại là 0. Ta có thể biểu diễn nó thành công thức như sau:

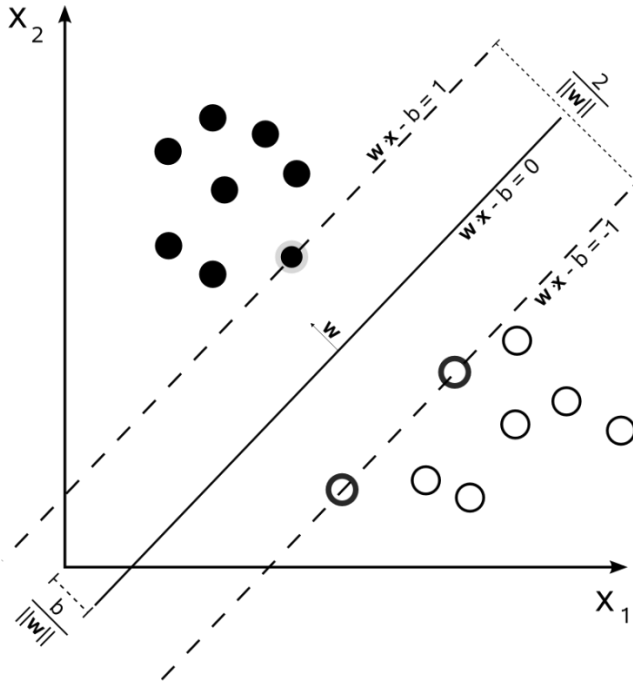
$$w_{ij} = \begin{cases} Tf - Idf(t_i, d_j), & t_i \in d_j \\ 0, & t_i \notin d_j \end{cases}$$

2.3 Các phương pháp phân loại văn bản bằng máy học

2.3.1 Phương pháp SVM (Support vector machine)

Máy học véc-tơ hỗ trợ (SVM) là một giải thuật học máy được xây dựng dựa trên lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng. Có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn.

Ý tưởng của phương pháp này là cho trước một tập huấn luyện được biểu diễn trong không gian véc-tơ, trong đó mỗi văn bản được xem như là một điểm trong không gian này. Phương pháp này tìm ra một mặt siêu phẳng h quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng, tạm gọi là lớp + (dương) và lớp - (âm). Như vậy, bộ phân loại SVM là một mặt siêu phẳng tách các mẫu thuộc lớp dương ra khỏi các mẫu thuộc lớp âm với độ chênh lệch lớn nhất. Độ chênh lệch này hay còn gọi là khoảng cách biên được xác định bằng khoảng cách giữa mẫu dương và mẫu âm gần mặt siêu phẳng nhất (hình). Khoảng cách này càng lớn các mẫu thuộc hai lớp càng được phân chia rõ ràng, nghĩa là sẽ đạt được kết quả phân loại tốt. Mục tiêu của thuật toán SVM là tìm được khoảng cách biên lớn nhất để tạo được kết quả phân loại tốt.



Hình 2.1 Siêu phẳng với lề cực đại cho một SVM phân tách dữ liệu thuộc hai lớp

Phương trình mặt siêu phẳng chứa véc-tơ x trong không gian đối tượng như sau:

$$w \cdot x + b = 0$$

Trong đó w là véc-tơ trọng số, b là độ dịch. Hướng và khoảng cách từ gốc tọa độ đến mặt siêu phẳng thay đổi khi thay đổi w và b . Bộ phân loại SVM được định nghĩa như sau:

$$f(x) = \text{sign}(w \cdot x + b)$$

Trong đó:

$$\begin{cases} \text{sign}(z) = +1 & z \geq 0 \\ \text{sign}(z) = -1 & z < 0 \end{cases}$$

Gọi y_i mang giá trị $+1$ hoặc -1 . Nếu $y_i = +1$ thì x thuộc về lớp dương, ngược lại $y_i = -1$ thì x thuộc về lớp âm.

Hai mặt siêu phẳng tách các mẫu thành hai phần được mô tả bởi các phương trình:

$$w \cdot x + b = 1$$

$$w \cdot x + b = -1$$

Bằng hình học ta có thể tính khoảng cách giữa hai mặt siêu phẳng này là $\frac{2}{\|w\|}$. Để khoảng cách biên là lớn nhất, ta phải tìm giá trị nhỏ nhất của $\|w\|$. Đồng thời ngăn chặn các điểm dữ liệu rơi vào vùng bên trong biên, chúng ta thêm ràng buộc sau:

$$w \cdot x_i + b \geq 1 \quad \text{với mẫu dương}$$

$$w \cdot x_i + b \leq -1 \quad \text{với mẫu âm}$$

Có thể gộp lại thành:

$$y_i(w \cdot x_i + b) > 1, \text{ với } i \in (1, n)$$

Khi đó để tìm mặt siêu phẳng h ta sẽ giải bài toán tìm $\min \|w\|$ với w và b thỏa điều kiện sau:

$$\forall i \in (1, n): y_i(w \cdot x_i + b) \geq 1$$

2.3.2 Phương pháp Naive Bayes

Naive Bayes đã được nghiên cứu rộng rãi từ những năm 1950. Được dùng lần đầu tiên trong lĩnh vực phân loại vào đầu những năm 1960. Sau đó nó trở nên phổ biến và được sử dụng rộng rãi trong lĩnh vực này cho đến ngày nay. Ý tưởng cơ bản của cách tiếp cận này là sử dụng xác suất có điều kiện giữa từ hoặc cụm từ và chủ đề để dự đoán xác suất chủ đề của một tập tin cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong tập tin đều độc lập với nhau. Ví dụ một loại trái cây có thể được cho là quả táo nếu nó đỏ, tròn và đường kính là 10cm. Giải thuật Naive Bayes sẽ cho rằng mỗi tính năng này đều đóng góp một cách độc lập để xác suất trái cây này là quả táo bất kể sự hiện diện hay vắng mặt của các tính năng khác.

Thuật toán Naive Bayes dựa trên định lý Bayes được phát biểu như sau:

$$P(Y|X) = \frac{P(XY)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

Trong đó:

$P(Y|X)$ là xác suất X thuộc lớp Y.

$P(X|Y)$ là xác suất một phần tử thuộc lớp Y, và phần tử đó có đặc điểm X.

$P(Y)$ xác suất xảy ra lớp Y, mức độ thường xuyên lớp Y xuất hiện trong tập dữ liệu

$P(X)$ xác suất xảy ra lớp X

Ví dụ: Giả sử ta có hai lớp $Y_1 = \text{Nam}$, $Y_2 = \text{nữ}$. Và một người không biết giới tính là Phương, $X = \text{Phương}$. Việc xác định Phương là Nam hay Nữ tương đương với việc so sánh xác suất $P(\text{Nam}/\text{Phương})$ và $P(\text{Nữ}/\text{Phương})$. Theo thuật toán Naive Bayes ta có công thức như sau:

$$P(\text{Nam}|\text{Phương}) = \frac{P(\text{Phương}|\text{Nam})P(\text{Nam})}{P(\text{Phương})}$$

Trong đó : $P(\text{Nam}|\text{Phương})$: xác suất Phương là nam

$P(\text{Nam}|\text{Phương})$: xác suất những người phái nam có tên Phương.

$P(\text{Nam})$: xác suất phái nam trong tập dữ liệu.

$P(\text{Phuong})$: xác suất tên Phuong trong tập dữ liệu.

Tương tự ta có :

$$P(\text{Nữ}|\text{Phuong}) = \frac{P(\text{Phuong}|\text{Nữ})P(\text{Nữ})}{P(\text{Phuong})}$$

Giả sử ta có bảng dữ liệu tên và giới tính như sau :

Bảng 2.1 Dữ liệu tên và giới tính

Tên	Giới Tính
Phuong	Nam
Nga	Nữ
Hồng	Nữ
Nam	Nam
Phuong	Nữ
Phuong	Nữ
Tiến	Nam
Giang	Nữ
Tùng	Nam
Đài	Nữ

$$P(\text{Nam}|\text{Phuong}) = \frac{\frac{1}{4} * \frac{4}{10}}{\frac{3}{10}} = \frac{1}{3}$$

$$P(\text{Nữ}|\text{Phuong}) = \frac{\frac{2}{6} * \frac{6}{10}}{\frac{3}{10}} = \frac{2}{3}$$

Như vậy Phuong là Nữ có xác suất cao hơn nên Phuong được phân vào lớp nữ khi phân loại.

Áp dụng vào bài toán phân loại văn bản :

Tập dữ liệu đã được véc-tơ hóa $D = (d_1, d_2, \dots, d_n)$

Tập các lớp $C = (C_1, C_2, \dots, C_m)$

Các thuộc tính độc lập đôi một với nhau.

Khi đó ta có :

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

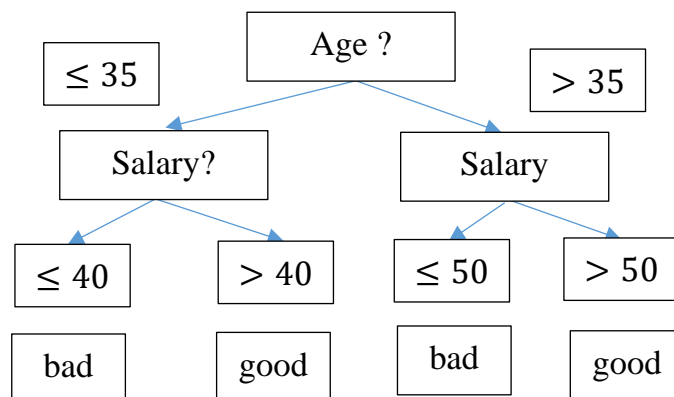
$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$$

Với $P(x_k|C_i)$ xác suất thuộc tính thứ k mang giá trị x_k khi đã biết X thuộc phân lớp i. Ưu điểm của giải thuật này đó là việc giả định rằng sự xuất hiện của tất cả các từ trong tập tin đều độc lập với nhau làm cho việc tính toán Naive Bayes hiệu quả và nhanh chóng vì không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề. Một lợi thế nữa của Naive Bayes là nó chỉ đòi hỏi một lượng nhỏ dữ liệu huấn luyện để ước lượng các tham số cần thiết để phân loại. Bởi vì các biến được giả định độc lập với nhau, nên chỉ có các phương sai của các biến cho mỗi lớp cần phải được xác định và không phải là toàn bộ ma trận hiệp phương sai. Tuy nhiên nhược điểm của phương pháp này cũng chính là giả định đó, vì nó rất khó xảy ra trong thực tế.

2.3.3 Phương pháp cây quyết định (Classification and regression trees)

Học máy cây quyết định là sử dụng mô hình cây quyết định để dự đoán kết quả về giá trị mục tiêu của một sự vật, hiện tượng. Nghĩa là ánh xạ từ các quan sát của một sự vật, hiện tượng đến các kết luận về giá trị mục tiêu của sự vật, hiện tượng. Cây quyết định là một trong những cách tiếp cận được sử dụng rộng rãi trong thống kê, khai phá dữ liệu và học máy. Nó là một trong những kỹ thuật thành công nhất trong việc học máy phân loại. Trong mô hình cây phân loại, các nút lá là các phân lớp, các nhánh là các liên từ, tính năng dẫn đến các lớp đó.

Ví dụ: cây quyết định phân lớp mức lương



Hình 2.2 Cây quyết định mức lương

Có nhiều thuật toán xây dựng cây quyết định như CLS, ID3, C4.5, CART... nhưng nhìn chung quá trình xây dựng cây quyết định đều được chia thành ba giai đoạn cơ bản:

Xây dựng cây: thực hiện chia một cách đệ quy tập mẫu dữ liệu huấn luyện cho đến khi các mẫu ở mỗi nút lá thuộc cùng một lớp.

Cắt tỉa cây: nhằm tối ưu hóa cây. Công việc chính là trộn một cây con vào trong một nút lá.

Đánh giá cây: đánh giá độ chính xác của cây kết quả. Tiêu chí đánh giá là phần trăm số mẫu phân lớp đúng trên tổng số mẫu đưa vào.

Việc chọn thuật toán nào để có hiệu quả phân lớp cao tuy thuộc vào rất nhiều yếu tố, trong đó cấu trúc dữ liệu ảnh hưởng rất lớn đến kết quả của các thuật toán. Chẳng hạn như thuật toán ID3 và CART cho hiệu quả phân lớp rất cao đối với các trường dữ liệu số (quantitative value) trong khi đó các thuật toán như J48, C4.5 có hiệu quả hơn đối với các dữ liệu Qualitative value (ordinal, binary, nominal). Trong phần này chúng ta chỉ đi vào tìm hiểu giải thuật CART (Classification and Regression Tree).

Giải thuật CART [15] chấp nhận sự tham lam (nonbacktracking) cách tiếp cận cây quyết định được xây dựng từ trên xuống một cách đệ quy, bắt đầu với một bộ dữ liệu huấn luyện tập và các nhãn lớp của họ. Hầu hết giải thuật cây quyết định đều theo cách tiếp cận từ trên xuống. Tập dữ liệu huấn luyện được phân vùng một cách đệ quy thành tập hợp con nhỏ hơn trong lúc cây được xây dựng.

Đối với các phương pháp phân loại khác, cây quyết định tương đối dễ hiểu, đòi hỏi mức tiền xử lý dữ liệu đơn giản. Tuy nhiên hiệu quả phân lớp của cây quyết định, phụ thuộc rất nhiều vào huấn luyện (training) data.

