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: ỨNG DỤNG THUẬT TOÁN CÂY QUYẾT ĐỊNH, SVM, NAVIE BAYES ĐỂ PHÂN LOẠI THƯ RÁC

Giảng viên hướng dẫn: PHẠM THỊ MIÊN

Sinh viên thực hiện: ĐẶNG VĂN THỌ

Lớp:CQ.CNTT.K59

Khoá:59

Tp. Hồ Chí Minh, năm 2022

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: ỨNG DỤNG THUẬT TOÁN CÂY QUYẾT ĐỊNH, SVM, NAVIE BAYES ĐỂ PHÂN LOẠI THƯ RÁC

Giảng viên hướng dẫn: PHẠM THỊ MIÊN

Sinh viên thực hiện: ĐẶNG VĂN THỌ

Lớp:CQ.CNTT.K59

Khoá:59

Tp. Hồ Chí Minh, năm 2022

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**
PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH Độc lập – Tự do – Hạnh phúc

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

***

Mã sinh viên: 5951071103 **Họ tên SV:** Đặng Văn Thọ
Khóa: 59 **Lớp:** CQ.59.CNTT

1. **Tên đề tài:** ứng dụng thuật toán cây quyết định, SVM, Navie Bayes để phân loại thư rác

2. Mục đích, yêu cầu

a. Mục đích: : Ứng dụng thuật toán cây quyết định, SVM, Navie Bayes để phân loại thư rác

b. Yêu cầu:

- Tìm hiểu 3 thuật toán cây quyết định, Navie Bayes, SVM
- Xây dựng : ứng dụng thuật toán cây quyết định, SVM, Navie Bayes để phân loại thư rác

3. Nội dung và phạm vi đề tài

a. Nội dung đề tài

- Tìm hiểu ngôn ngữ lập trình Python ứng dụng vào việc xây dựng chương trình.
- Sử dụng các thư viện NumPy , Pandas ,Scikit-learning áp dụng vào ứng dụng
- Hỗ trợ phân loại thư rác

4. **Phạm vi đề tài:** ứng dụng thuật toán cây quyết định, SVM, Navie Bayes để phân loại thư rác

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

a. Công nghệ sử dụng

- HTML
- CSS
- Python

b. Công cụ

- IDE: Pycharm, Google Colab

c. Ngôn ngữ lập trình

- Python

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

- Hoàn chỉnh cuốn báo cáo đề tài.
- Xây dựng ứng dụng thuật toán cây quyết định, SVM, Navie Bayes để phân loại thư rác
- Phân loại được thư rác

7. Giảng viên hướng dẫn:

- Họ tên: Phạm Thị Miên
- Đơn vị công tác: Trường Đại học Nông Lâm Thành phố Hồ Chí Minh

Điện thoại: **0931 741 860** Email: nvdu@hcmuaf.edu.vn

Ngày….tháng….năm 2022
Trưởng BM Công nghệ Thông tin

Đã giao nhiệm vụ TKTN
Giảng viên hướng dẫn

ThS. Trần Phong Nhã

ThS. Phạm Thị Miên

Đã nhận nhiệm vụ TKTN
Sinh viên: Hoàng Đình Thiên Đông
Điện thoại: 0398294385

Ký tên:
Email: 5951071017@st.utc2.edu.vn

LỜI CẢM ƠN

Lời nói đầu tiên, tôi 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 tôi thực hiện đề tài tốt nghiệp "ứng dụng thuật toán cây quyết định, SVM, Navie Bayes để phân loại thư rác

Trong thời gian làm đồ án tốt nghiệp vừa qua là khoảng thời gian khó quên trong quãng đời sinh viên của tôi cũng như là quãng thời gian quý báu để tôi có thể vận dụng những kiến thức mà thầy cô đã truyền dạy trong gần suốt 4 năm tại trường.

Tôi muốn gửi lời cảm ơn chân thành nhất đến toàn thể quý thầy cô trong bộ môn Công nghệ thông tin Trường Đại học Giao thông Vận tải Phân hiệu tại thành phố Hồ Chí Minh, các thầy cô đã giảng dạy các môn học đến từ các trường lân cận, và đặc biệt sự giúp đỡ nhiệt tình của ThS.Phạm Thị Miên, hướng dẫn đồ án tốt nghiệp cho tôi, để tôi có thể hoàn thành xuất sắc nhất đồ án tốt nghiệp.

Tôi mong sau khi hoàn thành đồ án tốt nghiệp tôi sẽ có thể bước ra ngoài xã hội với một công việc ổn định, đúng ngành nghề đã theo học và không ngừng phát triển hoàn thiện bản thân trên con đường sự nghiệp của mình.

Mặc dù bản thân đã rất cố gắng nhưng do thời gian, kiến thức và kinh nghiệm có hạn nên bài làm của em còn có nhiều thiếu sót trong việc trình bày, đánh giá và đề xuất ý kiến. Em rất mong nhận được sự thông cảm và đóng góp ý kiến của quý thầy cô và các bạn.

Trong suốt quá trình làm đồ án, với điều kiện thời gian cũng như kinh nghiệm còn hạn chế, chắc chắn không thể tránh khỏi những thiếu sót, tôi mong thầy cô đóng góp ý kiến để tôi có thể bổ sung, hoàn thiện đồ án tốt nghiệp tốt hơn.

Tôi xin chân thành cảm ơn!

TP. Hồ Chí Minh, ngày … tháng … năm 2022

Sinh viên thực hiện

Đặng Văn Thọ

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. Hồ Chí Minh, ngày …… tháng …… năm 2022

Giảng viên hướng dẫn

ThS.Phạm Thị Miên

MỤC LỤC

DANH MỤC CHỮ VIẾT TẮT

STT	Mô tả	Ý nghĩa	Ghi chú
1	Tf	Tf- term frequency	
2	IDF	Inverse Document Frequency	
3	SVM	Support Vector Machine	

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

GIỚI THIỆU

Hiện tại thời gian internet _ mở nhiều kênh truyền hình kết nối và mới dịch vụ vì người dùng . Một từ dịch vụ , mà còn lại phổ biến , là điện tử mail (thư điện tử). Cô ấy là hình như chính bạn rất giản dị có nghĩa vì thông tin liên lạc . Nhờ vào lợi ích , số lượng tin nhắn mà _ trao đổi trực tuyến _ mỗi ngày tăng lên .

Hệ thống điện tử thư là người mẫu lưu trữ và chuyển tiếp . May chủ điện tử thư nhận , chuyển tiếp , giao hàng và lưu trữ tin nhắn . Qua đo lường sự phát triển Internet điện tử thư phát triển và nhu cầu sử dụng tăng lên . Trước bây giờ từ điện tử thư Là nổi tiếng có nghĩa giao tiếp trực tuyến , nhưng _ to lớn phần bức thư Là thư rác .

Trong gần đây năm Thư rác hoặc không mong muốn danh sách mail trở thành vấn đề và bị đe dọa Nhân loại giao tiếp . Thư rác biến thành một hình dạng chuyên nghiệp quảng cáo , phân phối vi rút và trộm cắp thông tin sử dụng _ bộ vô cùng tinh vi thủ thuật . Người dùng chiếm dùng rất nhiều thời gian trên xóa ' không được mời nhưng _ đến » tin nhắn . họ đang không cố ý có thể được bị lây nhiễm vi rút và thậm chí mất thông tin như vậy _ thế nào tín dụng thẻ , ngân hàng hóa đơn , v.v.

Dựa theo nghiên cứu Các phòng thí nghiệm Kaspersky (Kaspersky) , vào năm 2020 trung bình mức độ thư rác trên thế giới là 50,37. Vĩ đại nhất số lượng thư rác ở Nga - 21,27%, hơn anh ta theo dõi Đức (10,97%), Mỹ (10,47%) và Trung Quốc (6,21%). Theo quốc gia , mục tiêu độc hại thư từ , là Tây Ban Nha với 8,48%. Trên thứ hai nơi Đức với 7,28%, xếp sau anh ta Nên Nga (6,29%).

Đến ngăn ngừa thư rác , nhiều tổ chức và cá nhân những khuôn mặt nghiên cứu và phát triển phương pháp sự phân loại tin nhắn trên các nhóm vì nhận biết thư rác và thường xuyên thư từ . Tuy nhiên người sáng tạo Thư rác luôn luôn tìm kiếm cách đi xung quanh này phân loại và phân phối thư rác . Do đó , nên được tốt hệ thống sự phân loại thư rác và email . tiến hành từ mô tả Các vấn đề là xây dựng và giải quyết nhiệm vụ lọc thư rác có mục đích định nghĩa hiệu quả phương pháp lọc tin nhắn rác .

Vì thành tựu bàn thắng làm việc , cần thiết quyết định sau đây nhiệm vụ :

- Sưu tầm bộ dụng cụ dữ liệu vì lọc thư rác và làm sơ bộ chế biến ;

- Xác định bộ dụng cụ dấu hiệu vì sự sáng tạo mô hình ;
- Xây dựng người mẫu lọc sử dụng thư rác _ các thuật toán ;
- Thực hiện nhận kết quả và tính toán sự chính xác sự phân loại ;
- Phát triển các ứng dụng có chức năng lọc thư rác .

CHƯƠNG 1. PHÂN TÍCH TÊN MIỀN

1. Khái niệm học máy

Một nhiệm vụ phân loại thư rác thực sự là một nhiệm vụ phân loại văn bản trong đó dữ liệu nguồn bao gồm thư rác và email thông thường. Ý tưởng chính của phương pháp này là tìm cách xây dựng một bộ phân loại để lọc một mẫu mới bằng cách sử dụng đào tạo dựa trên các mẫu hiện có.

Để tạo ra một bộ lọc thư rác, cần phải làm quen với trí tuệ nhân tạo, học máy và cách xử lý ngôn ngữ tự nhiên. Và cũng để có được kiến thức đầu tiên về việc gửi thư hàng loạt (spam) và tác động của nó đối với người dùng.

1.1. Trí tuệ nhân tạo

Trước đây, khi nói đến trí tuệ nhân tạo (AI), mọi người thường quan tâm đến việc xây dựng các máy tính có khả năng "suy nghĩ" ngay cả trong một số khu vực hẹp có thể cạnh tranh hoặc vượt qua khả năng của bộ não con người. Những ý tưởng lâu đời này đã ảnh hưởng đến nghiên cứu trong phòng thí nghiệm. Mặc dù các mô hình như máy tính thông minh đã được giới thiệu nhiều năm trước chỉ sau khi Alan Turing công bố kết quả của nghiên cứu lớn đầu tiên, mọi người thực sự bắt đầu nghiêm túc nghiên cứu vấn đề AI. Phát hiện của Turing cho thấy một chương trình có thể được lưu trữ trong bộ nhớ và sau đó được thực hiện dựa trên các hoạt động toán học cơ bản bằng cách sử dụng các bit 0, 1. Điều này tạo thành nền tảng của các máy tính hiện đại. Lưu các chương trình vào máy một cách nhanh chóng và dễ dàng thay đổi chức năng bằng cách tải một chương trình mới vào bộ nhớ. Điều này có nghĩa là một máy tính có thể học hỏi và suy nghĩ, và cũng là một trong những biểu hiện quan trọng đầu tiên của máy tính được trang bị AI.

Trí tuệ nhân tạo là một hướng khoa học riêng biệt chuyên nghiên cứu các phương pháp làm cho trí thông minh máy móc tương tự như trí thông minh của con người. Trí tuệ nhân tạo có chủ đề nghiên cứu cụ thể của riêng mình. Trong trí tuệ nhân tạo, cả ba loại phương pháp nghiên cứu cổ điển đều được sử dụng: suy luận, thực nghiệm và mô tả [4].

Một số định nghĩa về trí tuệ nhân tạo là:

• Một hệ thống có thể suy nghĩ như một con người.

Một hệ thống biết cách hành động như một con người.

Để một hệ thống suy nghĩ và hành động như một con người, nó phải được trang bị các công cụ như thính giác, kiến thức, giải thích tự động, học tập, nhìn và di chuyển như một con người.

Trí tuệ nhân tạo được sử dụng trong nhiều lĩnh vực của nền kinh tế:

1. Điều khiển học, robot, giao diện thông minh của robot.

2. Trò chơi trên máy tính.

3. Các thiết bị điện tử thông minh sử dụng logic mờ.

4. Hệ thống chuyên gia về giáo dục, y tế, địa chất, quản lý.

5. Xử lý ngôn ngữ tự nhiên.

6. Nhận dạng hình ảnh và âm thanh.

7. Hệ thống xử lý dữ liệu và kiến thức tích hợp.

8. Mô hình giải quyết vấn đề.

1.2. Học máy

Con người có nhiều cách học, chẳng hạn như học thông qua trí nhớ, học thông qua quan sát và nghiên cứu, học thông qua thực hành, học thông qua sự phát triển của hệ thần kinh sinh học, hoặc học qua gen - các thế hệ trước.

Bất kể phương thức học tập là gì, mục tiêu của việc học là thu thập kiến thức mới và sau đó xử lý kiến thức đó để nó có thể thích nghi với các tình huống và sự kiện mới.

