TRƯỜNG ĐẠI HỌC NGOẠI NGỮ-TIN HỌC TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

**KHÓA LUẬN TỐT NGHIỆP**

**ỨNG DỤNG KHAI THÁC DỮ LIỆU VĂN BẢN TRONG VIỆC PHÂN LOẠI NỘI DUNG WEB**

GIẢNG VIÊN HƯỚNG DẪN: TS. TRẦN MINH THÁI

SINH VIÊN THỰC HIỆN:

La Thị Cẩm Hồng - 20DH110783

Trần Phạm Hà Phương - 20DH111828

**TP. HỒ CHÍ MINH – THÁNG 05 NĂM 2024**

TRƯỜNG ĐẠI HỌC NGOẠI NGỮ-TIN HỌC TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

**KHÓA LUẬN TỐT NGHIỆP**

**ỨNG DỤNG KHAI THÁC DỮ LIỆU VĂN BẢN TRONG VIỆC PHÂN LOẠI NỘI DUNG WEB**

GIẢNG VIÊN HƯỚNG DẪN: TS. TRẦN MINH THÁI

SINH VIÊN THỰC HIỆN:

La Thị Cẩm Hồng - 20DH110783

Trần Phạm Hà Phương - 20DH111828

**TP. HỒ CHÍ MINH – THÁNG 05 NĂM 2024**

# Lời cảm ơn

Chúng em xin gửi lời cảm ơn sâu sắc đến Thầy Trần Minh Thái, người đã tận tình hướng dẫn và động viên trong suốt quá trình thực hiện bài khoá luận này. Sự am hiểu và kiến thức sâu rộng của Thầy đã giúp chúng em vượt qua những khó khăn và hoàn thành công việc một cách xuất sắc.

Chúng em muốn bày tỏ lòng biết ơn đặc biệt đến Ban giám hiệu và các thầy cô trong Khoa Công nghệ Thông tin, Trường Đại học Ngoại ngữ - Tin học TPHCM, đã tạo điều kiện tốt nhất để chúng em có thể thực hiện và hoàn thiện bài khoá luận này. Sự hỗ trợ và sự quan tâm chân thành từ phía quý thầy cô luôn là nguồn động viên lớn lao đối với chúng em.

Cuối cùng, chúng em cũng muốn gửi lời tri ân tới tất cả các giảng viên, đồng nghiệp và bạn bè đã đồng hành và ủng hộ trong suốt thời gian qua. Sự giúp đỡ và động viên từ mọi người đã là động lực lớn giúp chúng em vượt qua mọi khó khăn và hoàn thành bài khoá luận một cách thành công.

Chúng em xin chân thành cảm ơn!

# Lời cam đoan

Chúng em xin cam đoan rằng toàn bộ nội dung trong bài khoá luận với đề tài "Ứng dụng khai thác dữ liệu văn bản trong việc phân loại nội dung trang web" được trình bày dưới đây là thành quả của công sức nghiên cứu và làm việc của chúng em dưới sự hướng dẫn của Thầy TS. Trần Minh Thái. Chúng em cam đoan rằng:

* Mọi thông tin, ý kiến, và kết quả nghiên cứu được trình bày trong bài khoá luận là hoàn toàn chính xác và trung thực.
* Chúng em đã tuân thủ mọi nguyên tắc và quy định về nghiên cứu đạo đức và trí tuệ, không có bất kỳ hành vi vi phạm nào trong quá trình thực hiện bài khoá luận.
* Mọi tài liệu và nguồn trích dẫn từ các tác giả khác đều được ghi rõ và tham khảo đầy đủ theo quy định.
* Bài khoá luận này không được sử dụng cho bất kỳ mục đích gian lận nào, và mọi thông tin và kết quả được trình bày nhằm mục đích nghiên cứu và học tập.

Chúng em xin cam đoan trách nhiệm với những điều trên và sẵn sàng chịu trách nhiệm về bất kỳ nội dung nào trong bài khoá luận này.

# MỤC LỤC

[Lời cảm ơn i](#_Toc168603883)

[Lời cam đoan ii](#_Toc168603884)

[MỤC LỤC iii](#_Toc168603885)

[Danh mục các ký hiệu và chữ viết tắt vi](#_Toc168603886)

[Danh mục bảng biểu vii](#_Toc168603887)

[Danh mục hình ảnh viii](#_Toc168603888)

[Danh mục biểu đồ viii](#_Toc168603889)

[MỞ ĐẦU 1](#_Toc168603890)

[CHƯƠNG 1: TỔNG QUAN 5](#_Toc168603891)

[1.1. Trang web 5](#_Toc168603892)

[1.2. Phân loại trang web và ứng dụng 6](#_Toc168603893)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 8](#_Toc168603894)

[2.1. Tổng quan khai phá dữ liệu văn bản 8](#_Toc168603895)

[2.1.1. Khai phá dữ liệu văn bản 8](#_Toc168603896)

[2.1.2. Các kỹ thuật khai phá dữ liệu 10](#_Toc168603897)

[2.2. Khái niệm phân loại văn bản 11](#_Toc168603898)

[2.3. Quy trình phân loại văn bản 13](#_Toc168603899)

[2.3.1. Tiền xử lý 13](#_Toc168603900)

[2.3.2. Rút trích đặc trưng 13](#_Toc168603901)

[2.3.3. Xây dựng và huấn luyện mô hình phân loại 17](#_Toc168603902)

[2.3.4. Các chỉ số đánh giá mô hình 27](#_Toc168603903)

[2.4. Kết luận 31](#_Toc168603904)

[CHƯƠNG 3: WEB CRAWLING 33](#_Toc168603905)

[3.1. Khái niệm 33](#_Toc168603906)

[3.2. Frontier 38](#_Toc168603907)

[3.3. Fetching 40](#_Toc168603908)

[3.4. Parsing 40](#_Toc168603909)

[3.5. Quá trình lấy và chuẩn hóa các URL 41](#_Toc168603910)

[3.6. Xây dựng cây HTML 43](#_Toc168603911)

[3.7. Spider traps 45](#_Toc168603912)

[3.8. Page Repository 46](#_Toc168603913)

[3.9. Crawler đa luồng (Multi-threaded crawler) 46](#_Toc168603914)

[3.10. Các thuật toán crawling 47](#_Toc168603915)

[3.11. Kết luận 48](#_Toc168603916)

[CHƯƠNG 4: XÂY DỰNG MÔ HÌNH PHÂN LOẠI NỘI DUNG WEBSITE TIẾNG VIỆT 50](#_Toc168603917)

[4.1. Bài toán lấy dữ liệu 50](#_Toc168603918)

[4.1.1. Duyệt URL theo chiều sâu (DFS): 54](#_Toc168603919)

[4.1.2. Lấy dữ liệu từng trang web và gán nhãn 57](#_Toc168603920)

[4.2. Xây dựng và đánh giá các mô hình phân lớp 61](#_Toc168603921)

[4.2.1. Xây dựng tập dữ liệu 63](#_Toc168603922)

[4.2.2. Tiền xử lý 63](#_Toc168603923)

[4.2.3. Vector hoá dữ liệu và trích xuất đặc trưng 66](#_Toc168603924)

[4.2.4. Áp dụng các mô hình phân loại 67](#_Toc168603925)

[4.2.5. So sánh và đánh giá các mô hình 68](#_Toc168603926)

[KẾT LUẬN 80](#_Toc168603927)

[TÀI LIỆU THAM KHẢO 81](#_Toc168603928)

[PHỤ LỤC 89](#_Toc168604021)

[Phụ lục 1: Tập dữ liệu thô (chưa xử lý) được biểu diễn như sau: 89](#_Toc168604022)

[Phụ lục 2: Nhãn tương ứng với thứ tự trong ma trận nhầm lẫn 91](#_Toc168604023)

# Danh mục các ký hiệu và chữ viết tắt

|  |  |
| --- | --- |
| Từ viết tắt | Nội dung |
| AST | Abstract Syntax Tree |
| BFS | Breadth-First Search |
| CSS | Cascading Style Sheets |
| DFS | Depth-First Search |
| FN | False Negative |
| FP | False Positive |
| HTML | Hypertext Markup Language |
| LWP | Libwww-perl |
| MAP | Maximum A Posteriori |
| MLE | Maximum Likelihood Estimation |
| ML | Machine Learning |
| NLP | Natural Language Processing |
| SVM | Support Vector Machine |
| STT | Số thứ tự |
| TFIDF | Term Frequency-Inverse Document Frequency |
| TMDT | Thương mại điện tử |
| TN | True Negative |
| TPR | True Positive Rate |
| TP | True Positive |
| URL | Uniform Resource Locator |
| WWW | World Wide Web |

# Danh mục bảng biểu

[Bảng 2.1. Ưu và khuyết điểm của dữ liệu văn bản 9](#_Toc168603709)

[Bảng 2.2. Tập dữ liệu các từ trong tin nhắn 22](#_Toc168603710)

[Bảng 3.1. Chuẩn hoá biến thể một URL 42](#_Toc168603711)

[Bảng 3.2. Chuẩn hoá URL 43](#_Toc168603712)

[Bảng 4.1. Thể loại các website 52](#_Toc168603713)

[Bảng 4.2. Số lượng links URL từng thể loại 56](#_Toc168603714)

[Bảng 4.3. Các tập URL từng thể loại 59](#_Toc168603715)

[Bảng 4.4. Số lượng nội dung đã lấy 61](#_Toc168603716)

[Bảng 4.5. Các phương pháp gõ dấu 64](#_Toc168603717)

[Bảng 4.6. So sánh các phương pháp gõ dấu 64](#_Toc168603718)

[Bảng 4.7. Tách từ tiếng Việt 65](#_Toc168603719)

[Bảng 4.8. Các mô hình để áp dụng để phân loại văn bản 68](#_Toc168603720)

[Bảng 4.9. Giá trị F1-score 76](#_Toc168603721)

# Danh mục hình ảnh

[Hình 2.1. Mô hình phân loại văn bản tiếng Việt tự động với Machine learning [7] 12](#_Toc168603864)

[Hình 2.2. Quy trình khai phá dữ liệu văn bản [8] 13](#_Toc168603865)

[Hình 2.3. Biểu diễn văn bản [10] 14](#_Toc168603866)

[Hình 2.4. Có vô số đường thẳng phân tách hai lớp dữ liệu 17](#_Toc168603867)

[Hình 2.5. (a) Một mặt phẳng phân chia 2 lớp dữ liệu; (b) biên độ (Margin) 2 lớp phải bằng nhau và lớn nhất 18](#_Toc168603868)

[Hình 2.6. Ma trận nhầm lẫn 29](#_Toc168603869)

[Hình 2.7. Precision và Recall [20] [21] 30](#_Toc168603870)

[Hình 3.1. Mô hình WebCrawler (en.ryte.com/wiki/Crawler, n.d.) 34](#_Toc168603871)

[Hình 3.2. FlowChart WebCrawling Program (Bing, 2011) 37](#_Toc168603872)

[Hình 4.1. Mô hình của hệ thống phân loại trang web 62](#_Toc168603873)

# Danh mục biểu đồ

[Biểu đồ 4.1. Thống kê dữ liệu được sử dụng trên mỗi nhãn 66](#_Toc168603874)

[Biểu đồ 4.2. Ma trận nhầm lẫn của mô hình Multinomial Naive Bayes 69](#_Toc168603875)

[Biểu đồ 4.3. Kết quả kiểm thử với mô hình Multinomial Naive Bayes 70](#_Toc168603876)

[Biểu đồ 4.4. Ma trận nhầm lẫn của mô hình hồi quy logistic 71](#_Toc168603877)

[Biểu đồ 4.5. Kết quả kiểm thử với mô hình hồi quy Logistic 72](#_Toc168603878)

[Biểu đồ 4.6. Ma trận nhầm lẫn của mô hình SVM 73](#_Toc168603879)

[Biểu đồ 4.7. Kết quả kiểm thử với mô hình SVM 74](#_Toc168603880)

[Biểu đồ 4.8. F1-score các mô hình theo từng thể loại 77](#_Toc168603881)

[Biểu đồ 4.9. So sánh thời gian huấn luyện mô hình 78](#_Toc168603882)

# MỞ ĐẦU

Chúng ta đang sống trong thời đại bùng nổ công nghệ thông tin, đặc biệt là sự bùng nổ của mạng internet, các phương thức hoặc giao dịch sử dụng giấy tờ đã dần được số hóa chuyển sang các dạng văn bản lưu trữ trên máy tính, đám mây,…Với các tính năng ưu việt của tài liệu số như các lưu trữ gọn nhẹ, dễ dàng, không tốn quá nhiều thời gian, tiện dụng trao đổi nhiều mục đích, dễ dàng truy xuất, đối soát và sửa đổi… Chính vì thế, các tài liệu, dữ liệu được truyền tải lên mạng internet ngày càng nhiều, tốc độ tăng chóng mặt và chưa thấy dấu hiệu giảm. Internet trở thành kho kiến thức khổng lồ về mọi lĩnh vực, là phương tiện giúp truyền đạt, trao đổi thông tin, giao lưu… giữa mọi người trên toàn cầu, ngoài ra giúp cho nhân loại học tập, nghiên cứu những kiến thức mới một cách dễ dàng hơn.

Các phần mềm, công cụ, hệ thống được sinh ra để hỗ trợ người dùng và trong đó đã sản sinh ra loại hình tài liệu mới. Đó là tài liệu điện tử, tài liệu dưới dạng số hóa. Theo nguồn thông tin từ thư viện quốc gia, tài liệu điện tử là vật mang tin được tạo ra, lưu trữ hay truyền nhận dựa trên nền tảng công nghệ, điện tử, kỹ thuật số… chúng ta có thể đưa ra quan điểm về tài liệu điện tử như sau: “Tài liệu điện tử là tài liệu dưới dạng thông điệp dữ liệu được tạo lập định dạng điện tử hoặc được số hóa từ tài liệu có vật mang tin khác”. Nội dung rất đa dạng và phạm vi phong phú. Trong cuốn sách “Quản lý tài liệu điện tử và công tác lưu trữ” được xuất bản năm 1990 tại Mỹ, giáo sư luật học Henry H. Perritt. Jr, Trưởng Khoa Luật, trường Đại học Luật Chicago - Kent có quan niệm như sau: “Tài liệu điện tử là toàn bộ tài liệu do các cơ quan, tổ chức tạo ra dưới định dạng điện tử, chúng được xem như một hệ thống thông tin điện tử và được hỗ trợ bằng các phương tiện kỹ thuật điện tử”. Năm 2011, tiêu chuẩn quốc tế ISO 15489 đưa ra khái niệm “tài liệu” với cách nhìn nhận và quan niệm mới mẻ như sau: “Tài liệu là thông tin được tạo ra, nhận được và duy trì như là các bằng chứng và thông tin được tạo ra bởi các cá nhân, tổ chức trong quá trình thực hiện nghĩa vụ pháp lý hoặc trong giao dịch công việc”.

Như vậy, tất cả các bản ghi thông tin ở dạng truyền thống hay dạng điện tử đều được gọi chung là tài liệu. Những bản ghi thông tin điện tử có thể được gọi là tài liệu điện tử.

Với sự phát triển của khoa học và công nghệ, nhu cầu của con người về các ứng dụng của công nghệ thông tin vào thực tiễn đời sống cũng tăng dần. Cùng với sự gia tăng về số lượng văn bản và thông tin trên trên internet, nhu cầu tìm kiếm văn bản của con người ngày càng tăng theo. Nhưng với lượng thông tin khổng lồ như vậy thì phải làm thế nào? Có cách nào để giúp con người tổ chức thông tin và tìm kiếm thông tin đạt hiệu quả cao nhất không? Internet và tài liệu điện tử mang đến cho con người nguồn thông tin dồi dào nhưng trong hàng triệu, hàng tỷ nguồn thông tin thì việc lọc trích, tìm kiếm đúng thông tin hữu ích là một thách thức, một bài toán khó với con người.

Do đó, các phương pháp phân loại văn bản tự động được ra đời. Số lượng văn bản đồ sộ như vậy thì việc phân loại văn bản tự động là vô cùng bức thiết. Việc phân loại nội dung văn bản tự động sẽ giúp con người tiết kiệm nhiều thời gian và công sức.

**Mục tiêu nghiên cứu:**

Đề tài "Ứng dụng khai thác dữ liệu văn bản trong việc phân loại nội dung trang web" tập trung vào khám phá các phương pháp phân loại văn bản khác nhau và phát triển một chương trình thử nghiệm để khai thác dữ liệu và phân loại nội dung trang web một cách tự động. Bằng cách áp dụng thuật toán Naive Bayes để xây dựng chương trình dựa trên dữ liệu huấn luyện, mục tiêu là phân loại nội dung văn bản của trang web. Nghiên cứu này sẽ phân tích các vấn đề liên quan và các phương pháp tiếp cận vấn đề phân loại nội dung trang web tiếng Việt như tách từ và phân loại văn bản.

Tuy nhiên, trước khi có thể thực hiện bất kỳ nhiệm vụ nào, máy tính cần hiểu cách ngôn ngữ hoạt động. Điều này được thực hiện thông qua quá trình gọi là học máy. Tuy nhiên, việc dạy cho máy tính một số kiến thức về ngôn ngữ của con người không phải là một công việc đơn giản. Dữ liệu đào tạo lớn hoặc các ví dụ về cách ngôn ngữ được sử dụng trong các ngữ cảnh khác nhau cần được cung cấp. Với kiến thức đó, vấn đề chính mà luận văn này giải quyết là làm thế nào để phân loại nội dung văn bản theo chủ đề.

**Đối tượng nghiên cứu:**

* Trang web: Nghiên cứu tập trung vào việc phân loại nội dung của các trang web.
* Dữ liệu văn bản: Thu thập và phân tích các dữ liệu văn bản từ các trang web để xác định và phân loại chúng.
* Các phương pháp khai thác dữ liệu: Sử dụng các kỹ thuật và công cụ khai thác dữ liệu văn bản để tự động phân loại nội dung trang web.

**Phạm vi nghiên cứu:**

* Phân loại nội dung trang web: Xác định và phân loại các trang web thành các nhóm hoặc danh mục dựa trên nội dung của chúng.
* Ứng dụng của machine learning: Sử dụng các phương pháp và mô hình machine learning để xây dựng các hệ thống phân loại dựa trên dữ liệu văn bản từ các trang web.
* Đánh giá hiệu suất: Đánh giá hiệu suất của các phương pháp phân loại trang web dựa trên các tiêu chí như độ chính xác, độ phủ, và độ nhớ.

**Kết quả nghiên cứu dự kiến:**

Phát triển và triển khai một hệ thống tự động có khả năng phân loại nội dung của các trang web thành các nhóm hoặc danh mục dựa trên dữ liệu văn bản.

Nghiên cứu và áp dụng các phương pháp và mô hình khai thác dữ liệu văn bản, bao gồm machine learning và xử lý ngôn ngữ tự nhiên, để xây dựng các hệ thống phân loại hiệu quả.

Tiến hành đánh giá hiệu suất của hệ thống phân loại dựa trên các tiêu chí như độ chính xác, độ phủ và độ nhớ, để đảm bảo rằng hệ thống hoạt động hiệu quả trong thực tế.

Kết quả nghiên cứu dự kiến sẽ cung cấp một bước tiến quan trọng trong việc tự động hóa quá trình phân loại nội dung trang web và cung cấp giải pháp hiệu quả cho nhiều ứng dụng thực tiễn trên internet.

**Bố cục của khoá luận:**

Nội dung của luận văn tốt nghiệp này được trình bày thành 6 chương. Trong đó chương đầu giới thiệu đề tài, các khái niệm tổng quan. 2 chương tiếp theo trình bày cơ sở lý thuyết và các khái niệm liên quan đến đề tài. Chương kế tiếp xây dựng ứng dụng phân loại nội dung web sử dụng các kỹ thuật, phương pháp học máy, trình bày kết quả và đánh giá. Kế đến là kết luận và cuối cùng là chương tài liệu tham khảo.