2.3.4 K- Nearest Neighbor(KNN)

K- Nearest Neighbor hay còn gọi là K láng giềng gần nhất là phương pháp phân loại dựa trên hướng tiếp cận thống kê. Nó là một trong những phương pháp tốt nhất từ thời kì đầu của phân loại văn bản.

Ý tưởng của phương pháp này là khi cần phân loại một văn bản mới, thuật toán sẽ tính toán khoảng cách của tất cả các văn bản trong tập huấn luyện đến văn bản này để tìm ra tập K láng giềng gần nhất. Sau đó dùng khoảng cách này đánh trọng số cho tất cả các chủ đề. Khi đó trọng số của một chủ đề chính là tổng khoảng cách tất cả các văn bản nằm trong tập K láng giềng có cùng chủ đề. Những chủ đề mà không xuất hiện bất

kì văn bản nào trong tập K láng giềng thì có trọng số bằng 0. Sau đó các chủ đề sẽ được sắp xếp theo giá trị trọng số giảm dần, chủ đề có trọng số cao sẽ được chọn làm chủ đề của văn bản phân loại.

Để tính khoảng cách có thể áp dụng các công thức như độ đo Cosin(4.8), độ đo Euclid (4.9), hay công thức Manhattan:

$$d(A, j) = |x_{j1} - A_{j1}| + |x_{j2} - A_{j2}| + \dots + |x_{jn} - A_{jn}|$$

Trong đó:

$j = (x_{j1}, x_{j2}, \dots, x_{jn})$ là véc-tơ đặc trưng đại diện cho văn bản thứ j trong tập huấn luyện.

$A = (A_{j1}, A_{j2}, \dots, A_{jn})$ là véc-tơ đặc trưng của văn bản mới cần phân loại.

2.3.5 Linear Least Square Fit(LLSF)

Linear Least Square Fit là phương pháp phân loại dựa trên cách tiếp cận ánh xạ. LLSF sử dụng phương pháp hồi quy để học từ tập huấn luyện và các chủ đề có sẵn. Tập huấn luyện được biểu diễn dưới dạng một cặp véc-tơ đầu vào và đầu ra như sau:

Véc-tơ đầu vào một văn bản bao gồm các từ và trọng số

Véc-tơ đầu ra gồm các chủ đề cùng với trọng số nhị phân của văn bản ứng với véc-tơ đầu vào.

Giải phương trình các cặp véc-tơ đầu vào/ đầu ra, ta sẽ được ma trận đồng hiện của hệ số hồi quy của từ và chủ đề (matrix of wordcategory regression coefficients). Phương pháp này sử dụng công thức:

$$F_{LS} = \arg_{\mathbf{F}} \min \|\mathbf{FA} - \mathbf{B}\|^2$$

Trong đó:

A, B là ma trận đại diện tập dữ liệu huấn luyện (các cột trong ma trận tương ứng là các véc-tơ đầu vào và đầu ra).

FLS là ma trận kết quả chỉ ra một ánh xạ từ một văn bản bất kỳ vào véc-tơ của chủ đề đã gán trọng số.

Nhờ vào việc sắp xếp trọng số của các chủ đề, ta được một danh sách chủ đề có thể gán cho văn bản cần phân loại. Nhờ đặt ngưỡng lên trọng số của các chủ đề mà ta tìm được chủ đề thích hợp cho văn bản đầu vào. Hệ thống tự động học các ngưỡng tối ưu cho từng chủ đề, giống với KNN. Mặc dù LLSF và KNN khác nhau về mặt thống kê, nhưng ta vẫn tìm thấy điểm chung ở hoạt động của hai phương pháp là việc học ngưỡng tối ưu.

2.3.6 Các thông số đánh giá giải thuật

Một số chỉ số thông dụng được dùng để đánh giá một giải thuật máy học. Giả sử để đánh giá một bộ phân loại hai lớp tạm gọi là dương và âm:

- + Số đúng dương (TP - True positive): số phần tử dương được phân loại dương.
- + Số sai âm (FN – False negative): số phần tử dương được phân loại âm.
- + Số đúng âm (TN – True negative): số phần tử âm được phân loại âm.
- + Số sai dương (FP – False positive): số phần tử âm được phân loại dương.
- + TP Rate: tỉ lệ những phần tử được phân loại lớp x mà đúng trên tổng số những phần tử thực sự thuộc lớp x. Cho biết tỉ lệ x được phân loại đúng là bao nhiêu. Tương tự với recall.

$$TP\ rate = \frac{TP}{TP + FN}$$

- + FP rate: tỉ lệ những phần tử được phân loại lớp x, nhưng mà nó không thuộc lớp x (phân loại sai) chia cho tổng số những phần tử không thuộc lớp x. Cho biết lớp x bị phân loại sai bao nhiêu.

$$FP\ rate = \frac{FP}{FP + TN}$$

- + Độ chính xác (precision): tỉ lệ những phần tử thật sự là lớp x trên tổng số những phần tử được phân loại vào lớp x. Số kết quả chính xác chia cho số kết quả trả về.

$$Precision = \frac{TP}{TP + FP}$$

- + Độ bao phủ (recall): có ý nghĩa tương tự như TP rate.

$$Recall = \frac{TP}{TP + FN}$$

- + Độ đo F1: chỉ số cân bằng giữa độ chính xác (precision) và độ bao phủ (recall). Nếu độ chính xác và độ bao phủ cao và cân bằng thì độ đo F1 lớn, còn độ chính xác và hồi tưởng nhỏ và không cân bằng thì độ đo F1 nhỏ.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Các chỉ số này sẽ được dùng để đánh giá hiệu quả cây quyết định và máy học SVM trong [2].

2.4 Độ tương đồng giữa các véc-tơ (Cosine Similarity)

Sự tương đồng là đại lượng phản ánh cường độ mối quan hệ giữa hai đại lượng hay hai đặc trưng. Trong không gian véc-tơ, mỗi tin tức được biểu diễn thành một véc-tơ. Vì thế để tính độ tương đồng giữa các tin tức ta đi tính độ tương đồng của 2 véc-tơ được chuẩn hóa từ hai tập tin đó. Phát biểu bài toán tính độ tương đồng như sau: Xét 2 văn bản di và dj. Tính độ tương đồng giữa hai văn bản đó là tìm ra một giá trị của hàm $S(di, dj)$. Hàm $S(di, dj)$ được gọi là độ tương đồng giữa hai văn bản di và dj. Trên thực tế rất khó để tính được độ tương đồng có độ chính xác cao vì ngữ nghĩa chỉ có thể được hiểu đầy đủ trong một ngữ cảnh cụ thể. Trong luận văn này chúng tôi trình bày hai phương pháp đo độ tương đồng giữa các véc-tơ là Cosin và Euclid.

Giả sử ta đi ta có hai véc-tơ cần tính độ tương đồng $v1$ và $v2$

$$V1 = (k11, k12, k13, \dots, k1n)$$

$$V2 = (k21, k22, k23, \dots, k2n)$$

+ Tính theo độ đo Cosin

$$Sim(v1, v2) = \frac{v1 * v2}{|v1| * |v2|}$$

Với $v1 * v2$ là tích vô hướng hai véc-tơ $v1, v2$.

$|v1| * |v2|$ là tích độ dài các véc-tơ $v1, v2$.

Giá trị sim có giá trị là -1 nghĩa là hai véc-tơ hoàn toàn khác nhau và càng gần về một thì độ tương đồng giữa hai véc-tơ càng cao.

+ Tính theo độ đo Euclid

$$Sim(v1, v2) = \sqrt{(k21 - k11)^2 + (k22 - k12)^2 + \dots + (k2n - k1n)^2}$$

Giá trị sim theo độ đo Euclid có giá trị từ 0 đến 1. Khi giá trị sim càng nhỏ thì hai véc-tơ có độ tương đồng càng cao. Độ đo Euclid đánh giá độ tương đồng giữa hai véc-tơ bằng việc sử dụng khoảng cách giữa hai véc-tơ. Điều đó dẫn đến một nhược điểm đó là khi mà độ dài các véc-tơ quá lớn thì độ lệch giữa các véc-tơ có thể bị sai dẫn đến việc tính toán độ tương đồng không chính xác.

2.5 Các công trình nghiên cứu liên quan

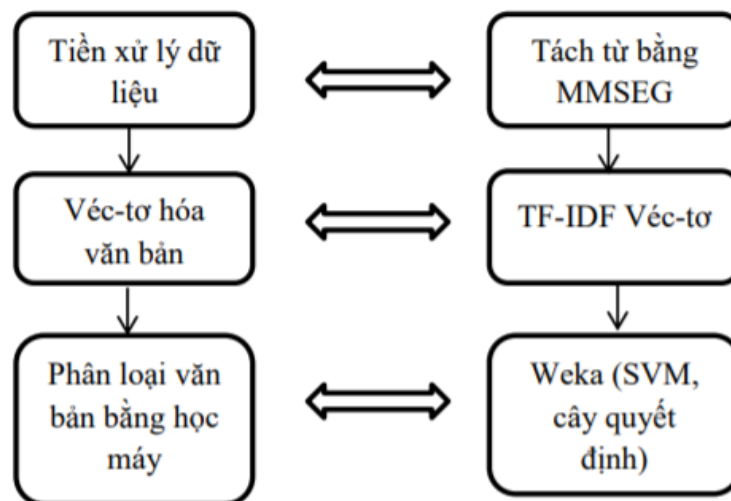
Phân loại văn bản nói chung hay phân loại tin tức nói riêng là một lĩnh vực đã có rất nhiều công trình nghiên cứu, bài báo, luận văn, đồ án, đề cập đến. Các công trình đó đều đạt được những kết quả hết sức khả quan và có nhiều điểm để học hỏi.

Trong chương này chúng tôi sẽ trình bày ba công trình nghiên cứu mà theo đánh giá chủ quan của chúng tôi là tương đối đơn giản, và chúng tôi tìm hiểu nhiều nhất để hoàn thành đồ án này.

2.5.1 Phân loại văn bản với máy học véc-tơ hỗ trợ và cây quyết định

Công trình nghiên cứu đầu tiên tôi tham khảo là bài báo nghiên cứu khoa học của hai tác giả Trần Cao Đệ và Phạm Nguyên Khang công tác tại Đại học Cần Thơ, được đăng trên Tạp chí Khoa học 2012:21a 52-63. Như tác giả đã viết trong [2]: *Bài viết này nghiên cứu máy học véc-tơ hỗ trợ (SVM), áp dụng nó vào bài toán phân loại văn bản và so sánh hiệu quả của nó với hiệu quả của giải thuật phân lớp cổ điển, rất phổ biến đó là cây quyết định*. Ngoài ra để áp dụng có hiệu quả giải thuật SVM, tác giả đã sử dụng kỹ thuật phân tích giá trị đơn Singular Value Decomposition để rút ngắn số chiều của không gian đặc trưng, từ đó giảm nhiều quá trình phân loại.

Tác giả tiến hành phân loại văn bản theo trình tự như sau:



Hình 2.3 Quy trình phân loại văn bản của [2]

Trong giai đoạn tiền xử lý dữ liệu, tác giả sử dụng giải thuật MMSEG để tiến hành tách từ. Đây là giải thuật được dùng phổ biến để tách từ tiếng Trung Quốc với độ chính xác 99%^[2] và đã được áp dụng vào tách từ Tiếng Việt thành công trong nhiều công trình. Theo nghiên cứu của tác giả giải thuật này khi áp dụng vào tách từ tiếng Việt sẽ cho độ chính xác trên 95%^[2]. Sau khi tách từ tác giả tiến hành mô hình hóa văn bản thành dạng véc-tơ, sử dụng TF-IDF véc-tơ hóa; tiến hành phân loại văn bản với hai giải thuật SVM và cây quyết định trong phần mềm Weka. Với tập dữ liệu là 7842 văn bản thuộc 10 chủ đề khác nhau, ứng với mỗi chủ đề, tác giả chọn ra 500 văn

bản một cách ngẫu nhiên để tiến hành huấn luyện, số văn bản còn lại để kiểm chứng độc lập. Để huấn luyện SVM, tập ngữ liệu đang xét sẽ được phân tích giá trị đơn và rút ngắn số chiều. Kết quả về hiệu quả phân loại văn bản với cây quyết định và máy học SVM được tác giả thể hiện qua bảng 2.2:

Bảng 2.2 Kết quả phân loại của [2]

Tên lớp	Cây quyết định			Máy học SVM		
	Precision	Recall	F1	Precision	Recall	F1
CNTT	84.5%	87.4%	85.9%	89.5%	92.7%	91.1%
ĐTVT	81.9%	80.5%	81.2%	88.2%	87.2%	87.7%
Giáo dục	83.4%	77.3%	80.2%	90.2%	92.3%	91.2%
Âm thực	83.8%	86.9%	85.3%	93.2%	93.8%	93.5%
Bất động sản	81.5%	84.9%	83.2%	91.9%	94.0%	92.9%
Khoa học	84.3%	80.1%	82.2%	90.0%	89.0%	89.5%
Kinh tế	86.2%	83.5%	84.8%	91.0%	87.3%	89.1%
Y học	84.9%	89.9%	87.3%	91.2%	89.9%	90.5%
Thể thao	84.3%	94.8%	89.2%	91.8%	93.4%	92.6%
Giải trí	85.5%	78.6%	81.9%	92.8%	90.0%	91.4%
	Trung bình		84.1%	Trung bình		91.0%

Qua kết quả thực nghiệm có thể thấy phân lớp với SVM thực sự tốt hơn phân lớp bằng cây quyết định. Ngoài ra, việc dùng SVD để phân tích và rút gọn số chiều của không gian đặc trưng đã nâng cao hiệu quả phân lớp SVM.