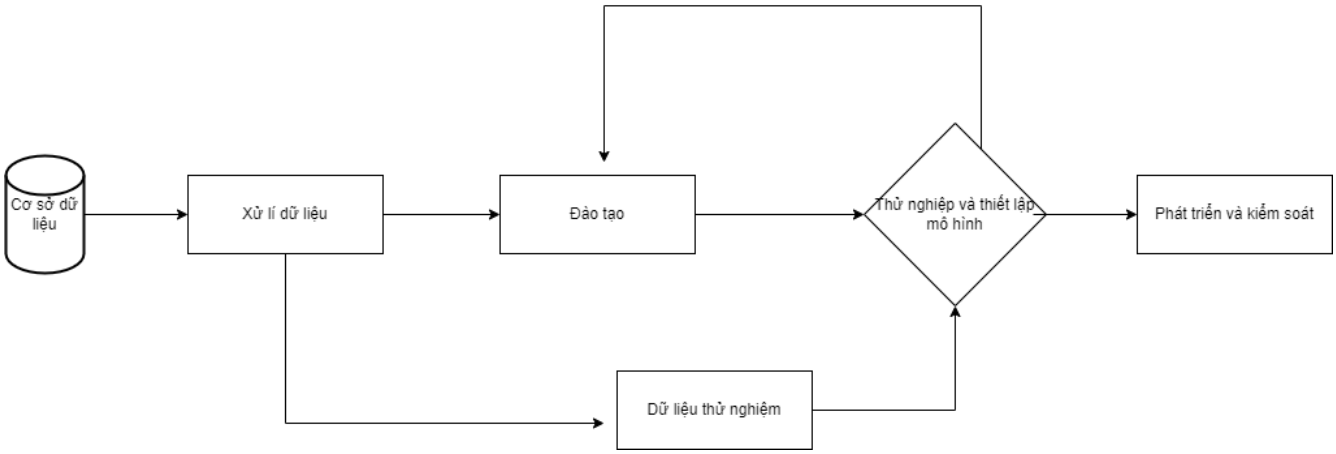
Tương tự như phương pháp này, mọi người cũng muốn tạo ra các chương trình trong máy tính để máy tính có thể tự thu thập kiến thức mới, xử lý nó và sau đó thích nghi với các tình huống cụ thể.

Đây là lý do mà học máy ngày nay đang trở thành một trong những nhiệm vụ chính trong khoa học máy tính.

Học máy là một lĩnh vực nghiên cứu trong lĩnh vực trí tuệ nhân tạo, nghiên cứu các phương pháp để xây dựng các thuật toán có thể học.[8]

Ý tưởng chính của học máy là dạy một máy tính để "học", tức là máy tính không chỉ sử dụng một thuật toán được viết sẵn, mà chính nó học cách giải quyết nhiệm vụ với sự trợ giúp của

kiến thức hữu ích từ bất kỳ dữ liệu nào.



Hình 1. 1 – Các bước chính của việc tạo ra một mô hình học máy.

Học máy bao gồm 5 giai đoạn:

- Thu thập dữ liệu: bộ dữ liệu được tải và lưu trữ.

- Xử lý dữ liệu: dữ liệu thu được cần phải được xử lý để cải thiện chất lượng của chúng.

- Xây dựng mô hình: ở giai đoạn này bạn cần xác định nên chọn mô hình nào và làm thế nào để cải thiện mô hình.

- Thử nghiệm và thiết lập mô hình: mô hình kết quả được kiểm tra bằng cách sử dụng một bộ dữ liệu thử nghiệm. Kết quả thử nghiệm được sử dụng để thiết lập một mô hình mới, và để tính đến các mô hình trước đó, tức là mô hình "học hỏi".

- Phát triển mô hình và kiểm soát: ở giai đoạn này, mô hình tốt nhất để phát triển và triển khai được lựa chọn.

Các loại máy học

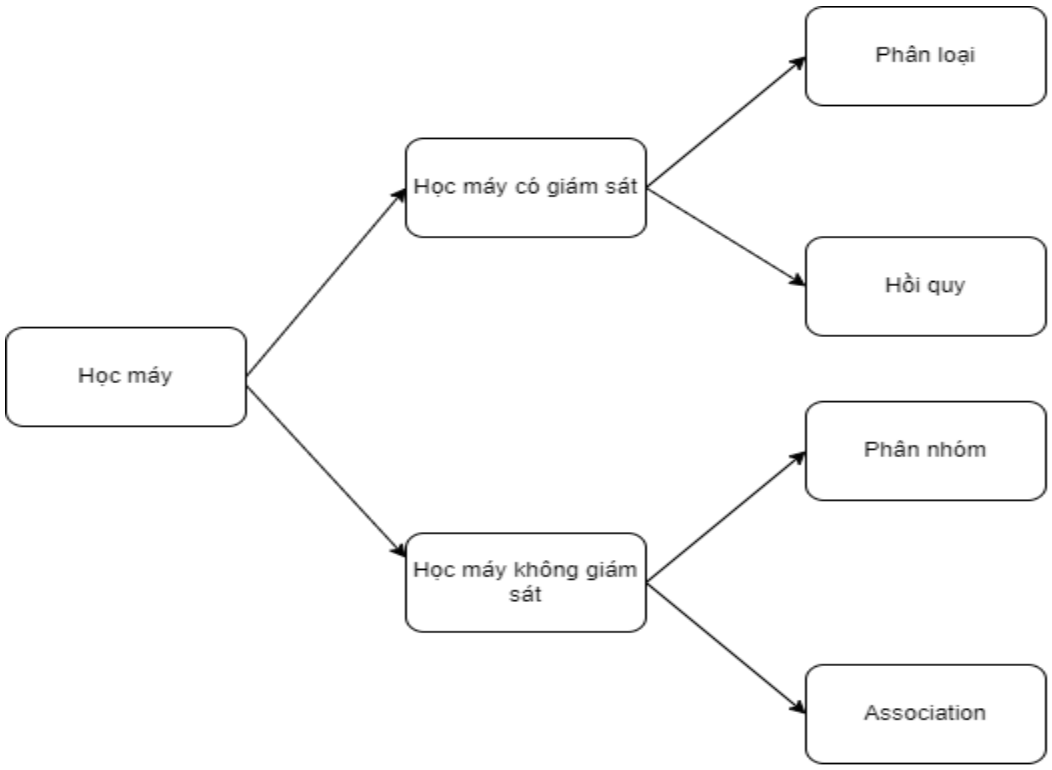
Học máy được chia thành 2 loại chính: có giám sát học tập hoặc "học tập có giám sát" và học tập không giám sát.

Học tập có giám sát là học tập với câu trả lời đúng. Trong trường hợp này, có bộ dữ liệu ban đầu. Dữ liệu như vậy được tạo ra theo mô hình "đối tượng - câu trả lời". Mô hình được đào tạo trên bộ dữ liệu này, nơi nó "biết" kết quả chính xác và bắt đầu dự đoán đầu ra cho đầu vào mới.

Không giống như học tập do giáo viên hỗ trợ, không có đào tạo giáo viên trong học tập không giám sát.

câu trả lời đúng. Học tập không giám sát cũng đòi hỏi dữ liệu để "học tập", nhưng không phải có bất kỳ kết quả cụ thể nào. Thuật toán không tìm kiếm các cặp phản hồi đối tượng, mà cho các kết nối giữa các đối tượng.

Trên Hình 1.2 cho thấy chính các loại cỗ máy sự học hỏi .



Hình 1. 2 *Hai chính loại hình cỗ máy học tập*

Sử dụng _ đây phương pháp học tập một trải nghiệm rõ ràng xác định thế nào hình thức đầu vào và đầu ra Mục tiêu các chức năng , ví dụ bộ dụng cụ các mẫu giống nhau danh tính .

Học với giáo viên bao gồm phân loại và hồi quy ; trong đó sự phân loại kiểm soát định hình với mục tiêu hàm số phân phối giá trị và hồi quy sử dụng _ Mục tiêu chức năng tiếp diễn các giá trị .

Giáo dục không có giáo viên là _ đào tạo , với cái mà một trải nghiệm bao gồm chỉ có từ mẫu không có liên quan, thích hợp Mục tiêu nhãn mác hoặc các giá trị . Ví dụ , chỉ xem quy mô mọi người và dần dần học hỏi khái niệm đầy đủ , bình thường và mỏng

con người .

Hai phần lớn nổi tiếng loại hình học tập không có giáo viên là _ phân cụm và liên kết . Trong trường hợp nhóm lại các đối tượng đăng lại trên một số nhóm , vì vậy Gì mỗi Tập đoàn bao gồm từ giống hệt nhau các đối tượng . Lên đỉnh mặt hàng đang ở trong một nhóm . Và các hiệp hội là _ đường dò tìm các mặt hàng _ thường gặp cùng nhau ví dụ bánh mì và sữa _ _ thường được mua cùng nhau .

1.3. Phương pháp xử lý ngôn ngữ tự nhiên

Lọc thư rác hoạt động theo nguyên tắc phân loại tin nhắn thành hai nhóm "spam" và "bình thường" bằng cách phân tích nội dung của tin nhắn. Lúc đầu, nội dung của thông điệp được thể hiện dưới dạng các hàm hoặc thuộc tính (đây là những tính năng để học tập), mỗi chức năng thường là một từ hoặc cụm từ trong thư.

Để hiểu và thực hiện một nhiệm vụ như vậy, máy tính sử dụng một phương pháp đặc biệt gọi là Xử lý ngôn ngữ tự nhiên (NLP). Chúng có liên quan đến ngôn ngữ của con người, chẳng hạn như: dịch thuật, phân tích dữ liệu văn bản, nhận dạng giọng nói, truy xuất thông tin, nâng cao hiệu quả giao tiếp hoặc đơn giản là tăng hiệu quả của trình soạn thảo văn bản.

Ngôn ngữ tự nhiên phát sinh từ cảm xúc, vì vậy thường không có quy tắc hoặc sự tương ứng với logic, bao gồm cả về cú pháp, ngữ nghĩa và biểu hiện ngôn ngữ. Nó có độ mơ hồ cao ở tất cả các cấp độ, bao gồm cấp độ từ vựng, cấp độ cú pháp, cấp độ ngữ nghĩa và cấp độ văn bản. Người ta nói rằng một ngôn ngữ là mơ hồ nếu có nhiều cấu trúc ngôn ngữ khác nhau phù hợp với nó. Sự mơ hồ của ngôn ngữ tự nhiên gây khó khăn cho việc xử lý trên máy tính.

Xem xét trên ví dụ :

Ví dụ 1:

They *book* that hotel. (S1)

They read that *book*. (S2)

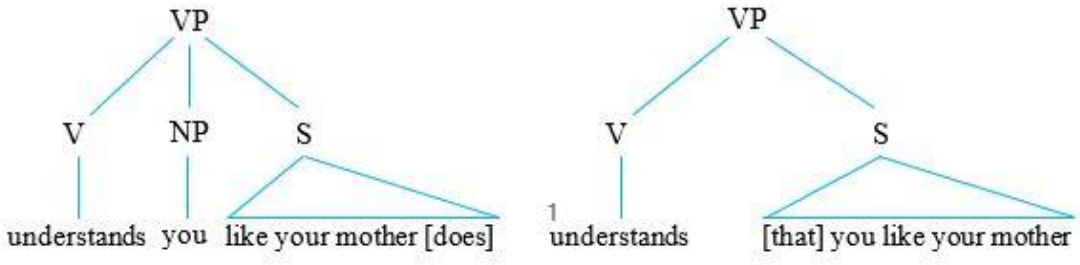
Đầu tiên, từ "book" là mơ hồ về thuật ngữ này. Tùy thuộc vào ngữ cảnh, "book" có thể là một động từ (trong câu S1) hoặc một danh từ (trong câu S2). Hiện tượng này là một vấn đề với việc xử lý cú pháp đánh dấu kiểu. "Sách" có thể là hiệu ứng của việc đặt hàng một cái gì đó (trong ưu đãi S1) hoặc văn bản bằng văn bản được xuất bản ở dạng in hoặc điện tử (trong ưu đãi S2). Hiện tượng này làm phức tạp nhiệm vụ xác định một từ, đó là

một giai đoạn xử lý ngữ nghĩa.

Ví dụ 2:

A computer understands you like your mother. (S3)

Hình vẽ cho thấy các cây cú pháp của câu.



Hình 1. 3 *Cú pháp cây gợi ý*

Từ quan điểm ngữ pháp, câu này có thể được giải thích bởi hai cây cú pháp, như được thể hiện trong hình. Các cấu trúc khác nhau dẫn đến các cách giải thích khác nhau: "máy tính hiểu rằng bạn giống như mẹ của bạn" hoặc "máy tính hiểu rằng bạn thích mẹ của bạn". Hiện tượng này làm cho nó khó khăn không chỉ để phân tích mà còn về mặt ngữ nghĩa.

1.4. Các tác vụ cơ bản trong xử lý ngôn ngữ tự nhiên

1.4.1.Phân tích tin nhắn đại chúng và tác động của nó với người dùng

Xử lý ngôn ngữ tự nhiên bao gồm hiểu ngôn ngữ tự nhiên (NLU) và tạo ngôn ngữ tự nhiên (NLG). [7] Hiểu ngôn ngữ tự nhiên (NLU) bao gồm bốn bước chính sau:

- Phân tích chẩn đoán: đây là sự công nhận, phân tích và mô tả cấu trúc của các con số trong một ngôn ngữ nhất định và các đơn vị ngôn ngữ khác như từ gốc, từ, phụ gia, lớp con, v.v.

- Phân tích cú pháp: Đây là quá trình phân tích một loạt các ký tự dưới dạng ngôn ngữ tự nhiên hoặc máy tính theo ngữ pháp chính thức.

