**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

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



**TRƯƠNG CÔNG HẢI**

**DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI**

**DỰA TRÊN NỘI DUNG BÀI VIẾT**

**LUẬN VĂN THẠC SĨ KỸ THUẬT**

***(Theo định hướng ứng dụng)***

HÀ NỘI - 2016

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

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



**TRƯƠNG CÔNG HẢI**

**DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI**

**DỰA TRÊN NỘI DUNG BÀI VIẾT**

|  |  |  |
| --- | --- | --- |
| **CHUYÊN NGÀNH** | **:** | **KHOA HỌC MÁY TÍNH** |
| **MÃ SỐ** | **:** | **60.48.01.01** |

**LUẬN VĂN THẠC SĨ KỸ THUẬT**

***(Theo định hướng ứng dụng)***

NGƯỜI HƯỚNG DẪN KHOA HỌC

**PGS.TS. TỪ MINH PHƯƠNG**

HÀ NỘI - 2016

LỜI CAM ĐOAN

Luận văn này là thành quả của quá trình học tập nghiên cứu của tôi cùng sự giúp đỡ, khuyến khích của các quý thầy cô sau 2 năm tôi theo học chương trình đào tạo Thạc sĩ, chuyên ngành Khoa học máy tính của trường Học viện Công nghệ Bưu chính Viễn thông.

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. Nội dung của luận văn có tham khảo và sử dụng một số thông tin, tài liệu từ các nguồn sách, tạp chí được liệt kê trong danh mục các tài liệu tham khảo và được trích dẫn hợp pháp.

|  |  |
| --- | --- |
|  | Tác giả  *(Ký và ghi rõ họ tên)*  **Trương Công Hải** |

LỜI CÁM ƠN

Tôi xin gửi lời cảm ơn và tri ân tới các thầy cô giáo, cán bộ của Học viện Công nghệ Bưu chính Viễn thông đã giúp đỡ, tạo điều kiện tốt cho tôi trong quá trình học tập và nghiên cứu chương trình Thạc sĩ.

Tôi xin gửi lời cảm ơn sâu sắc tới PGS.TS. Từ Minh Phương đã tận tình hướng dẫn, giúp đỡ và động viên tôi để hoàn thành tốt nhất Luận văn “DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI DỰA TRÊN NỘI DUNG BÀI VIẾT“.

Do vốn kiến thức lý luận và kinh nghiệm thực tiễn còn ít nên luận văn không tránh khỏi những thiếu sót nhất định. Tôi xin trân trọng tiếp thu các ý kiến của các thầy, cô để luận văn được hoàn thiện

Trân trọng cám ơn.

|  |  |
| --- | --- |
|  | Tác giả  *(Ký và ghi rõ họ tên)*  **Trương Công Hải** |

MỤC LỤC

[MỞ ĐẦU 1](#_Toc482277124)

[1 Tính cấp thiết của đề tài 1](#_Toc482277125)

[2 Tổng quan về vấn đề nghiên cứu 1](#_Toc482277126)

[3 Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu 2](#_Toc482277127)

[4 Cấu trúc luận văn 3](#_Toc482277128)

[Chương 1 - GIỚI THIỆU BÀI TOÁN XÁC ĐỊNH GIỚI TÍNH 6](#_Toc482277129)

[1.1 Giới thiệu bài toán xác định giới tính. 6](#_Toc482277130)

[1.2 Các phương pháp xác định giới tính 8](#_Toc482277131)

[1.3 Các phương pháp xác định giới tính dựa trên các bài biết của người dùng 10](#_Toc482277132)

[*1.3.1* *Dự đoán giới tính dựa trên nội dung bình luận trên Youtube* 10](#_Toc482277133)

[*1.3.2* *Dự đoán giới tính sử dụng bài viết từ blog* 15](#_Toc482277134)

[*1.3.3* *Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy* 15](#_Toc482277135)

[1.4 Kết luận chương 16](#_Toc482277136)