2.5.2 Xây dựng hệ thống phân loại tài liệu Tiếng Việt

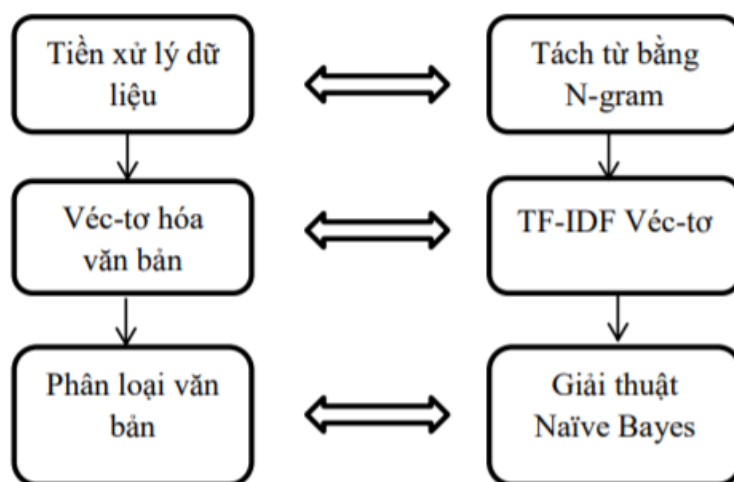
Công trình thứ hai mà tôi tham khảo là Bài báo Nghiên cứu Khoa học của hai tác giả Trần Thị Thu Thảo và Vũ Thị Chinh công tác tại Khoa Công nghệ thông tin, trường Đại học Lạc Hồng[4]. Đề tài này áp dụng phương pháp Naïve Bayes thực hiện phân loại trên đối tượng là các bài báo khoa học thuộc 9 chuyên ngành trong lĩnh vực Công nghệ thông tin.

Trong bài báo tác giả đưa ra các bước xử lý chung của quy trình phân loại văn bản qua sơ đồ:



Hình 2.4 Quy trình phân loại văn bản theo [4]

Sau đó tác giả tiến hành phân loại văn bản theo trình tự hình III.3:



Hình 2.5 Trình tự phân loại văn bản của [4]

Tác giả tiến hành xây dựng module tách từ theo mô hình N-gram, sau đó mô hình hóa văn bản đã được tách từ bằng véc-tơ TF-IDF. Với tập dữ liệu đã được mô hình hóa thành véc-tơ, tác giả tiến hành phân loại dựa trên phương pháp Naïve Bayes. Tác giả xây dựng phần mềm phân loại, tích hợp thêm các chức năng quản lý, sửa, xóa bài báo để tiến hành thử nghiệm trên tập dữ liệu là 281 bài báo khoa học thuộc các chuyên ngành của lĩnh vực CNTT.

Kết quả phân loại được thể hiện trong bảng 2.3:

Bảng 2.3 Kết quả phân loại văn bản [4]

STT	Tập dữ liệu	Phân loại tay	Phân loại máy	Phân loại sai chuyên ngành	Tỉ lệ (%)
1	Các hệ thống tính toán đi động	23	20	3	86.95
2	Công nghệ đa phương tiện	34	30	4	88.23

3	Công nghệ phần mềm	32	28	4	87.5
4	Cơ sở toán học của công nghệ thông tin	25	22	3	88
5	Hệ thống thông tin	40	35	5	87.5
6	Khoa học máy tính	26	23	3	88.46
7	Mạng máy tính và truyền thông	31	27	4	87.09
8	Trí tuệ nhân tạo	28	23	5	82.14
9	Xử lý ngôn ngữ tự nhiên và tiếng nói	42	38	4	90.47
Trung bình					87.37

Tuy kết quả phân loại đạt được khá khả quan tuy nhiên đề tài còn hạn chế về tập dữ liệu thử nghiệm và chưa có những so sánh đánh giá phương pháp Naïve Bayes với các phương pháp phân loại khác.

2.5.3 Phân loại email spam bằng matlab áp dụng 6 giải thuật

Đây là đồ án cuối khóa Phân Loại Email Spam môn Học Máy của hai tác giả Shahar Yifrah và Guy Lev[5]. Toàn bộ dữ liệu và source code của đồ án được đưa lên và tải miễn phí http://www.academia.edu/download/32673211/ML_Project_doc.pdf.

Trong đồ án này, tác giả sử dụng 6 giải thuật để tiến hành phân loại email spam bằng Matlab và so sánh hiệu suất giữa chúng. Sáu giải thuật đó bao gồm:

- Adaboost
- Naive Bayes
- Perceptron
- Winnow
- SVM (Support Véc-tơ Machine)
- KNN (K-Nearest Neighbors)

Tác giả đánh giá hiệu suất phân loại của các giải thuật dựa trên hai khía cạnh:

- Error rate: tỉ lệ để sót email spam.
- False positive ratio: (number of false positives) / (number of ham messages).

Tỉ lệ phân loại sai email bình thường thành email spam trên tổng số email bình thường.

Với tập dữ liệu là 5172 email, trong đó 29% email spam và 71% email thường. Tác giả xây dựng một Python script tiến hành tiền xử lý dữ liệu và véc-tơ hóa các email. Tiền xử lý dữ liệu hay tách từ là công việc khá đơn giản với ngôn ngữ tiếng Anh, bởi các từ được phân biệt với nhau bằng khoảng trắng. Cách véc-tơ hóa email: tác giả chọn ra 100 từ trong tập dữ liệu. Mỗi email được biểu diễn bằng tần suất xuất hiện của 100 từ đó trong chính nó. Những từ được chọn có thể là một trong những kí tự đặc biệt như ;[!\$#. Cách chọn ra 100 từ để xây dựng véc-tơ của tác giả: Đầu tiên với mỗi từ riêng biệt xuất hiện trong tập dữ liệu. Tác giả xếp hạng chúng theo thuộc tính Spamicity. Được tính theo công thức:

$$Spamcity(w) = \frac{Pr(w|S)}{Pr(w|S) + Pr(w|H)}$$

Trong đó: $(|)$: Xác suất từ w xuất hiện trong email spam. $(|)$: Xác suất từ w xuất hiện trong email thường. Các xác suất này được ước lượng thông qua tập huấn luyện. Những từ có Spamcity càng gần 1 thì càng là những từ đại diện cho email thường và ngược lại, spamcity của từ đó càng gần 0 thì từ đó thường đại diện cho email spam. Do đó những từ có giá trị spamcity càng xa 0.5 (lớn hơn hoặc nhỏ hơn) thì xếp hạng càng cao. Tuy nhiên trong trường hợp các xác suất $(|)$ $(|)$ quá nhỏ thì những từ đó rất khó được chọn cho dù nó có giá trị spamcity phù hợp. Để giải quyết vấn đề đó tác giả đưa ra trình tự lựa chọn 100 từ như sau:

- Lọc ra những từ có $|spamcity - 0.5| < 0.05$
- Lọc ra những từ hiếm, tần suất hiện nhỏ hơn 1% trong toàn bộ tập dữ liệu
- Với những từ chưa được lọc ra, tính giá trị $|Pr(w|S) - Pr(w|H)|$
- Chọn 100 từ có giá trị $|Pr(w|S) - Pr(w|H)|$ lớn nhất.

Ngoài việc véc-tơ hóa các email thì Python script còn tạo ra 90 cặp tập huấn luyện và tập kiểm thử. Tác giả chọn tỉ lệ kích thước tập huấn luyện so với kích thước tập dữ liệu tăng dần từ 0.1, 0.2, 0.3... đến 0.9. Ứng với mỗi tỉ lệ, lại chọn ra ngẫu nhiên 10 cặp tập huấn luyện và tập kiểm thử để tiến hành kiểm thử.

Kết quả đạt được thể hiện trong bảng 2.4:

Bảng 2.4 Kết quả phân loại theo [5]

	Minimizing Error (%)		Minimizing False Positive Ratio (%)	
	Error Rate	FP Ratio	Error Rate	FP Ratio
AdaBoost	8.7	4	NA	NA
Naive Bayes	7.6	3	20	0.5
Perceptron	18.5	12.5	NA	NA
Winnow	16.9	6.8	19.3	4.4
SVM	5	4.5	NA	NA
KNN	8.8	9.5	20	0.7

Chúng ta có thể thấy rằng hai giải thuật tốt nhất là SVM và Naive Bayes. Naive Bayes là giải thuật dễ hiện thực và thời gian chạy ngắn, lại cho hiệu suất cao không khó hiểu khi nó được sử dụng rộng rãi trong thực tiễn.

2.6 Nhận xét

Qua các công trình nguyên cứu và tạp chí khoa học thì hai giải thuật thường cho kết quả tốt nhất là SVM và Naive Bayes.

Yêu cầu thuật toán dùng trong bài toán phân loại tag câu hỏi:

- Phù hợp với tập dữ liệu nhỏ
- Tiết kiệm bộ nhớ : Do chỉ có một tập hợp con của các điểm được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ khi ra quyết định
- Dễ sử dụng và nhanh: Là 2 thuật toán dễ học và khá nhanh
- Thời gian chạy phân loại nhanh: Thích hợp để xây dựng phân loại tag trên diễn đàn ở chương 3.

Với thuật toán Naive Bayes thích hợp:

- Bài toán phân loại văn bản thuộc nhiều lớp (multi class prediction)
- Bài toán văn bản có 2 lớp như phân loại email spam thì Naive Bayes rất hiệu quả và chính xác
- Bài toán với dữ liệu dưới 1000 văn bản

➔ Thuật toán Naive Bayes không thích hợp với phân loại tag câu hỏi.

Chọn thuật toán Support Vector Machine để phân loại tag câu hỏi trên website hỗ trợ sinh viên vì:

- Xử lý trên không gian số chiều cao: SVM là một công cụ tính toán hiệu quả trong không gian chiều cao, trong đó đặc biệt áp dụng cho các bài toán phân loại văn bản và phân tích quan điểm nơi chiều có thể cực kỳ lớn.
- Tính linh hoạt - phân lớp thường là phi tuyến tính. Khả năng áp dụng Kernel mới cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn.
- Dữ liệu tương đối (1800 câu) nên cho độ chính xác cao.

CHƯƠNG 3: 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">
```

Câu hỏi
Và
Trả lời

```
      <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,  
      hỏi 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>
```

```
    </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.

```

1  [
2  [
3    "#10316: WIFI KỶ TỨC XÁ",
4    "Thưa nhà trường, chất lượng wifi ở ký túc xá quá tệ, em kính mong nhà trường xem xét. Em cảm ơn ạ",
5    " ",
6    "AD: Ban quản lý KTX đã yêu cầu Công ty iNet kiểm tra và thông tin trả lời các Bạn như sau:",
7    "Thời gian qua chất lượng wifi khá ổn - KTX thường xuyên thăm dò ý kiến sinh viên về vấn đề này. Chất lượng yếu chỉ mới xuất hiện
8    "Hiện nay công ty iNet có đặt văn phòng để xử lý về vấn đề wifi tại phòng 6-D31. Khi có vấn đề về wifi, các bạn vui lòng liên hệ
9    "Các bạn cũng có thể phản ánh với BQL KTX, để chúng tôi yêu cầu phía cung cấp dịch vụ xử lý.",
10   "Trân trọng!"
11  ],
12  [
13    "#10315: TỔNG HỢP",
14    "1. Em đi thực tập full tuần nên chỉ đánh giá vào buổi trưa hay buổi tối, vào buổi trưa ngày 4/9 em có đánh giá mà tại sao lúc gi
15    "2. Hiện tại em là sinh viên năm cuối và em không còn nợ môn gì hết và đủ điểm để nhận đồ án. Em muốn hỏi là\ em làm xong đồ án
16    "3. Ad cho em hỏi : Như kì chính mình học 14 chỉ mà hè mình trả nợ 3 chỉ vậy khi xét học bổng mình được xét khi kết quả học tập mĩ
17    "Với ad cho em hỏi ví dụ điểm tích lũy mình thấp hơn mấy bạn nhưng điểm rèn luyện mình cao hơn vậy khi xét học bổng sẽ dựa vào đi
18    "4. Dạ e kẻo cho e hỏi trường có tổ chức đăng kí học phần đợt 2 không a tại vì chỉ tiêu môn học đã đủ r ",
19    " ",
20    "AD:"
21  ]
22 ]