*Chương 1: Trình bày tổng quan các kiến thức cơ bản về khai phá dữ liệu, các kiến thức tổng quan liên quan đề tài. Vai trò quan trọng của khai phá dữ liệu trong việc khám phá thông tin ẩn trong dữ liệu. Giới thiệu các phương pháp phổ biến trong khai phá dữ liệu như phân loại, gom cụm, học tập từ dữ liệu, khám phá mẫu, và dự đoán.*

*Chương 2: Cơ sở lý thuyết, giới thiệu các kỹ thuật khai thác, phân tích dữ liệu liên quan đến đề tài. Đề cập đến các kỹ thuật như học máy, học sâu, học tăng cường, và kỹ thuật khai thác thông tin từ dữ liệu không có giám sát.*

*Chương 3: Các khái niệm cơ bản về trình thu thập web (web crawler), các chiến lược thu thập, trình bóc tách,… Từ những khái niệm đưa ra cái nhìn chung về Web Crawling.  
Chương 4: Xây dựng ứng dụng, tiến vào hiện thực chương trình, có được thông tin đầu vào, tiến hành phân tích dữ liệu theo thuật toán, kết quả kiểm thử chương trình.  
Kết Luận: Nhìn lại kết quả đã làm được, đưa ra những đánh giá và hướng phát triển.*

Ngoài ra báo cáo còn có các phần tóm tắt, phụ lục, tài liệu tham khảo và các phần còn lại khác.

# TỔNG QUAN

## Trang web

Theo Kajal Kumari [1], một trang web là một tài liệu hoặc nguồn thông tin có thể truy cập thông qua World Wide Web. Thông thường, nó được tạo thành từ HTML (Hypertext Markup Language), cung cấp cấu trúc và nội dung của trang, và CSS (Cascading Style Sheets), cung cấp thông tin về cách trang web nên được trình bày cho người dùng.

Một trang web có thể chứa văn bản, hình ảnh, video, âm thanh và nội dung đa phương tiện khác. Nó cũng có thể chứa các yếu tố tương tác, như các biểu mẫu, liên kết và nút, cho phép người dùng tương tác với trang và truy cập thông tin hoặc dịch vụ bổ sung.

Trang web bao gồm nhiều chức năng theo từng lĩnh vực, có thể kể đến:

* Phân phối thông tin: Trang web được sử dụng để cung cấp thông tin cho một đối tượng lớn người dùng. Điều này bao gồm thông tin giáo dục, tin tức và nội dung khác dành cho việc thông tin hoặc giáo dục người dùng.
* Thương mại điện tử: Các doanh nghiệp sử dụng trang web để bán các sản phẩm và dịch vụ trực tuyến. Điều này bao gồm cả hàng hóa vật lý, như sách và điện tử, và hàng hóa số, như phần mềm và nhạc.
* Giao tiếp: Trang web được sử dụng cho giao tiếp cá nhân, như email và tin nhắn tức thời, cũng như cho giao tiếp chuyên nghiệp, như hội nghị video và các công cụ hợp tác nhóm.
* Giải trí: Trang web được sử dụng cho giải trí, như xem video và âm nhạc trực tuyến, chơi game và truy cập các trang web mạng xã hội.
* Dịch vụ của chính phủ: Các cơ quan chính phủ sử dụng trang web để cung cấp thông tin và dịch vụ cho công dân, như nộp thuế và gia hạn giấy phép.
* Tìm kiếm việc làm: Người tìm việc sử dụng trang web để tìm cơ hội việc làm, nghiên cứu về nhà tuyển dụng và nộp đơn xin việc trực tuyến.
* Tạo thương hiệu cá nhân: Cá nhân sử dụng trang web để xây dựng thương hiệu cá nhân và thiết lập sự hiện diện trực tuyến của mình.

## Phân loại trang web và ứng dụng

Phân loại trang web là quá trình phân loại các trang web vào các lớp hoặc danh mục đã được xác định trước dựa trên nội dung và cấu trúc của chúng. Điều này có thể hữu ích cho nhiều mục đích, như tổ chức các trang web để tìm kiếm và duyệt web dễ dàng hơn, lọc ra các trang web không liên quan hoặc độc hại và cải thiện độ chính xác của kết quả tìm kiếm trên các công cụ tìm kiếm [1].

Phân loại các trang web có thể được thực hiện bằng các kỹ thuật học máy, như cây quyết định, rừng ngẫu nhiên, máy vector hỗ trợ và mạng nơ-ron. Những thuật toán này lấy làm đầu vào một tập hợp các đặc điểm được trích xuất từ các trang web, như tần suất của các thẻ HTML cụ thể, sự hiện diện của các từ khóa cụ thể trong văn bản, hoặc cấu trúc của các liên kết giữa các trang web. Sau đó, các thuật toán học cách ánh xạ các đặc điểm này thành các nhãn lớp dựa trên một tập dữ liệu huấn luyện của các trang web đã được đánh dấu thủ công với các nhãn lớp đúng của chúng [1].

Để đạt được độ chính xác cao trong việc phân loại trang web, quan trọng là phải có một tập dữ liệu huấn luyện lớn và đa dạng cùng kỹ thuật kỹ thuật đặc trưng hiệu quả để nắm bắt thông tin liên quan về các trang web. Lựa chọn thuật toán học máy cũng sẽ ảnh hưởng đến hiệu suất và nên được lựa chọn dựa trên tính chất của vấn đề và dữ liệu huấn luyện có sẵn.

Phân loại trang web có thể hữu ích cho nhiều mục đích khác nhau, có thể kể đến [1]:

* Tổ chức thông tin: Bằng cách phân loại các trang web vào các lớp khác nhau, việc tổ chức và duyệt qua lượng lớn nội dung trực tuyến trở nên dễ dàng hơn. Điều này có thể giúp người dùng tìm thông tin mình đang tìm kiếm một cách nhanh chóng và dễ dàng hơn.
* Tối ưu hóa công cụ tìm kiếm: Các công cụ tìm kiếm sử dụng phân loại trang web để xếp hạng và sắp xếp kết quả của truy vấn tìm kiếm của người dùng. Bằng cách phân loại trang web một cách chính xác, các công cụ tìm kiếm có thể trả lại kết quả liên quan hơn và cải thiện trải nghiệm của người dùng.
* Lọc nội dung: Phân loại trang web có thể được sử dụng để loại bỏ nội dung độc hại hoặc không liên quan, như thư rác hoặc trang web dành cho người lớn. Điều này có thể giúp bảo vệ người dùng khỏi việc tiếp xúc với tài liệu không phù hợp hoặc có hại và cải thiện chất lượng của nội dung trực tuyến.
* Định tuyến quảng cáo: Phân loại trang web cũng có thể được sử dụng để định tuyến quảng cáo một cách hiệu quả hơn bằng cách xác định chủ đề hoặc nội dung của một trang web và hiển thị quảng cáo phù hợp với nội dung đó
* Cá nhân hóa: Bằng cách phân tích lịch sử duyệt web của người dùng, phân loại trang web có thể được sử dụng để cá nhân hóa trải nghiệm trực tuyến của người dùng bằng cách đề xuất nội dung phù hợp với sở thích của họ.

Tóm lại, việc phân loại trang web là một phần quan trọng trong việc tổ chức, khám phá và phân tích nội dung trực tuyến và có nhiều ứng dụng trong các lĩnh vực như công cụ tìm kiếm, quảng cáo trực tuyến và cá nhân hóa người dùng.

# CƠ SỞ LÝ THUYẾT

## Tổng quan khai phá dữ liệu văn bản

### Khai phá dữ liệu văn bản

Theo IBM [2], khai phá dữ liệu văn bản là quá trình chuyển đổi văn bản không cấu trúc thành định dạng cấu trúc để nhận diện các mẫu có ý nghĩa và thông tin mới. Có thể sử dụng khai phá dữ liệu văn bản để phân tích các bộ sưu tập lớn của tài liệu văn bản để thu thập các khái niệm chính, xu hướng và mối quan hệ ẩn.

Cơ sở dữ liệu văn bản thường được phân thành hai loại chính:

* Dạng không có cấu trúc: Bao gồm các tài liệu văn bản thông thường được sử dụng hàng ngày như sách, báo, trên internet,... Đây là loại dữ liệu thể hiện ngôn ngữ tự nhiên của con người và không tuân theo bất kỳ cấu trúc định sẵn nào. Dữ liệu không có cấu trúc đòi hỏi các phương pháp xử lý đặc biệt để trích xuất thông tin và hiểu nội dung của văn bản.
* Dạng nửa cấu trúc: Bao gồm các văn bản được tổ chức dưới dạng cấu trúc lỏng lẻo nhưng vẫn chứa đựng nội dung chính của văn bản, như văn bản HTML, email,… Dữ liệu này thường có một số cấu trúc cơ bản hoặc siêu cấu trúc, điều này giúp trong việc ánh xạ dữ liệu và trích xuất thông tin hơn so với dạng không có cấu trúc hoàn toàn. Tuy nhiên, việc xử lý cũng đòi hỏi sự linh hoạt vì cấu trúc không cố định và đôi khi không tuân theo các quy tắc rõ ràng [3].

Hầu hết các nghiên cứu trước đây về khai thác dữ liệu đã tập trung vào dữ liệu có cấu trúc, chẳng hạn như dữ liệu quan hệ, giao dịch và kho dữ liệu. Tuy nhiên, một phần đáng kể thông tin có sẵn lại được lưu trữ trong cơ sở dữ liệu văn bản, bao gồm các tập lớn về tài liệu từ các nguồn khác nhau như bài báo, bài nghiên cứu, sách, thư từ điện tử, tin nhắn email và trang web.

|  |  |
| --- | --- |
| Ưu điểm | Khuyết điểm |
| Dữ liệu văn bản có thể truy cập dễ dàng nó có thể nằm trong sách, báo, trang web và các nguồn khác | Dữ liệu văn bản thường không có cấu trúc và dễ phân tích hơn bằng các công cụ chuyên dụng như xử lý ngôn ngữ tự nhiên (NLP) |
| Thu thập và lưu trữ dữ liệu văn bản tương đối rẻ so với hình ảnh và video | Dữ liệu văn bản dễ bị sai lệch do tính chất chủ quan của ngôn ngữ con người |
| Dữ liệu văn bản có thể cung cấp nhiều thông tin về chủ đề cụ thể hoặc nghiên cứu về một lĩnh vực nào đó | Sẽ tốn thời gian để trích xuất những hiểu biết có ý nghĩa từ các bộ sưu tập dữ liệu văn bản phong phú theo cách thủ công |
| Phân tích dữ liệu văn bản có thể giúp khám phá các mô hình, xu hướng và mối tương quan không thể nhìn thấy ngay lập tức | Dữ liệu văn bản có thể chứa sai sót hoặc thiếu chính xác, dẫn đến kết luận không chính xác hoặc sai lệch nếu không được phát hiện và sửa chữa |
| Dữ liệu văn bản có thể được sử dụng để hiểu rõ hơn về dữ liệu tình cảm của các nhóm cụ thể |  |

Bảng 2.1. Ưu và khuyết điểm của dữ liệu văn bản

Cơ sở dữ liệu văn bản đang phát triển nhanh chóng do lượng thông tin ngày càng tăng thông qua hình thức điện tử, chẳng hạn như các xuất bản điện tử, các loại tài liệu điện tử khác nhau, email và World Wide Web. Ngày nay, hầu hết thông tin trong chính phủ, ngành công nghiệp, doanh nghiệp và các tổ chức khác đều được lưu trữ điện tử, dưới dạng cơ sở dữ liệu văn bản. Dữ liệu được lưu trữ trong hầu hết cơ sở dữ liệu văn bản là dữ liệu nửa cấu trúc. Ví dụ, một tài liệu có thể chứa một số trường có cấu trúc như tiêu đề, tác giả, ngày xuất bản và danh mục, nhưng cũng chứa một số thành phần văn bản không cấu trúc lớn như tóm tắt và nội dung. Có rất nhiều nghiên cứu đã được thực hiện về việc mô hình hoá và triển khai dữ liệu nửa cấu trúc trong nghiên cứu cơ sở dữ liệu gần đây. Hơn nữa, các kỹ thuật truy xuất thông tin, như phương pháp lập chỉ mục văn bản, đã được phát triển để xử lý các tài liệu không có cấu trúc [4].

Các kỹ thuật truy xuất thông tin truyền thống trở nên không đủ cho lượng dữ liệu văn bản ngày càng lớn. Thường chỉ một phần nhỏ trong số các tài liệu có sẵn sẽ liên quan đến một người dùng cụ thể. Mà không biết được những gì có thể trong các tài liệu, thì khó khăn để đưa ra các truy vấn hiệu quả để phân tích và trích xuất thông tin hữu ích từ dữ liệu. Người dùng cần những công cụ để so sánh các tài liệu khác nhau, xếp hạng tầm quan trọng và tính liên quan của các tài liệu, hoặc tìm ra mẫu và xu hướng trên nhiều tài liệu. Do đó, khai thác văn bản đã trở thành một chủ đề ngày càng phổ biến và thiết yếu trong lĩnh vực khai thác dữ liệu.

### Các kỹ thuật khai phá dữ liệu

Các kỹ thuật khai phá dữ liệu có thể được phân loại dựa vào phương thức sử dụng học máy, hay còn được gọi là Machine Learning (ML). ML được áp dụng để tạo ra các mô hình và thuật toán phức tạp có khả năng tự dự đoán, đây là quá trình được gọi là phân tích dự đoán trong ngữ cảnh thương mại. Những mô hình này cung cấp khả năng "đưa ra quyết định và kết quả đáng tin cậy, có thể tái sử dụng" và khám phá "thông tin chi tiết tiềm ẩn" bằng cách học từ các mối quan hệ và xu hướng lịch sử trong dữ liệu đầu vào [5].

Triển khai học máy có thể phân loại thành nhiều loại dựa trên tính chất của "tín hiệu" hoặc "phản hồi" có sẵn cho một hệ thống học tập, có thể kể đến bao gồm:

* **Học có giám sát (Supervised learning)**

Học có giám sát là quá trình dự đoán kết quả của dữ liệu mới dựa trên các cặp dữ liệu đã biết trước đó, trong đó mỗi cặp bao gồm dữ liệu và nhãn tương ứng. Đây là một nhóm thuật toán phổ biến nhất trong Machine Learning [6].

Học có giám sát có 2 hình thức học tập dựa theo giá trị đầu ra. Nếu giá trị đầu ra là giá trị rời rạc thì sử dụng hình thức học tập phân loại, còn nếu giá trị đầu ra là một giá trị liên tục thì sử dụng hình thức học tập hồi quy:

* Phân loại: Một vấn đề phân loại là khi biến đầu ra là một danh mục, như "đỏ" hoặc "xanh" hoặc "bệnh" và "không bệnh".
* Hồi quy: Một vấn đề hồi quy là khi biến đầu ra là một giá trị thực, như "đô la" hoặc "cân nặng".
* **Học không giám sát (Unsupervised learning)**

Học không giám sát là một nhóm thuật toán sử dụng dữ liệu không được gán nhãn. Các thuật toán trong phương pháp này nhằm mô hình hóa cấu trúc hoặc thông tin ẩn trong dữ liệu. Nói cách khác, các phương pháp này tập trung vào việc mô tả các đặc điểm hoặc tính chất của dữ liệu. Giải quyết các vấn đề mối quan hệ tương tự giữa các ví dụ trong dữ liệu được xét trong các thuật toán phân cụm, xác suất đồng xuất hiện của các đối tượng được xem xét trong khai phá luật kết hợp, các phép biến đổi ma trận để trích xuất các đặc trưng [5].

## Khái niệm phân loại văn bản

Phân loại văn bản là quá trình tổ chức các tài liệu thành các nhóm dựa trên các tiêu chí nhất định, thường là các nhãn hoặc danh mục, sử dụng các mô hình học máy đã được huấn luyện từ các dữ liệu mẫu. Phương pháp tiếp cận học có giám sát để phân loại văn bản đã trở nên rất phổ biến và được coi là phương pháp tiêu biểu. Đây là quá trình tự động hóa việc học từ các ví dụ và sau đó áp dụng kiến thức đó để phân loại các tài liệu mới. Đầu vào của mô hình là một tập dữ liệu huấn luyện, mỗi mẫu dữ liệu trong đó đã được gán nhãn lớp (hoặc loại). Mỗi mẫu dữ liệu cũng có một tập các thuộc tính đã được xác định. Mục tiêu là xây dựng một mô hình hoặc một bản miêu tả cho mỗi lớp dựa trên các thuộc tính đó. Khi mô hình hoặc miêu tả đã được xác định, chúng được sử dụng để phân loại các mẫu dữ liệu mới vào các nhóm mà không biết trước lớp của chúng.

Hình ảnh dưới đây là cái nhìn tổng quát về cách hoạt động của một bài toán phân loại văn bản [7]:

A diagram of a training

Description automatically generated

Hình 2.1. Mô hình phân loại văn bản tiếng Việt tự động với Machine learning [7]

Giai đoạn (a): Huấn luyện là giai đoạn mà mô hình phân loại văn bản được đào tạo. Trong quá trình này, mô hình học từ dữ liệu đã được gán nhãn (trong hình ảnh, các nhãn là Positive, Negative, Neutral). Dữ liệu văn bản được chuyển đổi thành các vector đa chiều thông qua quá trình trích xuất đặc trưng, biến mỗi mẫu dữ liệu trong tập huấn luyện thành một vector đặc trưng. Các thuật toán máy học sẽ học và điều chỉnh các tham số để đạt được hiệu suất tốt trên tập dữ liệu này. Nhãn của dữ liệu được sử dụng để đánh giá việc mô hình học và tinh chỉnh dựa trên đó.

Giai đoạn (b): Dự đoán là giai đoạn mà mô hình máy học được sử dụng sau khi đã hoàn tất quá trình huấn luyện. Trong giai đoạn này, dữ liệu cần được dự đoán cũng phải trải qua quá trình trích xuất đặc trưng. Mô hình đã học sau đó nhận đặc trưng này làm đầu vào và tạo ra dự đoán tương ứng.

## Quy trình phân loại văn bản

Các bước trong quá trình tổng thể của việc khai thác văn bản được mô tả trong Hình 2.3 [8]:

A diagram of training and training

Description automatically generated with medium confidence

Hình 2.2. Quy trình khai phá dữ liệu văn bản [8]

### Tiền xử lý

Tiền xử lí dữ liệu là một bước quan trọng, thường được coi là việc làm sạch văn bản. Việc này giúp cải thiện khả năng của thuật toán để trích xuất các đặc trưng quan trọng, từ đó tăng cường hiệu suất và chất lượng của các mô hình và thuật toán.

### Rút trích đặc trưng

#### Phương pháp TF-IDF

[9] Đối với các dữ liệu dạng văn bản, ta cần các kĩ thuật trích lọc đặc trưng để biến dữ liệu từ dạng chưa mã hoá sang dạng số học thì mới có thể huấn luyện được mô hình. Hầu hết các mô hình phân loại đều sử dụng cách biểu diễn văn bản sử dụng vector đặc trưng.

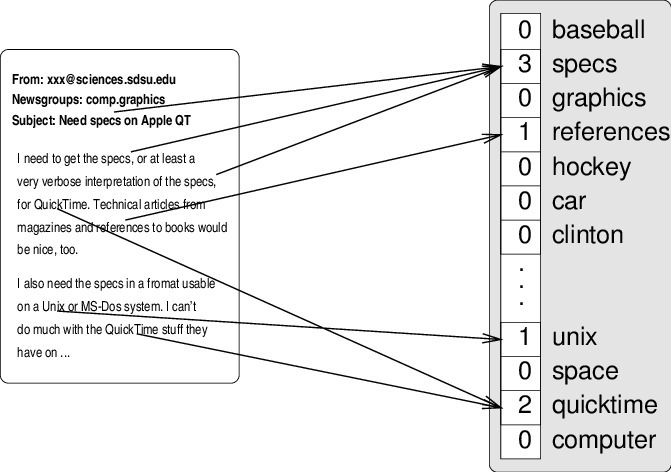
Ý tưởng chính là xem mỗi văn bản tương ứng là một vector đặc trưng

trong không gian các từ ( là một từ, một đặc

trưng, tương ứng một chiều của không gian). Giá trị của chính là số lần xuất

hiện của từ trong văn bản . Từ được chọn là một đặc trưng khi nó xuất hiện

trong ít nhất 3 văn bản. Để không bị phụ thuộc vào chiều dài văn bản vector đặc trưng sẽ được chuẩn hóa về chiều dài đơn vị:



Hình 2.3. Biểu diễn văn bản [10]

Trong thực tế để cải thiện tốc độ và kết quả người ta thường sử dụng hoặc thay cho :

Với:

* + là số văn bản huấn luyện
  + là số văn bản có chứa từ

Các đặc trưng của văn bản khi biểu diễn dưới dạng vector :

* Số chiều không gian đặc trưng thường rất lớn (trên 10000).
* Có các đặc trưng độc lập nhau, sự kết hợp các đặc trưng này thường không có ý nghĩa trong phân loại
* Đặc trưng rời rạc: vector có rất nhiều giá trị 0 do có nhiều đặc trưng không xuất hiện trong văn bản .
* Hầu hết các văn bản có thể được phân chia một cách tuyến tính bằng các hàm tuyến tính.