[Chương 2 - KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG MÃ HỘI 18](#_Toc482277137)

[2.1 Phạm vi áp dụng 18](#_Toc482277140)

[2.2 Các đặc trưng sử dụng 30](#_Toc482277141)

[*2.2.1* *Đặc trưng text* 30](#_Toc482277142)

[*2.2.2* *Đặc trưng ký hiệu đặc biệt* 31](#_Toc482277143)

[2.3 Mô hình phân loại SVM 31](#_Toc482277144)

[*2.3.1* *Giới thiệu kỹ thuật học máy SVM* 32](#_Toc482277145)

[*2.3.2* *Áp dụng kỹ thuật học máy SVM vào dự đoán giới tính* 33](#_Toc482277146)

[2.4 Kết luận chương 35](#_Toc482277147)

[Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ 36](#_Toc482277148)

[3.1 Thu thập và mô tả dữ liệu đầu vào 36](#_Toc482277149)

[3.2 Các tiêu chuẩn dùng để đánh giá 40](#_Toc482277150)

[3.3 Phương pháp thực nghiệm 40](#_Toc482277151)

[3.4 Kết quả thực nghiệm 41](#_Toc482277152)

[2.5 So sách với một số phương pháp khác 41](#_Toc482277153)

[2.6 Độ phức tạp và thời gian thực hiện phương pháp 41](#_Toc482277154)

[3.7 Kết luận chương 42](#_Toc482277155)

[KẾT LUẬN VÀ KIẾN NGHỊ 43](#_Toc482277156)

[DANH MỤC TÀI LIỆU THAM KHẢO 45](#_Toc482277157)

[PHỤ LỤC 46](#_Toc482277158)

DANH MỤC TỪ VIẾT TẮT

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Ý nghĩa tiếng Anh** | **Ý nghĩa tiếng Việt** |
| 1 | SVM | Support vector machine | Máy vector hỗ trợ | |
| 2 | NB | Naïve Bayes |  | |
| 3 | kNN | K–Nearest Neighbor | K – Láng giềng gần nhất | |
| 4 | TF | Term Frequency | Tần số xuất hiện của 1 từ | |
| 5 | IDF | Inverse Document Frequency | Tần số nghịch của 1 từ trong tập văn bản | |
| 6 | RSS | Really Simple Syndication | Định dạng tài liệu | |

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

1 Tính cấp thiết của đề tài

Ngày này, với sự phát triển của các mạng xã hội như: Facebook, Twitter, Youtube…Với số lượng lớn người dùng và liên tục cập nhật thông tin liên quan đến mọi vấn đề như đời sống, xã hội, kinh tế, giải trí… Việc xác định chính xác thông tin cá nhân của người dùng được nhiều tổ chức, công ty, cá nhân quan tâm tới. Trong nhiều trường hợp những thông tin người dùng không cập nhật vào hồ sơ cá nhân hay do người dùng không muốn người khác thấy được vì vậy chúng ta không có đủ thông tin cần thiết. Trong đó có thông tin quan trọng đó là giới tính người dùng. Dựa vào một số nghiên cứu đã có, chúng ta có thể xác định được giới tính người dùng dựa trên văn phong, cách dùng từ, diễn đạt trong các nội dùng bài viết cùng với việc áp dụng mô hình học máy được huấn luyện trên các bài viết đã biết giới tính của người dùng. Việc xác định rõ giới tính người dùng sẽ đưa ra các số liệu thông kế, các kế hoạch quảng cáo của các công ty, tổ chức cũng như cung cấp các dịch vụ phù hợp với giới tính người dùng trên mạng xã hội nói riêng và mạng internet nói chúng.