```

Hình 3.1 Dữ 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 dạng 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.

```

1  [
2  {
3    "question": "Thưa nhà trường, chất lượng wifi ở ký túc xá quá tệ, em kính mong nhà trường xem xét. Em cảm ơn ạ",
4    "answer": "AD: Ban quản lý KTX đã yêu cầu Công ty iNet kiểm tra và thông tin trả lời các Bạn như sau: Thời gian qua chất lượng wi
5    "content": "null"
6  },
7  {
8    "question": "1. Em đi thực tập full tuần nên chỉ đánh giá vào buổi trưa hay buổi tối, vào buổi trưa ngày 4/9 em có đánh giá mà tạ
9    "answer": "1. Có thể bạn ko để ý lúc lưu không thành công rồi, bạn lên Phòng 6 nhà D3 để được hỗ trợ nhé.",
10   "content": "null"
11  },
12  {
13    "question": "2. Hiện tại em là sinh viên năm cuối và em không còn nợ môn gì hết và đủ điểm để nhận đồ án. Em muốn hỏi là\ em là
14    "answer": "2. Bạn phải hoàn thành đồ án được Bộ môn đồng ý cho bảo lưu thì mới được bảo lưu nhưng chỉ được bảo lưu 6 tháng thời n
15    "content": "null"
16  },
17  {
18    "question": "3. Ad cho em hỏi : Như kì chính mình học 14 chỉ mà hè mình trả nợ 3 chỉ vậy khi xét học bổng mình được xét khi kết qu
19    "answer": "3. Bạn xem số tay Khóa 59 Trang 55 nhé: \'- Điều kiện để sinh viên được xét cấp học bổng: + Mức học bổng loại Khá: C
20    "content": "null"
21  },
22  {
23    "question": "4. Dạ e kẻo cho e hỏi trường có tổ chức đăng kí học phần đợt 2 không a tại vì chỉ tiêu môn học đã đủ r ",
24    "answer": "4. Bạn trao đổi với Phòng Đào tạo 10 nhà D3 xem sao nhé. Tùy vào số lượng Sinh viên Khóa 60 không đăng ký được nhiều h
25    "content": "null"
26  },
27  {
28    "question": "Mình là nữ k58 cần tìm phòng ở ghép. Vì chưa có xe nên cần phòng gần gần trường một chút ạ . Ai cần người hay biết c
29    "answer": "AD: Bạn gửi đến diễn đàn thì AD hỗ trợ đăng vậy thôi chứ chuyện ở ghép cũng có nhiều rủi ro nhất định, nhớ sàng lọc kỹ
30    "content": "null"
31  },
32  {
33    "question": "1. Cho em hỏi kì I em được xuất sắc và kì II được học sinh giỏi thì cả năm được xếp loại gì vậy ạ?",
34    "answer": "1. Còn tùy điểm của Bạn bao nhiêu và tổng số tính chỉ Bạn học nữa ạ, trong số tay Sinh viên có chỉ cách tính đó ạ, tra
35    "content": "null"
36  },
37  {
38    "question": "2. Cho e hỏi là thi Quốc phòng hơi có đc mang cuốn tài liệu phở ở bên ngoài vô đc ko a hay chỉ đc mang vở mình viết

```

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.

Phân loại câu hỏi giúp Nguyễn Minh Mẫn

Đã làm 201/1753

Câu hỏi:

4. chào admin, em là sinh viên k55, kì trước em có bảo lưu đồ án, theo lịch đồ án ngày 8/6 là kết thúc nhưng hiện tại chưa có thông báo lịch bảo vệ. admin cho em hỏi lịch bảo vệ đợt này ạ. em cảm ơn admin.

Trả lời:

4. em liên hệ với bộ môn nhé

Thuộc ban:

null

Đồng ý

Ghi chú: Những câu trả lời khó phân loại thì cho vào mục **KHÁC** hoặc xem lại các ví dụ phân loại mẫu ở [trang chủ](#).

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

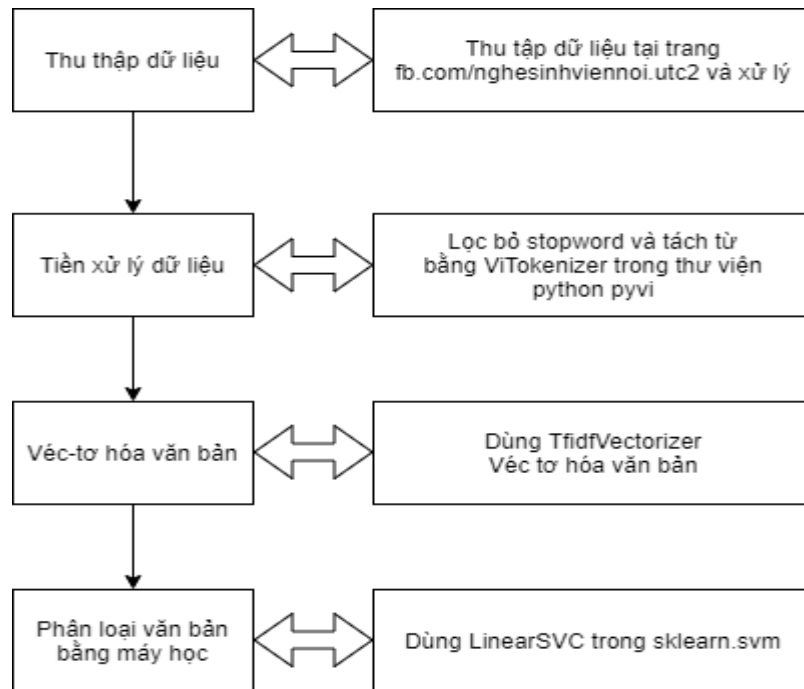
a) Đặt vấn đề

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 để 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 trúc 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:

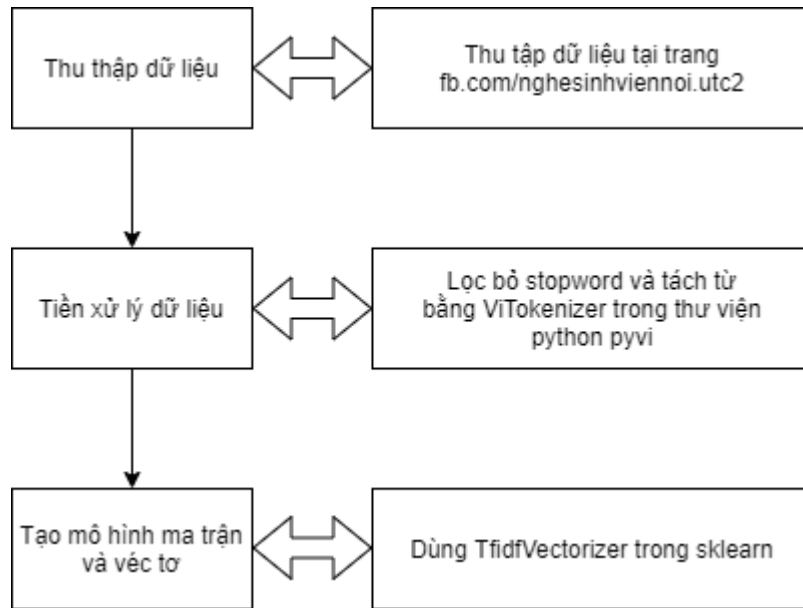


Hình 3.4 Quy trình hóa phân loại văn bản

QA Diễn đàn Cấu hình				
Tag				
Tên	Diễn đàn	Parent Tag	Subtag	Câu trả lời tùy chỉnh
<input type="checkbox"/> admin	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> ban quản lý ktx	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> ban thanh tra đào tạo	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> ban thông tin thư viện	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> ban công tác chính trị và sinh viên	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> ban khảo thí và đảm bảo chất lượng	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> ban đào tạo	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> ban thiết bị quản trị	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> ban tổ chức hành chính	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> trung tâm đào tạo thực hành	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> giảng viên	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> cố vấn	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> khác	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi
<input type="checkbox"/> đoàn và hội	Diễn đàn hỗ trợ sinh viên		Không có bảng ghi	Không có bảng ghi

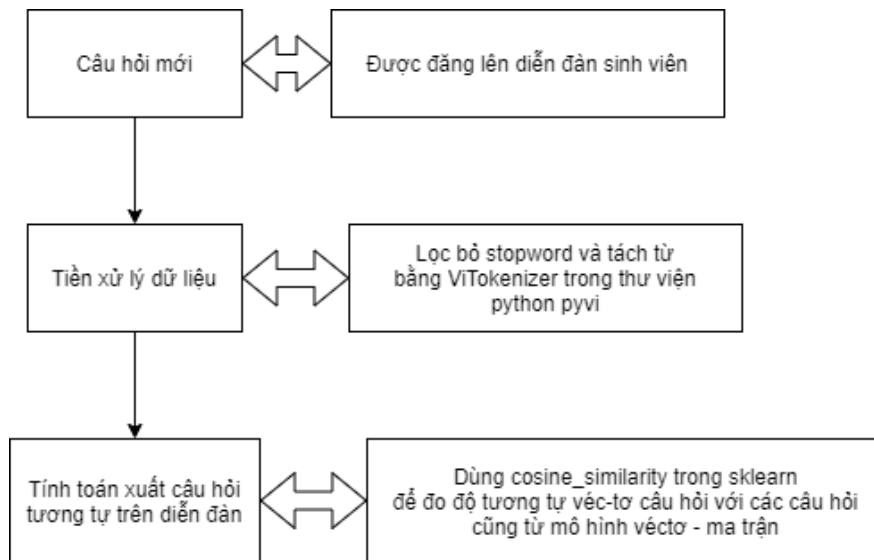
Hình 3.5 Các tag phân loại trên diễn đàn

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

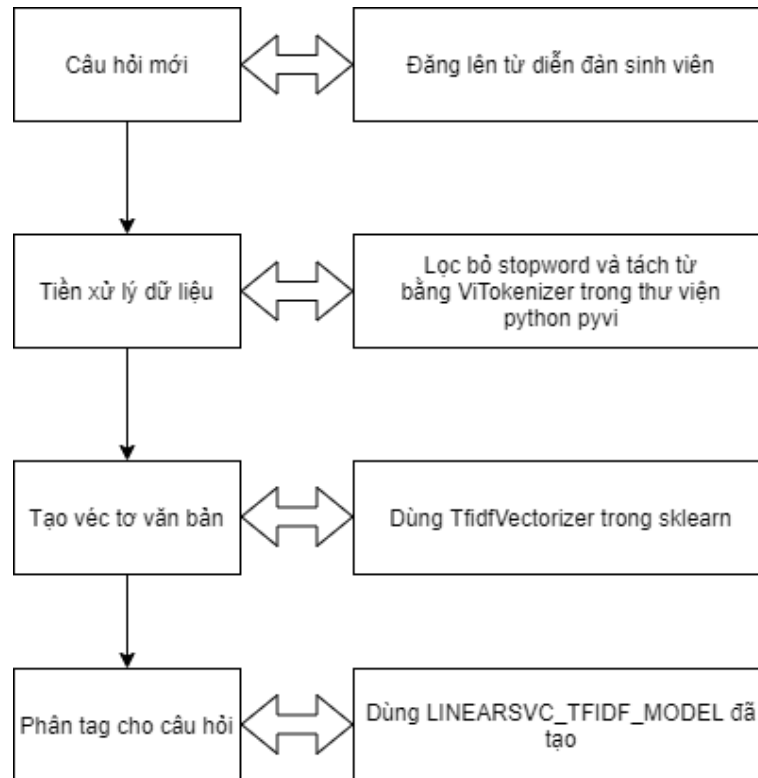
Câu hỏi mới được đưa vào hệ thống sẽ được tách câu, tách từ, loại bỏ các hư từ và loại bỏ các cụm từ xuất hiện nhiều nhưng không có ý nghĩa, để còn lại các từ cần thiết (từ khóa) cho việc phân loại và tìm câu tương tự.

Các từ khóa thu được trong giai trước sẽ được sử dụng để xây dựng 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 câu 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 đưa 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

```

1  [
2  {
3    "content": " 4. ad có thể vui lòng giúp em hỏi ban đào tạo giữ được ko ạ. do hồi đợt quên khảo sát nên em chưa đăng kí được đồ án tốt
4    "answer": " 4. bạn phải đi đến cùng thôi ad chúc bạn thành công",
5    "category": "ban đào tạo"
6  },
7  {
8    "content": " 4. em chào thầy/cô ạ! thầy/cô cho em hỏi là: sinh viên k50 có được phép chuyển qua hệ vừa học vừa làm đợt này kg ạ? vì thế
9    "answer": " 4. em đã liên hệ phòng đào tạo trả lời rồi, giờ thầy cũng không biết trả lời lại làm sao. nếu em chưa thuyết phục câu trả
10   "category": "ban đào tạo"
11 },
12 {
13   "content": " 4. chào ad (thầy, cô ạ.) em là sinh viên k55. cho em hỏi là hôm bữa em có đăng ký học ôn tiếng anh ở trường và chưa kịp đ
14   "answer": " 4. em liên hệ trực tiếp phòng đào tạo 10d3 để hướng dẫn đóng tiền em trực tiếp",
15   "category": "ban đào tạo"
16 },

```

Hình 3.9 Tổ chức tập dữ liệu train và test

+ Thực hiện train dữ liệu vào tạo mô hình LinearSVC

```

app.log x
log > app.log
1 INFO:root:(Training by SVM classification ... ', '2020-08-19 12:17:11.840679')
2 INFO:root:
3 Model: LinearSVC
4 C = 10
5 Penalty='l2'
6 Max_iter = 10
7 INFO:root:(Accuracy: ', 0.868421052631579)
8 INFO:root: precision recall f1-score support
9
10 admin 0.73 0.80 0.76 10
11 ban công tác chính trị và sinh viên 0.90 0.90 0.90 10
12 ban khảo thí & đảm bảo chất lượng 1.00 0.77 0.87 13
13 ban quản lý kỷ túc xá 1.00 1.00 1.00 15
14 ban thanh tra đào tạo 0.82 0.90 0.86 10
15 ban thiết bị quản trị 1.00 0.50 0.67 4
16 ban thông tin thư viện 1.00 1.00 1.00 2
17 ban tổ chức hành chính 0.91 1.00 0.95 10
18 ban đào tạo 0.81 1.00 0.89 25
19 giảng viên 1.00 0.60 0.75 5
20 khác 0.00 0.00 0.00 0
21 trung tâm đào tạo thực hành 1.00 0.50 0.67 8
22 đoàn và hội 0.67 1.00 0.80 2
23
24 accuracy 0.87 114
25 macro avg 0.83 0.77 0.78 114
26 weighted avg 0.90 0.87 0.87 114
27
28 INFO:root:(Training by SVM Done! ', '2020-08-19 12:17:26.776647')
29

```

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 dựa 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.