Việc phân loại sẽ tốt hơn nếu các thuật toán tận dụng được những đặc trưng này.

#### Phương pháp TF-IDF trong việc phân loại văn bản

Khi phân loại văn bản, việc chọn kỹ thuật rút trích đặc trưng phụ thuộc vào nhiều yếu tố, bao gồm kích thước của tập dữ liệu, đặc tính của văn bản, và mục tiêu của bài toán phân loại. Lý do chúng em chọn phương pháp TF-IDF để rút trích đặc trưng thay vì chọn các phương pháp khác bao gồm:

* Đặc tính của dữ liệu:
  + Tính chất của từ hoặc cụm từ quan trọng: TF-IDF đặc biệt hữu ích khi các từ hoặc cụm từ quan trọng trong việc phân loại văn bản xuất hiện ở một số văn bản nhưng không phổ biến trong toàn bộ tập dữ liệu. Khi các từ này xuất hiện, chúng mang thông tin đặc biệt về nội dung của văn bản. TF-IDF giúp tăng trọng số cho những từ này, giúp mô hình phân loại nhận biết được sự quan trọng của chúng trong việc xác định chủ đề hoặc ý nghĩa của văn bản.
  + Số lượng từ vựng lớn: Khi có một lượng lớn từ vựng đa dạng, việc sử dụng các phương pháp như Word Embeddings hoặc BERT có thể gặp khó khăn trong việc xử lý và huấn luyện mô hình. Đặc biệt là khi số lượng từ vựng lớn này tăng cường thêm độ phức tạp cho quá trình tính toán và yêu cầu nhiều tài nguyên tính toán hơn.
* Số lượng lớn dữ liệu cần xử lý: Khi có một lượng lớn dữ liệu, việc sử dụng TF-IDF có thể nhanh chóng và hiệu quả hơn so với việc sử dụng các phương pháp như Word Embeddings hoặc BERT. TF-IDF không đòi hỏi quá nhiều tài nguyên tính toán và có thể được áp dụng một cách hiệu quả trên lượng lớn văn bản.
* Mục tiêu cụ thể của việc phân loại: Nếu mục tiêu của việc phân loại là tập trung vào việc xác định chủ đề hoặc ý nghĩa của văn bản dựa trên từ vựng của nó, thì TF-IDF có thể là lựa chọn phù hợp. TF-IDF tập trung vào sự xuất hiện và tần suất của các từ trong văn bản, giúp nhận biết các từ khóa quan trọng và xác định mức độ quan trọng của chúng trong việc phân loại.

Tóm lại, TF-IDF thường được ưa chuộng khi dữ liệu văn bản có tính chất đặc biệt như mô tả trên và khi việc xử lý một lượng lớn dữ liệu hoặc đòi hỏi tính toán hiệu quả.

#### Các cách tiếp cận TF-IDF

Các cách tiếp cận để sử dụng TF-IDF Vectors như các đặc trưng (features) cho mô hình học máy trong việc phân loại văn bản:

* Cấp độ từ (Word level): Ở cấp độ này, mỗi đặc trưng đại diện cho một từ riêng lẻ trong văn bản. TF-IDF sẽ tính toán tần suất xuất hiện của từ đó trong văn bản và nhân với nghịch đảo của tần suất xuất hiện của từ đó trong toàn bộ tập dữ liệu. Điều này tạo ra một vectơ đặc trưng cho mỗi văn bản, với mỗi thành phần của vectơ là TF-IDF của một từ cụ thể.
* Cấp độ N-gram (N-Gram level): Trong cấp độ này, các đặc trưng được tạo ra từ các cụm từ có kích thước N liên tiếp trong văn bản. Ví dụ, với N=2, các đặc trưng sẽ đại diện cho các cặp từ (bigrams) trong văn bản. Điều này có thể giúp mô hình học máy hiểu được mối quan hệ giữa các từ không chỉ dựa trên từng từ một mà còn dựa trên các cặp từ liên tiếp.
* Cấp độ ký tự (Character level): Ở cấp độ này, mỗi đặc trưng đại diện cho một ký tự hoặc các cụm ký tự trong văn bản. Điều này có thể hữu ích trong việc xử lý các ngôn ngữ không có từ ngữ rõ ràng hoặc trong các trường hợp cần phân loại văn bản dựa trên các đặc điểm về ngữ pháp hoặc ngữ âm.

Việc lựa chọn cấp độ thích hợp phụ thuộc vào bản chất của dữ liệu và mục tiêu của bài toán. Trong một số trường hợp, có thể kết hợp cả ba cấp độ này để tận dụng các thông tin khác nhau từ văn bản.

### Xây dựng và huấn luyện mô hình phân loại

Trong phần này, chúng em sẽ đề cập sơ lược đến các khái niệm về các mô hình và phương pháp được sử dụng để giải quyết các bài toán phân loại văn bản bao gồm: Support Vector Machine, Multinomial Naive Bayes, Logistic Regression.

#### Support Vector Machine

[11] Support Vector Machine - SVM được xem là một trong các thuật toán phân loại phổ biến và hiệu quả. Ý tưởng cơ bản của SVM liên quan đến lý thuyết hình học về cách các điểm được phân bố đến một siêu mặt phẳng trong không gian đa chiều.

A white rectangular object with red and blue squares

Description automatically generated

Hình 2.4. Có vô số đường thẳng phân tách hai lớp dữ liệu

Hai lớp dữ liệu ở Hình 2.5 được giả thiết là có thể tách rời tuyến tính (linearly separable). Khi đó, khoảng cách từ một điểm có tọa độ tới siêu mặt phẳng chiều có phương trình được xác định bởi:

với = , .

A diagram of a line with red dots and blue squares

Description automatically generated

Hình 2.5. (a) Một mặt phẳng phân chia 2 lớp dữ liệu; (b) biên độ (Margin) 2 lớp phải bằng nhau và lớn nhất

Từ Hình 2.6 (a) và 2.6 (b), ta có thể nhận thấy rằng biên rộng hơn sẽ dẫn đến kết quả phân loại tốt hơn, vì sự phân chia sẽ trở nên rõ ràng hơn. Trong SVM, bài toán tối ưu là tìm đường phân chia sao cho biên giữa hai lớp là lớn nhất.

Ưu và khuyết điểm của SVM:

* Ưu điểm:
  + Hiệu suất cao trong không gian chiều cao: SVM hiệu quả khi số chiều dữ liệu lớn hơn số lượng mẫu.
  + Ít yêu cầu về dữ liệu: SVM chỉ cần một phần của dữ liệu để xác định siêu phẳng, nên rất ít bị ảnh hưởng bởi dữ liệu nhiễu.
  + Khả năng làm việc với dữ liệu phi tuyến: Sử dụng các hàm nhân, SVM có thể phân loại dữ liệu không phân chia tuyến tính.
  + Khả năng xử lý overfitting: SVM có khả năng điều chỉnh độ phức tạp của mô hình bằng cách chọn tham số C và kernel.
* Khuyết điểm:
  + Phức tạp tính toán: Việc huấn luyện một mô hình SVM có thể tốn nhiều thời gian khi kích thước của dữ liệu lớn.
  + Nhạy cảm với lựa chọn kernel và tham số: Việc chọn kernel và tham số tối ưu có thể ảnh hưởng đến hiệu suất của mô hình.
  + Khó khăn trong việc xử lý dữ liệu lớn: SVM không hiệu quả khi dữ liệu lớn và không thể lưu trữ dữ liệu training trong bộ nhớ.
  + Không thích hợp cho dữ liệu không cân bằng: SVM có thể không hiệu quả khi số lượng mẫu của các lớp không cân bằng.
  + Khó hiểu và khó giải thích: Siêu phẳng phân loại tạo ra bởi SVM có thể khó giải thích trong một số trường hợp.

Tóm lại, mặc dù SVM là một phương pháp phân loại mạnh mẽ, nhưng cũng cần một số điều chỉnh thích hợp của siêu tham số để đạt được hiệu suất tốt nhất trên dữ liệu cụ thể. Đồng thời, việc tiền xử lý dữ liệu cẩn thận cũng là một yếu tố quan trọng đối với việc áp dụng SVM trong phân loại văn bản.

#### Naive Bayes

[12] Phương pháp Naive Bayes là nhóm các thuật toán học có giám sát dựa trên việc áp dụng định lý Bayes với giả định "ngây thơ" về sự độc lập có điều kiện giữa mỗi cặp đặc trưng đối với giá trị của biến lớp. Định lý Bayes mô tả mối quan hệ giữa biến lớp đã biết y và vectơ đặc trưng phụ thuộc đến theo công thức:

Trong đó:

* + là vector đầu vào, n là số từ trong từ điển
  + : xác suất để đầu vào X rơi vào lớp y
  + : xác suất một điểm bất kỳ thuộc vào lớp y
  + : phân phối các điểm dữ liệu trong lớp y

Sử dụng giả định độc lập có điều kiện “ngây thơ” cho rằng:

Đối với tất cả , mối quan hệ này được đơn giản hóa thành:

Từ là hằng số với đầu vào, chúng ta có thể sử dụng quy tắc phân loại sau:

Chúng ta có thể sử dụng ước lượng xác suất lớn nhất (MLE: Maximum Likelihood Estimation) hoặc ước lượng tối đa Posteriori (MAP: Maximum A Posteriori) để ước lượng và ; sau đó, ta có thể tính toán tần suất tương đối của lớp y trong tập huấn luyện. Các bộ phân loại Naive Bayes khác nhau chủ yếu khác nhau bởi các giả định liên quan đến phân phối .

Mặc dù được đơn giản hóa trong giả định, các bộ phân loại Naive Bayes đã cho thấy hiệu suất tốt trong nhiều tình huống thực tế, như phân loại tài liệu và lọc thư rác. Naive Bayes yêu cầu một lượng nhỏ dữ liệu huấn luyện để ước tính các tham số cần thiết. Nó cũng học và phân loại nhanh chóng so với các phương pháp phức tạp khác. Việc tách và phân bố các đặc trưng có điều kiện của lớp giúp giảm các vấn đề của dữ liệu đa chiều, vì mỗi phân bố có thể được ước tính độc lập như phân bố một chiều.

Tuy nhiên, mặc dù Naive Bayes được công nhận là một công cụ phân loại tốt, nó vẫn được coi là một công cụ ước lượng tồi vì kết quả xác suất dự đoán không được đánh giá cao (do giảm số chiều của dữ liệu). Việc tính toán phụ thuộc vào loại dữ liệu, Các dạng phổ biến của thuật toán Naive Bayes có thể kể đến:

* Multinomial Naive Bayes: Phù hợp cho dữ liệu đếm như số lần xuất hiện của từ trong văn bản.
* Bernoulli Naive Bayes: Sử dụng cho dữ liệu nhị phân, nghĩa là từ có xuất hiện trong văn bản hay không.
* Gaussian Naive Bayes: Sử dụng cho dữ liệu liên tục, không phổ biến trong phân loại văn bản.

Chúng em sử dụng dạng Multinomial Naive Bayestrong đề tài này. [13] Trong Multinomial Naive Bayes, văn bản được biểu diễn dưới dạng túi từ (bag of words), trong đó mỗi từ được coi là một đặc trưng và tần suất xuất hiện của từ đó trong văn bản là giá trị của đặc trưng đó. Mỗi văn bản sẽ được biểu diễn dưới dạng vector đặc trưng với các giá trị tần suất của các từ.

Khi huấn luyện, Multinomial Naive Bayes sử dụng tần suất của các từ trong từ điển để ước lượng xác suất của mỗi lớp (phân loại). Khi dự đoán, nó sử dụng nguyên lý Bayes để tính toán xác suất của mỗi lớp dựa trên vector đặc trưng của văn bản và chọn lớp có xác suất cao nhất làm dự đoán cho văn bản đó.

Phân phối được tham số hóa bởi vectơ cho mỗi lớp , trong đó là số đặc trưng và là xác suất của đặc trưng xuất hiện trong một mẫu thuộc lớp . Các tham số được ước tính bằng một phiên bản làm mịn của xác suất tối đa (Maximum Likelihoods), tức là đếm tần suất tương đối theo công thức:

Trong đó:

* : là số lần đặc trưng xuất hiện trong một mẫu thuộc lớp trong tập huấn luyện
* : là tổng số lần xuất hiện của tất cả các đặc trưng trong lớp
* : là thông số làm mịn (), được sử dụng để điều chỉnh cho các đặc trưng không xuất hiện trong các mẫu huấn luyện và để tránh xác suất bằng không trong các tính toán tiếp theo. Khi , phương pháp này được gọi là làm mịn Laplace, còn khi , nó được gọi là làm mịn Lidstone.

Chứng minh công thức trên bằng ví dụ sau, cho tập dữ liệu về các từ xuất hiện trong một tin nhắn bình thường (B) và tin nhắn spam (S) như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Từ** | Thân | Bạn | Ăn | Tiền |
| **Số lần xuất hiện trong tin nhắn bình thường** | 8 | 5 | 3 | 1 |
| **Số lần xuất hiện trong tin nhắn spam** | 2 | 1 | 0 | 4 |

Bảng 2.2. Tập dữ liệu các từ trong tin nhắn

Phân loại đoạn tin nhắn “Ăn tiền tiền tiền tiền”, giả sử tập dữ liệu huấn luyện có 8 tin nhắn bình thường, 4 tin nhắn spam.

Đầu tiên, tính xác suất xuất hiện của mỗi từ trong 2 trường hợp tin nhắn “bình thường” và “spam” theo công thức (1):

|  |  |
| --- | --- |
| *;* | *;* |

Nhận xét: Có thể thấy từ “tiền” xuất hiện tới 4 lần và > nên kết quả có lẽ sẽ được dự đoán vào lớp “S”.

Tiếp tục, vì tập dữ liệu huấn luyện có 8 tin nhắn bình thường và 4 tin nhắn spam nên xác suất rơi vào mỗi trường hợp như sau:

|  |  |
| --- | --- |
|  |  |

Áp dụng quy tắc phân loại (), ta có:

|  |
| --- |
|  |
|  |

Nhận xét: Khác với dự đoán ban đầu, kết quả lại được dự đoán vào lớp “B”, như vậy có thể thấy dù từ “Tiền” xuất hiện bao nhiêu lần đi nữa thì nếu đi chung từ “Ăn” (xuất hiện 0 lần trong “S”) thì kết quả phân loại sẽ luôn là “N”. Để giải quyết vấn đề này người ta thường tăng số lượng mỗi từ lên 1, hay còn gọi là làm mịn Laplace.

Tính xác suất xuất hiện của mỗi từ trong 2 trường hợp tin nhắn “bình thường” và “spam” theo công thức (4), với :

|  |  |
| --- | --- |
| *;* | *;* |

Việc tăng số lượng mỗi từ lên 1 sẽ không làm thay đổi số lượng tin nhắn mỗi lớp, áp dụng quy tắc phân loại (3), ta có:

|  |
| --- |
|  |
|  |

Suy ra ta phân loại đoạn tin nhắn “Ăn tiền tiền tiền tiền” là tin nhắn spam.

Thuật toán Multinomial Naive Bayes giả định rằng các đặc trưng (từ) là độc lập với nhau, điều này có nghĩa là sự xuất hiện của một từ không phụ thuộc vào sự xuất hiện của các từ khác trong văn bản. Mặc dù giả định này không hoàn toàn chính xác trong thực tế, nhưng nó vẫn mang lại kết quả tốt trong nhiều trường hợp và làm giảm đáng kể độ phức tạp tính toán. Phương pháp Multinomial Naive Bayes giúp cải thiện bài toán phân loại văn bản khi mà tập dữ liệu không cân đối giữa các lớp [13].

Các ưu và khuyết điểm của Multinomial Naive Bayes:

* Ưu điểm
* Multinomial Naive Bayes rất đơn giản để triển khai và nhanh chóng khi huấn luyện và dự đoán. Điều này làm cho nó trở thành một lựa chọn lý tưởng cho các bài toán yêu cầu tốc độ xử lý cao.
* Thuật toán này đặc biệt hiệu quả trong việc xử lý dữ liệu văn bản, nơi mà các đặc trưng thường được biểu diễn dưới dạng số lượng (count) hoặc tần suất (frequency) của từ.
* Do tính chất đơn giản của mô hình, Multinomial Naive Bayes thường yêu cầu ít bộ nhớ hơn so với nhiều thuật toán phức tạp khác.
* Mô hình có thể xử lý tốt các bộ dữ liệu lớn, đặc biệt là khi kết hợp với các kỹ thuật tiền xử lý như TF-IDF để giảm kích thước của không gian đặc trưng.
* Multinomial Naive Bayes hoạt động tốt cho cả phân loại nhị phân và phân loại đa lớp.
* Khuyết điểm
* Một trong những nhược điểm lớn nhất của Naive Bayes là giả định rằng các đặc trưng độc lập với nhau, điều này hiếm khi đúng trong thực tế. Giả định này có thể dẫn đến hiệu suất kém nếu các đặc trưng thực sự phụ thuộc vào nhau.
* Mô hình có thể gặp khó khăn khi xử lý các đặc trưng hiếm gặp trong dữ liệu vì nó có thể không được mô hình hóa đúng cách.
* Khi số lượng đặc trưng tăng lên rất lớn, mô hình có thể trở nên kém hiệu quả và khó khăn trong việc duy trì độ chính xác cao.
* Multinomial Naive Bayes hoạt động tốt nhất với dữ liệu tuân theo phân phối đa biến (multinomial distribution). Đối với các dạng dữ liệu khác, hiệu suất có thể giảm.

Multinomial Naive Bayes là một thuật toán phân loại mạnh mẽ và hiệu quả, đặc biệt hữu ích cho các bài toán liên quan đến xử lý ngôn ngữ tự nhiên và phân loại văn bản. Tuy nhiên, nó cũng có một số hạn chế, đặc biệt liên quan đến giả định độc lập của các đặc trưng và khả năng xử lý các dữ liệu không gian cao. Việc hiểu rõ các ưu và khuyết điểm này giúp áp dụng mô hình một cách hiệu quả và biết khi nào cần xem xét các phương pháp thay thế khác.

#### Hồi quy logistic

Một trong những phương pháp phân loại sớm nhất là hồi quy logistic. Hồi quy logistic được giới thiệu và phát triển bởi nhà thống kê David Cox vào năm 1958 [14]. Hồi quy logistic là một bộ phân loại tuyến tính với ranh giới quyết định là θTx = 0. Hồi quy logistic dự đoán xác suất thay vì các lớp [15, 16].

Mục tiêu của hồi quy logistic là huấn luyện từ xác suất của biến là 0 hoặc 1 khi biết . Giả sử chúng ta có dữ liệu văn bản là . Nếu chúng ta có các bài toán phân loại nhị phân, hàm của mô hình hỗn hợp Bernoulli nên được sử dụng [17] như sau:

Trong đó và là một hàm sigmoid được định nghĩa như trong phương trình dưới đây:

Trường hợp nhị phân có thể được mở rộng thành các lớp dẫn đến hồi quy logistic đa thức [12]. Phân loại logistic đa thức (hoặc đa nhãn) [18] sử dụng xác suất của thuộc về lớp , được định nghĩa trong phương trình:

Trong đó là vector trọng số tương ứng với lớp .

Đối với phân loại nhị phân (m = 2), đây được gọi là hồi quy logistic cơ bản, nhưng đối với hồi quy logistic đa thức (m > 2), thường sử dụng hàm softmax:

Trong một nhiệm vụ phân loại trong bối cảnh học có giám sát, thành phần của được tính từ tập con của dữ liệu huấn luyện thuộc về lớp , với . Để thực hiện ước lượng xác suất tối đa (Maximum Likelihoods) của , chúng ta cần tối đa hóa hàm log-likelihood như sau:

Áp dụng ước lượng tối đa hậu nghiệm (Maximum A Posteriori) như sau:

Bộ phân loại hồi quy logistic hoạt động tốt trong việc dự đoán các kết quả phân loại. Tuy nhiên, dự đoán này yêu cầu mỗi điểm dữ liệu phải độc lập [19], trong đó nó cố gắng dự đoán các kết quả dựa trên một tập hợp các biến độc lập.