- Phân tích ngữ nghĩa: Đây là quá trình tương quan các cấu trúc ngữ nghĩa từ cấp độ của một cụm từ, câu, câu và đoạn văn đến mức độ của toàn bộ bài viết với ý nghĩa độc lập của chúng.

- Phân tích diễn ngôn: Thực dụng là nghiên cứu về mối quan hệ giữa ngôn ngữ và bối cảnh sử dụng.

Nhiệm vụ của bộ phân loại thư rác là một trong những nhiệm vụ quan trọng trong xử lý ngôn ngữ tự nhiên.

1.4.2. Định nghĩa về thư gửi hàng loạt và ý nghĩa của nó

Email đã trở thành một ứng dụng không thể thiếu khi Internet và các công nghệ mạng đã phát triển. E-mail, còn được gọi là thư, là một hệ thống để gửi thư qua mạng máy tính. Email là một phương tiện giao tiếp rất hữu ích. Nó có thể đồng thời truyền thông tin từ một máy nguồn đến một hoặc nhiều máy.

Tuy nhiên, trong những năm gần đây, một hình thức email mới đã xuất hiện với số lượng lớn, gây ra vấn đề cho người nhận và gây ra nhiều thiệt hại cho nền kinh tế, được gọi là thư rác hoặc thư rác.

Spam là một thư gửi hàng loạt các tin nhắn không có ý nghĩa và thể hiện mong muốn nhận được chúng. [9] Thư rác thường được gửi với số lượng lớn. Spam thường đi kèm với virus gây phiền toái cho người dùng, làm giảm tốc độ truyền Internet và tốc độ xử lý máy chủ email.

Hiện tại, không có định nghĩa đầy đủ và nhất quán về spam. Spam được hiểu là email thương mại không mong muốn (UCE), và rộng hơn, thư rác bao gồm quảng cáo, xúc phạm và thư rác (Unsolicited Bulk Emai -UBE).

1.4.3. Các loại thử gửi hàng loại

1.Thư quảng cáo

Email quảng cáo là hình thức spam phổ biến nhất. Thông thường chúng nằm trong thư mục spam, nhưng thường chiếm một lượng không gian rất lớn. Hàng trăm tỷ tin nhắn quảng cáo không mong muốn được gửi đi mỗi ngày, chủ yếu là thông tin về các loại thuốc thần kỳ để giảm cân, các chương trình du lịch, chương trình khuyến mãi hoặc các chương trình đào tạo, các khóa học trực tuyến.

2.Email lừa đảo

Email lừa đảo là loại thư rác phổ biến thứ hai trên nền tảng email, cũng như một trong những loại thư rác khó phát hiện nhất. Email lừa đảo thường được tạo ra để mô phỏng một nguyên mẫu email thực được gửi bởi các tổ chức và doanh nghiệp có uy tín (như Amazon, Google, Microsoft hoặc Facebook) để lừa người dùng nhấp vào liên kết hoặc tải xuống tệp đính kèm.

3.Trojan Điện tử

Trojan horse là một loại phần mềm độc hại mà thoạt nhìn trông giống như bình

thường, nhưng thực sự có thể điều khiển hoàn toàn một máy tính. Mục đích của Trojan là làm hỏng, đánh cắp thông tin hoặc thực hiện các hành động độc hại trên dữ liệu hoặc mạng.

4.Email

Điều này rất giống với một email lừa đảo, nhưng thay vì sử dụng một kỹ thuật để làm cho các phương pháp spam đáng tin cậy hơn, nhiều người gửi thư rác có xu hướng gửi tin nhắn xuất hiện từ một địa chỉ email tương tự như địa chỉ thực. Theo quy định, anh ta cố gắng lừa dối người nhận để đánh lừa anh ta về nguồn gốc của tin nhắn. Phương pháp đánh cắp danh tính này tạo ấn tượng rằng email lừa đảo đến từ một nguồn, công ty hoặc tổ chức đáng tin cậy.

1.4.4. Tác động của việc gửi thư hàng loạt đến người dùng

Spam hiện chiếm một tỷ lệ rất lớn trong tổng số lượng email được gửi qua Internet. Theo các thống kê khác nhau, tỷ lệ thư rác trong lưu lượng thư thế giới là khoảng 80% [10]. Số lượng thư rác là quá nhiều và gây ra tác hại lớn cho sự phát triển của Internet. Spam có tác dụng có hại:

- Thư rác dẫn đến thiệt hại kinh tế cho người nhận thư trong trường hợp người nhận phải trả tiền cho lượng thông tin được truyền qua mạng.
- Thư rác có thể điền vào hộp thư đến của người nhận và do đó các email khác đến sau sẽ bị mất.
- Spam lãng phí rất nhiều thời gian vì người nhận phải mở tin nhắn và xóa nó khỏi hộp thư đến của họ.
- Thư rác khiến người dùng email lo lắng. Theo thống kê trên http://www.pewinternet.org, 25% người dùng email coi spam là trở ngại chính khi sử dụng dịch vụ Internet.
- Thư rác chiếm một phần kết nối Internet và lãng phí thời gian xử lý máy chủ.

1.4.5. Một số phương pháp lọc Thư rác

Lọc thư rác với các quy tắc hạn chế thư rác Kể từ khi spam gây ra nhiều vấn đề trên thế giới, nhiều quốc gia đã đưa ra luật để ngăn chặn thư rác.

Lọc Thư rác bằng địa chỉ IP

Danh sách đen

Một danh sách các địa chỉ spam được biên soạn. Các nhà cung cấp dịch vụ email (ISP) sẽ dựa vào danh sách này để loại bỏ thư khỏi danh sách này. Danh sách này thường xuyên được cập nhật và chia sẻ bởi các nhà cung cấp dịch vụ.

Danh sách trắng

Danh sách những người gửi an toàn có thể được cung cấp cho các nhà cung cấp dịch vụ cụ thể. Các địa chỉ trong danh sách được truyền qua bộ lọc. Người dùng phải được đăng ký để có mặt trong danh sách.

Bộ lọc dựa trên bộ lọc yêu cầu/phản hồi

Một tính năng của cách tiếp cận này là khả năng tự động gửi phản hồi cho người gửi với yêu cầu thực hiện hành động. Chương trình thử nghiệm này được đặt tên là "Bài kiểm tra Turing" sau một số bài kiểm tra được phát triển bởi nhà toán học Alan Turing.

Phương thức Id Người gửi

Khóa tên miền là một phương pháp mã hóa danh tính được đề xuất bởi Yahoo vào tháng 5 năm 2004. Khóa tên miền không chỉ cho phép bạn xác định tên miền của người gửi mà còn cho phép bạn xác minh tính toàn vẹn của nội dung của email. Khóa tên miền sử dụng mật mã khóa công khai RSA để xác minh tính toàn vẹn của email ở cấp tên miền.

Kết quả:

Chương này bao gồm các định nghĩa về trí tuệ nhân tạo, học máy và xử lý ngôn ngữ tự nhiên. Cũng như kiến thức lý thuyết về thư rác, đặc điểm của nó và một số kỹ thuật lọc thư rác được mô tả ở trên. Tuy nhiên, các phương pháp lọc thư rác được mô tả ở trên có nhiều nhược điểm: chúng rất khó cập nhật danh sách địa chỉ IP vì người dùng thường thay đổi địa chỉ IP và dễ dàng bị phân loại do nhầm lẫn. Hoặc phương pháp yêu cầu / phản hồi mất rất nhiều. hoặc phương pháp yêu cầu / phản hồi mất rất nhiều. thời gian và cản trở người dùng vì cần phải liên tục thực hiện một số hành động.

Do đó, việc lọc thư rác bằng cách sử dụng nội dung của tin nhắn trở nên thú vị, được nghiên cứu và áp dụng nhiều nhất. Và nhiệm vụ lọc thư rác dựa trên học máy, sẽ được điều tra trong phần tiếp theo, được liên kết với phương pháp này.

CHƯƠNG 2. HỌC CÁCH LỌC THƯ RÁC

2.1. Thay thế nhiệm vụ

Thư rác được coi là một vấn đề lớn trên Internet. Thư rác tràn lan trên Internet, đó là điều chắc chắn.

Theo Dataprot: Số lượng email hợp pháp trung bình được gửi qua Internet mỗi ngày: 22,43 tỷ. Gần 85% tất cả các email là thư rác. (Quảng cáo chiếm 36% tổng số thư rác trên toàn thế giới.)

Tỷ lệ thư rác trong lưu lượng email toàn cầu năm 2020 giảm 6,14 điểm phần trăm so với kỳ báo cáo trước đó và đạt mức trung bình 50,37%. (dựa trên nghiên cứu của Kaspersky Lab) Một vấn đề quan trọng phát sinh trong lý thuyết thống kê về học máy là có bao nhiêu ví dụ ngẫu nhiên.

phải được sử dụng trong đào tạo để đảm bảo một lỗi phân loại đủ nhỏ với một mức độ chắc chắn nhất định.

Nhiệm vụ phân loại thư rác thực sự là nhiệm vụ phân loại thư nhận được thành hai nhóm chính: thư rác và thư thông thường.

Việc phân loại như sau. Đầu tiên, thông điệp được thể hiện dưới dạng hàm hoặc thuộc tính. Mỗi hàm thường là một từ hoặc cụm từ xuất hiện trong một thông điệp. Sau đó, nó sử dụng một tập hợp các tệp có nhãn {spam, ham}, được gọi là dữ liệu đào tạo.

Sau khi hoàn thành khóa đào tạo, người phân loại xác định loại nào trong hai loại mà các thông điệp mới rơi vào.

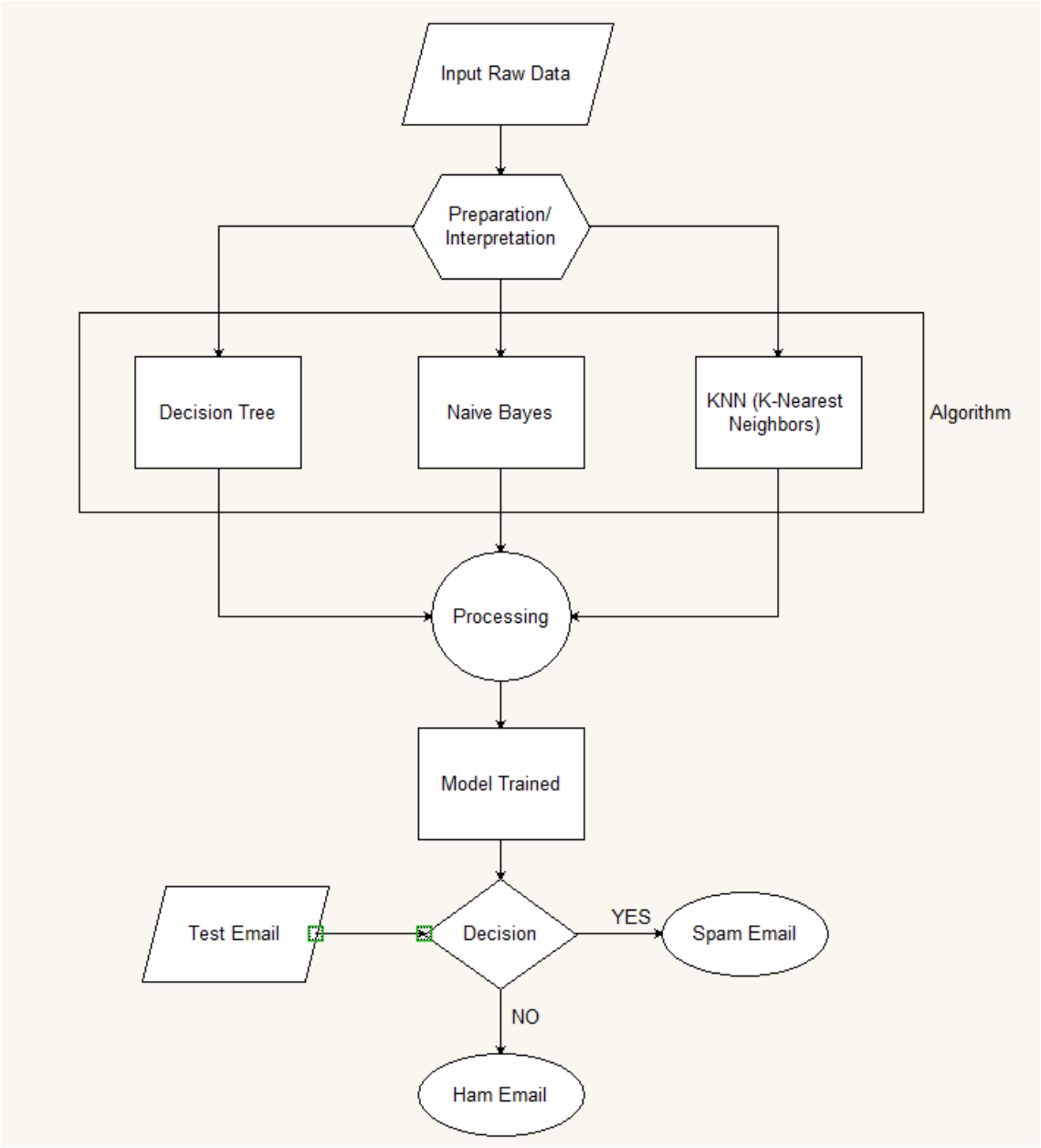
Chúng tôi xem xét nhiệm vụ phân loại thư rác với hai lớp văn bản, trong đó: thu thập dữ liệu nguồn là thư rác và thư thông thường (tin nhắn không phải là thư rác)

Kết quả của quá trình sắp xếp này là hai lớp văn bản: spam và ham.
Chúng ta có thể xây dựng nhiệm vụ như sau:

1. Mô tả vấn đề: Xác định (phân loại) email là thư rác.
2. Đầu vào: đại diện cho nội dung của thông điệp (dưới dạng vector).
3. Đầu ra : thư rác (tin nhắn hàng loạt) hoặc ham (tin nhắn thông thường)
4. Phương pháp học máy: cây quyết định, naive bayes và SVM

Bộ dữ liệu: nội dung của hơn 6000 tin nhắn và nhãn lớp ("spam" hoặc "ham").