Mã môn	Đồ án tốt nghiệp<-Thực tập chuyên môn + Thực tập tốt nghiệp
Môn dự đoán	CNT04.10 / Đồ án tốt nghiệp
Chế độ	
Môn liên quan 1	CNT01.2 / Thực tập chuyên môn
Môn liên quan 2	CNT03.2 / Thực tập tốt nghiệp

Hình 3.11 Công thức dự đoán điểm

QL dự đoán điểm				
Quản lý dự đoán điểm Quản lý Công cụ				
Môn dự đoán				
<div> <div>Tạo</div> <div>Nhập</div> </div> <div> <div>Bộ lọc</div> <div>Nhóm theo</div> <div>Yêu thích</div> </div> <div>1/8 / 8</div>				
<input type="checkbox"/> Mã môn	Môn dự đoán	Chế độ	Môn liên quan 1	Môn liên quan 2
<input type="checkbox"/> Phân tích thiết kế thuật toán<Hệ điều hành + Cơ sở dữ liệu	MHT08.3 / Phân tích thiết kế thuật toán		MHT04.3 / Hệ điều hành	MHT05.3 / Cơ sở dữ liệu
<input type="checkbox"/> Hệ điều hành Unix<Cơ sở dữ liệu + Hệ điều hành	MHT10.2 / Hệ điều hành Unix		MHT05.3 / Cơ sở dữ liệu	MHT04.3 / Hệ điều hành
<input type="checkbox"/> Đồ án tốt nghiệp<Thực tập chuyên môn + Thực tập tốt nghiệp	CNT04.10 / Đồ án tốt nghiệp		CNT01.2 / Thực tập chuyên môn	CNT03.2 / Thực tập tốt nghiệp
<input type="checkbox"/> Thực tập tốt nghiệp<Thực tập chuyên môn	CNT03.2 / Thực tập tốt nghiệp		CNT01.2 / Thực tập chuyên môn	
<input type="checkbox"/> Trí tuệ nhân tạo<Phân tích thiết kế thuật toán + Thiết kế cơ sở dữ liệu	MHT07.3 / Trí tuệ nhân tạo		MHT08.3 / Phân tích thiết kế thuật toán	MHT09.2 / Thiết kế cơ sở dữ liệu
<input type="checkbox"/> Phân tích thiết kế hướng đối tượng<Phân tích thiết kế hệ thống + Lập trình trực quan	CPM07.3 / Phân tích thiết kế hướng đối tượng		CPM06.3 / Phân tích thiết kế hệ thống	CPM211.3 / Lập trình trực quan
<input type="checkbox"/> Khai phá dữ liệu<Phân tích thiết kế thuật toán + Lập trình Web	MHT12.3 / Khai phá dữ liệu		MHT08.3 / Phân tích thiết kế thuật toán	MHT208.3 / Lập trình Web
<input type="checkbox"/> Lập trình Web<Phân tích thiết kế thuật toán + Cơ sở dữ liệu	MHT208.3 / Lập trình Web		MHT08.3 / Phân tích thiết kế thuật toán	MHT05.3 / Cơ sở dữ liệu

Hình 3.12 Quản lý công thức dự đoán điểm

3.5 Sơ đồ usecase của hệ thống

Biểu đồ usecase biểu diễn sơ đồ chức năng của hệ thống. Từ các yêu cầu của hệ thống, biểu đồ usecase chỉ ra hệ thống cần thực hiện những điều gì để đáp ứng nhu cầu của người sử dụng hệ thống.

Người quản trị trang web (admin) thông qua đăng nhập để thực hiện các chức năng quản lý, thống kê báo cáo và các chức năng khác trên website.

Giảng viên đăng nhập vào hệ thống có thể quản lý diễn đàn, quản lý nhắn tin trực tuyến, dự đoán điểm.

Khách là những người không có gmail sinh viên của trường truy cập vào trang web có thể thực hiện các chức năng như: Xem câu hỏi trên diễn đàn, tra cứu câu hỏi, đóng góp ý kiến

Đối với thành viên là sinh viên có tài khoản gmail do trường đại học Giao Thông Vận Tải Phân Hiệu Tại Tp. Hồ Chí Minh cung cấp sau khi đăng nhập có thể thực hiện đăng các câu hỏi lên diễn đàn, tương tác với bài viết, thực hiện chức năng chat trực tuyến, xem điểm dự đoán. Cũng như có thể cập nhật thông tin tài khoản.

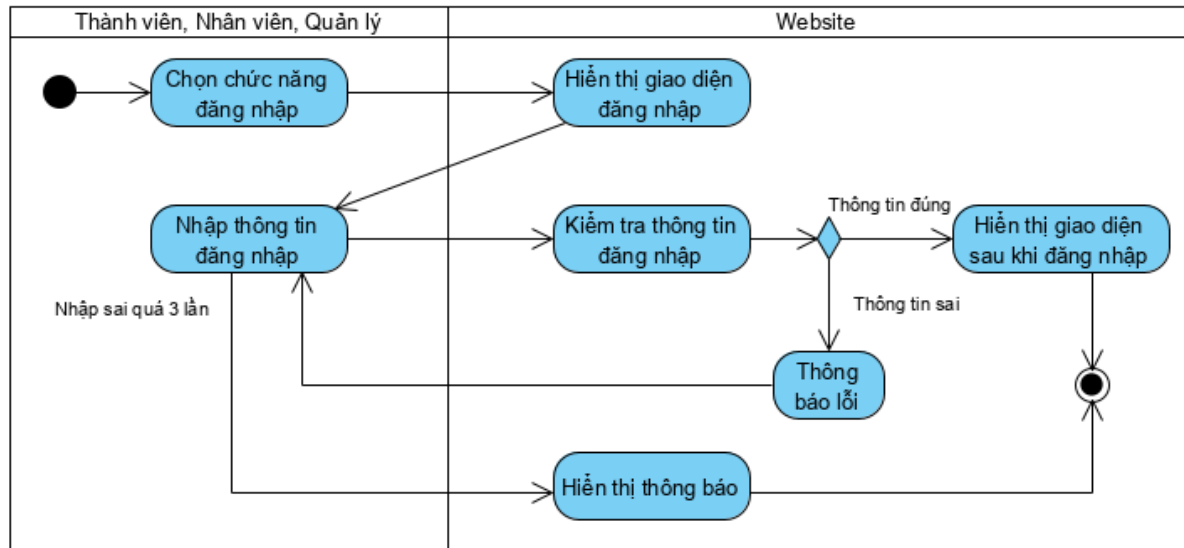


Hình 3.13 Sơ đồ usecase tổng quan của hệ thống

3.6 Sơ đồ hoạt động

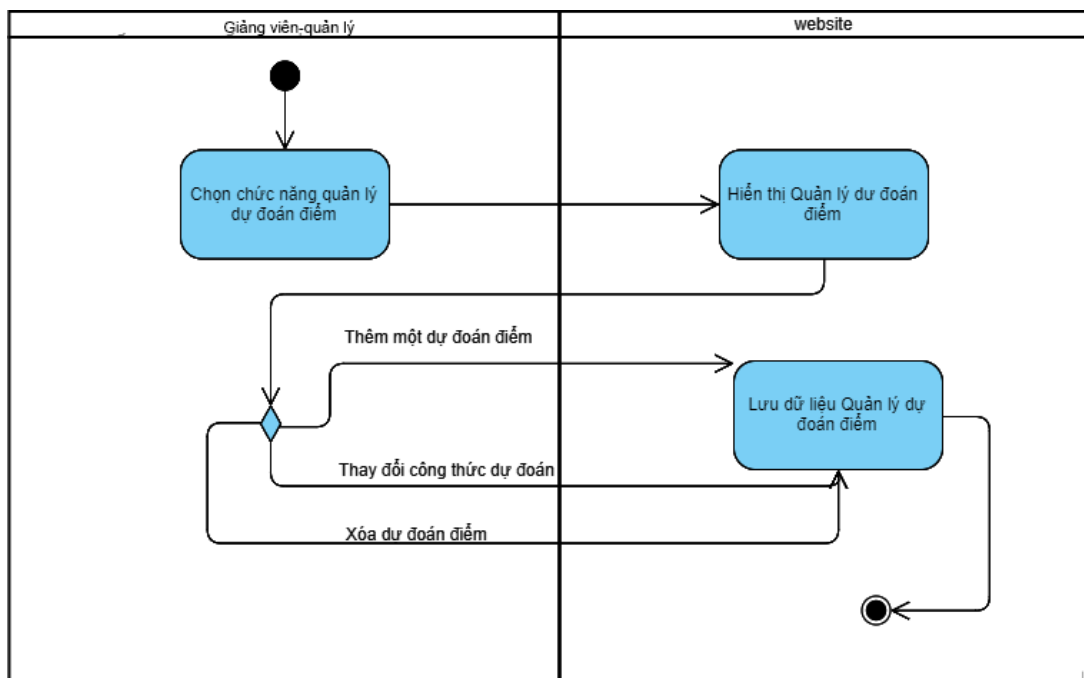
Sơ đồ hoạt động mô tả các bước thực hiện, các hành động, các nút quyết định và điều kiện rẽ nhánh để điều khiển luồng thực hiện của hệ thống. Nhờ vào sơ đồ hoạt động, có thể xác định được các ca sử dụng thông qua các quy trình nghiệp vụ.