Ưu điểm và khuyết điểm của hồi quy logistic như sau:

* Ưu điểm:
  + Khả năng mô hình hóa xác suất: Logistic Regression không chỉ dự đoán lớp của một điểm dữ liệu mà còn cung cấp xác suất rơi vào mỗi lớp.
  + Ít ảnh hưởng bởi nhiễu: Logistic Regression ít bị ảnh hưởng bởi dữ liệu nhiễu hơn so với các mô hình phân loại khác như Decision Trees.
  + Khả năng mở rộng: Logistic Regression có thể được mở rộng để xử lý các trường hợp có nhiều hơn hai lớp bằng cách sử dụng các biến độc lập phức tạp hơn.
* Khuyết điểm:
  + Giả định tuyến tính: Logistic Regression giả định rằng mối quan hệ giữa các biến độc lập và biến phụ thuộc là tuyến tính, điều này không phù hợp cho mọi tình huống.
  + Nhạy cảm với biến nhiễu: Logistic Regression có thể không hoạt động tốt nếu có quá nhiều nhiễu hoặc dữ liệu không đủ.
  + Khó xử lý các tập dữ liệu lớn với nhiều biến: Khi có quá nhiều biến độc lập, việc tính toán và đánh giá mô hình Logistic Regression có thể trở nên phức tạp và tốn kém.

### Các chỉ số đánh giá mô hình

Khi ứng dụng các mô hình phân loại, việc đánh giá hiệu suất của chúng là vô cùng quan trọng để đảm bảo tính chính xác và đáng tin cậy của các dự đoán.

Khi thực hiện dự đoán phân loại, có bốn kết quả chính có thể xảy ra:

* True Positive (TP): Dự đoán một quan sát thuộc về một lớp và quan sát đó thực sự thuộc về lớp đó.
* True Negative (TN): Dự đoán một quan sát không thuộc về một lớp và quan sát đó thực sự không thuộc lớp đó.
* False Positive (FP): Dự đoán một quan sát thuộc về một lớp trong khi thực tế không phải vậy.
* False Negative (FN): Dự đoán một quan sát không thuộc về một lớp trong khi thực tế là thuộc về lớp đó.

Các chỉ số để đánh giá mô hình phân loại bao gồm: accuracy, precision và recall, F1-score,..

* **Accuracy**

Accuracy là một phép đo để đánh giá hiệu suất của các mô hình phân loại. Nó đơn giản là tỷ lệ giữa số lần dự đoán chính xác và tổng số lần dự đoán.

Tỷ lệ True Positive (TPR) được tính bằng phần trăm của số lượng trường hợp tích cực được dự đoán chính xác so với tổng số trường hợp tích cực thực tế. Trong công thức này, mẫu số (TP + FN) biểu thị số lượng trường hợp thực tế dương có trong tập dữ liệu.

* **Ma trận nhầm lẫn**

Cách tính sử dụng accuracy như ở trên chỉ cho chúng ta biết được bao nhiêu phần trăm lượng dữ liệu được phân loại đúng mà không chỉ ra được cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất, và dữ liệu thuộc lớp nào thường bị phân loại nhầm vào lớp khác. Để có thể đánh giá được các giá trị này, chúng ta sử dụng một ma trận được gọi là ma trận nhầm lẫn.

Về cơ bản, confusion matrix thể hiện có bao nhiêu điểm dữ liệu thực sự thuộc vào một class, và được dự đoán là rơi vào một class:

**A diagram of a positive values

Description automatically generated**

Hình 2.6. Ma trận nhầm lẫn

* **Precsion và Recall**

Precision (hay còn gọi là positive predictive value) đo lường tỷ lệ của các trường hợp dự đoán chính xác trong số các trường hợp được dự đoán là positive, trong khi Recall (hay còn gọi là sensitivity) đo lường tỷ lệ của các trường hợp đã được dự đoán chính xác trong số tất cả các trường hợp positive thực sự. Vì vậy, cả Precision và Recall đều tập trung vào việc đánh giá độ chính xác của các dự đoán positive.

A diagram of different colors and shapes

Description automatically generated

Hình 2.7. Precision và Recall [20] [21]

Trong phân loại, Precision là tỷ lệ số lượng trường hợp thực sự tích cực (TP) so với tổng số trường hợp được gắn nhãn là thuộc lớp tích cực.

Recall là tỷ lệ số lượng các trường hợp dương thực sự so với tổng số lượng trường hợp dương thực sự thuộc vào lớp tích cực.

* **F1-score**

F1 Score là trung bình điều hòa giữa độ chuẩn xác và độ phủ. Do đó nó là chỉ số đại diện trong việc đánh giá tỷ lệ dự báo đúng của các trường hợp mẫu dương tính.

Tóm lại các chỉ số accuracy, precision, recall, F1-score,... là những thước đo quan trọng để đánh giá hiệu suất của mô hình phân loại. Mỗi chỉ số có vai trò riêng và giúp đánh giá các khía cạnh khác nhau của mô hình

Việc chọn chỉ số nào phụ thuộc vào mục tiêu cụ thể của bài toán. Nếu muốn tránh dự đoán sai dương tính, precision là quan trọng hơn. Nếu cần đảm bảo phát hiện đầy đủ các mẫu dương tính, recall là chỉ số cần được chú trọng hơn. F1-score là thước đo cân bằng khi cần xem xét cả hai yếu tố này. Khi dữ liệu bị mất cân bằng nghiêm trọng, việc sử dụng độ chính xác (accuracy) để đánh giá mô hình thường không hiệu quả vì mô hình có thể đạt độ chính xác rất cao một cách giả tạo. Ví dụ, một mô hình đơn giản dự đoán tất cả các mẫu thuộc nhóm đa số cũng có thể đạt độ chính xác gần 100%. Trong tình huống này, các chỉ số thay thế như precision, recall, và F1-score có thể được sử dụng. Những chỉ số này không dễ dẫn đến ngộ nhận về độ chính xác, vì chúng tập trung vào việc đánh giá hiệu suất của mô hình đối với nhóm thiểu số, nhóm mà chúng ta muốn dự đoán chính xác hơn so với nhóm đa số.

Một mô hình có các chỉ số trên đều cao thì mô hình có chất lượng dự báo càng tốt.

## Kết luận

Trong chương này, chúng em đã trình bày khá chi tiết về lĩnh vực khai phá dữ liệu văn bản, nắm bắt được các khái niệm cơ bản của phân loại văn bản cùng với quy trình thực hiện nó. Dưới đây là một tổng kết của những điểm chính đã được thảo luận:

* Tổng quan khai phá dữ liệu văn bản: hiểu rõ về dữ liệu văn bản và tầm quan trọng của việc khai phá dữ liệu văn bản trong thế giới kỹ thuật số ngày nay. Các ứng dụng của việc khai phá dữ liệu văn bản đã được phổ biến rộng rãi từ phân tích cảm xúc đến tóm tắt tự động và gợi ý nội dung.
* Khái niệm phân loại văn bản: Một phương pháp học máy phổ biến để tự động phân loại các tài liệu văn bản vào các danh mục cụ thể dựa trên nội dung của chúng. Các thuật toán phân loại, bao gồm Naive Bayes, Logistic Regression, và Support Vector Machines (SVM), đã được giới thiệu.
* Quy trình phân loại văn bản: Xem xét quy trình phân loại văn bản từ việc chuẩn bị dữ liệu, trích xuất đặc trưng, huấn luyện mô hình, đánh giá và tinh chỉnh mô hình. Mỗi bước trong quy trình này đóng vai trò quan trọng trong việc đảm bảo hiệu suất tốt của hệ thống phân loại.

Tóm lại, chương này đã cung cấp một cái nhìn tổng quan về lý thuyết và phương pháp cơ bản liên quan đến phân loại văn bản. Kiến thức này sẽ là nền tảng cho việc triển khai và tối ưu hóa các hệ thống phân loại văn bản trong các ứng dụng thực tế.

# WEB CRAWLING

## Khái niệm

Web Crawler, hay còn được gọi là spider hoặc robot, là một loại chương trình máy tính tự động được thiết kế để tự động tải xuống các trang web trên Internet. Chức năng chính là thu tập thông tin từ các trang web bằng cách theo dõi và tải về các liên kết trên các trang web. Các thông tin được thu thập sẽ được sử dụng trong nhiều mục đích khác nhau như là chỉ mục các công cụ tìm kiếm, thống kê, phân tích thị trường, nghiên cứu thị trường…

*Định nghĩa: Web crawler là các chương trình khai thác sơ đồ cấu trúc của Web bằng cách chuyển từ trang web này sang trang web khác.* [22]

A diagram of a cloud

Description automatically generated

Hình 3.1. Mô hình WebCrawler (en.ryte.com/wiki/Crawler, n.d.)

Một trình thu thập web có thể truy cập nhiều trang web để thu thập thông tin, được phân tích và khai thác tại một vị trí trung tâm, trực tuyến (khi được tải xuống) hoặc ngoại tuyến (sau khi được lưu trữ). Tuy nhiên, Web là một thực thể động tiến triển với tốc độ nhanh chóng. Do đó, có một nhu cầu liên tục cho các trình thu thập để giúp các ứng dụng duy trì tính hiện đại khi các trang và liên kết được thêm mới, xóa bỏ, di chuyển hoặc sửa đổi.

Hình trên mô tả một quy trình tổ chức và xử lý dữ liệu từ việc duyệt web. Mỗi phần trong quy trình có vai trò cụ thể và đóng góp vào việc thu thập và xử lý dữ liệu từ các trang web khác nhau. Dưới đây là ý nghĩa của từng phần:

linksToCrawl: Đây là nơi chứa các liên kết web cần được duyệt. Nó giữ quản lý các URL mà hệ thống sẽ truy cập để thu thập dữ liệu.

Fetcher: Fetcher trích xuất thông tin từ các trang web đã tải, bao gồm cả các liên kết khác và nội dung của trang web. Điều này giúp tạo ra một nguồn dữ liệu sạch và đầy đủ cho quy trình tiếp theo.

linkExtractor: linkExtractor có trách nhiệm lọc ra các liên kết quan trọng từ dữ liệu đã tải và chuyển chúng đến linkFilter. Điều này giúp tối ưu hóa việc tìm kiếm và thu thập dữ liệu từ các nguồn trực tuyến.

linkFilter: linkFilter loại bỏ các liên kết không cần thiết hoặc không mong muốn, giúp tăng hiệu suất của quá trình duyệt web bằng cách chỉ chọn những liên kết có giá trị.

VisitedLinks: VisitedLinks giúp tránh việc lặp lại việc duyệt các trang web đã được duyệt trước đó, tối ưu hóa thời gian và tài nguyên.

RecordProducer: RecordProducer xử lý dữ liệu trước khi lưu trữ vào Data Store, đảm bảo rằng chỉ dữ liệu có giá trị và đáng tin cậy mới được lưu trữ.

Data Store: Data Store là nơi lưu trữ toàn bộ dữ liệu đã thu thập từ việc duyệt web. Nó là nguồn tài nguyên quan trọng cho việc phân tích và sử dụng dữ liệu thu thập được từ internet.

Quy trình trích xuất và xử lý dữ liệu từ việc duyệt web là một quy trình phức tạp và có ý nghĩa quan trọng trong việc thu thập thông tin từ internet. Các phần trong quy trình này đóng vai trò quan trọng trong việc thu thập, lọc và lưu trữ dữ liệu một cách hiệu quả. Từ việc xác định các liên kết cần duyệt đến việc loại bỏ các liên kết không mong muốn và lưu trữ dữ liệu thu thập được, mỗi bước đều đóng góp vào việc tạo ra một nguồn dữ liệu sạch và có ích. Quy trình này cung cấp cơ sở dữ liệu cần thiết cho việc phân tích và tìm hiểu thông tin từ internet, hỗ trợ các ứng dụng trong nhiều lĩnh vực như trí tuệ nhân tạo, khoa học dữ liệu và tìm kiếm thông tin.

Nếu môi trường Web là một tập các trang web tĩnh cố định, thì chúng ta sẽ ít khi phải sử dụng các chương trình Crawler. Một khi các trang web đã được lưu vào kho chứa (ví dụ như một cơ sở dữ liệu của hệ thống tìm kiếm), ta sẽ chẳng còn lý do nào để sử dụng modul crawler. Tuy nhiên, môi trường Web là một thực thể động, với các không gian con thay đổi theo các xu hướng khác nhau và thường là với tốc độ rất nhanh. Do đó chúng ta luôn cần sử dụng các crawler để giúp các ứng dụng được cập nhật bằng cách cập nhật nội dung mới của các trang web, xóa bỏ hoặc sửa đổi nội dung cũ. Các hệ thống tìm kiếm thường cố gắng thu thập được càng nhiều trang web càng tốt. Các hệ thống này thường sử dụng Web crawler để bảo trì cơ sở dữ liệu được đánh chỉ mục của chúng, cân bằng cái giá của quá trình crawling và đánh chỉ mục với hàng triệu truy vấn mà hệ thống nhận được. Module crawler của các hệ thống này thường có xu hướng và mục tiêu chính là download hết các trang web mà nó gặp. Ngược lại, các crawler khác lại chỉ chọn một số trang web để tải và duyệt trong số rất nhiều các trang web nó gặp, các crawler này được gọi là các crawler có lựa chọn preferential crawler hoặc crawler dựa trên kinh nghiệm. Chúng được sử dụng để xây dựng các kho dữ liệu có chủ điểm, tự động hóa các nguồn lực khai phá và đáp ứng cho các đại lý phần mềm. Các crawler có lựa chọn được xây dựng để lấy ra các trang web theo một chủ đề xác định được gọi là các crawler theo chủ đề topic crawler hoặc crawler tập trung focused crawler.

Cũng có các ứng dụng độc hại của các trình thu thập, ví dụ như thu thập địa chỉ email để được sử dụng bởi các spammer hoặc thu thập thông tin cá nhân để sử dụng trong các cuộc tấn công phishing và đánh cắp danh tính khác.

A diagram of a computer

Description automatically generated

Hình 3.2. FlowChart WebCrawling Program (Bing, 2011)

Các trình thu thập web cơ bản hoạt động bằng cách tải xuống từng trang một lần, điều này thường dẫn đến sử dụng tài nguyên không hiệu quả. Trung tâm hoạt động của một crawler là việc duy trì một danh sách các URL chưa được truy cập, được gọi là “frontier”. Danh sách này được khởi tạo với các URLs seed, có thể do người dùng hoặc một chương trình khác cung cấp. Trong mỗi lần lặp của vòng lặp chính, crawler chọn URL tiếp theo từ “frontier”, truy cập trang tương ứng thông qua HTTP, phân tích cú pháp của trang để trích xuất các URL, thêm các URL mới phát hiện được vào “frontier”, và lưu trữ trang (hoặc thông tin khác được trích xuất) trong một kho dữ liệu cục bộ.

Quá trình thu thập có thể kết thúc sau khi một số lượng nhất định các trang đã được thu thập, hoặc nếu “frontier” rỗng. Tuy nhiên, cần lưu ý rằng trường hợp sau hiếm khi xảy ra trong thực tế do số lượng liên kết thường có trên các trang web.

Về bản chất, một web crawler hoạt động như một thuật toán tìm kiếm đồ thị. Web có thể được coi như một đồ thị lớn, với từng trang là các nút và các liên kết siêu văn bản là các cạnh. Crawler bắt đầu hoạt động từ một số nút hạt giống và sau đó đi theo các cạnh để đến các nút khác. Quá trình truy cập trang và trích xuất các liên kết trong đó tương tự như việc mở rộng một nút trong tìm kiếm đồ thị.

## Frontier

Trong crawler, Frontier là danh sách các URL chưa được truy cập. Trong ngữ cảnh của tìm kiếm đồ thị, “frontier” là một danh sách các nút chưa được mở rộng (chưa được thăm). Mặc dù có thể có nhu cầu lưu trữ “frontier” trên đĩa cho các crawler lớn, tuy nhiên để đơn giản hóa, chúng ta chỉ giới thiệu “frontier” là các cấu trúc dữ liệu trong bộ nhớ. Dựa trên dung lượng bộ nhớ hiện có, chúng ta có thể quyết định kích thước tối đa của “frontier”. Với sự tiến bộ trong công nghệ bộ nhớ máy tính, việc giữ “frontier” với hàng chục nghìn URL không còn là điều hiếm. Do “frontier” chỉ có dung lượng hạn chế, chúng ta cần có một cơ chế để quyết định xem URL nào sẽ bị loại bỏ khi “frontier” đầy. Lưu ý rằng “frontier” có thể nhanh chóng đầy hơn so với số lượng trang web đã truy cập. Có thể có tới 60,000 URL trong “frontier” khi mới chỉ thăm khoảng 10,000 trang web do trung bình mỗi trang có khoảng 7 liên kết.

“Frontier” có thể được triển khai dưới dạng hàng đợi FIFO nếu muốn xây dựng một crawler theo chiến lược duyệt chiều rộng mà không cần quan tâm nhiều đến chiến lược duyệt. URL tiếp theo cần thăm được lấy từ đỉnh của hàng đợi và các URL mới được thêm vào cuối hàng đợi. Do “frontier” có kích thước hạn chế, chúng ta cần đảm bảo không thêm các URL trùng lặp vào hàng đợi. Một cơ chế tìm kiếm tuyến tính để tìm ra một URL mới từ nội dung của URL đang được thăm là rất cần thiết. Một giải pháp có thể là sử dụng một bảng băm để lưu trữ mỗi URL trong “frontier” để thuận tiện cho việc tìm kiếm. Bảng băm này cần được đồng bộ với “frontier” thực sự. Một giải pháp khác là duy trì hàng đợi như một bảng băm. Điều này cung cấp một cách tìm kiếm nhanh chóng để tránh việc lưu lặp lại các URLs. Tuy nhiên, mỗi lần crawler cần một URL để thăm, nó cần phải tìm kiếm và lấy ra URL mới từ “frontier” gần đây nhất. Nếu tốc độ xử lý không quan trọng hơn, giải pháp đầu tiên có thể là lựa chọn tốt hơn. Khi “frontier” đạt đến kích thước tối đa, crawler chỉ có thể thêm duy nhất một URL chưa được thăm từ mỗi trang web đã được thăm.

Nếu “frontier” được triển khai dưới dạng hàng đợi ưu tiên, chúng ta có một loại crawler gọi là crawler ưu tiên. Hàng đợi ưu tiên có thể là một mảng động luôn được sắp xếp theo độ đo được đánh giá của các URL chưa được thăm. Ở mỗi bước, URL tốt nhất được lấy ra từ đầu hàng đợi. Khi trang web tương ứng được tải, các URL ra ngoài từ nó được lấy ra và được đánh giá dựa trên một số tiêu chí. Sau đó, chúng được thêm vào “frontier” tại các vị trí phụ thuộc vào đánh giá đó. Chúng ta có thể tránh việc thêm các URL trùng lặp vào “frontier” bằng cách giữ một bảng băm riêng biệt để tìm kiếm. Khi “frontier” đạt đến kích thước tối đa, chỉ có MAX URL tốt nhất được giữ lại trong “frontier”.

Nếu chương trình crawler phát hiện “frontier” trống khi cần URL tiếp theo để thăm, quá trình crawling sẽ dừng lại. Với một giá trị MAX lớn và một số URL hạt giống thường xuyên, “frontier” hiếm khi trở thành trống.

## Fetching

Để lấy một trang web, crawler hoạt động như một trình duyệt Web, sẽ gửi yêu cầu HTTP đến máy chủ và đọc phản hồi. Để tránh chờ đợi quá lâu, tránh tải quá nhiều dữ liệu thường được đặt thời gian chờ kết nối cho trình duyệt.

Thường sẽ được tải xuống một lượng dữ liệu nhất định cho mỗi trang. Trình duyệt phân tích phản hồi từ máy chủ để kiểm tra mã trạng thái và xác định các chuyển hướng. Vòng lặp chuyển hướng cần được phát hiện và dừng bằng cách lưu trữ URL từ chuỗi chuyển hướng trong một bảng băm và dừng lại nếu gặp cùng một URL. Việc kiểm tra lỗi và xử lý ngoại lệ là rất quan trọng trong quá trình lấy trang, đặc biệt là khi cùng một mã phải xử lý với hàng triệu máy chủ từ xa. Đồng thời, việc thu thập thống kê về thời gian chờ kết nối và các mã trạng thái có thể giúp xác định vấn đề hoặc điều chỉnh tự động thời gian chờ. Các ngôn ngữ lập trình như Java, Python và Perl cung cấp các giao diện lập trình đơn giản để lấy các trang web từ Internet. Tuy nhiên, cần phải cẩn thận khi sử dụng các giao diện cao cấp vì có thể khó khăn hơn trong việc phát hiện các vấn đề ở mức thấp hơn. Ví dụ, một crawler mạnh mẽ trong Perl nên sử dụng mô-đun Socket để gửi các yêu cầu HTTP thay vì sử dụng thư viện cấp cao hơn như LWP (thư viện Web toàn cầu cho Perl), vì LWP không cho phép kiểm soát chính xác thời gian chờ kết nối.