Hình 5 cho thấy sơ đồ các bước để giải quyết vấn đề lọc thư rác dựa trên học máy.



Hình 2. 1 Sơ đồ giải quyết vấn đề lọc thư rác

2.2. Mô hình đại diện tin nhắn để áp dụng thuật toán máy học

Để sử dụng máy học và kỹ thuật xác suất thống kê, tin nhắn phải được trình bày dưới dạng thuận tiện cho việc áp dụng các thuật toán học máy. Phương pháp lọc thư sắp xếp các đại diện vector của thư theo nội dung. Mặc dù có nhiều cách để xây dựng vectơ, đơn giản nhất là mô hình "túi từ". Các nguyên tắc cơ bản của định hướng là sắp xếp các từ hoặc cụm từ trong một thông điệp, và các chữ cái được coi là một tập hợp các từ không được sắp xếp. Mỗi chữ

cái được đại diện bởi một vector. Số lượng các yếu tố trong một vector bằng với số lượng các từ khác nhau trong bộ dữ liệu đào tạo.

Nó có nhiều cách để tính toán giá trị của các yếu tố của vector. Cách dễ nhất là sử dụng giá trị nhị phân: mỗi yếu tố của vector là 1 hoặc 0, tùy thuộc vào việc từ tương ứng có xuất hiện trong thư hay không.

Ví dụ: xem xét trên hai tài liệu văn bản đơn giản:

(1) “Hi. How are you?”

(2) “How old are you”

Dựa trên hai câu văn bản này, mỗi câu được tạo ra

từ điển: {hi, how, are, you, old}

Để tạo ra một mô hình Bag-of-Word của hai câu văn bản, được coi là số lần xuất hiện của mỗi từ trong mỗi câu

Trong câu 1: "how", "are", "you" xảy ra 2 lần, và mỗi từ "hi", "old" - một lần, vì vậy vectơ của các tính năng cho câu 1 sẽ là:

{hi, how, are, you, old}

Bag-of-Word Model for Proposal 1: {1,1,1,1,0}

Cũng cho câu 2: {0,1,1,1,1,1}

Quá trình chuyển đổi văn bản:

- Đếm bao nhiêu lần mỗi từ xuất hiện trong tài liệu.
- tính toán tần suất xuất hiện của từng từ trong tài liệu từ tất cả các từ trong tài liệu.

Có : P (fi = " how " | C = ham) = 2/8 = 0,25

Chúng ta có thể viết phương pháp này nói chung:

trong đó q là tổng số từ duy nhất trong từ điển.

Trong hầu hết các ngôn ngữ, có một số từ xảy ra thường xuyên, ví dụ, trong tiếng Anh có "is", "the". Do đó, nếu chúng ta chỉ xem xét tần suất xuất hiện của mỗi từ, việc phân loại văn bản có khả năng cho kết quả sai, điều này sẽ dẫn đến mức độ chính xác thấp.

Để giải quyết vấn đề này, hãy sử dụng phương pháp TF-IDF (TF - Tần số thuật ngữ - tần số từ, IDF - Tần số tài liệu nghịch đảo - tần số tài liệu nghịch đảo). IDF được

sử dụng để đánh giá tầm quan trọng của một từ trong văn bản. Giá trị cao có tầm quan trọng cao, và nó phụ thuộc vào số lần một từ xuất hiện trong văn bản, nhưng bù đắp cho tần suất của từ đó trong bộ dữ liệu. Một số biến thể của TF-IDF thường được sử dụng trong các công cụ tìm kiếm như là công cụ chính để đánh giá và sắp xếp văn bản dựa trên truy vấn của người dùng. TF-IDF cũng được sử dụng để lọc các từ bị bỏ qua trong các nhiệm vụ như tóm tắt văn bản và phân loại văn bản.

Tần số từ (TF) đưa ra tần số của từ trong mỗi tài liệu. Đây là

Tỷ lệ bao nhiêu lần một từ xuất hiện trong một tài liệu, so với tổng số từ trong tài liệu. Nó tăng lên khi số lượng tăng lên.

sự xuất hiện của từ này trong tài liệu. Mỗi tài liệu có giá trị TF riêng.

Công thức Tính TF :

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

trong đó: f (t, d): đây là số lần xuất hiện của từ t trong tài liệu d

max ({f (w, d): w ∈ d}): số lần xuất hiện của từ có số lần xuất hiện cao nhất trong tài liệu d.

Document Back frequency (IDF) được sử dụng để tính toán trọng lượng của các từ hiếm trong tất cả các tài liệu trong corpus. Những từ hiếm khi được tìm thấy trong corpus có giá trị IDF cao. Mục đích của máy tính IDF là để giảm lợi thế của các từ thường xuất hiện như "là", "the"... Bởi vì những từ này không có tầm quan trọng lớn trong việc phân loại văn bản.

Công thức IDF như sau:

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{ d_i \in D \mid t \in d_i \}|}$$

nơi | D| là số lượng tài liệu trong Gói D và | {di D | t di} | là số lượng tài liệu trong Corpus D trong đó t xảy ra (khi n_t≠0).

Do đó, giá trị TF-IDF của bất kỳ từ nào trong tài liệu được tính là:

$$\text{tf-idf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

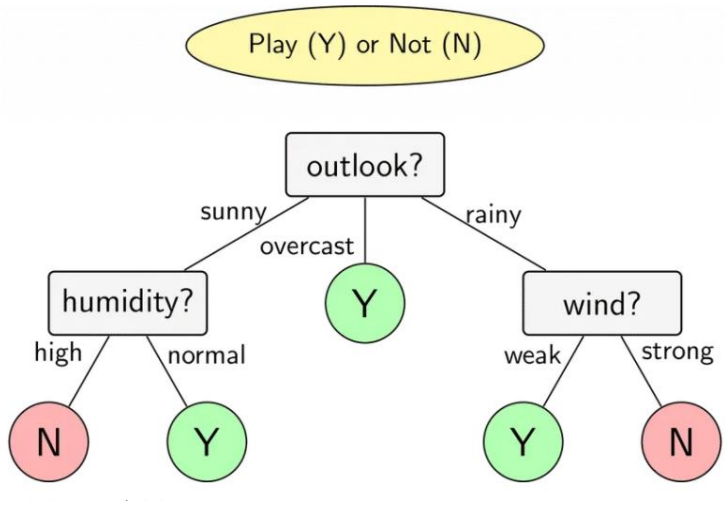
Các từ có giá trị TF-IDF cao là những từ phổ biến trong tài liệu này và ít phổ biến hơn trong các tài liệu khác. Điều này giúp lọc ra các từ phổ biến và lưu các từ quan trọng (từ khóa của văn bản này)..

2.3. Mô hình phân loại máy học và spam

2.3.1. Cây quyết định

Cây quyết định là một trong những loại cấu trúc quan trọng và hữu ích nhất cho học máy. Cây quyết định là một thuật toán học máy phổ quát có thể đối phó với các nhiệm vụ phân loại và hồi quy [2]. Một bộ dữ liệu đào tạo được sử dụng để tạo ra một cây mà sau này được sử dụng để dự đoán dữ liệu thử nghiệm. Trong thuật toán này, mục tiêu là đạt được kết quả chính xác nhất với số lượng quyết định ít nhất cần được đưa ra. Cây quyết định có thể được sử dụng cho cả nhiệm vụ phân loại và hồi quy.

Ví dụ: giả sử điều kiện thời tiết để đi ra ngoài, các chàng trai quyết định chơi bóng đá hay không. Mô hình được thể hiện trong hình 6.



Hình 2. 2 Mô hình cây quyết định sử dụng thuật toán ID3.

Thuật toán Entropy và ID3

Entropy ở đây đề cập đến mức độ không chắc chắn trong nội dung của dữ liệu. Phân bố xác suất của một biến rời rạc x, có thể lấy n các giá trị khác nhau. Giả sử rằng xác suất mà x sẽ chấp nhận các giá trị này là $p_i = p\ (x = x_i)$ *npu* $0 \leq p_i \leq 1$, $\sum_{i=1}^{n} p_i = 1$ thì công thức toán học cho entropy như sau

$$H(p) = -\sum_{i=1}^{n} p_i \, log(p_i)$$

Ví dụ: giả sử bạn có một bảng dữ liệu bữa tối tại một nhà hàng và có các biến khác nhau trong bảng đó. Những biến số này có thể là: "Đói?", "Thời tiết có tốt không?", "Tôi có muốn ở nhà không?", "Tôi có muốn đặt hàng thứ gì đó không?", v.v... Những biến số này dẫn đến hai kết quả khác nhau, đó là có hay không. Vì vậy, nó có 2 lớp: tích cực hoặc tiêu cực.

Đó là, nếu có 8 kết quả tích cực trong bảng và 4 kết quả tiêu cực, entropy có thể được tính bằng công thức:

$$H(p) = -\frac{8}{12}\log\frac{8}{12} - \frac{4}{12}\log\frac{4}{12} = 0{,}636$$

Một thuật toán phổ biến để xây dựng một cây quyết định là ID3 (Iterative Dichotomiser 3), áp dụng cho một vấn đề phân loại trong đó tất cả các thuộc tính ở dạng danh mục. Nó dựa trên các khái niệm về entropy và lợi ích thông tin. Trong ID3, tổng entropy có trọng số ở các nút cuối sau khi cây quyết định được xây dựng được coi là chức năng mất mát của cây quyết định đó. Công việc của ID3 là tìm các phân chia logic (thứ tự trong đó các

thuộc tính logic được chọn) để chức năng mất cuối cùng đạt kích thước nhỏ nhất có thể.

Nói một cách đơn giản, thuật toán ID3 có thể được mô tả như sau: bắt đầu từ nút gốc, ở mỗi giai đoạn chia dữ liệu thành các bộ dữ liệu đồng nhất. Đặc biệt, tìm thuộc tính sẽ dẫn đến việc truy xuất thông tin lớn nhất, tức là sẽ trả lại các nhánh đồng nhất nhất .

Tóm tắt thuật toán:

- Cuộc họp bộ dữ liệu
- Xác định thuộc tính tốt nhất về mặt nhận thông tin
- Chia một tập dữ liệu thành các tập con liên quan đến thuộc tính tốt nhất
- Tạo một cây quyết định, gốc của nó có thuộc tính tốt nhất
- Thực hiện các bước tương tự cho gốc mới và các tập con mới của nó.

Ưu điểm của cây quyết định như sau:

- Chỉ để hiểu và giải thích, thuật toán rất dễ giải thích;
- Thuật toán tuân theo cách tiếp cận tương tự mà con người thường làm theo khi đưa ra quyết định;
- Quá trình học nhanh.

Nhược điểm: Mô hình cây quyết định phụ thuộc rất nhiều vào dữ liệu ban đầu. Ngay cả với một thay đổi nhỏ trong bộ dữ liệu, cấu trúc của mô hình cây quyết định có thể thay đổi hoàn toàn.

2.3.2. Thuật toán Naive Bayes

Thuật toán Naive Bayes là một thuật toán học máy

để giải quyết vấn đề phân loại dựa trên định lý Bayes. Nó có thể được sử dụng cho cả hai nhiệm vụ phân loại nhị phân và nhiều lớp. Điểm chính là dựa trên ý tưởng xem xét độc lập của từng chức năng. Phương pháp ngây thơ của Bayes ước tính xác suất của từng đặc điểm một cách độc lập, bất kể mối tương quan nào và đưa ra dự đoán dựa trên định lý Bayes. .

Để hiểu Thuật toán Naive Bayes, trước tiên người ta phải giới thiệu các khái niệm.

xác suất của các lớp và xác suất có điều kiện.

- Xác suất lớp là xác suất của một lớp trong bộ dữ liệu. Nói cách khác, nếu bạn chọn một mục ngẫu nhiên từ một bộ dữ liệu, đó là xác suất mà nó thuộc về một lớp cụ thể;

- xác suất có điều kiện là xác suất của giá trị của một tính năng được chỉ định bởi một lớp.