Vì vậy, tôi đã lựa chọn đề tài luận văn thạc sỹ là ***“Dự đoán giới tính người dùng mạng xã hội dựa trên nội dung bài viết”.***

2 Tổng quan về vấn đề nghiên cứu

Sự ra đời và phát triển nhanh chóng của World Wide Web đã tạo điều kiện thuận lợi cho việc phân phối và chia sẻ thông tin trên Internet, do đó dẫn đến bùng nổ thông tin cả về số lượng, chất lượng và các chủ đề thông tin trên đó. Trong luận văn này, em tập trung nghiên cứu về phương pháp phân loại văn bản tiếng Việt, cụ thể là thu thập tự động nội dung của Website, thực hiện và tối ưu các bước tiền xử lý, sau đó phân loại nội dung theo một số lĩnh vực nghề nghiệp bằng một kỹ thuật phân loại khá tốt hiện nay đó là máy vector hỗ trợ SVM.

Công tác phân loại luôn được các thư viện và cơ quan thông tin trên thế giới hết sức quan tâm. Theo Yang & Xiu, 1999, “Việc phân loại văn bản tự động là việc gán các nhãn phân loại lên một văn bản mới dựa trên mức độ tương tự của văn bản đó so với các văn bản đã được gán nhãn trong tập huấn luyện”.

Từ trước đến nay, phân loại văn bản tự động trong tiếng Anh đã có rất nhiều công trình nghiên cứu và đạt được kết quả đáng khích lệ. Dựa trên các thống kê của Yang & Xiu (1999), một số phương pháp phân loại thông dụng hiện nay là: Support Vector Machine - Joachims, 1998; k-Nearest Neighbor - Yang, 1994; Linear Least Squares Fit - Yang and Chute, 1994; Neural Network - Wiener et al, 1995; Naïve Bayes - Baker and Mccallum, 2000. Các phương pháp trên đều dựa vào xác suất thống kê hoặc thông tin về trọng số của từ trong văn bản.

3 Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu

***Mục đích nghiên cứu:***

* Tìm hiểu về bài toán xác định giới tính người dùng.
* Các phương pháp dự đoán giới tính đã có.
* Tìm hiểu các kỹ thuật học máy và học máy SVM.
* Áp dụng và thực nghiệm dự đoán giới tính người dụng dựa trên nội dung bài viết.

***Đối tượng nghiên cứu:***

* Nội dung bài viết của người dùng trên mạng xã hội ở Việt Nam.
* Bài toán tiền xử lý văn bản.
* Bài toán phân lớp.
* Phương pháp SVM trong phân loại.

***Phạm vi nghiên cứu:***

* Nghiên cứu tổng quan về xử lý ngôn ngữ tự nhiên và bài toán dự đoán giới tính.
* Sau đó, luận văn đi sâu tìm hiểu ứng dụng phương pháp học máy SVM trong việc dự đoán giới tính người dùng trên mạng xã hội Facebook.

***Phương pháp nghiên cứu:***

* Tìm hiểu các phương pháp dự đoán giới tính hiện này đang có để xác định những điểm mạnh và hạn chế của các phương pháp đó.
* Các đặc trưng của nội dung bài viết ảnh hưởng đến việc xác định giới tính người dùng mạng xã hội.
* Tìm hiểu các kỹ thuật học máy hiện nay, nhưng ưu nhược điểm của các kỹ thuật học máy đó trong việc dự đoán giới tính người dùng mạng xã hội.
* Tìm kiểu chi tiết kỹ thuật học máy SVM và ứng dụng hiện này của phương pháp này.
* Tìm hiểu về mạng xã hội Facebook. Xác định những yếu tốt ảnh hưởng đến giới tính người dùng và các cách để lấy dữ liệu để nghiên cứu và thực nghiệm.
* Tìm hiểu phương pháp phân loại văn bản – Text categorization sẽ được sử dụng để phân tích nội dung bài viết hỗ trợ dự đoán giới tính người dùng trên mạng xã hội Facebook.

4 Cấu trúc luận văn

Với mục tiêu đặt ra như vậy, nội dung và kết quả của luận văn được trình bày qua 3 chương như sau:

CHƯƠNG 1: GIỚI THIỆU BÀI TOÁN XÁC ĐỊNH GIỚI TÍNH

***Giới thiệu chương:***

Giới thiệu về bài toán xác định giới tính và áp dụng để xác định giới tính người dùng trên các mạng xã hội hiện nay như: Facebook, Twitter, Youtube... Phần này cũng đưa ra các phương pháp xác định giới tính đã có trong đó chú ý đến phương pháp dựa trên nội dung bài viết.