Quá trình Crawling không thể kết thúc nếu không đề cập đến giao thức loại trừ Robot Exclusion Protocol. Giao thức này cung cấp một cơ chế cho các nhà quản trị Web server để thông báo về các quyền truy nhập file trên server, đặc biệt là để chỉ định các file không được truy cập bởi một crawler. Được thực hiện bằng cách lưu một file có tên robots.txt dưới thu mục chủ web server. File này cung cấp các chính sách truy cập cho các User-agent khác nhau trong file.

## Parsing

Parsing là quá trình phân tích dữ liệu đã thu thập để trích xuất thông tin cụ thể. Trong ngữ cảnh của web crawler, parsing module thực hiện việc trích xuất văn bản và liên kết từ trang web đã thu thập. Các thông tin quan trọng như tiêu đề, nội dung, liên kết, và các phần khác của trang web được trích xuất để lập chỉ mục và hiển thị trong kết quả tìm kiếm.

Dưới đây là một số khái niệm liên quan đến parsing:

Abstract Syntax Tree (AST): AST là một cây biểu diễn cấu trúc ngữ pháp của mã nguồn hoặc dữ liệu. AST giúp biểu diễn cấu trúc logic của chương trình hoặc tài liệu. Các thành phần của AST bao gồm các nút (nodes) biểu diễn các phần tử như biến, hàm, toán tử,… AST thường được sử dụng trong các trình biên dịch, transpiler, và các công cụ xử lý ngôn ngữ tự nhiên 1.

Parser là một công cụ dùng để chuyển đổi chuỗi các token (đơn vị cơ bản của mã nguồn) thành một AST. Có hai loại parser phổ biến:

* Bottom-up parser: Xây dựng AST từ dưới lên.
* Top-down parser: Xây dựng AST từ trên xuống.

## Quá trình lấy và chuẩn hóa các URL

Quá trình lấy ra và chuẩn hóa các URL là bước quan trọng trong việc thực hiện của một crawler. Trước khi truy cập vào các trang web, crawler phải tìm ra các liên kết trong trang gốc và chuyển đổi chúng thành các URL hợp lệ và chuẩn hóa để tiện cho việc truy cập sau này.

Đầu tiên, các liên kết trong trang web được tìm thấy thông qua việc phân tích nội dung HTML của trang. Sau đó, các liên kết này được trích xuất và kiểm tra để đảm bảo rằng chúng là các URL hợp lệ. Các URL có thể được viết dưới nhiều dạng khác nhau và có thể chứa các thông tin không cần thiết hoặc không phù hợp, vì vậy việc chuẩn hóa các URL là rất cần thiết để đảm bảo tính nhất quán và hiệu quả của quá trình lấy dữ liệu.

**Một số biến thể của cùng một URL có thể được chuẩn hóa thành một dạng cụ thể. Ví dụ:**

|  |  |
| --- | --- |
| **Trước khi chuẩn hoá** | **Sau khi chuẩn hoá** |
| "https://langhuunghi.vn/gioi-thieu-ve-lang?product\_id=418" | "https://langhuunghi.vn/gioi-thieu-ve-lang" |
| "https://tuyensinh247.com/combo.html?year=2025" | "https://tuyensinh247.com/combo.html" |

Bảng 3.1. Chuẩn hoá biến thể một URL

Khi đã có các URL, quá trình chuẩn hóa bao gồm việc loại bỏ các phần không cần thiết như các tham số truy vấn, phần mở rộng file không cần thiết, và việc đảm bảo rằng mỗi URL được biểu diễn một cách chính xác và duy nhất. Điều này giúp crawler tránh lặp lại việc truy cập vào cùng một nội dung thông qua các URL khác nhau.

|  |  |  |
| --- | --- | --- |
| **Loại chuẩn hoá** | **URL ban đầu** | **Sau khi chuẩn hoá** |
| Loại bỏ phần anchor hoặc reference của URL | http://myspider.biz.uiowa.edu/faq.html#what | http://myspider.biz.uiowa.edu/faq.html |
| Mã hóa URL bằng các ký tự thông dụng như '~' | http://dollar.biz.uiowa.edu/~pant | [http://dollar.biz.uiowa.edu /%7Euser/profile](http://example.com/%7Euser/profile) |
| Thêm vào dấu '/' cho các URL | https://sta.huflit.edu.vn/huong-dan/dang-ky-de-tai-khcn | [https://sta.huflit.edu.vn/huong-dan/dang-ky-de-tai-khcn/](http://example.com/path/) |
| Nhận ra các trang web mặc định và loại bỏ các tên file như index.html hay index.htm | <https://moon.vn/home.html> | <https://moon.vn/> |
| Loại bỏ ký tự '..' và thư mục cha khỏi đường dẫn URL | /%7Epant/BizIntel/Seeds/../ODPSeeds.dat | /%7Epant/BizIntel/ODPSeeds.dat |
| Để lại các số hiệu cổng trong các URL ngoại trừ đó là cổng 80 | [http://www.sistemaexclie.com:8080](http://www.sistemaexclie.com:8080/) | http://www.sistemaexclie.com:8080 |

Bảng 3.2. Chuẩn hoá URL

Sau khi các URL đã được chuẩn hóa, chúng được thêm vào danh sách chờ xử lý để crawler tiếp tục truy cập vào các trang liên kết từ đó và lặp lại quá trình lấy dữ liệu. Quá trình này là một phần hoạt động của crawler và đóng vai trò quan trọng trong việc thu thập thông tin trên Internet. Đảm bảo tính nhất quán trong khi áp dụng các luật chuẩn hóa. Có thể có hai luật khác nhâu lại cho kết quả tốt như nhau miễn là áp dụng luật chuẩn hóa URL một cách nhất quán.

## Xây dựng cây HTML

HTML (HyperText Markup Language) là ngôn ngữ đánh dấu được sử dụng để tạo nên cấu trúc của các trang web.

Trong HTML, các phần tử được đánh dấu bằng cặp thẻ mở <tag> và thẻ đóng </tag>, ví dụ: <div>, <p>, <a>,...

Cấu trúc của một trang web có thể được trình bày dưới dạng cây DOM (Document Object Model). Parser thực hiện việc phân tích cú pháp, kiểm tra tính hợp lệ của mã nguồn và tạo ra cấu trúc cây tương ứng. Nhưng chúng ta cần phải xem xét và đưa ra phương pháp kĩ càng vì rất nhiều trang web có cấu trúc HTML không chuẩn. Ví dụ, một thẻ bắt đầu có thể không có thẻ đóng, hoặc các thẻ không được lồng nhau một cách hợp lý. Trong nhiều trường hợp, thẻ hoặc đều bị thiếu trong trang HTML. Do đó các tiêu chuẩn dựa trên cấu trúc (structure-based criteria) thường cần có một bước tiền xử lý để chuẩn hóa. Nó bao gồm cả việc chèn thêm các thẻ bị thiếu và sắp xếp lại thứ tự các thẻ trong trang. Việc làm sạch một trang HTML là cần thiết để ánh xạ nội dung của trang vào trong một cấu trúc cây để đảm bảo tính toàn vẹn, mỗi nút có một cha duy nhất, từ đó phân tích nên cấu trúc cây của các thẻ. Ví dụ bên dưới chỉ ra cấu trúc cây của các thẻ tương ứng với văn bản HTML nguồn:

<html lang="vi">

<head>

    <meta charset="UTF-8">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">

    <title>Trang web của tôi</title>

</head>

<body>

    <h1>Chào mừng đến với trang web của tôi</h1>

    <p>Đây là nội dung chính của trang web

        <a href="https://example.com">Link</a>

</p>

    <ul>

        <li>Thẻ danh sách 1</li>

        <li>Thẻ danh sách 2</li>

        <li>Thẻ danh sách 3</li>

    </ul>

    <img src="image.jpg" alt="Hình ảnh minh họa">

</body>

</html>

A screenshot of a computer

Description automatically generated

## Spider traps

Spider traps là các cấu trúc hoặc điều kiện trong một trang web mà có thể làm cho các web crawler (hay spider) mắc kẹt hoặc bị vòng lặp vô hạn khi thu thập thông tin. Các spider traps thường gây ra những vấn đề cho các công cụ tìm kiếm và các robot web bằng cách tạo ra các liên kết không hợp lý hoặc lặp lại mà dẫn đến việc không thể kết thúc quá trình thu thập thông tin.

**Có một số loại spider traps phổ biến, bao gồm:**

Redirect loops (vòng lặp chuyển hướng): Trang web được cấu hình để tự động chuyển hướng từ một URL này sang URL khác, nhưng các chuyển hướng này tạo ra một vòng lặp vô hạn.

Dynamic parameters (tham số động): Các trang web có thể tạo ra hàng loạt các URL với các tham số động, dẫn đến việc tạo ra nhiều bản sao của cùng một nội dung.

Session IDs (ID phiên): Các trang web sử dụng session IDs trong các URL để theo dõi phiên làm việc của người dùng, nhưng điều này có thể dẫn đến việc tạo ra các URL khác nhau cho cùng một nội dung.

Spider traps gây ra vấn đề cho các công cụ tìm kiếm bởi vì chúng có thể làm tăng thời gian và tài nguyên cần thiết cho việc thu thập dữ liệu, và có thể làm cho các robot web mắc kẹt trong các vòng lặp không kết thúc. Đối với các webmaster, việc tránh tạo ra các spider traps giúp cải thiện hiệu suất của trang web và đảm bảo rằng các công cụ tìm kiếm có thể thu thập thông tin một cách hiệu quả.

## Page Repository

Chức năng chính của Page Repository là lưu trữ và quản lý các trang web thu thập từ Internet. Kho lưu trữ này thường bao gồm các phiên bản đã tải xuống của các trang web, cùng với các siêu dữ liệu như URL, thời gian thu thập, và các chỉ số khác liên quan đến nội dung của trang web. Page Repository cung cấp một nguồn tài nguyên quan trọng cho việc phân tích dữ liệu web, khai thác thông tin, và xây dựng các chỉ mục cho các công cụ tìm kiếm.

## Crawler đa luồng (Multi-threaded crawler)

Mô hình Crawler đa luồng (hay còn gọi là multithreaded crawler) là một phương pháp cho phép đồng thời thu thập thông tin từ nhiều nguồn trang web một cách hiệu quả.

Khái niệm thread: Trong lập trình, thread là một luồng thực thi độc lập, giúp chúng ta thực hiện nhiều công việc cùng lúc. Mỗi thread có thể xem như một công nhân chuyên thực hiện các nhiệm vụ mà chương trình giao.

Thread và Process: Một process có thể chứa nhiều thread. Process là không gian độc lập trong bộ nhớ, còn thread chia sẻ tài nguyên và trạng thái của process. Thread chính tạo ra giao diện ứng dụng.

**Mô hình đa luồng (Multithreading):**

* Đa luồng là mô hình cho phép tạo và sử dụng nhiều thread trong một process.
* Đa luồng thường được sử dụng trong các ứng dụng đa nhiệm, như web crawlers.

**Các đặc điểm chính của crawler đa luồng bao gồm:**

* Song song hóa: Crawler đa luồng chia công việc thành nhiều nhiệm vụ nhỏ và thực hiện chúng đồng thời bằng cách sử dụng nhiều luồng hoạt động.
* Tăng hiệu suất: Bằng cách hoạt động đồng thời, crawler đa luồng có thể thu thập dữ liệu nhanh chóng hơn so với crawler tuần tự truyền thống.
* Quản lý tài nguyên: Crawler đa luồng cần quản lý tài nguyên như bộ nhớ và băng thông một cách hiệu quả để tránh quá tải hệ thống.
* Xử lý lỗi và xung đột: Crawler đa luồng cần xử lý các vấn đề như lỗi kết nối, xung đột dữ liệu và xử lý ngoại lệ một cách an toàn để đảm bảo tính ổn định của quá trình thu thập dữ liệu.
* Phân chia công việc: Crawler đa luồng thường phân chia các URL hoặc tác vụ thu thập theo các phần nhỏ và phân phối chúng cho các luồng khác nhau để tối ưu hóa quá trình thu thập.

Tuy nhiên, việc thiết kế và triển khai crawler đa luồng cũng đặt ra một số thách thức như đồng bộ hóa dữ liệu, quản lý luồng và xử lý lỗi một cách hiệu quả. Điều này đòi hỏi sự kiểm soát cẩn thận và kiến thức chuyên sâu về lập trình đa luồng và công nghệ mạng.

## Các thuật toán crawling

Có nhiều thuật toán được sử dụng trong quá trình crawling, nhằm tối ưu hóa việc thu thập dữ liệu từ Internet. Dưới đây là một số thuật toán crawling phổ biến:

* Breadth-First Search (BFS): Thuật toán này tập trung vào việc thu thập các trang web từ "gần" đến "xa" theo cấp độ trang web. Nó bắt đầu từ các seed URLs và sau đó duyệt các trang con liên kết trực tiếp từ các trang gốc.
* Depth-First Search (DFS): Ngược lại với BFS, DFS tập trung vào việc duyệt sâu vào các nhánh của cây trang web trước khi quay lại các nhánh khác. Điều này có thể dẫn đến việc tiêu tốn nhiều tài nguyên hơn do số lượng trang con có thể rất lớn.
* Focused Crawling: Thuật toán này tập trung vào việc thu thập các trang web liên quan đến một chủ đề hoặc lĩnh vực cụ thể bằng cách sử dụng các chỉ mục hoặc từ khoá để xác định các trang liên quan.
* Politeness Policies: Đây không phải là một thuật toán cụ thể mà là một chiến lược quản lý việc truy cập trang web. Nó bao gồm các quy tắc về tần suất và cách thức tiếp cận các trang web để tránh làm quá tải máy chủ hoặc bị chặn.
* PageRank Algorithm: Thuật toán này không chỉ là một thuật toán tìm kiếm mà còn được sử dụng trong quá trình crawling. Nó ưu tiên việc thu thập các trang web được coi là quan trọng dựa trên sự phổ biến của chúng trên Internet.
* Incremental Crawling: Thuật toán này tập trung vào việc cập nhật các trang đã thu thập thay vì thu thập lại từ đầu. Điều này giúp tiết kiệm thời gian và tài nguyên khi thu thập dữ liệu trên một khoảng thời gian dài.
* Machine Learning-based Crawling: Các thuật toán học máy có thể được sử dụng để dự đoán các trang web tiềm năng để thu thập dựa trên dữ liệu thu thập được từ các trang web trước đó.

Các thuật toán crawling có thể được kết hợp và tinh chỉnh để tạo ra các chiến lược tối ưu cho việc thu thập dữ liệu từ Internet, tùy thuộc vào mục tiêu và yêu cầu cụ thể của ứng dụng.

## Kết luận

Web Crawling không chỉ là việc đơn giản là tải các trang web về máy chủ mà còn là quá trình tổ chức, lọc và lấy dữ liệu từ các trang web một cách tự động và hiệu quả. Dưới đây là một số điểm chi tiết về phần crawling web:

Nguồn tài nguyên lớn: Internet chứa một lượng lớn thông tin đa dạng trên hàng tỷ trang web. Crawling web cho phép trích xuất thông tin từ những nguồn tài nguyên này để phục vụ cho nhu cầu nghiên cứu, kinh doanh và giải quyết vấn đề trong thực tế.

Dữ liệu phong phú: Thông qua việc thu thập dữ liệu từ nhiều nguồn khác nhau trên internet, quá trình crawling web cung cấp cho các nhà nghiên cứu và doanh nghiệp một tập dữ liệu phong phú và đa dạng, giúp họ hiểu rõ hơn về xu hướng, hành vi và ý kiến trên internet.

Các Thuật Toán Duyệt Web: Có nhiều thuật toán và phương pháp được phát triển để duyệt web một cách hiệu quả, bao gồm phân tích cú pháp HTML, điều chỉnh tốc độ duyệt web và quản lý độ ưu tiên của các trang web.

Công Nghệ Web Scraping: Scraping web là một phương pháp phổ biến để trích xuất thông tin từ các trang web. Nó sử dụng các thư viện và công cụ để tự động lấy dữ liệu từ HTML của các trang web và chuyển đổi nó thành định dạng dễ đọc và sử dụng.

Robots.txt và Crawl Delay: Robots.txt là một tệp văn bản đặc biệt trên máy chủ web của một trang web, giúp quy định các quy tắc và hạn chế cho việc crawling web. Crawl delay cho phép người quản trị máy chủ thiết lập một thời gian chờ giữa các yêu cầu từ các crawler, giúp giảm áp lực lên máy chủ và tăng hiệu suất crawling web.

Quản Lý Độ Sâu và Phạm Vi: Cần xác định rõ phạm vi và độ sâu của quá trình crawling web để tránh lặp lại và tiêu tốn tài nguyên không cần thiết.

Tóm lại, phần crawling web không chỉ là một phần quan trọng trong quy trình thu thập dữ liệu từ internet mà còn là một quá trình phức tạp yêu cầu sự kỹ thuật và quản lý chặt chẽ để đảm bảo tính hiệu quả và tuân thủ các quy định.

# XÂY DỰNG MÔ HÌNH PHÂN LOẠI NỘI DUNG WEBSITE TIẾNG VIỆT

*Dựa trên sự tích lũy kiến thức và kinh nghiệm thu được từ những nghiên cứu liên quan, chúng em đã tiến hành triển khai một chương trình cụ thể nhằm phân loại nội dung văn bản trang web tiếng Việt. Trong phần này của dự án, chúng em sẽ mô tả chi tiết các hoạt động cụ thể mà chúng em đã thực hiện theo từng bước của quy trình phân loại văn bản. Điều này bao gồm việc xác định, thu thập dữ liệu, xử lý và đánh giá văn bản từ các nguồn khác nhau, cùng với việc áp dụng các kỹ thuật và công nghệ phân loại nhằm tạo ra hệ thống phân loại nội dung văn bản hiệu quả và chính xác.*

## Bài toán lấy dữ liệu

Việc thu thập dữ liệu là một bước quan trọng và tiền đề không thể thiếu để huấn luyện một mô hình phân loại nội dung văn bản trang web. Dữ liệu càng nhiều và đa dạng, mô hình sẽ có cơ hội học được nhiều điều hơn và trở nên bao quát hơn, giúp cho việc phân loại trở nên chính xác và hiệu quả hơn.

Đa dạng hóa dữ liệu: Dữ liệu đa dạng giúp mô hình hiểu được nhiều loại văn bản khác nhau, từ đó tăng cường khả năng tổng quát hóa và áp dụng mô hình vào các trường hợp mới.

Phát hiện đặc trưng tốt hơn: Bằng cách thu thập dữ liệu đa dạng, mô hình có thể học được nhiều đặc trưng và mẫu dữ liệu phong phú, giúp cải thiện hiệu suất của mô hình trong việc phân loại các trang web.

Phản ánh đời sống thực tế: Dữ liệu đa dạng và phong phú thường phản ánh đời sống thực tế và đa dạng của nội dung trên Internet, giúp mô hình có khả năng phân loại tốt hơn trong các tình huống thực tế.

Hạn chế overfitting: Một mô hình huấn luyện trên một lượng dữ liệu lớn và đa dạng sẽ ít có khả năng bị overfitting, tức là mô hình ít có xu hướng học "quá mức" từ dữ liệu huấn luyện và áp dụng những kiến thức học được vào các trường hợp mới một cách hiệu quả hơn.

Nâng cao độ chính xác và độ tin cậy: Dữ liệu đa dạng giúp mô hình phân loại nội dung văn bản trang web đạt được độ chính xác và độ tin cậy cao hơn trong các nhiệm vụ phân loại thực tế. Thu thập dữ liệu đa dạng và phong phú là một yếu tố quan trọng để huấn luyện một mô hình phân loại nội dung văn bản trang web hiệu quả và bao quát.