Xác suất của một lớp được tính đơn giản là số lượng mẫu trong lớp,

Chia cho tổng số mẫu:

$$P(C) = \frac{N(C)}{N(total)}$$

Xác suất có điều kiện được tính là tần số của mỗi giá trị thuộc tính chia cho tần suất của các phiên bản của lớp đó:

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

Xem xét công thức xác định xác suất có điều kiện:

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

Nó cho thấy xác suất có điều kiện của sự kiện A, miễn là kết quả của sự kiện B đã được biết đến, được biểu thị là P(A| B) và có tên thứ hai "xác suất posteriori". Đồng thời, bạn nên biết:

P(A) là xác suất tiên nghiệm của sự kiện A, nghĩa là xác suất thể hiện giả định về sự kiện A trước khi tính đến kết quả của thí nghiệm;

P(A|B) là xác suất của sự kiện A khi sự kiện B xảy ra (xác suất posteriori là xác suất của một sự kiện trong một số điều kiện);

- P(B|A) – xác suất của sự kiện B với sự thật của giả thuyết A (có điều kiện hàm xác suất hoặc khả năng);

- P(B) – tổng xác suất xảy ra sự kiện B;

Một mô hình đơn giản hơn là mô hình Naive Bayes[12], trong đó nếu xác suất A phụ thuộc vào xác suất của nhiều lý thuyết khác B, C, D... sau đó xác suất cho một số sự kiện được

$$P(A|B, C, D, \dots) = \frac{P(B,C,D,\dots|A) * P(A)}{P(B,C,D,\dots)},$$

tính theo quy tắc của Bayes như sau:

Theo một cách đơn giản hơn, trong đó mối quan hệ giữa sự kiện và thuộc tính hoặc các sự kiện khác không được biết đến :

$$P(A|B,C,D,\dots) = P(A|B) * P(A|C) * P(A|D) * \dots$$

Trong Gaussian Naive Bayes, các giá trị liên tục liên quan đến mỗi

vật thể dự kiến sẽ được phân phối theo Gaussian

Phân phối. Phân bố Gaussian còn được gọi là bình thường

Các classifiers Naive Bayes nổi tiếng khác bao gồm:

- Multinomial Naive Bayes: : vectơ tính năng đại diện cho tần số, với

theo đó một số sự kiện nhất định được tạo ra bởi một đa thức

Phân phối. Đây là mô hình sự kiện thường được sử dụng cho

phân loại tài liệu.

- Bernoulli Naive Bayes : trong một mô hình đa chiều của các sự kiện Bernoulli, các chức năng là Boolean độc lập (biến nhị phân),

mô tả dữ liệu đầu vào. Giống như mô hình đa thức, mô hình này

phổ biến cho các tác vụ phân loại tài liệu sử dụng nhị phân

sự xuất hiện của các thuật ngữ (tức là từ có xảy ra trong tài liệu hay không), không phải tần số

thuật ngữ (nghĩa là tần số của một từ trong tài liệu).

Những ưu điểm của phân loại Bayesian ngây thơ như sau:

- Rất đơn giản, dễ thực hiện và nhanh chóng;
- cần ít dữ liệu hơn để đào tạo;
- có khả năng mở rộng cao;
- có thể được sử dụng cho cả nhiệm vụ nhị phân và nhiều lớp

phân loại;

- xử lý dữ liệu liên tục và rời rạc.

Trong nhiệm vụ phân loại thư rác, chúng tôi coi mỗi mẫu là một tập hợp các chữ cái và bạn có thể thuộc các lớp C = {spam, ham}.

Khi bạn nhận được một tin nhắn, nếu bạn không biết gì về nó, thật khó để quyết định xem đó có phải là thư rác hay không.

Nếu bạn có thêm thông tin hoặc thuộc tính tin nhắn, bạn có thể cải thiện hiệu quả nhận thư dưới dạng thư rác. Tin nhắn có nhiều tính năng như tiêu đề, nội dung, tệp đính kèm,... Bạn có thể dựa vào thông tin này để cải thiện hiệu quả của bạn trong việc phân loại thư rác. Một ví dụ đơn giản: nếu bạn phát hiện ra rằng 95% tin nhắn html là thư rác, thì khi bạn nhận được tin nhắn html, thì để tính toán xác suất chúng ta có thể dựa vào xác suất được xác định trước rằng một tin nhắn như vậy là thư rác.

Sử dụng phương pháp phân loại Bayesian đơn giản, mỗi chữ cái được thể hiện bằng vector $\vec{x}$ = (x1, x2, ..., xn) trong đó x_1, x_2, ..., x_n là giá trị của đặc điểm của X_1, X_1, X_2, ..., X_n. Ở đây n là số lượng các tính năng được xác định từ bộ dữ liệu đào tạo, nghĩa là số lượng từ / cụm từ khác nhau trong bộ dữ liệu đào tạo. Mỗi thư được gán một nhãn phân loại Y, có thể lấy một trong hai giá trị: Y = 1 cho thư rác và Y = 0 cho thư thông thường.

Để xác định nhãn phân loại cho thư, bộ phân loại Bayesian tính xác suất có điều kiện:

$$P\,(Y = y \mid X_1 = x_1),\ \ldots,\ X_n = x_n)$$

$$P(Y = y \mid X_1 = x_1,\ldots,X_n = x_n) = \frac{P(X_1 = x_1,\ldots,X_n = x_n \mid Y = y)\cdot P(Y = y)}{P(X_1 = x_1,\ldots,X_n = x_n)}$$

đó là xác suất một thư có nội dung (x_1, x_2, ..., x_n) sẽ nhận được nhãn thể loại y, y $\in$ {1,0}. sử dụng công thức bayes xác suất trên được tính như sau

Do đó, trong trường hợp lọc thư rác, nhãn tin nhắn được xác định bằng cách tính giá trị của biểu thức:

$$\frac{P(Y = 1 \mid X_1 = x_1,\ldots,X_n = x_n)}{P(Y = 0 \mid X_1 = x_1,\ldots,X_n = x_n)} = \frac{P(X_1 = x_1,\ldots,X_n = x_n \mid Y = 1)\cdot P(Y = 1)}{P(X_1 = x_1,\ldots,X_n = x_n \mid Y = 0)\cdot P(Y = 0)}$$

Giá trị biểu thức lớn hơn 1 có nghĩa là xác suất thư là thư rác cao hơn tin nhắn thông thường và tin nhắn sẽ được đánh dấu là thư rác. Một giá trị biểu thức nhỏ hơn 1 tạo ra kết quả ngược lại.

2.3.3. SVM

Định nghĩa

Là phương pháp dựa trên nền tảng của lý thuyết thống kê nên có một nền tảng toán học chặt chẽ để đảm bảo rằng kết quả tìm được là chính xác.

Là thuật toán học giám sát (supervied learning) được sử dụng cho phân lớp dữ liệu.

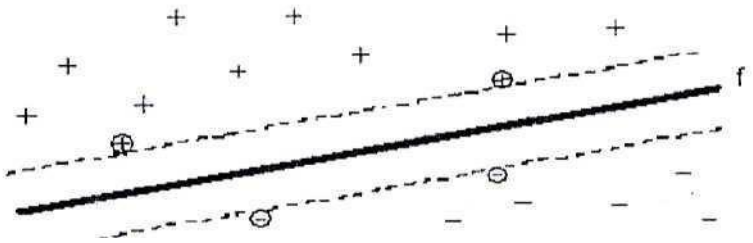
Là 1 phương pháp thử nghiệm, đưa ra 1 trong những phương pháp mạnh và chính xác nhất trong số các thuật toán nổi tiếng về phân lớp dữ liệu.

SVM là một phương pháp có tính tổng quát cao nên có thể được áp dụng cho nhiều loại bài toán nhận dạng và phân loại

Ý tưởng của phương pháp

Cho trước một tập huấn luyện, được biểu diễn trong không gian vector, trong đó mỗi tài liệu là một điểm, phương pháp này tìm ra một siêu phẳng quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp + và lớp -. Chất lọợng của siêu phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó, khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác.

Mục đích của phương pháp SVM là tìm được khoảng cách biên lớn nhất, điều này được minh họa như sau:



Hình 2. 3 Siêu phẳng phân chia dữ liệu học thành 2 lớp + và - với khoảng cách biên lớn nhất. Các điểm gần nhất (điểm được khoanh tròn) là các Support Vector.

Nội dung Phương pháp

Cơ sở lý thuyết

SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp

nhất.

Bài toán SVM có thể giải bằng kỹ thuật sử dụng toán tử Lagrange để biến đổi về thành dạng đẳng thức. Một đặc điểm thú vị của SVM là mặt phẳng quyết định chỉ phụ thuộc các Support Vector và nó có khoảng cách đến mặt phẳng quyết

Cho dù các điểm khác bị xóa đi thì thuật toán vẫn cho kết quả giống như ban đầu. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác vì tất cả các dữ liệu trong tập huấn luyện đều được đùng để tối ưu hóa kết quả.

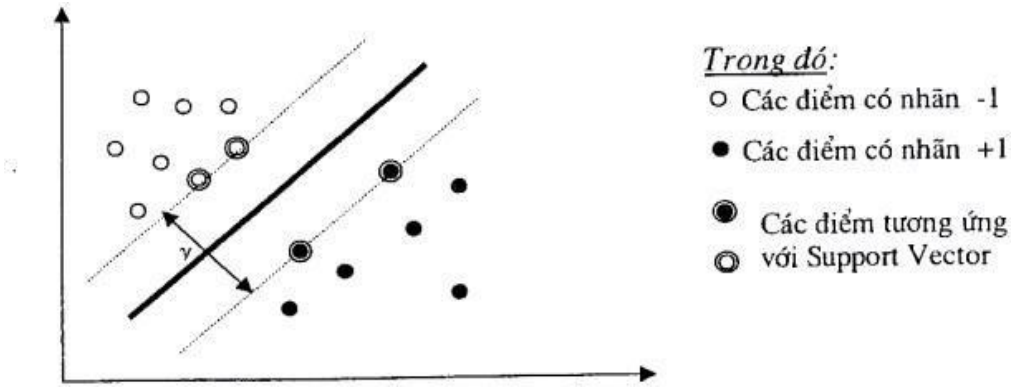
TÓM LẠI: trong trường hợp nhị phân phân tách tuyến tính, việc phân lớp được thực hiện qua hàm quyết định f(x) = sign(<w.x> + b), hàm này thu được bằng việc thay đổi vectơ chuẩn w, đây là vectơ để cực đại hóa viền chức năng

Việc mở rộng SVM để phân đa lớp hiện nay vẫn đang được đầu tơị nghiên cứu. Có một phương pháp tiếp cận để giải quyết vấn để này là xây dựng và kết hợp nhiều bộ phân lớp nhị phân SVM (Chẳng hạn: trong quá trình luyện với SVM, bài toán phân m lớp có thể được biến đổi thành bài toán phân 2*m lớp, khi đó trong mỗi hai lớp, hàm quyết định sẽ được xác định cho khả năng tổng quát hóa tối đa). Trong phương pháp này có thể đề cập tới hai cách là một-đổi-một, một-đối-tất cả

Bài toán phân 2 lớp với SVM

Bài toán đặt ra là: Xác định hàm phân lớp để phân lớp các mẫu trong tơịơng lai, nghĩa là với một mẫu dữ liệu mới xi thì cần phải xác định xi được phân vào lớp +1 hay lớp -1

Để xác định hàm phân lớp dựa trên phương pháp SVM, ta sẽ tiến hành tìm hai siêu phẳng song song sao cho khoảng cách y giữa chúng là lớn nhất có thể để phân tách hai lớp này ra làm hai phía. Hàm phân tách tơịơng ứng với phương trình siêu phẳng nằm giữa hai siêu phẳng tìm được



Hình 2. 4 Minh họa bài toán 2 phân lớp bằng phương pháp SVM

Các điểm mà nằm trên hai siêu phẳng phân tách được gọi là các Support Vector. Các điểm này sẽ quyết định đến hàm phân tách dữ liệu

Bài toán phân nhiều lớp với SVM

Để phân nhiều lớp thì kỹ thuật SVM nguyên thủy sẽ chia không gian dữ liệu thành 2 phần và quá trình này lặp lại nhiều lần. Khi đó hàm quyết định phân dữ liệu vào lớp thứ i của tập n , 2-Iớp sẽ là:

fi(x) = wiix + bi

Những phần tử x là support vector sẽ thỏa điều kiện

Như vậy, bài toán phân nhiều lớp sử dụng phương pháp SVM hoàn toàn có thể thực hiện giống như bài toán hai lớp. Bằng cách sử dụng chiến lọịợc "một- đối- một”(one – against - one).

Giả sử bài toán cần phân loại có k lớp (k > 2), chiến lọịợc "một-đối-một”sẽ tiến hành k(k-l)/2 lần phân lớp nhị phân sử dụng phương pháp SVM. Mỗi lớp sẽ tiến hành phân tách với k-1 lớp còn lại để xác định k-1 hàm phân tách dựa vào bài toán phân hai lớp bằng phương pháp SVM.

các bước chính của phương pháp SVM

Phương pháp SVM yêu cầu dữ liệu được diễn tả như các vector của các số thực. Như vậy nếu đầu vào chọịa phải là số thì ta cần phải tìm cách chuyển chúng về dạng số của SVM

Tiền xử lý dữ liệu: Thực hiện biến đổi dữ liệu phù hợp cho quá trình tính toán, tránh các số quá lớn mô tả các thuộc tính. Thọịờng nên co giãn (scaling) dữ liệu để chuyển về đoạn [-1, 1] hoặc [0, 1].

Chọn hàm hạt nhân: Lựa chọn hàm hạt nhân phù hợp tọơng ứng cho từng bài toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp.

Thực hiện việc kiểm tra chéo để xác định các thám số cho ứng đụng. Điều này cũng quyết định đến tính chính xác của quá trình phân lớp.