***Nội dung chương:***

* Giới thiệu bài toán dự đoán giới tính
* Các phương pháp dự đoán giới tính
* Các phương pháp xác định giới tính dựa trên các bài biết của người dùng

***Kết luận chương:***

Chương này đã giới thiệu tổng quan về bài toán dự đoán giới tính, các phương pháp dự đoán và một số phương pháp xác định giới tính đã được nghiên cứu hiện này là tiền đề để phát triền luận văn.

CHƯƠNG 2: KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG MÃ HỘI

***Giới thiệu chương:***

Trình bày tổng quan về phương pháp học máy, một số kỹ thuật đã và đang được sử dụng trong việc phân tích người dùng mạng mã hội hiện nay. Dựa vào những đặc trưng nội dung bài viết khác nhau, sử dụng phương pháp học máy SVM để dự đoán giới tính người dùng.

***Nội dung chương:***

* Chi tiết về phạm vi áp dụng.
* Đưa ra các đặc trưng sử dụng để dự đoán giới tính người dùng mạng xã hội.
* Trình bày chi tiết thuật toán Máy vector hỗ trợ SVM.

***Kết luận chương:***

Chương này đã giới thiệu chi tiết về phạm vi áp dụng thực nghiệm và đưa ra các đặc trưng sử dụng vào bài toán dự đoán giới tính. Sau đó, chương 2 cũng trình bày chi tiết về kỹ thuật SVM là cơ sở lý thuyết để áp dụng vào thực hiện việc huấn luyện và dự đoán dựa trên nội dung bài viết trên mạng xã hội Facebook.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

***Giới thiệu chương:***

Xây dựng các bước để thực nghiệm cho bài toán dự đoán giới tính người dùng mạng xã hội dựa trên nội dung bài viết. Lấy bộ dữ liệu đầu từ các bài viết trên mạng xã hội đã biết nhãn giới tính, sử dụng thư viện LibSVM có hỗ trợ kỹ thuật học máy SVM. Sau đó đưa bộ dữ liệu vào huấn luyện và sử dụng để dự đoán với bộ dữ liệu chưa có nhãn, đưa ra tỉ lệ và độ chính xác của phương pháp dự đoán dựa trên nội dung bài viết. Đánh giá kết quả so sách với các phương pháp dự đoán khác.

***Nội dung chương:***

* Xây dựng bộ dữ liệu huấn luyện và thử nghiệm
* Giai đoạn dự đoán
* Kết luận và hướng phát triển

***Kết luận chương:***

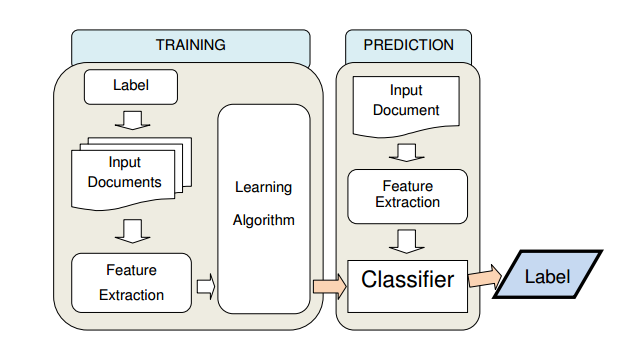
Dữ liệu huấn luyện và kiểm thử được xây dựng. Sau đó tiến hành giai đoạn dự đoán và đánh giá kết quả.