Với bài luận văn này, chỉ tiếng hành phân loại nội dung web tiếng việt trong 14 phân lớp (thể loại) dựa theo WebRank [23] bao gồm:

|  |  |  |
| --- | --- | --- |
| STT | Thể loại | Mô tả |
|  | Chính phủ | Trang web chính phủ cung cấp thông tin và dịch vụ công trực tuyến, với giao diện thân thiện, thông tin hành chính, tin tức, và tài liệu pháp luật |
|  | Tin tức | Cung cấp tin tức mới nhất về các sự kiện, vấn đề xã hội, chính trị và kinh doanh |
|  | Sức khỏe | Cung cấp thông tin về sức khỏe, y tế, bệnh lý và lối sống lành mạnh |
|  | Động vật | Thông tin về thế giới động vật, chăm sóc và bảo vệ động vật |
|  | Nghề nghiệp | Cung cấp thông tin, tư vấn và cơ hội việc làm trong các lĩnh vực nghề nghiệp khác nhau |
|  | Tmdt | Tin tức và thông tin về thị trường mua bán trực tuyến, công nghệ và xu hướng thương mại điện tử |
|  | Cộng đồng | Thảo luận và chia sẻ về các vấn đề xã hội, văn hóa, và cộng đồng |
|  | Du lịch | Hướng dẫn du lịch, điểm đến, khám phá văn hóa và ẩm thực địa phương |
|  | Thể thao | Tin tức, kết quả trận đấu, và thông tin về các môn thể thao và sự kiện thể thao |
|  | Giải trí | Cung cấp tin tức, bài viết, video và hình ảnh giải trí, phim ảnh, âm nhạc và trò chơi |
|  | Tài chính | Thông tin về tài chính, đầu tư, tiết kiệm, và quản lý tài sản |
|  | Xe cộ | Tin tức, đánh giá, và hướng dẫn về ô tô, xe máy và ngành công nghiệp xe hơi |
|  | Khoa học | Cung cấp thông tin và bài viết về khoa học tự nhiên, công nghệ và các phát hiện mới hay các thông tin liên quan đến giáo dục |
|  | Thực phẩm | Hướng dẫn nấu ăn, tin tức và xu hướng trong ngành ẩm thực và dinh dưỡng |
|  | Tài liệu | Cung cấp tài liệu học tập, sách điện tử, bài giảng, và tài liệu tham khảo trong nhiều lĩnh vực khác nhau |

Bảng 4.1. Thể loại các website

Để tạo sự đa dạng cho dữ liệu thu thập về, chúng em tiến hành lấy dữ liệu từ nhiều trang web tiếng Việt, mỗi trang thuộc các thể loại khác nhau. Mỗi trang web sẽ lấy về các thông tin: tiêu đề, nội dung và đường dẫn (URL), các dữ liệu trên sẽ được lưu trữ thành .xlsx hoặc .txt.

Trước tiên sẽ thực hiện lấy các URL từ các trang web được chọn, lưu thành các file .xlsx, quá trình này được thực hiện tự động. Sử dụng Scrapy, một thư viện Python mạnh mẽ để thực hiện web crawling, kết hợp với OpenPyXL để lưu các URL thu thập được vào một tệp Excel. Giả sử chúng ta có một trang web với cấu trúc sau:

* Trong đó: A là trang web gốc. B và C là các liên kết từ trang A. D, E, F và G là các liên kết từ trang B và C.

Quá trình lấy URL theo chiều sâu sẽ diễn ra như sau:

* Bắt đầu từ trang A: Duyệt trang A và lấy URL của nó. Đánh dấu A là đã duyệt. Tiếp tục duyệt các liên kết từ trang A (B và C).
* Duyệt trang B: Lấy URL của trang B. Đánh dấu B là đã duyệt. Tiếp tục duyệt các liên kết từ trang B (D và E).
* Duyệt trang D và E: Lấy URL của D và E. Đánh dấu D và E là đã duyệt.
* Quay lại trang B: Kiểm tra xem còn liên kết nào khác từ B không (không có).
* Duyệt trang C: Lấy URL của trang C. Đánh dấu C là đã duyệt. Tiếp tục duyệt các liên kết từ trang C (F và G).
* Duyệt trang F và G: Lấy URL của F và G. Đánh dấu F và G là đã duyệt.
* Kết thúc quá trình: Đã lấy được các URL từ trang web theo chiều sâu.

### Duyệt URL theo chiều sâu (DFS):

*Sử dụng Scrapy để duyệt URL theo chiều sâu từ một trang web cụ thể*

* **Import các thư viện:**

Mã import các thư viện cần thiết bao gồm scrapy, CrawlSpider, Rule, LinkExtractor từ Scrapy, và openpyxl để làm việc với file Excel.

* **Xác định URL ban đầu:** Bắt đầu với một hoặc nhiều URL ban đầu từ trang web cần thu thập dữ liệu.

Giả sử chúng ta có một số trang web gốc là <https://gearvn.com/>, <https://baomoi.com/>, https://vnexpress.net, <https://tiki.vn/>, ...

* **Tạo một Spider:**

Định nghĩa một class Spider kế thừa từ CrawlSpider trong Scrapy.

Class URLCrawler được định nghĩa và kế thừa từ CrawlSpider của Scrapy.

name của Spider được đặt là 'url-crawler'.

domain được chỉ định là [‘gearvn.com’].

start\_urls là URL ban đầu muốn bắt đầu duyệt.

max\_pages\_per\_domain là số lượng trang tối đa muốn thu thập từ mỗi miền.

current\_pages là biến để theo dõi số lượng trang đã thu thập từ miền.

allowed\_domains là danh sách các miền được phép duyệt.

urls là một set để lưu trữ các URL đã thu thập.

image\_extensions là tập hợp các phần mở rộng của các tệp hình ảnh, để loại bỏ khi duyệt.

rules chứa các quy tắc duyệt, ở đây chỉ có một quy tắc để duyệt các liên kết trong miền đã cho.

Thiết lập các thông số cơ bản như tên Spider, domain của trang web đang duyệt, và URL bắt đầu.

Bắt đầu từ URL gốc này, chúng ta lấy ra và duyệt qua các URL liền kề trên trang web (ví dụ: gearvn.com/pages/laptop-van-phong, gearvn.com/pages/huong-dan-tra-gop, gearvn.com/collections, tiki.vn/nha-sach-tiki/c8322, tiki.vn/dien-thoai-may-tinh-bang/c1789, vnexpress.net/kinh-doanh, vnexpress.net/the-gioi, baomoi.com/chu-de.epi, baomoi.com/tin-moi.epi,...).

* **Xác định quy tắc duyệt:**

Sử dụng Rule để xác định cách duyệt các trang và các liên kết trên trang đó.

Sử dụng LinkExtractor để xác định cách trích xuất các liên kết từ mỗi trang.

* **Thiết lập hàm callback:**

Định nghĩa một hàm callback để xử lý dữ liệu từ mỗi trang được trích xuất.

Trong hàm callback, thu thập các URL từ mỗi trang và lưu chúng vào một danh sách.

**Hàm callback parse\_item:**

Hàm này được gọi khi một trang web được tải xuống và được sử dụng để xử lý dữ liệu trên trang.

Nó kiểm tra nếu số trang đã thu thập chưa vượt quá max\_pages\_per\_domain, nếu có, thì thêm URL vào danh sách urls.

Nếu đã đạt đến giới hạn, nó lưu trữ danh sách URL vào một file Excel và dừng Spider.

**Hàm callback closed:**

Hàm này được gọi khi Spider kết thúc.

Nếu danh sách urls không rỗng, nó lưu trữ các URL vào file Excel trước khi kết thúc.

* **Lưu trữ vào file Excel:**

Sau khi thu thập đủ số lượng URL hoặc hết các URL có thể duyệt, lưu chúng vào một file Excel với tên tương ứng với domain của trang web.

Quá trình này sẽ tiếp tục cho tới khi không còn URL nào trong hàng đợi hoặc khi đạt được số lượng URL mong muốn. Sau đó tự động lưu trữ vào file xlsx đã được định nghĩa. Ví dụ: urls\_summary\_gearvn.xlsx, urls\_summary\_baomoi.xlsx, urls\_summary\_tiki.xlsx, urls\_summary\_vnexpress.xlsx.

Chúng em liệt kê và phân chia các trang web vào các thể loại, mỗi thể loại sẽ có khoảng 10-30 các trang web gốc khác nhau. Mỗi trang web gốc sẽ có khoảng 100-500 url được lấy. Khi đã có các url, việc liệt kê và phân chia trang web theo các thể loại sẽ được thực hiện thủ công.

|  |  |  |
| --- | --- | --- |
| STT | Tên thể loại | Số links URL |
|  | Chính phủ | 2149 |
|  | Tin tức | 1668 |
|  | Sức khỏe | 1296 |
|  | Động vật | 1086 |
|  | Nghề nghiệp | 1069 |
|  | Cộng đồng | 1129 |
|  | Tmdt | 1408 |
|  | Du lịch | 1198 |
|  | Thể thao | 780 |
|  | Giải trí | 1097 |
|  | Tài chính | 1117 |
|  | Xe cộ | 1275 |
|  | Khoa học | 1044 |
|  | Thực phẩm | 1265 |
|  | Tài liệu | 1098 |

Bảng 4.2. Số lượng links URL từng thể loại

### Lấy dữ liệu từng trang web và gán nhãn

Sau khi thu thập danh sách các liên kết URL từ các trang web gốc, tiếp theo là quá trình lấy dữ liệu từng trang web và gắn nhãn cho nội dung tương ứng. Quá trình này giúp tổ chức và phân loại dữ liệu thu thập được, tạo điều kiện cho việc phân tích và sử dụng dữ liệu một cách hiệu quả.

Đầu tiên, từng URL trong danh sách được truy cập một cách tuần tự. Sử dụng các công cụ BeautifulSoup, trang web được tải xuống và dữ liệu của nó được trích xuất.

Cho các file txt chứa danh sách url vào một folder tên label, các tệp bên trong được định nghĩa \_\_label\_[theloai].txt.

|  |  |  |
| --- | --- | --- |
| STT | Thể loại | Định nghĩa |
|  | Chính phủ |  |
|  | Tin tức |  |
|  | Sức khỏe |  |
|  | Động vật |  |
|  | Nghề nghiệp |  |
|  | Cộng đồng |  |
|  | Tmdt |  |
|  | Du lịch |  |
|  | Thể thao |  |
|  | Giải trí |  |
|  | Tài chính |  |
|  | Xe cộ |  |
|  | Khoa học |  |
|  | Thực phẩm |  |
|  | Tài liệu |  |

Bảng 4.3. Các tập URL từng thể loại

* **Truy cập và tải dữ liệu từ các URL:**

Sử dụng thư viện requests để truy cập và tải dữ liệu từ mỗi URL trong tập hợp. Đảm bảo kiểm tra xem trang web có phản hồi thành công không trước khi tiếp tục xử lý.

* **Xử lý HTML, JavaScript và CSS:**

Sử dụng BS4 để phân tích cú pháp HTML của trang web và trích xuất các thành phần cần thiết như tiêu đề, nội dung, hình ảnh, v.v. Nếu trang web sử dụng JavaScript để tạo nội dung, sử dụng công cụ như Selenium để tải trang và lấy dữ liệu đã được render.

* **Loại bỏ khoảng trắng và xuống dòng không cần thiết:**

Sau khi trích xuất dữ liệu, loại bỏ các khoảng trắng và ký tự xuống dòng không cần thiết để làm sạch nội dung. Có thể sử dụng các phương pháp như strip() để loại bỏ các khoảng trắng ở đầu và cuối chuỗi.

Sau khi thu thập được dữ liệu, quá trình tiếp theo là gắn nhãn cho nội dung đã thu thập. Quá trình này được thực hiện tự động, tự động gán nhãn dựa trên các từ khóa xuất hiện khớp với các từ khóa trong *‘label\_mapping’*, phương pháp gán nhãn cơ bản dựa trên quy tắc để dễ dàng sử dụng ở các bước phân tích và học máy sau này, các nhãn tương ứng với các thể loại đã liệt kê bên trên.

Cuối cùng, dữ liệu đã được gắn nhãn được lưu trữ vào một cơ sở dữ liệu hoặc file để sử dụng trong các mục đích tiếp theo như phân tích dữ liệu, đào tạo mô hình máy học, hoặc hiển thị thông tin cho người dùng. Quá trình này giúp tổ chức và tận dụng dữ liệu một cách hiệu quả, mang lại giá trị và thông tin có ích cho dự án. Nhưng không phải mọi bài viết đều có thể sử dụng được, kiểm duyệt lại các bài viết này là cần thiết.  Ngoài ra, chúng em còn phát hiện ra một số lỗi khác xảy ra trong quá trình phân tích html như sau:

* Thiếu nội dung
* Lỗi do mã hóa dữ liệu (encoding)
* Trùng dữ liệu
* Còn dính các thẻ code
* Các ký tự không mong muốn
* Dữ liệu dư thừa

Lỗi này xảy ra là do nguồn mã hóa không đồng nhất. Các bài viết phát sinh lỗi này có thể là trên một phần hoặc toàn bộ bài viết. Việc dùng cái này sẽ làm cho quá trình tách từ xuất hiện các từ không có trong tiếng Việt, gây nhiễu và không có giá trị gì trong phân loại bài viết.

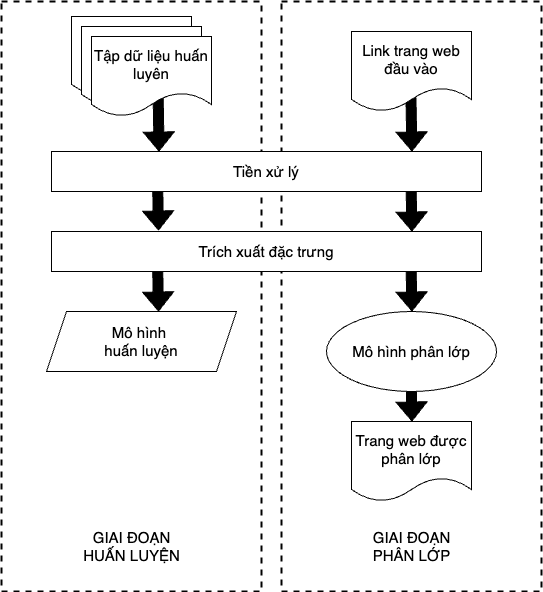
|  |  |  |
| --- | --- | --- |
| STT | Tên thể loại | Số nội dung thu thập |
|  | Chính phủ | 2142 |
|  | Tin tức | 2012 |
|  | Sức khỏe | 1951 |
|  | Động vật | 1240 |
|  | Nghề nghiệp | 1445 |
|  | Cộng đồng | 1668 |
|  | Tmdt | 1692 |
|  | Du lịch | 1303 |
|  | Thể thao | 1120 |
|  | Giải trí | 1123 |
|  | Tài chính | 1689 |
|  | Xe cộ | 1689 |
|  | Khoa học | 1528 |
|  | Thực phẩm | 1417 |
|  | Tài liệu | 1240 |
|  | **Tổng cộng** | **23688** |

Bảng 4.4. Số lượng nội dung đã lấy

Việc kiểm tra và tiền xử lý rất quan trọng. Nếu dữ liệu không chính xác, không đầy đủ hoặc không đại diện cho mục tiêu mong muốn, mô hình học máy sẽ không hoạt động hiệu quả.

## Xây dựng và đánh giá các mô hình phân lớp

Để tiến hành huấn luyện, dữ liệu đầu vào cần phải được tiền xử lý và trích xuất đặc trưng. Cấu trúc hệ thống phân loại được minh họa ở hình dưới đây:



Hình 4.1. Mô hình của hệ thống phân loại trang web

Dữ liệu dùng cho việc đào tạo mô hình chúng em sẽ chia thành 80:20. Có nghĩa là tập dữ liệu dành cho việc huấn luyện là 80%, dữ liệu để kiểm tra là 20%. Sau khi có dữ liệu sẽ tới bước rút trích, chọn lọc các đặc trưng tập dữ liệu huấn luyện kèm theo là nhãn (đầu ra mong muốn) và sử dụng thuật toán máy học để xây dựng mô hình dự đoán. Khi mô hình đã được huấn luyện đầy đủ, ta tiến hành việc kiểm tra. Sử dụng 20% dữ liệu còn lại mà mô hình chưa từng thấy để đưa vào mô hình dự đoán. Sau đó, so sánh các dự đoán này với giá trị thực tế và tính toán độ chính xác của mô hình.

### Xây dựng tập dữ liệu

Tập dữ liệu được xây dựng theo định dạng , trong đó là nội dung của các trang web và là nhãn của trang web đó, tương ứng với lớp của trong một tập hợp hữu hạn các lớp đã được định nghĩa trước. Hệ thống sẽ được huấn luyện trên tập dữ liệu này với một mô hình có khả năng phân loại một trang web mới vào một trong các nhãn thuộc tập . Trong nghiên cứu này, tập dữ liệu bao gồm cặp (nội dung trang web, nhãn) và số lượng các lớp , tương ứng với 15 thể loại website. Tập dữ liệu này là kết quả của bài toán lấy dữ liệu ở phần 4.1.

Bước tiếp theo, chúng em sẽ tiến hành tiền xử lý dữ liệu trước khi đưa vào huấn luyện mô hình phân loại văn bản. Việc tiền xử lý dữ liệu là hết sức quan trọng để đảm bảo mô hình đạt được kết quả tốt.

### Tiền xử lý

Các bước tiền xử lý được thực hiện để đảm bảo rằng dữ liệu đầu vào sạch sẽ, nhất quán và phù hợp với các yêu cầu của mô hình học máy như chúng em đã trình bày ở chương cơ sở lý thuyết bên trên. Việc tiền xử lý dữ liệu bao gồm các vấn đề sau:

* **Chuẩn hoá dữ liệu**

Chuẩn hoá dữ liệu là quá trình chuyển đổi văn bản thành một dạng cơ bản duy nhất mà trước đó có thể không có. Việc chuẩn hóa văn bản trước khi lưu trữ hoặc xử lý nó cho phép phân tách các mối quan tâm, vì đầu vào được đảm bảo là nhất quán trước khi thực hiện các hoạt động trên nó. Việc chuẩn hóa văn bản đòi hỏi phải nhận biết được loại văn bản nào sẽ được chuẩn hóa và cách xử lý nó sau đó; không có quy trình chuẩn hóa nào phù hợp cho mọi trường hợp [24].

Có thể kể đến các trường hợp cần chuẩn hoá sau:

* + Chuẩn hoá Unicode tiếng Việt

Unicode tiếng Việt là một phần của hệ thống mã hóa Unicode mà sử dụng để biểu diễn các ký tự, ký hiệu và dấu văn bản trong tiếng Việt. Unicode tiếng Việt bao gồm các ký tự chữ cái, số, dấu thanh, dấu ngã, dấu hỏi, và các ký tự đặc biệt khác được sử dụng trong viết tiếng Việt.

Trong hệ thống Unicode, các ký tự có thể được tạo ra từ các thành phần nhỏ hơn như ký tự ASCII, ký tự dấu nguyên âm và ký tự dấu thanh. Ví dụ, để hiển thị chữ "ễ", có thể áp dụng hai phương pháp sau:

|  |  |
| --- | --- |
| Phuơng pháp tổ hợp | Phương pháp dựng sẵn |
| Kết hợp ký tự "ê" với dấu “~” | Sử dụng ký tự "ễ" đã có sẵn trong bộ mã Unicode |

Bảng 4.5. Các phương pháp gõ dấu

Đa số người dùng sử dụng Unikey với bộ mã Unicode đã dựng sẵn, nhưng cũng có những trường hợp nơi mà các bộ mã Unicode tôt hợp vẫn được sử dụng. Kết quả là, trong các nguồn dữ liệu thu thập trên internet, thường gặp phải sự lẫn lộn giữa cả hai phương pháp này, điều đó dẫn đến trường hợp sau:

|  |
| --- |
| "thiếu" == "thiếu" ® True |
| "thiếu" == "thiếu" ® False |

Bảng 4.6. So sánh các phương pháp gõ dấu

Cách khắc phục vấn đề này là chuẩn hoá dữ liệu từ cách gõ Unicode tổ hợp thành cách gõ Unicode dựng sẵn do nó phổ biến hơn.