Sử dụng các tham số cho việc huấn luyện với tập mẫu. Trong quá trình huấn luyện sẽ sử dụng thuật toán tối ưu hóa khoảng cách giữa các siêu phẳng trong quá trình phân lớp, xác định hàm phân lớp trong không gian đặc trưng nhờ việc ánh xạ dữ liệu vào không gian đặc trưng bằng cách mô tả hạt nhân, giải quyết cho cả hai trường hợp dữ liệu là phân tách và không phân tách tuyến tính trong không gian đặc trưng.

CHƯƠNG 3. CHUẨN BỊ DỮ LIỆU VÀ THIẾT KẾ ỨNG DỤNG

3.1. Lựa chọn ngôn ngữ lập trình và chuẩn bị dữ liệu

Công việc được thực hiện trong Python, sử dụng PyCharm IDE. Các thư viện scikit-learn, NumPy và gấu trúc đã được sử dụng.

3.1.1. Python

Python là một ngôn ngữ lập trình hướng đối tượng mạnh mẽ, tiên tiến được tạo ra bởi Guido van Rossum. Nó rất dễ học, và nó trở thành một trong những ngôn ngữ lập trình giới thiệu tốt nhất cho người mới bắt đầu. Python có cấu trúc dữ liệu cấp cao mạnh mẽ, nhưng là một cách tiếp cận hiệu quả để lập trình hướng đối tượng. Các lệnh cú pháp Python là một điểm cộng rất lớn vì nó rõ ràng, chỉ để hiểu.

Python nhanh chóng trở thành một ngôn ngữ lý tưởng để viết kịch bản và phát triển các ứng dụng trong các lĩnh vực khác nhau trên tất cả các nền tảng.

Python được coi là ngôn ngữ ưa thích để học máy, đào tạo Vì:

Ngôn ngữ lập trình đơn giản, dễ dàng: Python có cú pháp rất đơn giản và dễ hiểu. Dễ đọc và viết hơn nhiều so với các ngôn ngữ lập trình khác như C++, Java, C#. Python làm cho lập trình trở nên thú vị bằng cách cho phép bạn tập trung vào các giải pháp thay vì cú pháp.

Ngôn ngữ lập trình nguồn mở: Sử dụng và phân phối Python miễn phí, ngay cả cho mục đích thương mại. Vì nó là mã nguồn mở, không chỉ có thể sử dụng phần mềm hoặc chương trình được viết bằng Python, mà còn có thể sửa đổi mã nguồn của nó. Python có một cộng đồng khổng lồ được cải thiện với mỗi bản cập nhật.

Tính di động: Các chương trình Python có thể di chuyển sang một nền tảng khác và chạy nó mà không cần bất kỳ sửa đổi nào. Nó hoạt động mà không gặp vấn đề gì trên hầu hết các nền tảng, chẳng hạn như Windows, macOS, Linux.

Thư viện tiêu chuẩn lớn: Python có một số lượng lớn các thư viện tiêu chuẩn đơn giản hóa việc lập trình vì bạn không phải tự viết tất cả mã

3.1.2. Pycharm

PyCharm là một môi trường phát triển tích hợp cho ngôn ngữ lập trình Python. PyCharm cung cấp một bộ công cụ hoàn chỉnh cho các nhà phát triển Python chuyên nghiệp. PyCharm được xây dựng trên cơ sở một trình chỉnh sửa và gỡ lỗi hiểu sâu sắc về mã và đưa ra một ý tưởng rõ ràng về hành vi của mã.

PyCharm cung cấp tích hợp với các công cụ cộng tác như hệ thống điều khiển phiên bản và trình theo dõi. PyCharm mở rộng các khả năng cốt lõi bằng cách tích hợp liền mạch với các khung web, công cụ JavaScript, ảo hóa và hỗ trợ container hóa.

Các tính năng chính của PyCharm
- Hỗ trợ Windows, macOS và Linux
Hỗ trợ hoàn thành mã thông minh, điều hướng một cú nhấp chuột và xác thực kiểu PEP8
- Tái cấu trúc an toàn và tự động trong dự án
Tự động phát hiện các vấn đề về mã: ví dụ: phân tích mã không sử dụng
- Trình gỡ lỗi hiệu suất cao
- Chế độ mô phỏng VIM

3.1.3. Thư viện Scikit-learning , NumPy , Pandas ,Scikit-learning

Scikit-learn (Sklearn) là thư viện mạnh nhất cho các thuật toán học máy được viết bằng Python. Thư viện cung cấp một bộ công cụ để xử lý các nhiệm vụ mô hình hóa máy học và thống kê, bao gồm classification, regression, clustering và dimensionality reduction**.**
Để cài đặt scikit-learn, trước tiên bạn phải cài đặt thư viện SciPy (Scientific Python). Thành phần bao gồm :
NumPy: Một gói thư viện xử lý các mảng và ma trận đa chiều. NumPy được sử dụng để thực thi số biến đổi và hoạt động của đại số tuyến tính.
SciPy: Một gói các chức năng tính toán logic khoa học.
Matplotlib: trình bày dữ liệu dưới dạng đồ thị 3 chiều 2 chiều
IPython: Một notepad được sử dụng để tương tác trực quan với Python.
SymPy: thư viện các biểu tượng toán học
Pandas : cũng là một thư viện rất phổ biến được sử dụng để trích xuất
dữ liệu từ các nguồn khác nhau (cơ sở dữ liệu SQL, tệp JSON, tệp CVS).
Thư viện có nhiều phương pháp để lọc, kết hợp và phân tích dữ liệu.
Scikit-learn cung cấp hỗ trợ mạnh mẽ cho việc viết thuật toán. Điều này có nghĩa là thư viện này được hỗ trợ sâu sắc: dễ sử dụng, dễ mã hóa và xử lý và hiệu quả.

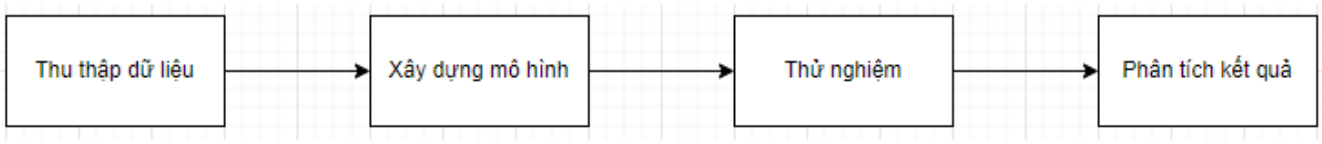
3.1.4. Microframework Flusk

Flask là một khung web nhỏ và nhẹ cho Python, cũng được coi là một microframework vì :
nó không yêu cầu các công cụ hoặc thư viện đặc biệt.
Ưu điểm của Flask là dễ sử dụng. Rất ít lỗi xuất hiện tại nơi làm việc do ít phụ thuộc hơn và dễ dàng phát hiện và loại bỏ các lỗ hổng bảo mật.
Flask cung cấp cho các nhà phát triển khả năng tùy chỉnh sự phát triển của các ứng dụng web và cung cấp các công cụ, thư viện và cơ chế cần thiết cho phép bạn tạo các ứng dụng web.

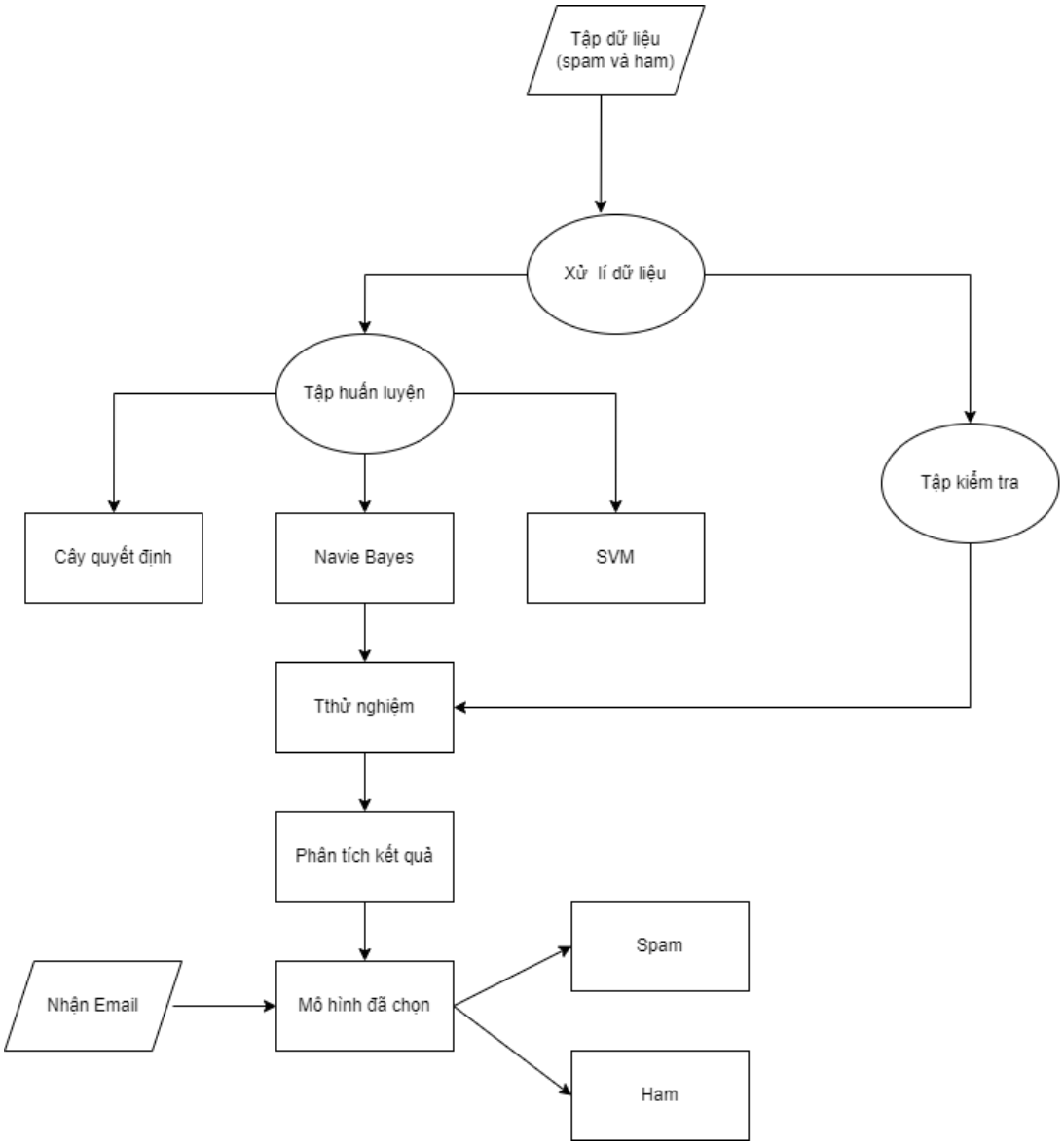
3.2. Chuẩn bị dữ liệu và xây dựng lược đồ mô hình

3.2.1. Các giai đoạn xây dựng mô hình

Các số liệu cho thấy các giai đoạn của quá trình xây dựng mô hình và sơ đồ của quá trình xây dựng mô hình.



Hình 3. 1 Các giai đoạn chính của việc xây dựng mô hình



Hình 3. 2 Sơ đồ chi tiết quy trình xây dựng mô hình

3.2.2. Tạo bộ dữ liệu và xử lý trước

Tập hợp các trang Internet độc hại được sử dụng được lấy từ các nguồn sau:

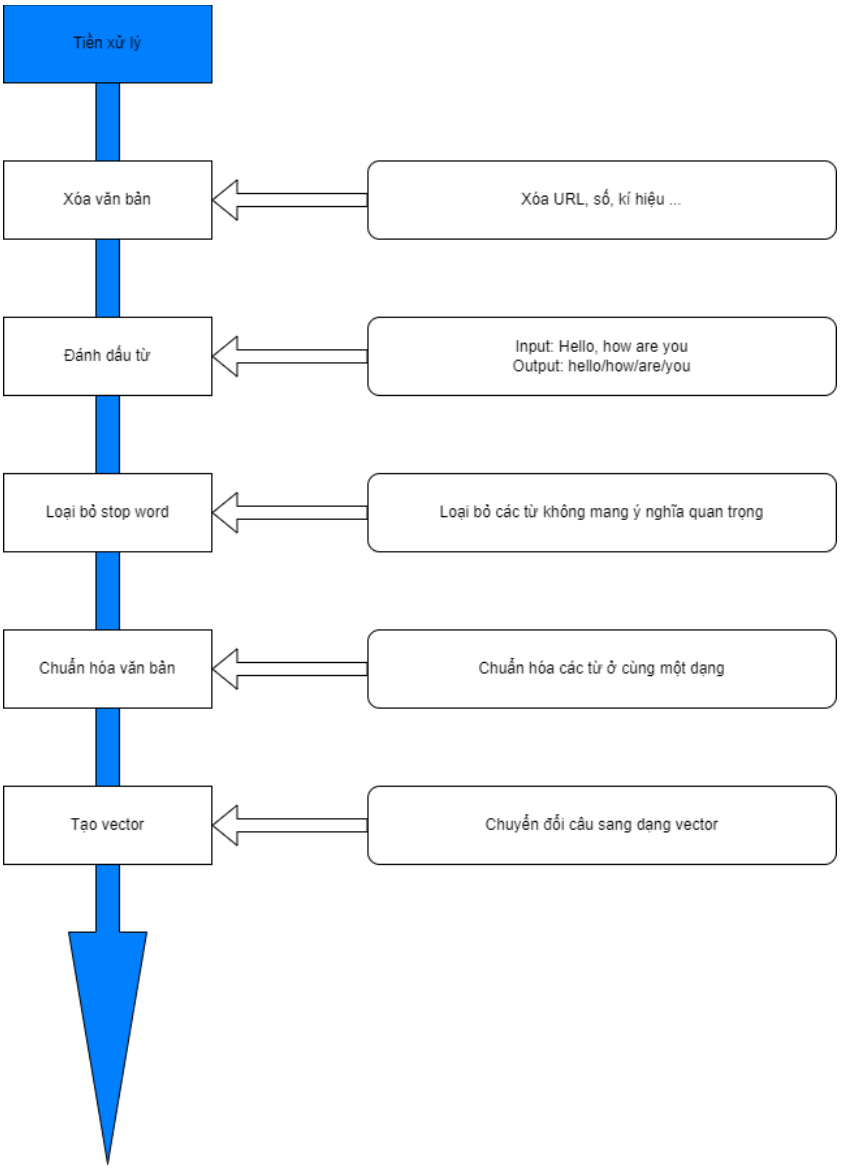
https://www.kaggle.com/veleon/ham-and-spam-dataset

https://www.kaggle.com/nitishabharathi/email-spam-dataset

Bộ dữ liệu bao gồm hơn 4300 thư rác thường xuyên và hơn 1500 thư rác.