3.6.2 Sơ đồ hoạt động đăng nhập



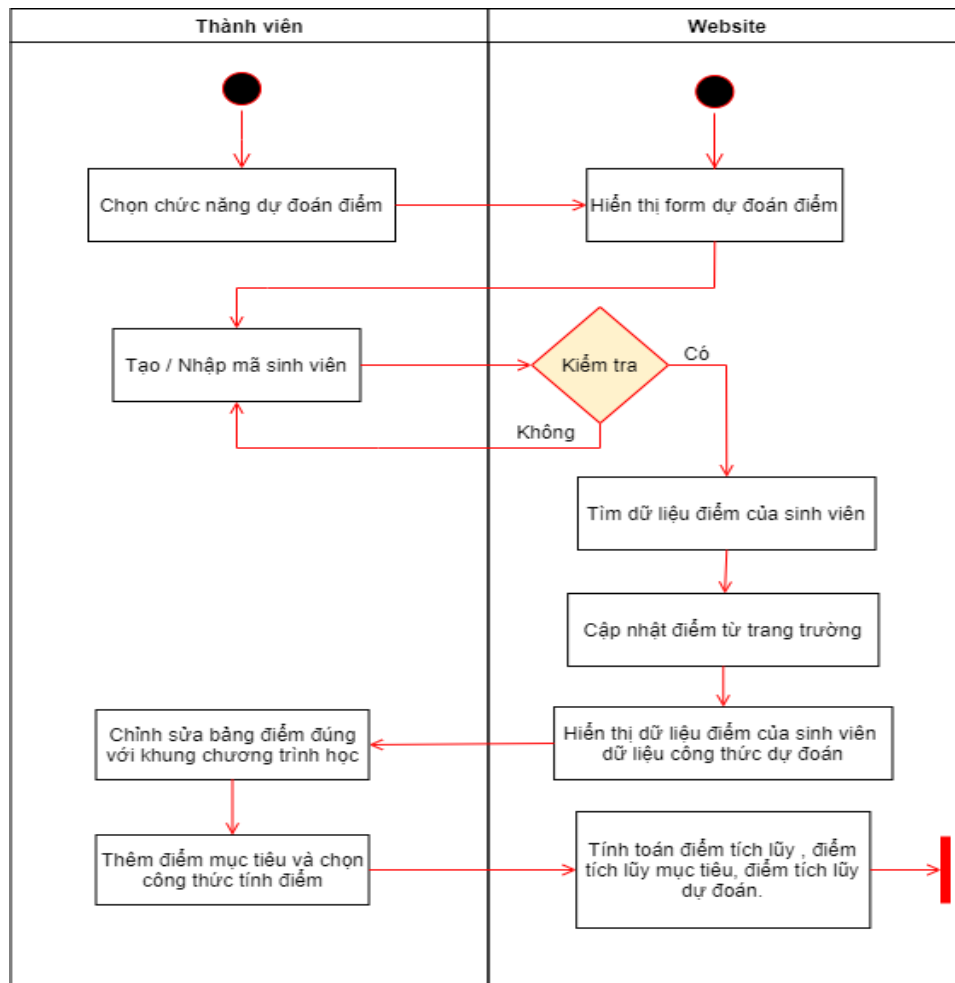
Hình 3.14 Sơ đồ hoạt động đăng nhập

3.6.11 Sơ đồ quản lý dự đoán điểm



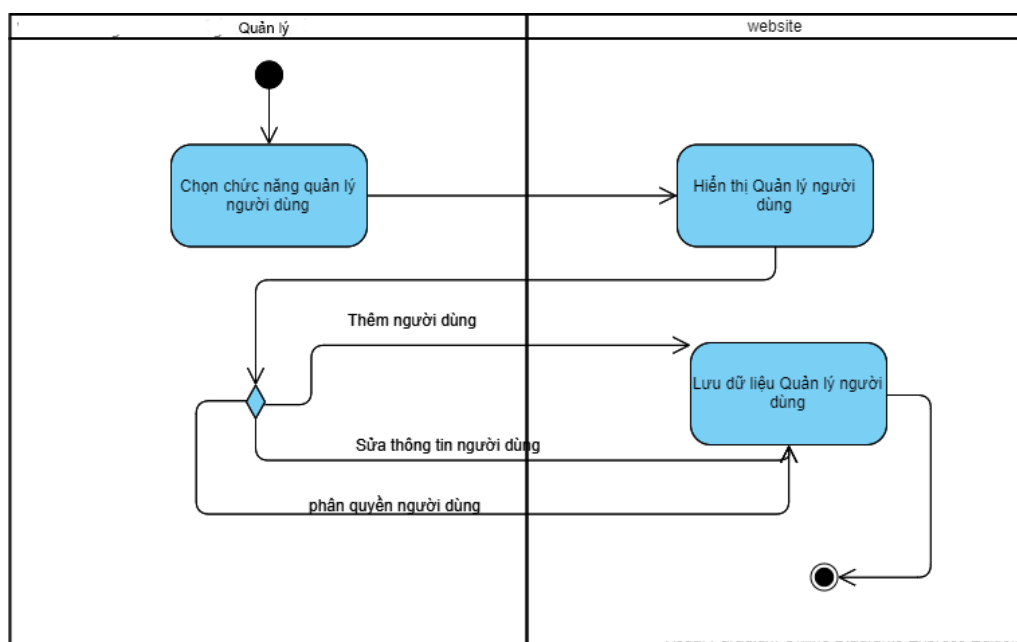
Hình 3.15 Sơ đồ hoạt động quản lý dự đoán điểm

3.5.6 Sơ đồ dự đoán điểm và tạo bảng điểm



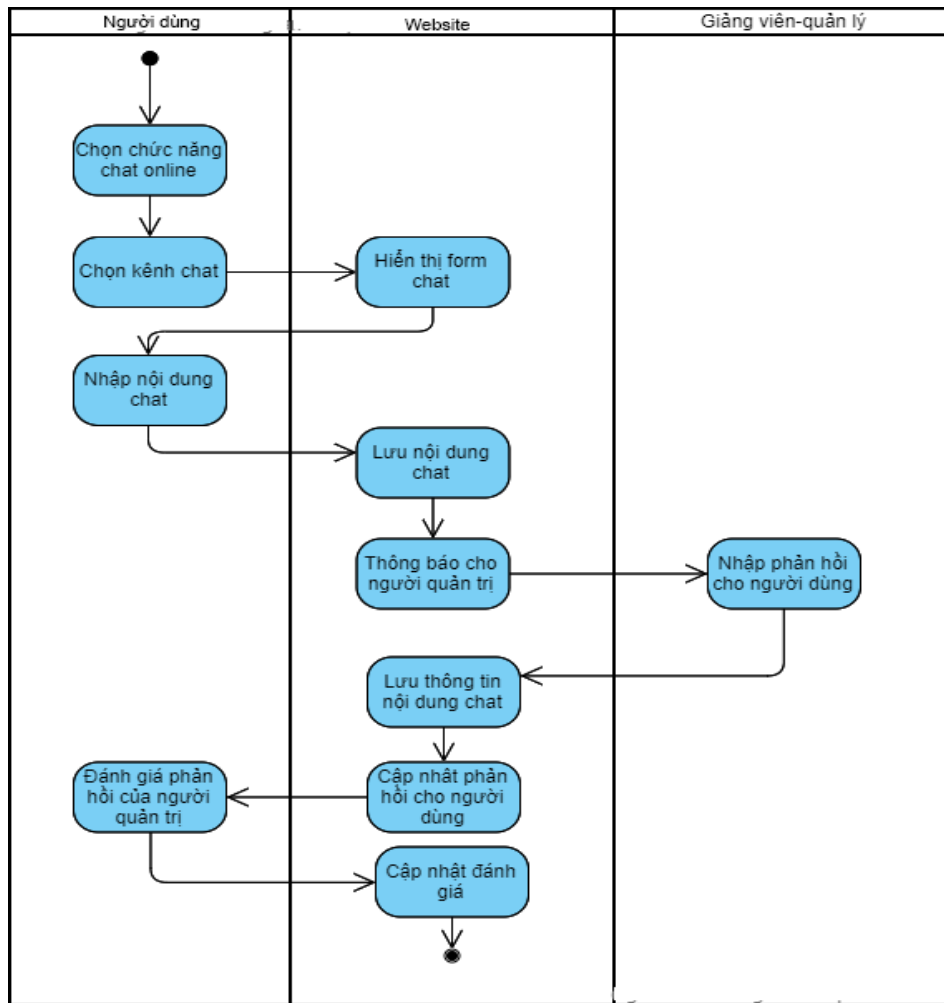
Hình 3.16 Sơ đồ hoạt động dự đoán điểm

3.6.8 Sơ đồ quản lý người dùng



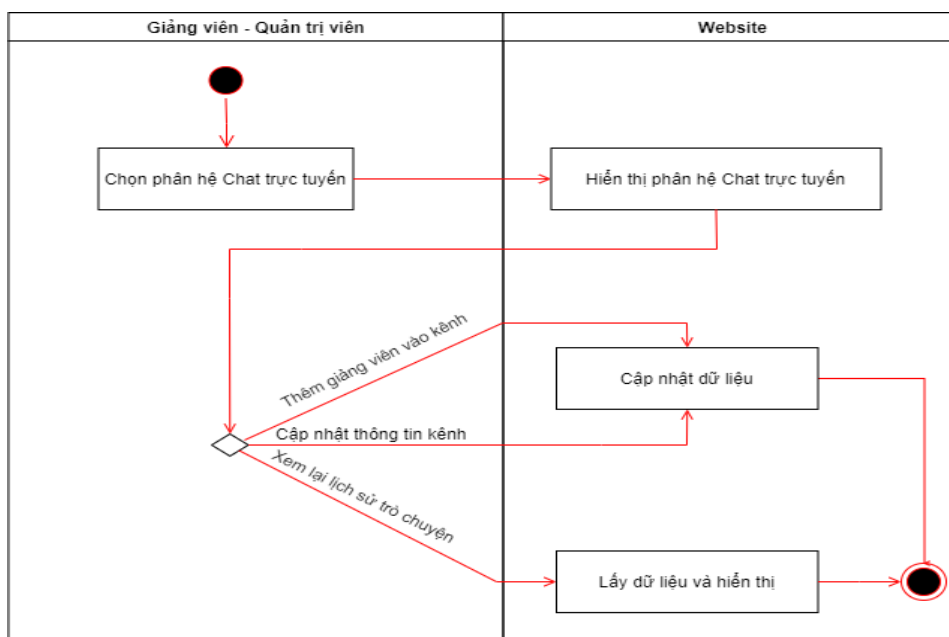
Hình 3.17 Sơ đồ hoạt động quản lý người dùng

3.6.7 Sơ đồ hỗ trợ chat trực tuyến



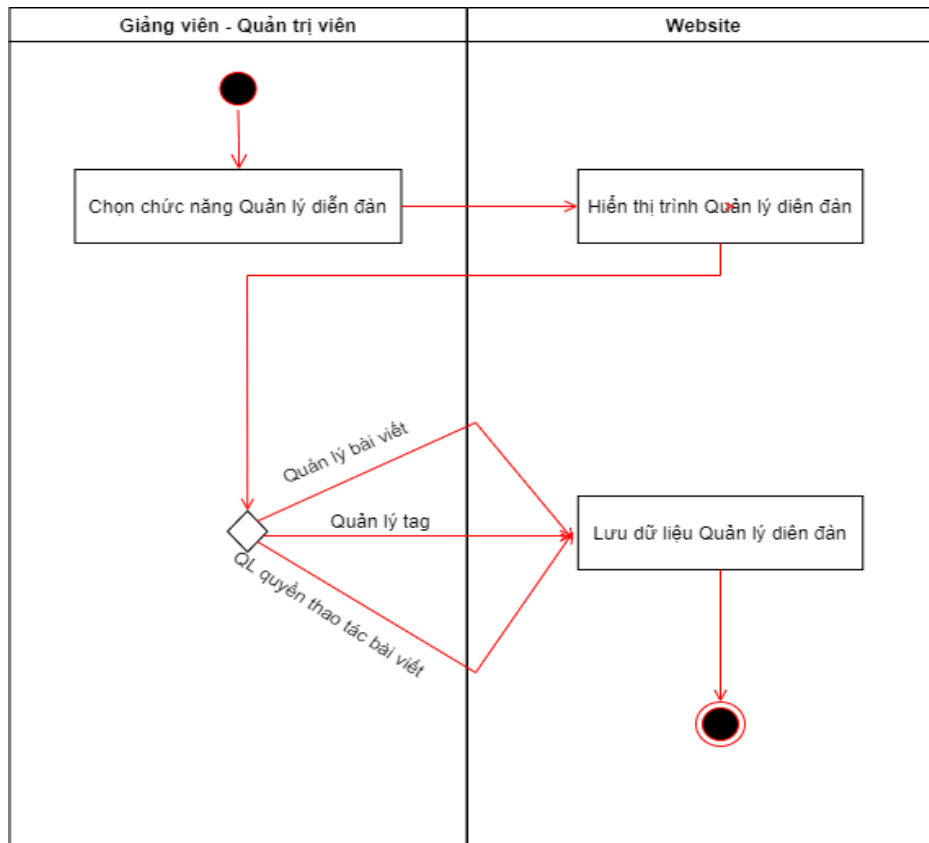
Hình 3.18 Sơ đồ hoạt động dự đoán điểm

3.6.9 Sơ đồ quản lý hỗ trợ trực tuyến



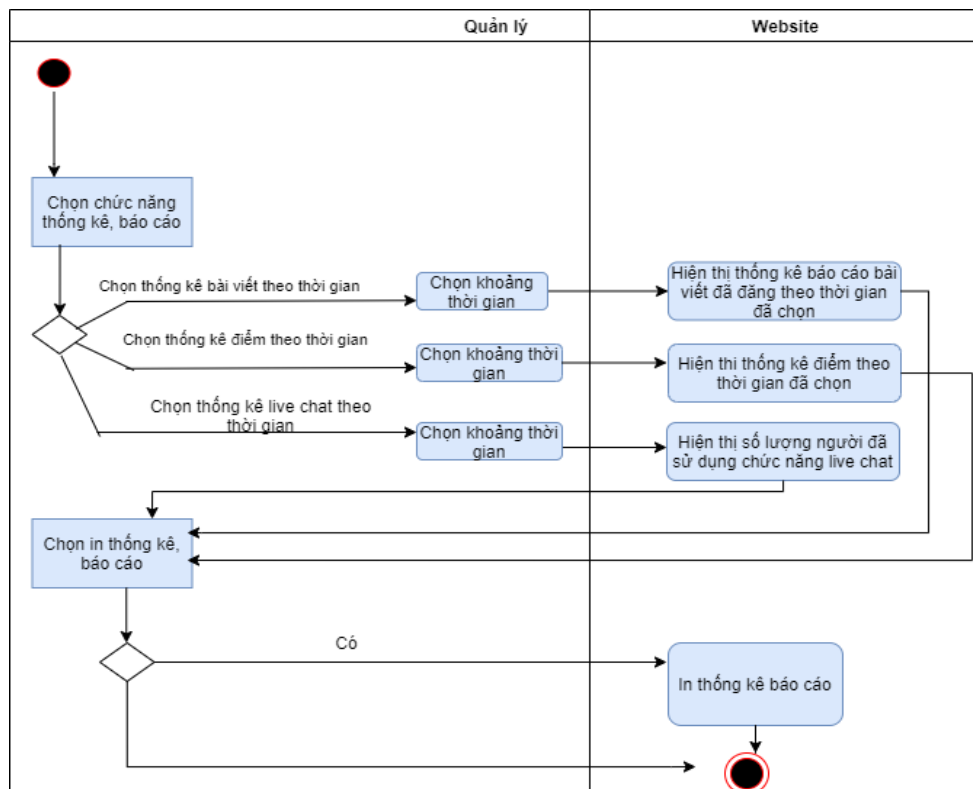
Hình 3.19 Sơ đồ hoạt động hỗ trợ trực tuyến