* + Chuẩn hoá kiểu gõ dấu

Hiện nay, trong tiếng Việt tồn tại hai kiểu gõ dấu khác nhau được gọi là kiểu mới (ví dụ: oà, uý) và kiểu cũ (ví dụ: òa, úy). Điều này dẫn đến việc các từ có cùng ý nghĩa lại có cách viết khác nhau, gây khó khăn trong việc sử dụng dữ liệu văn bản trong các bài toán học máy. Do đó, việc tiêu chuẩn hóa một bộ gõ là cần thiết.

* **Tách từ**

Bài toán tách từ (tokenization) trong việc khai phá dữ liệu văn bản là quá trình phân tách một đoạn văn thành các từ hoặc các đơn vị ngữ cảnh nhỏ nhất. Mục tiêu của bài toán này là chia văn bản thành các phần tử nhỏ hơn để dễ dàng xử lý và phân tích ngữ cảnh của từng từ hoặc đơn vị [25]. Ví dụ:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Tiền xử lý dữ liệu là một bước quan trọng | | | | | | |
| tiền | xử lý | dữ liệu | là | một | bước | quan trọng |

Bảng 4.7. Tách từ tiếng Việt

Khác với tiếng Anh và một số ngôn ngữ khác, các từ được phân cách nhau bằng khoảng trắng [25] như "use some examples", "i love you", trong văn bản tiếng Việt, một đặc điểm đặc biệt là một từ có thể được hình thành từ nhiều tiếng khác nhau, ví dụ như "sử\_dụng", "bắt\_đầu",... Do đó, để đảm bảo ý nghĩa của từ được bảo toàn thì cần phải thực hiện quá trình tách từ.

* **Loại bỏ từ dừng**

Từ dừng (StopWords) là những từ phổ biến và thường xuất hiện trong ngôn ngữ tự nhiên nhưng không đóng góp nhiều vào ý nghĩa của văn bản. Các từ này thường là những từ phụ, mệnh đề, hoặc từ chức năng như "is", "that", "this" trong tiếng Anh hoặc "là", "và", "của" trong tiếng Việt.

Việc loại bỏ từ dừng trong bài toán phân loại văn bản là cần thiết vì những từ này không cung cấp thông tin quan trọng cho việc phân loại chủ đề của văn bản. Thay vào đó, các từ khác như các từ khóa chính hoặc từ có liên quan đến chủ đề sẽ giúp mô hình phân loại văn bản hiểu rõ hơn về nội dung và ý nghĩa của văn bản.

Có 2 phương pháp chính để loại bỏ từ dừng. Một là là dùng từ điển để loại bỏ những từ xuất hiện trong từ điển StopWords. Hai là dựa theo tần suất xuất hiện của từ trong bộ dữ liệu: Đếm số lần xuất hiện của mỗi từ trong tập dữ liệu, sau đó loại bỏ những từ xuất hiện quá thường xuyên (hoặc quá ít).

Sau hoàn thành tất cả các bước tiền xử lý, chúng em có được số lượng dữ liệu cho mỗi nhãn như sau:



Biểu đồ 4.1. Thống kê dữ liệu được sử dụng trên mỗi nhãn

Nhãn có số lượng mẫu lớn nhất: label\_tmdt với 2069 mẫu. Nhãn có số lượng mẫu ít nhất: label\_nghề\_nghiệp với 1083 mẫu. Tỷ lệ giữa nhãn có số lượng mẫu lớn nhất và nhãn có số lượng mẫu ít nhất: 2069 / 1083 ≈ 1.91.

Từ tỷ lệ này có thể thấy rằng mặc dù có sự chênh lệch, nhưng nó không quá lớn (chưa đến 2 lần). Mặc dù dữ liệu không cực kỳ mất cân bằng, nhưng sự chênh lệch giữa các nhãn vẫn đáng kể. Khi sử dụng dữ liệu này để huấn luyện mô hình, cần chú ý đến việc điều chỉnh để đảm bảo mô hình không bị thiên vị đối với các nhãn có số lượng mẫu lớn hơn.

Chúng em sẽ sử dụng thư viện Scikit-learn trong Python để tách dữ liệu làm 2 tập train/test riêng biệt theo tỉ lệ 80/20 như đã trình bày bên trên.

### Vector hoá dữ liệu và trích xuất đặc trưng

Để áp dụng các thuật toán vào các bài toán học máy, dữ liệu văn bản cần được chuyển đổi thành các vector số học. Trong nghiên cứu này, phương pháp TF-IDF được sử dụng để trích xuất các đặc trưng từ văn bản. Quá trình này được thực hiện bằng cách sử dụng các công cụ từ thư viện Scikit-learn [12]. TF-IDF là phương pháp giúp xác định những từ quan trọng trong mỗi tài liệu, đồng thời giảm thiểu tác động của các từ thông dụng nhưng không mang nhiều giá trị thông tin.

Chúng em sẽ chọn phương pháp TF-IDF cấp độ từ (Word level) cho bài toán phân loại trang web dựa vào nội dung vì:

* Đại diện rõ ràng về ngữ nghĩa: Cấp độ từ là đơn vị ngữ nghĩa cơ bản trong ngôn ngữ tự nhiên. Việc sử dụng TF-IDF ở cấp độ từ giúp mô hình nắm bắt được ngữ nghĩa của nội dung trang web một cách trực tiếp và rõ ràng. Còn đối với cách tiếp cập N-gram có thể tạo ra nhiều đặc trưng không cần thiết, và cấp độ ký tự (character level) có thể làm mất đi ngữ nghĩa và tạo ra không gian đặc trưng quá lớn.
* Hiệu quả về mặt tính toán: Số lượng từ trong một tài liệu thường nhỏ hơn số lượng N-gram hoặc ký tự, do đó không gian đặc trưng sẽ nhỏ hơn và tiết kiệm tài nguyên tính toán.
* Khả năng tổng quát hóa tốt: TF-IDF ở cấp độ từ có thể giúp mô hình tổng quát hóa tốt hơn, vì nó tập trung vào tần suất và mức độ quan trọng của từng từ trong ngữ cảnh tài liệu. N-gram có thể tạo ra các đặc trưng cụ thể hơn nhưng cũng dễ dẫn đến overfitting nếu dữ liệu huấn luyện không đủ lớn. Cấp độ ký tự có thể dẫn đến việc mô hình hóa các chi tiết quá nhỏ và bỏ qua thông tin ngữ nghĩa quan trọng.
* Giảm thiểu nhiễu và đặc trưng không cần thiết: Tập trung vào từ giúp loại bỏ các chi tiết không cần thiết và nhiễu có thể xuất hiện ở cấp độ ký tự.

### Áp dụng các mô hình phân loại

Trong nghiên cứu này, chúng em sử dụng các mô hình Support Vector Machine, Naive Bayes, và Logistic Regression từ thư viện scikit-learn [12]. Các chi tiết về các thư viện này được trình bày cụ thể trong sau:

|  |  |  |
| --- | --- | --- |
| Mô hình | Phương pháp | Thư viện Scikit-learn |
| Support Vector Machine | Support Vector Classifier | sklearn.svm, SVC |
| Naive Bayes | Multinomial Naive Bayes | sklearn.naive\_bayes, MultinomialNB |
| Hồi quy Logistic | LogisticRegression | sklearn.linear\_model, LogisticRegression |

Bảng 4.8. Các mô hình để áp dụng để phân loại văn bản

Chúng em sử dụng các mô hình trên cho bài toán phân loại trang web dựa trên nội dung do một số lý do:

* Đơn giản và hiệu quả: Cả ba thuật toán đều đơn giản để triển khai và hiệu quả trong nhiều trường hợp. Điều này làm cho chúng trở thành lựa chọn phổ biến cho các bài toán phân loại cơ bản.
* Xử lý dữ liệu văn bản: Cả Naive Bayes và Logistic Regression đều làm việc tốt với dữ liệu văn bản, như là nội dung của trang web. Chúng có khả năng xử lý vector từ văn bản một cách hiệu quả.
* Khả năng diễn giải: Logistic Regression có thể dễ dàng diễn giải vì nó tạo ra các hệ số cho mỗi biến đầu vào, cho phép hiểu được tầm quan trọng của mỗi đặc trưng trong việc phân loại.
* Tính linh hoạt: SVM có thể được sử dụng để phân loại dữ liệu phi tuyến, thông qua việc sử dụng các hàm nhân. Điều này có thể hữu ích khi cần xử lý các trường hợp phức tạp hơn.
* Hiệu suất: Mặc dù đơn giản, các thuật toán này thường có hiệu suất tốt trên các tập dữ liệu có kích thước nhỏ đến trung bình và dữ liệu không phức tạp.

### So sánh và đánh giá các mô hình

Để tìm ra mô hình tốt nhất cho bài toán phân loại nội dung web, chúng em sẽ tiến hành so sánh các mô hình với nhau dựa vào các yếu tố như hiệu suất và thời gian huấn luyện của mỗi mô hình.

#### So sánh các mô hình với tập kiểm thử

Trong bài toán với nhiều lớp và phân phối dữ liệu không đồng đều, cũng như đã thống kê bên trên, có sự chênh lệch khá lớn giữa nhãn có số lượng mẫu lớn nhất và nhãn có số lượng mẫu nhỏ nhất, nếu đánh giá theo độ chính xác (accuracy) chưa đủ để đánh giá hiệu suất của một mô hình. Trong trường hợp này chúng em sử dụng ma trận nhẫm để biểu diễn kết quả cụ thể mỗi nhãn được phân loại như thế nào, nhãn nào được phân loại đúng nhiều nhất, và dữ liệu thuộc nhãn nào thường bị phân loại nhầm vào lớp khác.

Bên cạnh đó, cần sử dụng cặp tham số precision và recall. Và đôi khi cặp tham số precision và recall không đủ để đánh giá một cách toàn diện, cần sử dụng F1-score.

* **Mô hình Multinomial Naive Bayes**

Các số tương đương với nhãn của hình trên được để ở phần phụ lục 2 bên dưới.

****

Biểu đồ 4.2. Ma trận nhầm lẫn của mô hình Multinomial Naive Bayes

Ma trận nhầm lẫn này cho thấy mô hình Naive Bayes có hiệu suất phân loại tốt trên một số nhãn nhưng lại gặp khó khăn với một số nhãn khác. Các nhãn có nhiều trường hợp nhầm lẫn cần được xem xét lại, có thể do đặc điểm của các nhãn này không phân biệt rõ ràng hoặc có sự tương đồng cao giữa các nhãn, dẫn đến nhầm lẫn. Điều này gợi ý rằng việc cải thiện đặc trưng của dữ liệu hoặc sử dụng mô hình phân loại khác có thể giúp cải thiện hiệu suất tổng thể.

****

Biểu đồ 4.3. Kết quả kiểm thử với mô hình Multinomial Naive Bayes

* Precision cao (≥ 0.9): Các nhãn như "Giải trí", "Khoa học", "Nghề nghiệp", "Thực phẩm", "Nông nghiệp" có precision rất cao, cho thấy mô hình này ít đưa ra kết quả dương tính giả (false positives).
* Recall cao (≥ 0.9): Các nhãn như "Công nghệ", "Chính trị" có recall rất cao, cho thấy mô hình ít bỏ sót các kết quả dương tính thực (false negatives).
* F1-score cao (≥ 0.85): Nhãn "Khoa học", "Nông nghiệp", "Sức khỏe", và "Thực phẩm" có F1-score cao, chứng tỏ mô hình hoạt động tốt và cân bằng giữa precision và recall.
* F1-score thấp: Nhãn "Tài liệu" có F1-score thấp nhất, cho thấy sự không cân bằng giữa precision và recall, mặc dù precision là 1.000 nhưng recall chỉ là 0.474. Lý do là vì mô hình bỏ sót nhiều trường hợp thuộc nhãn này, có thể do đặc điểm của nhãn không rõ ràng hoặc không đủ dữ liệu.
* **Mô hình hồi quy logistic**



Biểu đồ 4.4. Ma trận nhầm lẫn của mô hình hồi quy logistic

Ma trận nhầm lẫn này cho thấy mô hình hồi quy logistic có hiệu suất phân loại tốt hơn trên một số nhãn so với mô hình Naive Bayes, với ít nhầm lẫn hơn ở các nhãn cụ thể. Tuy nhiên, một số nhãn như nhãn 2 (giải trí), nhãn 8 (tin tức), và nhãn 11(tài liệu) vẫn có nhiều nhầm lẫn. Để cải thiện hiệu suất, có thể cần phải tinh chỉnh mô hình hoặc cải thiện đặc trưng của dữ liệu đầu vào. Tổng quan, Logistic Regression có hiệu suất tốt và ít nhầm lẫn hơn trong việc phân loại so với Naive Bayes.



Biểu đồ 4.5. Kết quả kiểm thử với mô hình hồi quy Logistic

* Precision cao: Hầu hết các nhãn đều có precision cao (> 0.85), đặc biệt là các nhãn như "Giải trí", "Công nghệ", "Du lịch" và "Nông nghiệp" có precision trên 0.95. Điều này cho thấy mô hình Logistic Regression rất tốt trong việc đưa ra kết quả chính xác và ít có false positives.
* Recall cao: Nhãn "Công nghệ" có recall cao nhất với 0.956, cho thấy mô hình này rất tốt trong việc nhận diện các mẫu dương tính, ít bỏ sót các kết quả thực sự dương tính (false negatives).
* F1-score cao: Nhiều nhãn có F1-score rất cao, chứng tỏ sự cân bằng tốt giữa precision và recall. Các nhãn như "Công nghệ", "Giải trí", và "Khoa học" có F1-score gần hoặc trên 0.93.
* F1-score thấp nhất: Nhãn "Tin tức" có F1-score thấp nhất với 0.764, điều này cho thấy mô hình gặp khó khăn trong việc phân loại chính xác nhãn này, mặc dù precision và recall không quá thấp.
* **Mô hình SVM**



Biểu đồ 4.6. Ma trận nhầm lẫn của mô hình SVM

Ma trận nhầm lẫn này cho thấy, mô hình phân loại bằng SVM có hiệu suất khá tốt. Tuy nhiên có thể thấy nhãn 8 (tin tức) vẫn có nhiều nhầm lẫn.



Biểu đồ 4.7. Kết quả kiểm thử với mô hình SVM

* Precision cao: Hầu hết các nhãn đều có precision cao (> 0.90), đặc biệt là các nhãn như "Công nghệ", "Nghề nghiệp", và "Du lịch" có precision trên 0.97. Điều này cho thấy mô hình SVC rất chính xác trong việc phân loại đúng các nhãn.
* Recall cao: Các nhãn như "Công nghệ" và "Tin tức" có recall khá cao, đặc biệt là "Tin tức" với recall là 0.833, cho thấy mô hình này có khả năng nhận diện tốt các mẫu dương tính.
* F1-score cao: Nhiều nhãn có F1-score rất cao, ví dụ như "Công nghệ", "Du lịch", và "Nông nghiệp" đều có F1-score gần hoặc trên 0.95, cho thấy mô hình SVC hoạt động rất hiệu quả và cân bằng giữa precision và recall.
* F1-score thấp nhất: Nhãn "Tin tức" có F1-score thấp nhất với 0.711, điều này cho thấy mô hình có thể gặp khó khăn trong việc phân loại chính xác nhãn này, mặc dù recall khá cao.

Sau khi so sánh mô hình với các tập kiểm thử có thể thấy hiệu suất cả 3 mô hình khá tốt, tuy nhiên vẫn xảy ra dự đoán sai dựa trên các chỉ số đã thống kê. Nguyên nhân có thể kể đến là do:

* Sự cân bằng của dữ liệu: Các nhãn có số lượng mẫu không đều có thể dẫn đến các mô hình bị thiên vị về phía các nhãn có nhiều dữ liệu hơn. Ví dụ, nếu có nhiều mẫu thuộc nhãn "\_label\_\_chính\_phủ" nhưng ít mẫu thuộc nhãn "\_label\_\_tài\_liệu", mô hình có thể dễ dàng học được các đặc điểm của nhãn phổ biến và bỏ qua nhãn ít phổ biến.
* Đặc điểm của từng nhãn: Một số nhãn có các đặc điểm phân biệt rõ ràng, dễ dàng nhận biết, trong khi các nhãn khác có thể có các đặc điểm mơ hồ hoặc chồng chéo với các nhãn khác. Ví dụ, nhãn "\_label\_\_giải\_trí" có precision cao nhưng recall thấp, có thể do các mẫu thuộc nhãn này có các từ khóa đặc trưng dễ nhận biết, nhưng mô hình bỏ sót nhiều trường hợp khi các từ khóa không xuất hiện.
* Mô hình không đủ phức tạp: Mô hình không đủ phức tạp để nắm bắt được các đặc điểm của tất cả các nhãn, dẫn đến việc một số nhãn được phân loại tốt hơn các nhãn khác.
* Chất lượng của dữ liệu: Dữ liệu huấn luyện có thể chứa lỗi hoặc không đủ đại diện cho các nhãn khác nhau, dẫn đến việc mô hình không học được đầy đủ các đặc điểm cần thiết.

Hướng khắc phục các vấn đề trên:

* Cân bằng lại dữ liệu: Sử dụng các kỹ thuật như oversampling (tăng số lượng mẫu của các nhãn ít) hoặc undersampling (giảm số lượng mẫu của các nhãn nhiều) để cân bằng lại dữ liệu.
* Sử dụng mô hình phức tạp hơn: Thử nghiệm với các mô hình phức tạp hơn như ensemble methods (Random Forest, Gradient Boosting) hoặc deep learning models (LSTM, BERT) để cải thiện khả năng phân loại.
* Cải thiện chất lượng dữ liệu: Làm sạch dữ liệu để loại bỏ các mẫu không chính xác hoặc nhiễu. Thu thập thêm dữ liệu để đảm bảo đại diện đầy đủ cho tất cả các nhãn.

#### So sánh F1-score các mô hình

Chúng em sử dụng chỉ số F1-score vì nó là một chỉ số đáng tin cậy khi cần so sánh và lựa chọn giữa các mô hình học máy. Nó cung cấp một cái nhìn tổng quan và hài hòa về hiệu suất của mô hình bằng cách cân bằng giữa precision và recall. Giúp đánh giá hiệu suất tổng thể của mô hình mà không thiên vị bất kỳ chỉ số riêng lẻ nào.

Điểm F1 cao thường biểu thị hiệu suất cân bằng tốt, chứng minh rằng mô hình có thể đồng thời đạt được độ chính xác cao và độ thu hồi cao. Điểm F1 thấp thường biểu thị sự đánh đổi giữa độ thu hồi và độ chính xác, ngụ ý rằng mô hình gặp khó khăn trong việc đạt được sự cân bằng đó. Theo nguyên tắc chung, giá trị F1-score có thể được diễn giải như sau [26]:

|  |  |
| --- | --- |
| Chỉ số F1-score | Hiệu suất |
| >0.9 | Rất tốt |
| 0.8 - 0.9 | Tốt |
| 0.5 - 0.8 | Ổn |
| <0.5 | Không tốt |

Bảng 4.9. Giá trị F1-score



Biểu đồ 4.8. F1-score các mô hình theo từng thể loại

Nhận xét: Hiệu suất cao nhất là SVM, nó thể hiện tốt nhất trên hầu hết các loại dữ liệu. Tiếp đến là Logistic Regression, cũng có hiệu suất cao nhưng không bằng SVM.

Hiệu suất thấp nhất là Multinomial Naive Bayes, nó có F1-score thấp nhất trong hầu hết các trường hợp, điều này có thể do giả định độc lập giữa các đặc trưng không phù hợp với các loại dữ liệu được sử dụng.

#### So sánh thời gian huấn luyện

Chúng em sử dụng hàm thời gian trong Python để tính thời gian (giây) từ lúc bắt đầu đến lúc hoàn thành quá trình huấn luyện:



Biểu đồ 4.9. So sánh thời gian huấn luyện mô hình

Nhận xét: Mặc dù SVM có thể cung cấp hiệu suất tốt, nhưng thời gian huấn luyện của nó là rất lớn, đặc biệt trên các tập dữ liệu lớn. Trong khi đó, Naive Bayes và Logistic Regression có thời gian huấn luyện ngắn hơn và vẫn cung cấp hiệu suất tốt trên tập dữ liệu.

Như vậy để chọn ra mô hình nào có khả năng phân loại tốt nhất cần xét đến 2 tiêu chí là hiệu suất và thời gian. Nếu ưu tiên hiệu suất và thời gian không phải là yếu tố quan trọng thì mô hình SVM là lựa chọn tốt nhất. Nếu muốn cân bằng giữa hiệu suất và thời gian thì lựa chọn mô hình hồi quy logistic.