Sơ bộ xử lý :

Trước khi bạn bắt đầu đào tạo, bạn phải xử lý trước các tin nhắn. Xử lý trước dữ liệu được coi là một trong những phần quan trọng nhất của phân loại thư rác. Hình ảnh cho thấy sơ đồ các bước xử lý trước.



Hình 3. 3 Sơ đồ các bước xử lý trước

Loại bỏ URL

URL nên được xóa và kết quả phân tích không bị ảnh hưởng, vì nó không được tính là từ.

Loại bỏ tất cả các ký tự không cần thiết (số và dấu chấm câu)

Xóa các số nếu chúng không liên quan đến phân tích (0–9). Và dấu câu cũng sẽ được gỡ bỏ. Dấu câu về cơ bản là một tập hợp các ký tự [! "# $% &' () * +, -. / :; <=>? @ [] ^ _ `{|} ~]: Xóa bỏ từ dừng

"Stop words" là những từ phổ biến nhất trong các ngôn ngữ như "the", "a", "i", "is", "all". Những từ này không mang ý nghĩa quan trọng và thường bị loại bỏ khỏi các văn bản. Tôi sẽ trình bày kết quả của việc xử lý trước trong hình.

	Content	Label
1848	from yyyy example com mailto yyyy example com...	0
2243	url http boingboing net 85531557 date not supp...	0
1802	url http boingboing net 85519849 date not supp...	0
8	url http www aaronsw com weblog 000614 date 20...	0
120	how are you doing if you ve been like me you v...	1

Hình 3. 4 Kết quả lọc

Chuẩn hóa văn bản

Biểu tượng cảm xúc phải được loại bỏ, sau đó văn bản được chuẩn hóa, các khoảng trống trailing được loại bỏ và chữ thường được chuyển đổi. Bởi vì "điện thoại" và "TelePhone" sẽ được coi là hai từ riêng biệt nếu bước này không được thực hiện.

Tạo vector

Sử dụng TF-IDF Vectors làm phương pháp tính năng, bạn sẽ nhận được một ma trận trong đó mỗi hàng đại diện cho văn bản, mỗi cột đại diện cho một từ trong từ điển và mỗi ô sẽ chứa tần số của các từ trong văn bản tương ứng.

3.2.3. Thiết kế mô hình

Trên sân khấu học tập dữ liệu sẽ 3 thuật toán được sử dụng : cây nhận con nuôi quyết định , ngây thơ Bayesian bộ phân loại và k - phương thức gần nhất hàng xóm láng giềng . Ban đầu dữ liệu trên đây sân khấu là túi làm _ trên sân khấu sơ bộ quá trình xử lý . kết quả đây bươc Là tạo người mẫu cỗ máy đào tạo phù hợp _ vì bộ dữ liệu . Trên nền tảng đây người mẫu dự đoán sự phân loại khác đầu vào các tài liệu . Cũng thế Là nền tảng nhiệm vụ nghiên cứu trong này dự án .

Bộ dụng cụ ban đầu dữ liệu đã chia ra thành 2 tệp : đào tạo (gồm 4389 tin nhắn) và kiểm tra (gồm 1463 tin nhắn).

Sau học tập người mẫu là đã chọn tốt nhất người mẫu vì sự sáng tạo các ứng dụng .

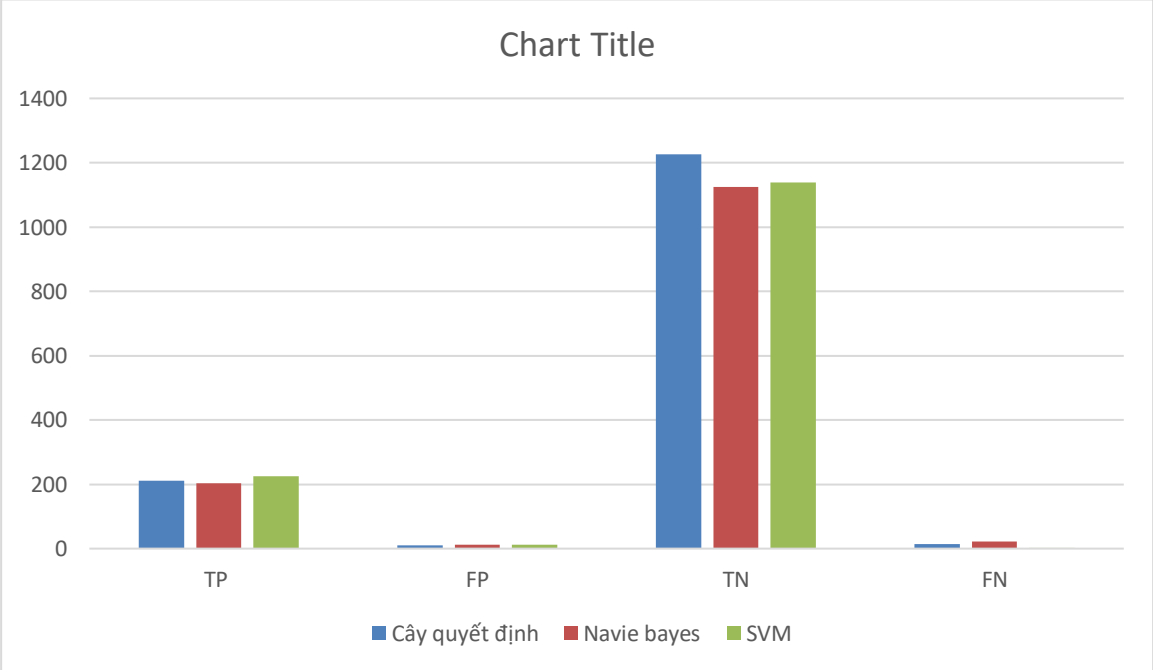
3.2.4. Tính toán nhu cầu sắc suất

Sau khi xử lý trước dữ liệu và đào tạo dữ liệu bằng 3 thuật toán. Kết quả khá khả quan. Bộ dữ liệu đào tạo cho kết quả kiểm tra với độ chính xác 100%.

Kết quả xác thực dữ liệu thử nghiệm được trình bày trong bảng sau:

	TP.	FP	TN	FN
quyết định cây	212	10	1226	15
Naïve-Bayes	204	12	1 224	23
SVM	225	98	1138	2

Bảng 3. 1 Thuật toán 3 Kết quả kiểm tra



Hình 3. 5 Đồ thị sự so sánh giữa 3 thuật toán

Bằng cách so sánh các kết quả thu được, thuật toán hóa ra cây quyết định có kết quả

phân loại chính xác nhất. Số lượng lỗi ít hơn những lỗi khác. Tiếp theo, một số số liệu được tính toán để đánh giá chất lượng của các thuật toán.

Độ chính xác :

- Cây quyết định:

Độ chính xác $= \frac{P}{N} = \frac{212+1226}{212+10+1226+15} = 0.983$

- Navie Bayes

Độ chính xác $= \frac{P}{N} = \frac{204+1224}{204+12+1224+23} = 0.976$

- SVM

Độ chính xác $= \frac{P}{N} = \frac{225+1338}{225+98+1338+2} = 0.940$

Hiệu quả của lọc thư được đánh giá bởi nhiều tiêu chí, chẳng hạn như, Thu hồi, Độ chính xác và độ chính xác tổng thể, có nghĩa là tỷ lệ phần trăm của các tin nhắn được phân loại chính xác, bất kể đó có phải là thư rác hay không. Trong dự án của mình, tôi tập trung vào việc đánh giá hiệu quả của việc lọc thư với Độ chính xác, được định nghĩa như sau:

Độ chính xác :

- Cây quyết định:

Độ chính xác $= \frac{TP}{TP + FP} = \frac{212}{212+10} = 0.955$

- Navie Bayes

Độ chính xác $= \frac{TP}{TP + FP} = \frac{204}{204+12} = 0.944$

- SVM

Độ chính xác $= \frac{TP}{TP + FP} = \frac{225}{225+98} = 0.697$

Độ chính xác nên quan trọng hơn khi chọn một mô hình, vì nhận dạng FP không đúng cách dẫn đến kết quả kém. Ví dụ, phân loại nhầm một lá thư thông thường là thư rác sẽ ảnh hưởng đến công việc của người dùng do thiếu thư quan trọng (ví dụ, hợp đồng là một triệu rúp).

Phương pháp hàng xóm gần nhất có độ chính xác thấp hơn vì có ít tin nhắn rác trong bộ

dữ liệu hơn các tin nhắn thông thường.

Do phân tích kết quả, một mô hình cây quyết định đã được chọn để thiết kế và thực hiện ứng dụng.

Kết luận: Trong chương này, các hành động sau đây đã được thực hiện:

- Ngôn ngữ lập trình và IDE đã được chọn.

- Một sơ đồ của các giai đoạn tạo mô hình đã được xây dựng, một tập dữ liệu đã được tạo ra và xử lý trước đã được thực hiện.

- 3 mô hình đã được thiết kế và các xác suất cần thiết đã được tính toán.

Kết quả chính của chương này là một mô hình tốt hơn để phát triển một ứng dụng.

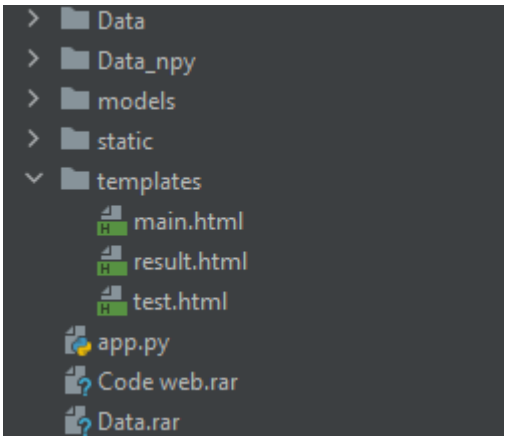
CHƯƠNG 4. TIỂN KHAI ỨNG DỤNG VÀ PHÂN BỔ KẾT QUẢ

4.1. Công cụ triển khai ứng dụng web

Để bắt đầu phát triển một dự án, bạn cần tạo một thư mục gốc với Tên dự án được chỉ định.

Dự án được thiết kế bằng cách sử dụng khung Flask cho hoạt động chính của ứng dụng.

Con số cho thấy thư mục ứng dụng.



Hình 4. 1 Thư mục gốc ứng dụng

Trong này thư mục là thư mục :

- Thư mục dữ liệu: chứa một bộ dữ liệu cho các mô-đun đào tạo và thử nghiệm.
- Thư mục tĩnh: chứa các tệp thống kê để thiết kế tệp html, bao gồm hình ảnh, tệp css, v.v. để phát triển giao diện người tiêu dùng.
- Thư mục mẫu: đây là các mẫu cho ứng dụng (bao gồm trang chính, trang để kiểm tra thư và kết quả).
- main.py: Thuật toán học máy để đánh giá chất lượng mô hình.
- app.py: hồ sơ chính để quản lý và thực hiện dự án.