3.6.10 Sơ đồ quản lý diễn đàn



Hình 3.20 Sơ đồ hoạt động quản lý diễn đàn

3.6.1 Sơ đồ báo cáo thống kê



Hình 3.21 Sơ đồ hoạt động quản lý báo cáo thống kê

3.7 Xây dựng giao diện chương trình

3.7.1 Tạo tài khoản và đăng nhập

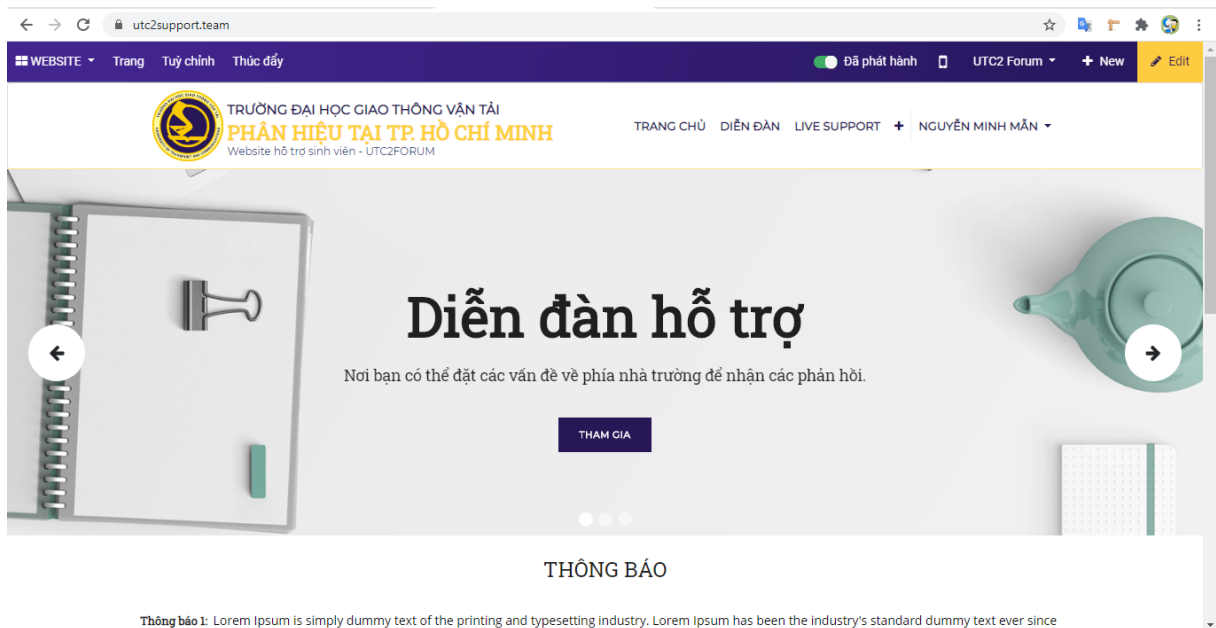
Hình 3.22 Giao diện tạo tài khoản đăng nhập

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

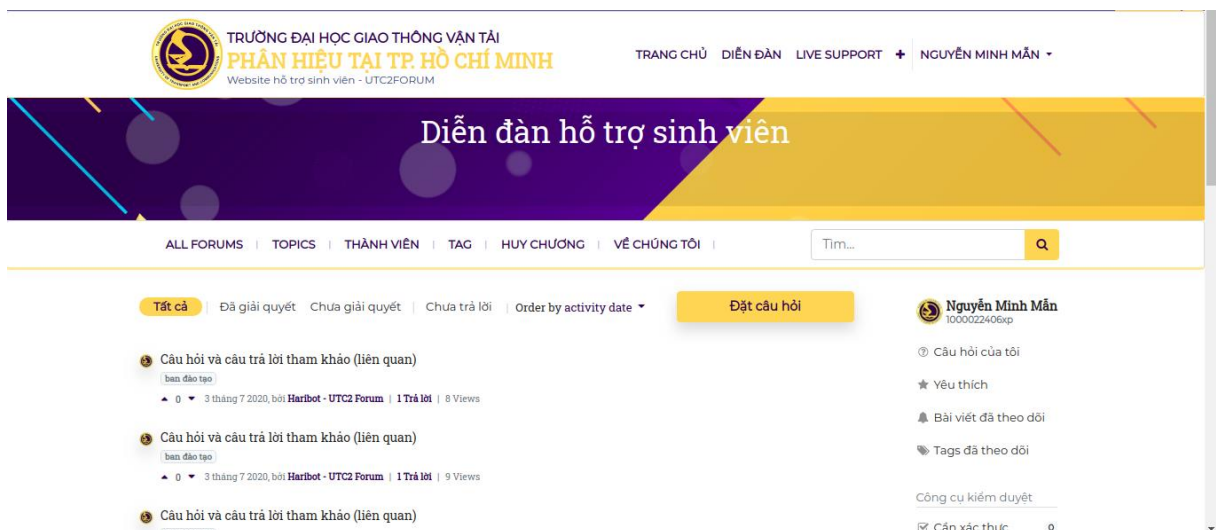
Sau khi đã có tài khoản, người dùng có thể đăng nhập để vào trang web, ngoài ra có thể đăng nhập với tài khoản google

3.7.2 Giao diện trang chủ website



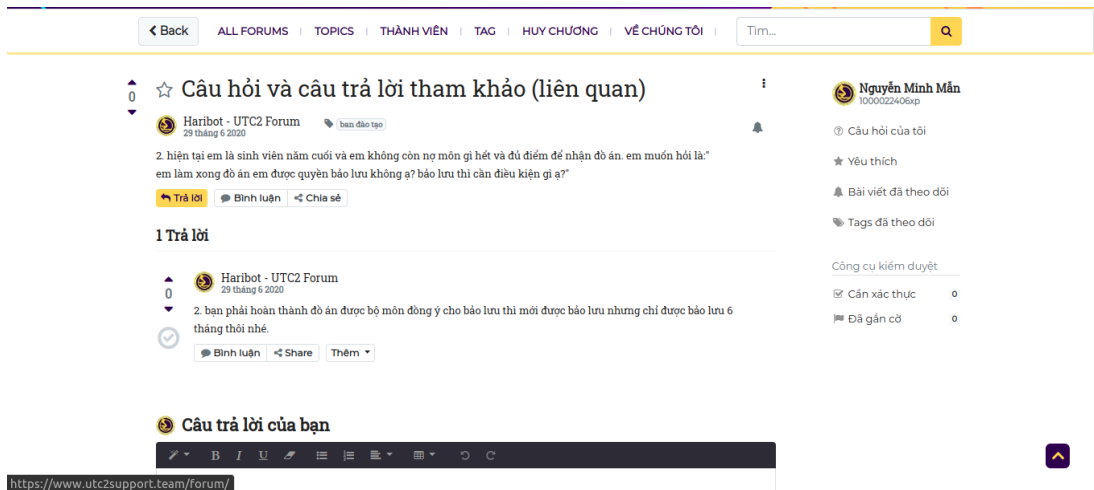
Hình 3.24 Giao diện trang chủ của website

3.7.3 Tương tác bài viết



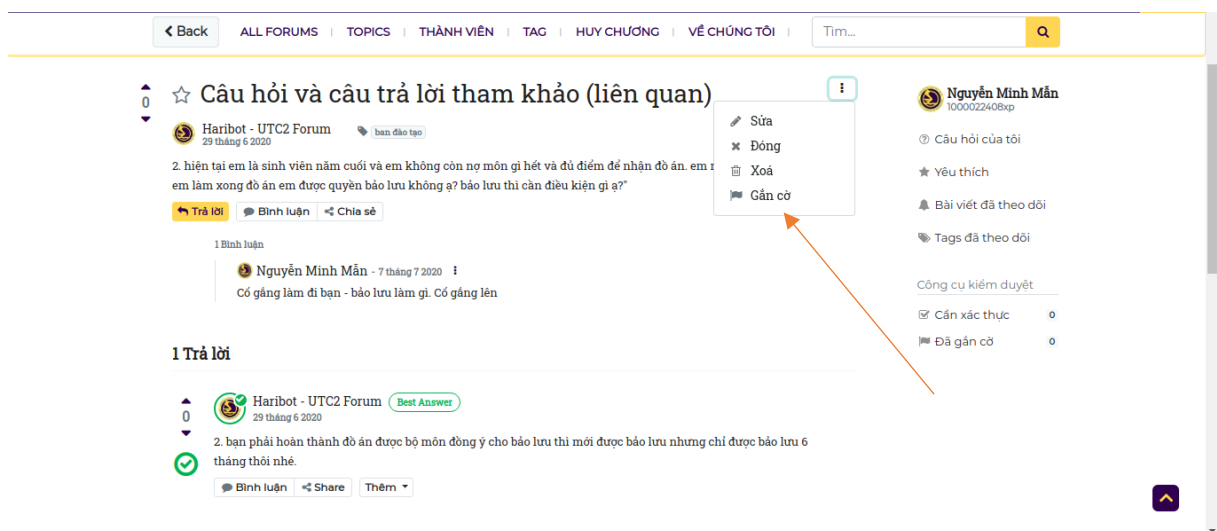
Hình 3.25 Giao diện đặt câu hỏi trên diễn đàn

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.



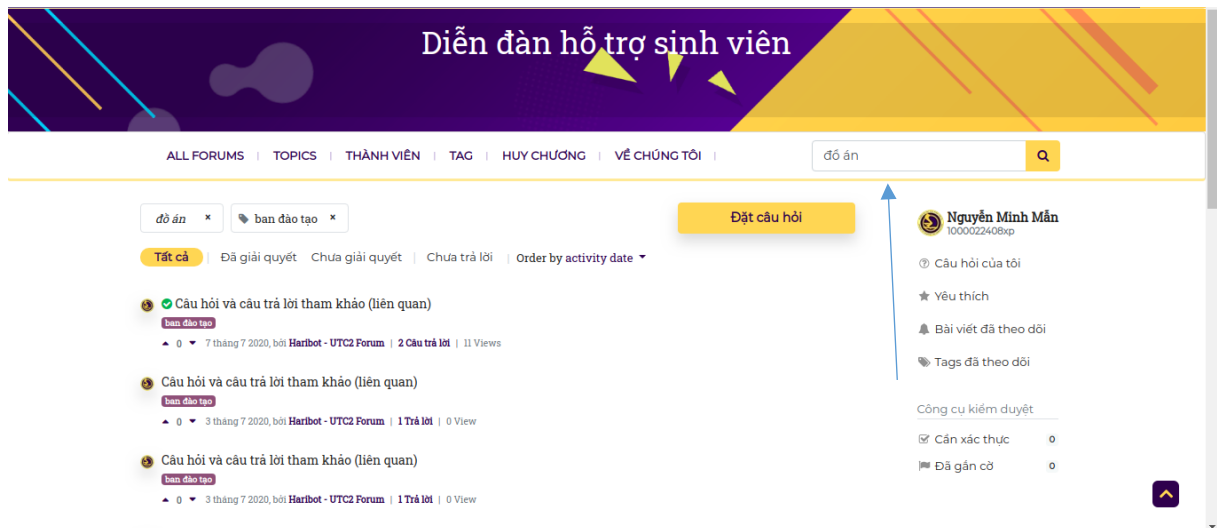
Hình 3.26 Giao diện trả lời câu hỏi

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

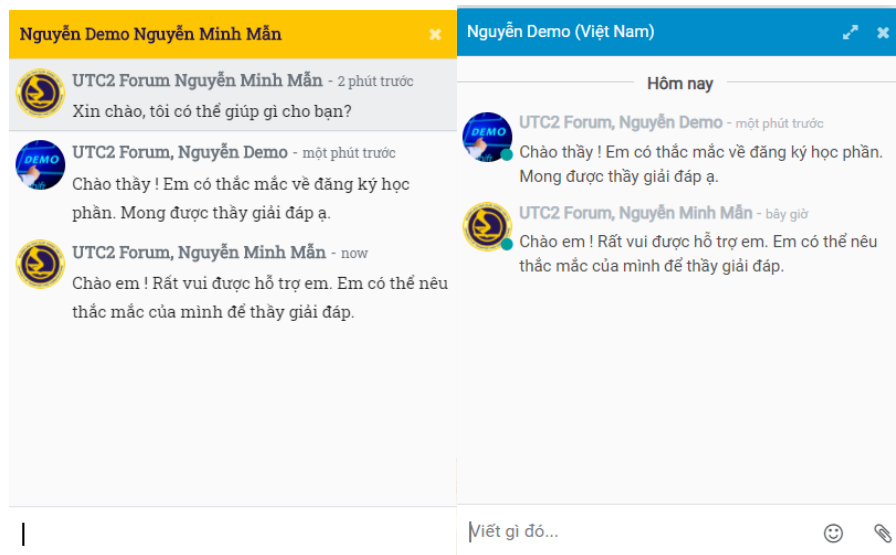
Người dùng chọn tìm kiếm câu hỏi và gõ các từ khóa cần tìm để tìm ra câu hỏi mà mình mong muốn