Chương 1 - GIỚI THIỆU BÀI TOÁN XÁC ĐỊNH GIỚI TÍNH

1.1 Giới thiệu bài toán xác định giới tính.

Phân loại văn bản (hay Text Categorization hoặc Document Classificant) là quá trình gán các văn bản vào một hay nhiều chủ đề đã biết trong một tập hữu hạn các chủ đề đã được xác định từ trước. Ví dụ một bài báo trong một trang web có thể thuộc một hoặc một vài chủ đề nào đó (như thể thao, giáo dục, pháp luật, công nghệ thông tin,…). Việc phân loại có thể được tiến hành một cách thủ công: đọc nội dung của từng văn bản và gán nó vào một lớp nào đó. Tuy nhiên, đối với hệ thống gồm rất nhiều văn bản thì phương pháp này sẽ tốn rất nhiều thời gian và công sức. Do vậy cần phải có phương pháp tự động để phân loại văn bản. Phương pháp này giúp cho việc lưu trữ và truy vấn tài liệu dễ dàng hơn.

Dưới đây là hình vẽ mô tả quy trình của bài toán phân loại văn bản:



Hình 1.1 Quy trình phân loại văn bản

Để tiến hành phân loại văn bản nói chung, chúng ta sẽ thực hiện các bước như sau:

* Bước 1: Xây dựng bộ dữ liệu huấn luyện dựa vào tài liệu văn bản đã được phân loại sẵn. Tiến hành học cho bộ dữ liệu, xử lý và thu thập được dữ liệu của quá trình học là các đặc trưng riêng biệt cho từng chủ đề.
* Bước 2: Dữ liệu cần phân loại được xử lý, rút ra đặc trưng kết hợp với đặc trưng được học trước đó để phân loại và rút ra kết quả.

Đặc điểm nổi bật của bài toán này là sự đa dạng của chủ đề văn bản và tính đa chủ đề của văn bản. Tính đa chủ đề của văn bản làm cho sự phân loại chỉ mang tính tương đối và có phần chủ quan, nếu do con người thực hiện có thể dễ bị nhập nhằng. Ví dụ có bài báo về tuyển dụng nhân viên bán thuốc tại một công ty dược phẩm, bài báo này có thể xếp vào chủ đề tuyển dụng lĩnh vực Y dược hoặc cũng có thể xếp vào chủ đề lĩnh vực Bán hàng.

Về bản chất, một văn bản là một tập hợp từ ngữ có liên quan với nhau tạo nên nội dung ngữ nghĩa của văn bản. Từ ngữ của một văn bản là đa dạng do tính đa dạng của ngôn ngữ (đồng nghĩa, đa nghĩa, từ vay mượn nước ngoài,…) và số lượng từ cần xét là lớn. Ở đây cần lưu ý rằng, một văn bản có thể có số lượng từ ngữ không nhiều, nhưng số lượng từ ngữ cần xét là rất nhiều vì phải bao hàm tất cả các từ của ngôn ngữ đang xét.

Trên thế giới đã có nhiều công trình nghiên cứu đạt những kết quả khả quan, nhất là đối với phân loại văn bản tiếng Anh. Tuy vậy, các nghiên cứu và ứng dụng đối với văn bản tiếng Việt còn nhiều hạn chế do khó khăn về tách từ và câu. Có thể liệt kê một số công trình nghiên cứu trong nước với các hướng tiếp cận khác nhau cho bài toán phân loại văn bản, bao gồm: phân loại với máy học vectơ hỗ trợ, cách tiếp cận sử dụng lý thuyết tập thô, cách tiếp cận thống kê hình vị, cách tiếp cận sử dụng phương pháp học không giám sát và đánh chỉ mục, cách tiếp cận theo luật kết hợp. Theo các kết quả trình bày trong các công trình đó thì những cách tiếp cận nêu trên đều cho kết quả khá tốt. Tuy nhiên khó có thể so sánh các kết quả ở trên với nhau vì tập dữ liệu thực nghiệm của mỗi phương pháp là khác nhau.