Các bộ dữ liệu để đào tạo mô hình được lấy từ một số nguồn. Chúng được lưu trữ dưới dạng tệp và csv. Hình này hiển thị các ví dụ về bộ dữ liệu:

Name	Date modified	Type	Size
0001.ea7e79d3153e7469e7a9c3e0af6a357e	10/17/2019 10:59 PM	EA7E79D3153E746...	5 KB
0002.b3120c4bcbf3101e661161ee7efcb8bf	10/17/2019 10:59 PM	B3120C4BCBF3101...	4 KB
0003.acfc5ad94bbd27118a0d8685d18c89dd	10/17/2019 10:59 PM	ACFC5AD94BBD2...	4 KB
0004.e8d5727378ddde5c3be181df593f1712	10/17/2019 10:59 PM	E8D5727378DDDE...	4 KB
0005.8c3b9e9c0f3f183ddaf7592a11b99957	10/17/2019 10:59 PM	8C3B9E9C0F3F183...	5 KB
0006.ee8b0dba12856155222be180ba122058	10/17/2019 10:59 PM	EE8B0DBA1285615...	4 KB
0007.c75188382f64b090022fa3b095b020b0	10/17/2019 10:59 PM	C75188382F64B09...	4 KB
0008.20bc0b4ba2d99aae1c7098069f611a9b	10/17/2019 10:59 PM	20BC0B4BA2D99A...	4 KB
0009.435ae292d75abb1ca492dcc2d5cf1570	10/17/2019 10:59 PM	435AE292D75ABB...	4 KB
0010.4996141de3f21e858c22f88231a9f463	10/17/2019 10:59 PM	4996141DE3F21E8...	9 KB
0011.07b11073b53634cff892a7988289a72e	10/17/2019 10:59 PM	07B11073B53634C...	6 KB
0012.d354b2d2f24d1036caf1374dd94f4c94	10/17/2019 10:59 PM	D354B2D2F24D103...	4 KB
0013.ff597adee000d073ae72200b0af00cd1	10/17/2019 10:59 PM	FF597ADEE000D07...	4 KB
0014.532e0a17d0674ba7a9baa7b0afe5fb52	10/17/2019 10:59 PM	532E0A17D0674BA...	6 KB
0015.a9ff8d7550759f6ab62cc200bdf156e7	10/17/2019 10:59 PM	A9FF8D7550759F6...	4 KB

Hình 4. 2 Bộ dữ liệu đào tạo

Ứng dụng được thể hiện bằng một ứng dụng web dựa trên HTML5 và CSS sử dụng Python làm ngôn ngữ lập trình.

4.2. Tạo mô hình và giao diện người dùng

Trước đây, bạn đã chọn phương pháp cây quyết định để tạo ra mô hình. Đầu tiên, bạn cần xử lý trước dữ liệu. Mã mẫu được hiển thị trong hình.

```
if email.get_content_type() == 'text/plain' or email.get_content_type() == 'text/html':
    l = email.get_content()
    l = cleanhtml(l)
    l = l.replace('\n', ' ')
    s = re.sub(r"[^a-zA-Z0-9]+", ' ', l)
    spam_df.append(s.lower())
```

Hình 4. 3 Xử lý trước dữ liệu

Trước khi bạn có thể tạo mô hình lọc thư rác, bạn phải tạo các tính năng bằng mô hình Bag-of-Word, sau đó tạo vector tính năng với TF-IDF:

```
df_list = email3k_df['Content'].to_list()
df_list.extend(email6k_df['email'].to_list())
vectorizer = TfidfVectorizer()
x = vectorizer.fit_transform(df_list)
x = x.toarray()
X = []  # Vector after encode email
for i in x:
    X.append(i.flatten())
Y = email3k_df['Label'].to_list()
Y.extend(email6k_df['label'].to_list())
```

Hình 4. 4 Tạo vector dữ liệu

Sau đó, mô hình cây quyết định đã được tạo ra. Nó được thể hiện trong hình 19.

```
# Decision tree
from sklearn.tree import DecisionTreeClassifier
clf_DT = DecisionTreeClassifier()
clf_DT = clf_DT.fit(X_train,y_train)
```

Biểu đồ 19 – Mô hình cây quyết định

Trong Python, biểu tượng @ đại diện cho các nhà trang trí cho Flask biết URL nào tương ứng với chức năng nào. Trong ví dụ trên, url '/kết quả.html' tương ứng với hàm dự đoán(), vì vậy khi khách hàng gửi yêu cầu, hãy làm theo http path:// <ip_or_domain>:// <port> / kết quả.html (ví dụ: http://127.0. 0.1:5000/result.html), máy chủ web sử dụng trình trang trí đường dẫn và tuyến đường để chuyển yêu cầu cho người xử lý dự đoán thích hợp (), hoạt động như một ứng dụng web.(Biểu đồ 20)

```python
@app.route('/')
def index():
    return render_template('main.html')

@app.route('/test.html')
def test():
    return render_template('test.html')

@app.route('/result', methods=['POST'])
def predict():

    if request.method == 'POST':
        message = request.form['message']
        demo_email = preprocess_input(message)
        my_prediction = clf_DT.predict(demo_email)
    return render_template('result.html', prediction=my_prediction)
```

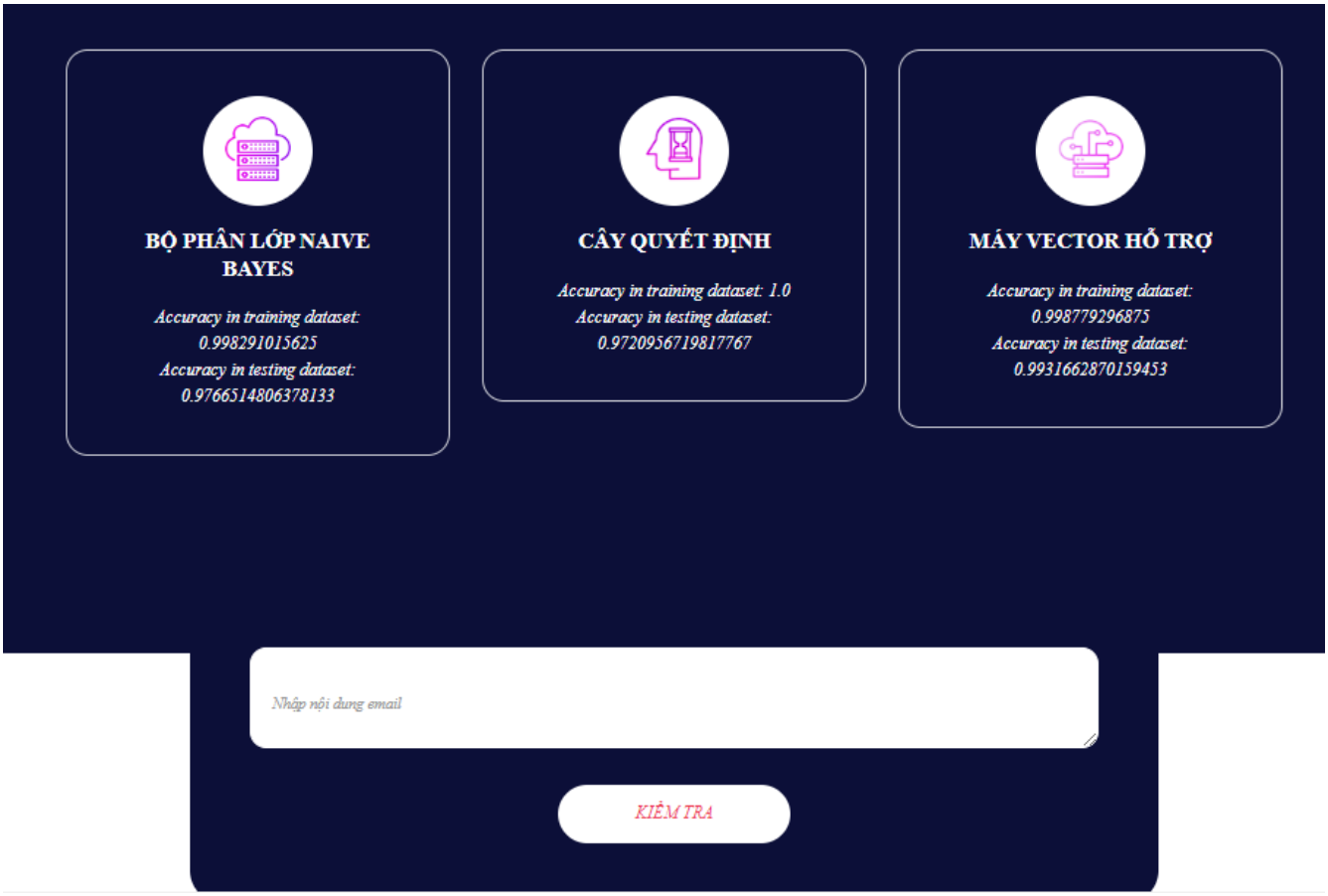
Hình 4. 5 Flask

Kết quả là tạo ruột thừa tung ra xác minh , kết quả của nó trình bày trên Hình 21. Tiếp theo , nó theo sau đi trên địa chỉ http://127.0.0.1:5000/ .

```
Python 3.8.6 (tags/v3.8.6:db45529, Sep 23 2020, 15:52:53) [MSC v.1927 64 bit (AMD64)] on win32
In[2]: runfile('C:/Users/Mmeo/PycharmProjects/SpamEmail/app.py', wdir='C:/Users/Mmeo/PycharmProjects/SpamEmail')
 * Serving Flask app "app" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Hình 4. 6 Kết quả khi ứng dụng

Để thực hiện các giao diện người dùng trước đó mục này đã được sử dụng bởi HTML5 và CSS. Hình hiển thị trang để xác thực thư. Và cũng mô tả phẩm chất của các thuật toán được đào tạo.



Hình 4. 7 Trang lọc thư

Sau khi người dùng nhập dữ liệu mới trong trường và nhấp vào nút "kiểm tra", một yêu cầu sẽ được gửi đến máy chủ. Nhờ mô hình học máy được tạo ra, kết quả kiểm tra sẽ được trả về trên trang kết quả (Hình 23).

Hình 4. 8 Kết quả sau khi lọc

Sử dụng các liên kết đã tạo trong ứng dụng, người dùng nhấp vào nút "lọc thư khác" để

trở lại trang kiểm tra.

4.3. Phân bổ kết quả

Mục tiêu của công việc là tạo ra một mô hình học máy có khả năng phân loại thư rác. Mô hình được xây dựng trên cơ sở thuật toán cây quyết định và được đào tạo bằng cách sử dụng một bộ dữ liệu chứa hơn 5750 tin nhắn.

Ứng dụng web được viết dựa trên HTML5 và CSS sử dụng khung Flask.

Dựa trên kết quả thu được, có thể kết luận rằng mô hình, trong

được sử dụng bởi thuật toán cây quyết định đảm bảo độ chính xác cao của phân loại. Điều này là do mô hình sử dụng đủ dữ liệu.

Kết quả của dự án là một ứng dụng có thể phân loại các tin nhắn đơn giản và đánh dấu chúng là thư rác (hoặc tin nhắn đơn giản).

Kết luận: Trong chương này, một ứng dụng được viết bằng thuật toán cây quyết định, đã có sự phát triển thiết kế và thực hiện các trang web. Một số trang của quá trình thực nghiệm làm việc trên dự án đã được hiển thị. Đồng thời, cũng có đánh giá kết quả thực nghiệm và xác minh một số báo cáo.

PHẦN KẾT LUẬN

Mục đích của công việc này là thiết kế một ứng dụng đơn giản để giải quyết vấn đề lọc thư rác dựa trên học máy. Để hoàn thành công việc, cần phải hiểu được lĩnh vực chủ đề, cụ thể là: có được kiến thức chính về thư rác, bao gồm định nghĩa về thư rác, các loại thư rác, tác hại và các phương pháp phân loại thư rác hiện đang được sử dụng, chẳng hạn như: lọc các thông điệp cảm giác bằng cách đưa ra các luật hạn chế và ngăn chặn e-mail, lọc theo địa chỉ IP, lọc dựa trên một loạt các yêu cầu và phản hồi ... Các phương pháp học máy, phương pháp tạo tập dữ liệu và xử lý trước dữ liệu cũng được nghiên cứu.

Vấn đề chính là bạn cần đào tạo một mô hình lọc thư rác. Đó là, tăng xác suất phân loại. Để giải quyết vấn đề này, một bộ dữ liệu lớn đã được tạo ra, một số thuật toán đã được sử dụng để đào tạo, bao gồm một cây quyết định, một phân loại Bayesian ngây thơ và một phương pháp của những người hàng xóm k gần nhất. Sau khi xác nhận, mô hình tốt nhất để triển khai ứng dụng đã được chọn. Các phương pháp Bag-of-Word và TF-IDF cũng được sử dụng để xử lý trước và cải thiện chất lượng dữ liệu.

Sự liên quan của nghiên cứu này là các hệ thống sử dụng dữ liệu thô bằng cách sử dụng các kỹ thuật học máy để quyết định xem một tin nhắn có phải là thư rác hay không. Hệ thống trình bày các kết quả thử nghiệm có tính đến lọc thư rác trong bộ dữ liệu mẫu và khách hàng.

Trong quá trình thực hiện ứng dụng, các trang đơn giản được viết dựa trên công việc được đề xuất trong phần thứ hai và thứ ba. Điều này cho phép người dùng tương tác thuận tiện với hệ thống. Kết quả của dự án là một ứng dụng đơn giản với chức năng lọc thư rác.

Công việc trong tương lai:

- lưu các tin nhắn mới mà khách hàng kiểm tra để tạo ra một bộ dữ liệu lớn hơn cho việc học;
- Nâng cao độ chính xác của phân loại khi sử dụng người khác
- thuật toán phân loại.
- phát triển ứng dụng để sử dụng thuận tiện.sự phát triển các ứng dụng vì tiện lợi sử dụng .

DANH SÁCH NGUỒN SỬ DỤNG

1. https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF_t%E1%BB%B1_nhi%C3%AAn
2. http://www.machinelearning.ru/wiki/index.php?%20title=Learning_from_precedents
3. https://ru.abcdef.wiki/wiki/Email_spam
4. https://trituenhantao.io/kien-thuc/decision-tree/
5. https://machinelearningcoban.com/2017/08/08/nbc/
6. https://machinelearningcoban.com/2017/04/09/smv/