3.7.4 Chat trực tuyến



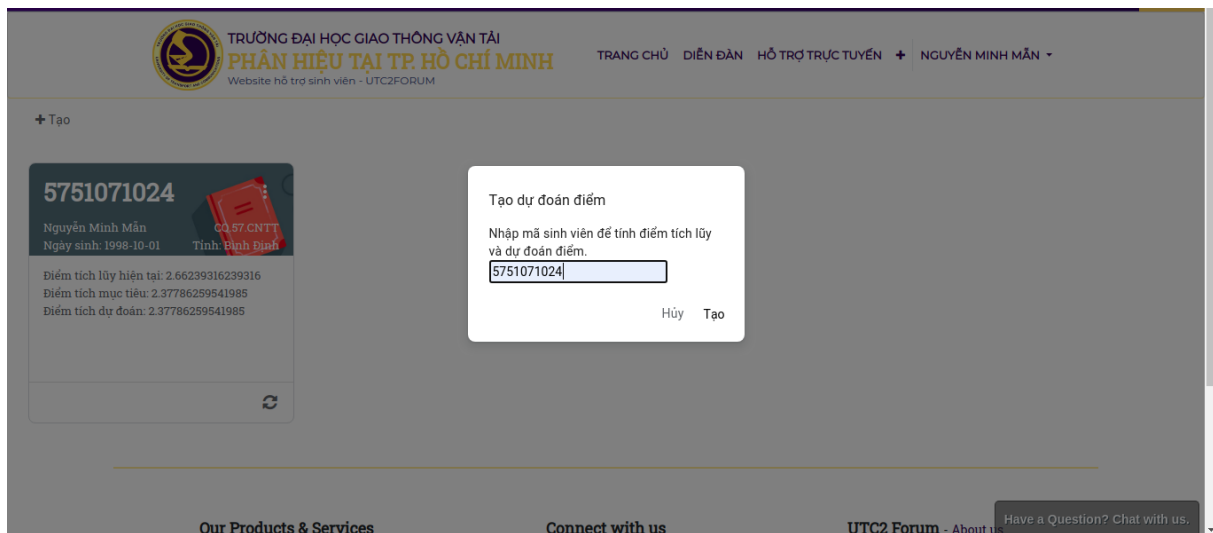
Hình 3.29 Giao diện chức năng chat trực tuyến

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



Mã môn	Tên môn	STC	Điểm chữ	Điểm số	Điểm mục tiêu	Điểm dự đoán	Công thức dự đoán
GQP20L...	Giáo dục QP-AN F1	3	0,0	6,7	6,7	6,7	
GQP20L...	Giáo dục QP-AN F2	2	0,0	6,6	6,6	6,6	
GQP20L...	Giáo dục QP-AN F3	3	0,0	6,6	6,6	6,6	

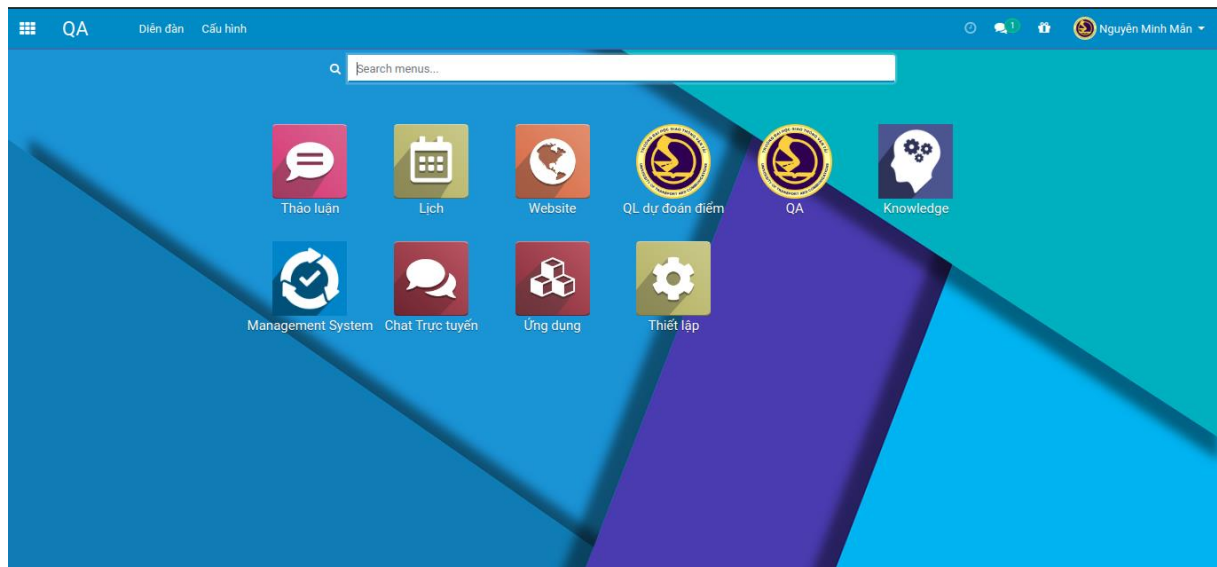
Điểm tích lũy hiện tại: 2.66239316239316
 Điểm tích mục tiêu: 2.45275590551181
 Điểm tích dự đoán: 2.45275590551181
 Đồ án tốt nghiệp: 0 -> 10

[Cập nhật](#) [Lưu mục tiêu](#)

<https://www.utc2support.team/sinhvien/92>

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

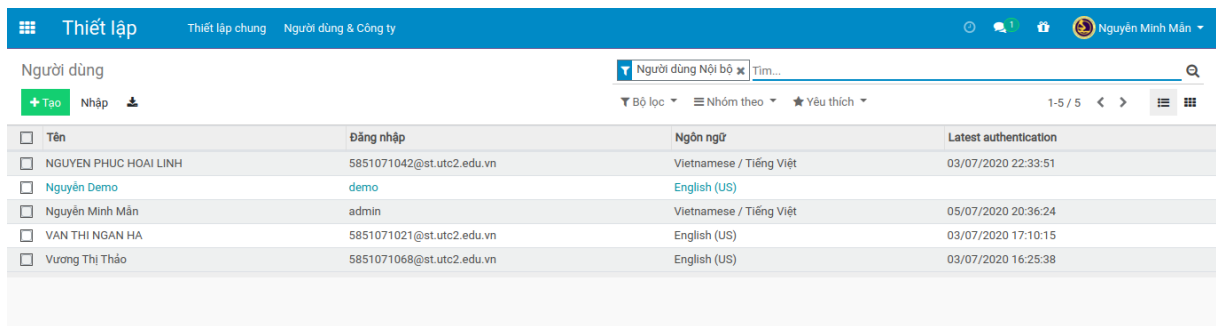
3.7.6 Nhóm chức năng quản lý



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, ...

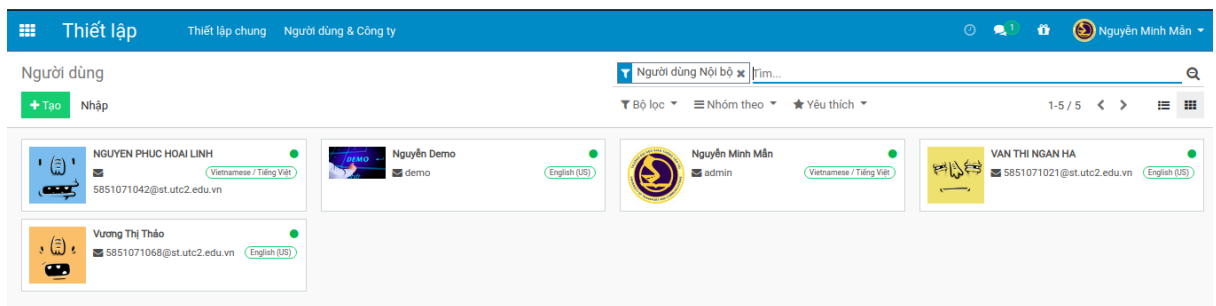
3.7.7 Quản lý người dùng



Tên	Đăng nhập	Ngôn ngữ	Latest authentication
NGUYEN PHUC HOAI LINH	5851071042@st.utc2.edu.vn	Vietnamese / Tiếng Việt	03/07/2020 22:33:51
Nguyễn Demo	demo	English (US)	
Nguyễn Minh Mẫn	admin	Vietnamese / Tiếng Việt	05/07/2020 20:36:24
VAN THI NGAN HA	5851071021@st.utc2.edu.vn	English (US)	03/07/2020 17:10:15
Vương Thị Thảo	5851071068@st.utc2.edu.vn	English (US)	03/07/2020 16:25:38

Hình 3.34 Giao diện quản lý người dùng

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

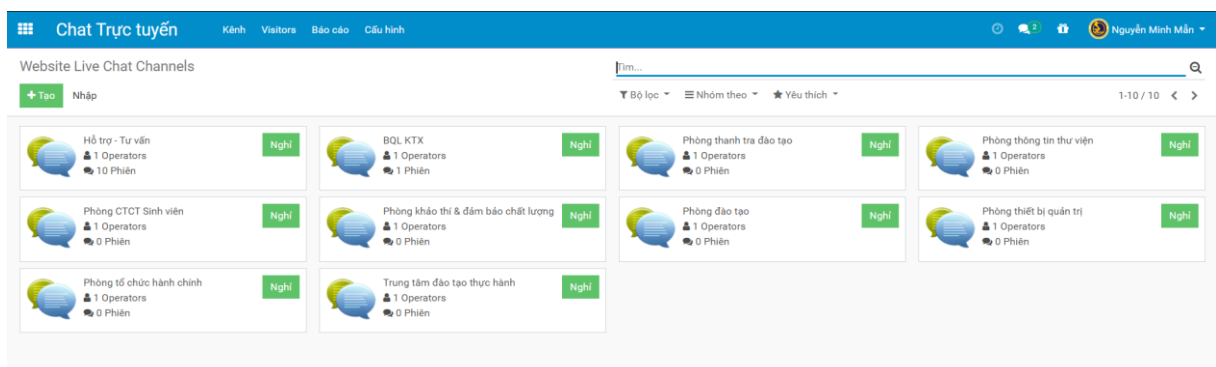


Hình 3.35 Giao diện hiện thị người dùng truy cập trang web

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 dõi có người đang hoạt động trong các phòng chat. Số lượng tài khoản nhắn tin, lịch sử chat, ...

Visitors

Tim...

▼ Bộ lọc ▾ ▮ Nhóm theo ▾ ★ Yêu thích ▾ 1-80 / 13 View kanban

	Nguyễn Minh Mẫn ●	13 phút trước Hành động cuối cùng	10 Khách ghé thăm	Home Last Page	7 Visited Pages	Nguyễn Minh Mẫn Speaking With	Email
	Website Visitor #144 ●	2 giờ trước Hành động cuối cùng	1 Khách ghé thăm	Website hỗ trợ sinh viên UTC2 Last Page	1 Visited Pages	- Speaking With	
	Website Visitor #143 ●	2 giờ trước Hành động cuối cùng	1 Khách ghé thăm	Website hỗ trợ sinh viên UTC2 Last Page	1 Visited Pages	- Speaking With	
	Website Visitor #142 ●	4 giờ trước Hành động cuối cùng	1 Khách ghé thăm	About us Last Page	1 Visited Pages	- Speaking With	
	Website Visitor #141 ●	4 giờ trước Hành động cuối cùng	1 Khách ghé thăm	Website hỗ trợ sinh viên UTC2 Last Page	1 Visited Pages	- Speaking With	
	Website Visitor #140 ●	5 giờ trước Hành động cuối cùng	1 Khách ghé thăm	Website hỗ trợ sinh viên UTC2 Last Page	1 Visited Pages	- Speaking With	
	Website Visitor #139 ●	5 giờ trước Hành động cuối cùng	1 Khách ghé thăm	Website hỗ trợ sinh viên UTC2 Last Page	1 Visited Pages	- Speaking With	

Hình 3.37 Giao diện theo dõi lịch sử chat

3.6.9 Quản lý dự đoán điểm

QL dự đoán điểm

Quản lý dự đoán điểm Quản lý Công cụ

Dự đoán điểm

Tim...

▼ Bộ lọc ▾ ▮ Nhóm theo ▾ ★ Yêu thích ▾ 1-16 / 16 < >

+ Tạo Nhập

<input type="checkbox"/> Mã dự đoán	Mã sinh viên	Tổng số tín chỉ	Điểm tích lũy	Điểm tích dự đoán	Điểm tích mục tiêu	Bảng điểm dự đoán
<input type="checkbox"/> DDD0006	5751071024	129	2,66	2,41	2,41	54 dữ liệu
<input type="checkbox"/> DDD0007	5851071068	146	3,16	1,78	1,66	61 dữ liệu
<input type="checkbox"/> DDD0010	5751071024	136	2,66	2,29	2,29	56 dữ liệu
<input type="checkbox"/> DDD0011	5751071024	146	2,66	2,13	2,13	61 dữ liệu
<input type="checkbox"/> DDD0012	5751071024	140	2,66	2,23	2,23	58 dữ liệu
<input type="checkbox"/> DDD0013	5851071068	140	3,16	1,74	1,74	58 dữ liệu
<input type="checkbox"/> DDD0015	5851071068	140	3,16	1,74	1,74	58 dữ liệu
<input type="checkbox"/> DDD0017	5851071068	140	3,16	1,80	1,74	58 dữ liệu
<input type="checkbox"/> DDD0018	5851071068	140	3,16	1,74	1,74	58 dữ liệu
<input type="checkbox"/> DDD0020	5851071068	136	3,16	1,79	1,79	56 dữ liệu
<input type="checkbox"/> DDD0021	5751071024	127	2,66	2,45	2,45	52 dữ liệu
<input type="checkbox"/> DDD0023	5451012044	132	1,85	1,85	1,90	63 dữ liệu
<input type="checkbox"/> DDD0024	5851071021	140	3,24	1,90	1,90	58 dữ liệu
<input type="checkbox"/> DDD0024	5851071021	140	3,24	1,90	1,90	58 dữ liệu
<input type="checkbox"/> DDD0025	5751071021	140	2,87	2,40	2,40	58 dữ liệu
<input type="checkbox"/> DDD0025	5751071021	140	2,87	2,40	2,40	58 dữ liệu

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 hỏi 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> , 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> ,
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
 , 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.