# KẾT LUẬN

* **Đánh giá**

Phần Web Crawling: Chúng em đã áp dụng thuật toán DFS để lấy dữ liệu theo chiều sâu từ các trang web. Mặc dù đã thành công trong việc lấy nội dung tự động từ URL bằng Python, nhưng còn tồn đọng một số hạn chế. Thời gian chờ vẫn còn lâu và có một số trang web mà không thể gửi request tới.

Phần mô hình phân loại: Kết quả khá tốt từ cả ba mô hình phân loại: Naive Bayes, Hồi quy Logistic và SVM, khi áp dụng các kỹ thuật tiền xử lý và rút trích đặc trưng. Mô hình đã cho kết quả phân loại khá chính xác. Tuy nhiên, vẫn còn hạn chế từ nguồn dữ liệu chưa được dồi dào và phong phú, cũng như hạn chế trong việc xử lý cho tiếng Việt.

* **Kết luận**

Dựa trên đánh giá trên, có thể kết luận rằng việc áp dụng các kỹ thuật khai thác dữ liệu văn bản trong phân loại nội dung trang web có tiềm năng và hiệu quả. Mặc dù còn một số hạn chế nhưng đã đạt được những kết quả khả quan, đặc biệt là trong việc phân loại nội dung trang web.

* **Hướng phát triển**

Các hướng phát triển tiếp theo có thể bao gồm:

* Nâng cao hiệu suất của phần web crawling bằng cách tối ưu hóa thời gian chờ và xử lý các trang web khó khăn hơn.
* Mở rộng nguồn dữ liệu để cải thiện hiệu suất và đa dạng hóa kết quả phân loại.
* Nghiên cứu và phát triển các kỹ thuật xử lý dữ liệu tiếng Việt để cải thiện khả năng phân loại cho ngôn ngữ này.
* Nghiên cứu các thuật toán giảm chiều dữ liệu để cải thiện tài nguyên cũng như thời gian phân loại.

Những hướng phát triển này sẽ giúp cải thiện và mở rộng ứng dụng của phương pháp khai thác dữ liệu văn bản trong việc phân loại nội dung trang web.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | K. Kumari, "How to Classify Web Pages Using Machine Learning?," 10 3 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2023/03/how-to-classify-web-pages-using-machine-learning/#Why\_do\_we\_need\_Web\_Page\_Classification?. [Accessed 30 4 2024]. |
| [2] | IBM, "What is text mining?," [Online]. Available: https://www.ibm.com/topics/text-mining. [Accessed 24 4 2024]. |
| [3] | D. Hoàng Văn , "Khai phá dữ liệu Web bằng kỹ thuật phân cụm," Hà Nội, 2007. |
| [4] | F. S. Gharehchopogh and Z. A. Khalifelu, "Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing," *International Conference on Application of Information and Communication Technologies,* pp. 1-4, 2011. |
| [5] | D. Xuân, "Tự học ML | Giới thiệu về Machine Learning(Máy học) cực chi tiết," 17 12 2020. [Online]. Available: https://cafedev.vn/tu-hoc-ml-gioi-thieu-ve-machine-learningmay-hoc-cuc-chi-tiet/. [Accessed 24 4 2024]. |
| [6] | Vũ Hữu Tiệp, "Phân nhóm các thuật toán Machine Learning," 27 12 2016. [Online]. Available: https://machinelearningcoban.com/2016/12/27/categories/. [Accessed 29 4 2024]. |
| [7] | Rachel Wolff, "Sentiment Analysis & Machine Learning," 20 4 2020. [Online]. Available: https://monkeylearn.com/blog/sentiment-analysis-machine-learning/. [Accessed 24 4 2024]. |
| [8] | A. Rashid, U. Shoaib and M. Shahzadsarfraz, "KNOWLEDGE DISCOVERY IN DATABASE USING INTENTION MINING," vol. 28, 2016. |
| [9] | Nguyễn Trần Thiên Thanh and Trần Khải Hoàng, "Tìm hiểu các hướng tiếp cận cho bài toán phân loại văn bản và xây dựng ứng dụng phân loại tin tức báo điện tử," 2005. [Online]. Available: https://www.slideshare.net/slideshow/phan-loai-tintucbaodientu-40099422/40099422. [Accessed 29 4 2024]. |
| [10] | A. Anforderungen and J. Thorsten, "Estimating the Generalization Performance of an SVM Efficiently," 2000. |
| [11] | N. T. Bằng, T. Q. H. Nguyên and P. T. Đ. Khoa, "NGHIÊN CỨU CÁC MÔ HÌNH PHÂN LOẠI VĂN BẢN ĐỂ XÂY DỰNG CHATBOT TƯ VẤN TUYỂN SINH," *TẠP CHÍ KHOA HỌC VÀ CÔNG NGHỆ - ĐẠI HỌC ĐÀ NẴNG,* vol. 18, pp. 40-45, 2020. |
| [12] | Scikit-learn: Machine Learning in Python, "Naive Bayes," [Online]. Available: https://scikit-learn.org/stable/modules/naive\_bayes.html. [Accessed 16 5 2024 ]. |
| [13] | R. R. B. E. Frank, "Naive bayes for text classification with unbalanced classes," in *Knowledge Discovery in Databases: PKDD*, 2006, pp. 503-510. |
| [14] | Cox, D.R, Analysis of Binary Data, Routledge: London, UK, 2018. |
| [15] | Fan, R.E., Chang, K.W, Hsieh, C.J., Wang, X.R. and Lin, C.J., "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res,* vol. 9, p. 1871–1874, 2008. |
| [16] | G. Alexander, D. Lewis and D. Madigan, "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics,* vol. 49, 2007. |
| [17] | A. Juan and E. Vidal, "On the use of Bernoulli mixture models for text classification," *Pattern Recognition,* vol. 35, no. 12, pp. 2705-2710, 2002. |
| [18] | B. Krishnapuram, L. Carin, M. Figueiredo and A. Hartemink, "Sparse multinomial logistic regression: fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, pp. 957-968, 2005. |
| [19] | K. Huang, Unconstrained Smartphone Sensing and Empirical Study for Sleep Monitoring and Self-Management, 2015. |
| [20] | W. contributors, "Precision and recall," 26 4 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Precision\_and\_recall&oldid=1220938692. [Accessed 30 4 2024]. |
| [21] | Walber, Artist, *Precision and recall.* [Art]. 2021. |
| [22] | N. T. H. Giang, "K45\_Nguyen\_thi\_Huong\_Giang\_Thesis". |
| [23] | WebRank, "Công cụ Phân tích Traffic Website," 3F Solutions, 2020. [Online]. Available: https://webrank.vn/. [Accessed 19 5 2024]. |
| [24] | Wikipedia contributors , "Text normalization," 8 12 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Text\_normalization&oldid=1188865739. [Accessed 29 4 2024]. |
| [25] | H. T. Nguyễn, "Phân loại văn bản tự động bằng Machine Learning như thế nào?," 11 10 2018. [Online]. Available: https://viblo.asia/p/phan-loai-van-ban-tu-dong-bang-machine-learning-nhu-the-nao-4P856Pa1ZY3. [Accessed 24 4 2024]. |
| [26] | Thái Học, "Thời đại của Dữ liệu lớn (Big Data)," 4 1 2021. [Online]. Available: https://vienthongke.vn/thoi-dai-cua-du-lieu-lon-big-data/. |
| [27] | L. Bing, "What is Data Mining?," in *Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data*, 851 S. Morgan St. Chicago, Springer, 2011, p. 6. |
| [28] | M. A. Hearst, " In Proceedings of the 37th Annual meeting of the Association for Computational Linguistics," in *Untangling text data mining*, 1999, pp. 3-10. |
| [29] | DrDiane McDonald, Ursula Kelly, "Intelligent Digital Options," 2012. |
| [30] | E. Griffing, "What is Textual Data and Why Does It Matter to Customer Support Teams?," 23 10 2023. [Online]. Available: https://www.dashbot.io/blog/textual-data. [Accessed 5 4 2024]. |
| [31] | Đinh Điền and Hồ Bảo Quốc, "Vấn đề về ranh giới từ trong ngữ liệu song ngữ Anh - Việt," in *Báo cáo Hội thảo Khoa học "Bảo vệ và Phát triền tiếng Việt"*, Viện ngôn ngữ học, Hội ngôn ngữ học TPHCM, ĐH KHXN&NV - TPHCM, 2002, pp. 70-78. |
| [32] | Lê Thanh Hương and Bộ môn Hệ thống Thông tin Viện CNTT &TT – Trường Đ, "Tách từ tiếng Việt," [Online]. Available: https://users.soict.hust.edu.vn/huonglt/UNLP/3\_wordsegmentation.pdf. [Accessed 5 4 2024]. |
| [33] | Nguyễn Phú Phong, "Các phương diện hình thái học tiếng Việt," in *Kỷ yếu hội thảo quốc tế Việt Nam học lần thứ nhất* , ĐHQGHN, 1998, pp. 139-150. |
| [34] | Bích Phương, "Khách Australia ấn tượng với 'cả thế giới gia vị' của bún riêu," 2 4 2024. [Online]. Available: https://vnexpress.net/khach-australia-an-tuong-voi-ca-the-gioi-gia-vi-cua-bun-rieu-4729105.html. [Accessed 6 4 2024]. |
| [35] | Nguyễn Linh Giang and Nguyễn Mạnh Hiển, "Phân loại văn bản tiếng Việt với bộ phân loại vectơ hỗ trợ SVM," *Tạp chí CNTT&TT,* 2006. |
| [36] | Trần Cao Đệ and Phạm Nguyên Khang, "PHÂN LOẠI VĂN BẢN VỚI MÁY HỌC VECTOR HỖ TRỢ VÀ CÂY QUYẾT ĐỊNH," *Tạp chí Khoa học Đại học cần Thơ, (21a), 52-63,* pp. 52-53, 2012. |
| [37] | "What is Website Categorization?," Threat Media, [Online]. Available: https://threat.media/definition/what-is-website-categorization/#:~:text=It%20aids%20in%20threat%20investigation.&text=Know%20which%20domains%20have%20ties,cybersecurity%20posture%20and%20brand%20reputation.. [Accessed 7 4 2024]. |
| [38] | Wikipedia, "Quy tắc đặt dấu thanh của chữ Quốc ngữ," 18 2 2024. [Online]. Available: https://vi.wikipedia.org/wiki/Quy\_t%E1%BA%AFc\_%C4%91%E1%BA%B7t\_d%E1%BA%A5u\_thanh\_c%E1%BB%A7a\_ch%E1%BB%AF\_Qu%E1%BB%91c\_ng%E1%BB%AF. [Accessed 8 4 2024]. |
| [39] | Rathanak, "Into to Machine Learning: Supervised learning," 27 3 2017. [Online]. Available: https://viblo.asia/p/into-to-machine-learning-supervised-learning-vyDZOyRk5wj. [Accessed 24 4 2024]. |
| [40] | Falguni N. Patel and Neha R. Soni, "Text mining: A Brief survey," *International Journal of Advanced Computer Research,* vol. 2, no. 6, pp. 243-246, 2012. |
| [41] | Do Dang Hung, "Tiền xử lí dữ liệu văn bản với NLTK," 14 9 2021. [Online]. Available: https://viblo.asia/p/tien-xu-li-du-lieu-van-ban-voi-nltk-Az45b0LgZxY. [Accessed 24 4 2024]. |
| [42] | Huynh Minh Tan, "Phân loại văn bản (Text Classification)," 2019. [Online]. Available: https://github.com/huynhminhtan/thuattoanthongminh?tab=readme-ov-file. [Accessed 24 4 2024]. |
| [43] | D. Cowan, "Confusion Matrix," [Online]. Available: https://www.ml-science.com/confusion-matrix. [Accessed 30 4 2024]. |
| [44] | G. Aimee, Using Demographic Variables and In-College Attributes to Predict Course-Level Retention for Community College Spanish Students, 2016. |
| [45] | "en.ryte.com/wiki/Crawler," [Online]. Available: https://en.ryte.com/wiki/Crawler. |

# PHỤ LỤC

## Phụ lục 1: Tập dữ liệu thô (chưa xử lý) được biểu diễn như sau:

*Vì tập đữ liệu lớn, nên chỉ có thể minh họa 1 phần*

|  |
| --- |
| \_\_label\_\_chính\_phủ THƯ VIỆN PHÁP LUẬT ...loại rủi ro pháp lý, nắm cơ hội làm giàu... Các gói dịch vụ Chính sách Pháp luật mới Pháp Luật Dân Luật Liên hệ Danh mục Văn bản và Tra cứu 1. Tra cứu Văn Bản 2. Tra cứu Dự thảo 3. Văn bản mới ban hành 4. Tra cứu Tiêu Chuẩn 5. ICS (Phân loại Quốc tế về Tiêu chuẩn) 6. Tra cứu Công Văn 7. Thuật ngữ pháp lý 8. Tra cứu Bản án 9. Luật sư toàn quốc 10. Website ngành luật 11. Hỏi đáp pháp luật 12. Tra cứu Xử phạt vi phạm hành chính 13. Bảng giá đất 14. Tra cứu mẫu hợp đồng 15. Tra cứu biểu mẫu 16. Tra cứu mức phí, lệ phí 17. Tra cứu diện tích tách thửa đất ở 18. Tra cứu xử phạt giao thông đường bộ 19. Biểu thuế WTO 20. Biểu thuế ASEAN 21. PHÁP LUẬT DOANH NGHIỆP Thông tin và bài viết 22. Pháp lý phát sinh do Covid-19 23. Hiệp định CPTPP 24. HTKK - Hỗ Trợ Kê Khai 25. Tính thuế thu nhập cá nhân Online 26. Văn bản TP. Hồ Chí Minh Sơ đồ WebSite Giới thiệu Hướng dẫn sử dụng … [còn tiếp]  \_\_label\_\_cộng\_đồng Thứ tư, 24/04/2019 11:03 (GMT +7) Kinh Phật Vật phẩm Phật giáo Mới nhất Đạo Phật Online | Hà Nội 34°C / 57% Tịnh Độ tông Chuyên mục Trang chủ Tin tức Trong nước Quốc tế Tin Phật sự Xiển dương Đạo pháp Đạo Phật trong trái tim tôi Media Ảnh Audio Môi trường Lời Phật dạy Sống an vui Thuần chay Đức Phật Nhân vật Phật giáo Sách Phật giáo Giáo hội Các kỳ đại hội Hệ thống tổ chức Văn bản VESAK 2019 40 năm … [còn tiếp]  \_\_label\_\_doanh\_nghiệp Giới thiệu Tin tức Hoạt động kinh doanh Tin tức chuyên ngành Tài chính doanh nghiệp Hoạt động trách nhiệm xã hội Dịch vụ Khai thác cảng Dịch vụ logistics Vận tải và các ngành … [còn tiếp]  \_\_label\_\_đời\_sống TRANG THÔNG TIN KIẾN THỨC VÀ KỸ NĂNG DÀNH CHO CHA MẸ #Fanpage #Diễn đàn #Tin nóng Diễn đàn Thời sự Đời sống Làm Cha Mẹ Giải trí Sức khỏe Tổ ấm Đẹp Tâm sự Ăn chơi Khám phá Clip Thời trang 10 cách mặc chân váy giúp phong cách đi du lịch nổi bật hơn 10:46 16/04/2024 (lamchame.vn) - Chị em nên tham khảo … [còn tiếp]  \_\_label\_\_động\_vật Mua Hàng Online Liên Hệ Ngay Hotline: 090.6789.543 Giới thiệu Tuyển dụng TƯƠI NGON TỪ NGUỒN Từ khóa Hot king crab tôm sú cá mú cá mú trân châu cua hoàng đế tôm hùm canada tôm hùm alaska ốc vòi voi cua tuyết cua nâu cua nâu ireland cua dungeness cua bắc mỹ cua gạch cua thịt cua … [còn tiếp]  \_\_label\_\_du\_lịch Vé máy bay Khách sạn Tour du lịch Combo khuyến mại Vinpearl Du thuyền Hạ Long Vé vui chơi 1900 2605 Tổng đài: 1900 2605 Hà Nội: 024 73072605 Hồ Chí Minh: 028 73072605 Khuyến mại Điểm đến Săn vé máy bay Kinh nghiệm du lịch Văn hóa-Ẩm thực Tin tức đó đây Tin tức BestPrice Trang … [còn tiếp]  \_\_label\_\_giải\_trí Game Vui ADVERTISEMENT Tiny Fishing Pixel Craft - Hide and Seek Parkour World World Challenge Cuộc Đua Vui Nhộn Mine … [còn tiếp]  \_\_label\_\_khoa\_học Lớp 12 Toán học 12 SGK Toán SGK Toán Nâng cao SBT Toán SBT Toán Nâng cao Trắc nghiệm Toán Đề thi, đề kiểm tra Toán SGK Toán - Kết nối tri thức SGK Toán - Chân trời sáng tạo >> Xem thêm Ngữ văn 12 Soạn văn siêu ngắn Soạn văn chi tiết Tác giả - Tác phẩm Văn mẫu Luyện dạng đọc hiểu … [còn tiếp]  \_\_label\_\_nghề\_nghiệp Skip to content Home Việc làm Hồ sơ xin việc Sinh ra từ Làng Tự làm chủ Menu Home Việc làm Hồ sơ xin việc Sinh ra từ Làng Tự làm … [còn tiếp]  \_\_label\_\_nhà\_cửa Vietceramic Catalogue 0797.555.299 Navigation Gạch Thiết Bị Phòng Tắm SÀN GỖ Vietceramics Dự Án Tin Tức SHOWROOM Ý Tưởng Khuyến mãi … [còn tiếp]  \_\_label\_\_sức\_khỏe Tải app Thuốc Dân Tộc MIỄN PHÍ × Chuyên mục 10 Tin mới Trung tâm Nghiên cứu và Ứng dụng Thuốc Dân Tộc Tìm kiếm Liên hệ Tìm kiếm Da liễu Mề đay … [còn tiếp]  \_\_label\_\_tài\_liệu Trang chủ Giới thiệu Khát Vọng Tiên Phong Giới Thiệu Dịch Thuật Pro Gặp Gỡ Giám Đốc Công Ty Dịch vụ Dịch Thuật Chuyên Nghiệp Phiên Dịch Chuyên … [còn tiếp]  \_\_label\_\_thể\_thao Trang chủ MỚI NHẤT Tin Chuyển Nhượng Bóng Đá Anh Premier League FA Cup EFL Cup Tin Khác Bóng đá Tây Ban Nha La … [còn tiếp]  \_\_label\_\_thực\_phẩm WEB NAUAN .VN Cách nấu Dịp nấu Địa điểm ăn uống Mẹo vặt Làm đẹp Đọc truyện Recipes Mẹ & Bé … [còn tiếp]  \_\_label\_\_tin\_tức Xã hội Tin nóng Đời sống Mua sắm Bảo vệ người tiêu dùng Chuyện lạ Tổ ấm Nhà đẹp Phong thủy Không gian xanh Mẹo vặt … [còn tiếp]  \_\_label\_\_TMDT Quan hệ cổ đông (IR) Cửa hàng 1800 54 54 57 Đặt lịch hẹn Tài khoản của tôi Giỏ hàng 0 Trang sức Chủng loại Nhẫn Dây chuyền Mặt dây chuyền Bông tai Lắc Vòng … [còn tiếp]  \_\_label\_\_xe\_cộ Giá xe ô tô Ô Tô Mới Xe Vinfast Xe Toyota Xe Hyundai Xe Kia Xe Honda Xe Ford Xe Mazda Xe Mercedes Đánh giá xe Ô tô cũ Phụ kiện ô tô An toàn ô tô Dụng … [còn tiếp] |

## Phụ lục 2: Nhãn tương ứng với thứ tự trong ma trận nhầm lẫn

|  |  |
| --- | --- |
| STT | Nhãn |
|  | \_label\_\_chính\_phủ |
|  | \_label\_\_cộng\_đồng |
|  | \_label\_\_giải\_trí |
|  | \_label\_\_khoa\_học |
|  | \_label\_\_nghề\_nghiệp |
|  | \_label\_\_sức\_khỏe |
|  | \_label\_\_thể\_thao |
|  | \_label\_\_thực\_phẩm |
|  | \_label\_\_tin\_tức |
|  | \_label\_\_tmdt |
|  | \_label\_\_tài\_chính |
|  | \_label\_\_tài\_liệu |
|  | \_label\_\_xe\_cộ |
|  | \_label\_\_động\_vật |
|  | \_lable\_\_du\_lịch |