Trong những năm gần đây, phương pháp phân loại sử dụng Máy vector hỗ trợ (SVM) được quan tâm và sử dụng nhiều trong những lĩnh vực nhận dạng và phân loại. SVM là một họ các phương pháp dựa trên cơ sở các hàm nhân (kernel) để tối thiểu hóa rủi ro ước lượng. Phương pháp SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn. Các thử nghiệm thực tế cho thấy, phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân loại văn bản cũng như trong nhiều ứng dụng khác (như nhận dạng chữ viết tay, phát hiện mặt người trong các ảnh, ước lượng hồi quy, ...). So sánh với các phương pháp phân loại khác, khả năng phân loại của SVM là tương đương hoặc tốt hơn đáng kể. Vì những lý do đó mà tôi đã chọn phương pháp này cho việc phân loại văn bản tiếng Việt, cụ thể thuật toán và ứng dụng sẽ được trình bày trong các chương sau.

1.2 Các phương pháp xác định giới tính

Hiện nay trên thế giới đã có rất nhiều công trình nghiên cứu về các phương pháp phân loại văn bản. Một số phương pháp cần kể đến là: Naïve Bayes, Support Vector Machine, K–Nearest Neighbor, Linear Least Squares Fit, Neural Network, Naïve Bayes, Centroid–Based… Điểm chung của các phương pháp này đều dựa vào xác suất thống kê hoặc dựa vào trọng số của các từ, cụm từ trong văn bản. Trong mỗi phương pháp đều có cách tính toán khác nhau, tuy nhiên các phương pháp này đều phải thực hiện một số bước chung như: mỗi phương pháp sẽ dựa vào thông tin về sự xuất hiện của các từ trong văn bản (tần số xuất hiện trong tập văn bản…) để biểu diễn thành dạng vector, sau đó tùy từng bài toán cụ thể mà chúng ta sẽ quyết định chọn áp dụng phương pháp nào, công thức tính toán nào cho phù hợp để phân loại tập văn bản dựa trên tập các vector đã xây dựng được ở bước trên, nhằm mục đích đạt được kết quả phân loại tốt nhất.

* + 1. ***Thuật toán Naïve Bayes***

1. Giới thiệu

Naïve Bayes (NB) là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học (Mitchell trình bày năm 1996, Joachims trình bày năm 1997 và Jason năm 2001) được sử dụng lần đầu tiên trong lĩnh vực phân loại bởi Maron vào năm 1961, sau đó trở nên phổ biến dùng trong nhiều lĩnh vực như trong các công cụ tìm kiếm (được mô tả năm 1970 bởi Rijsbergen), các bộ lọc mail (mô tả năm 1998 bởi Sahami)

1. Ý tưởng

Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Với giả định này NB không sử dụng sự phụ thuộc của nhiều từ vào một chủ đề, không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề và do đó việc tính toán NB chạy nhanh hơn các phương pháp khác với độ phức tạp theo hàm số mũ.

1. Thuật toán

*Công thức*

* Dựa trên định lý Bayes:
* Trong đó:
* H (Hypothesis) là giả thuyết và E (Evidence) là chứng cứ hỗ trợ cho giả thuyết H
* P(E|H): xác xuất E xảy ra khi H xảy ra (xác suất có điều kiện, khả năng của E khi H đúng).
* P(H|E): xác suất hậu nghiệm của H nếu biết E.

*Áp dụng trong bài toán phân loại*

* Các dữ kiện cần có:
* Một tập dữ liệu D, trong đó mỗi ví dụ học x đã được vector hóa dưới dạng một vector n chiều (x1, x2, ..., xn)
* Một tập xác định các nhãn lớp: C={c1, c2, ..., cm}
* Với một ví dụ (mới) z\*, z\* sẽ được phân vào lớp nào?

🡪 Áp dụng định lý Bayes để xác định phân lớp phù hợp nhất đối với z:

cMAP =argmax P(ci|z) = argmax P(ci | z1, z2,..., zn) = argmax

Vì P(z1, z2,…, zn) là như nhau với các lớp nên chỉ cần tìm phân lớp với:

cMAP =argmax P(z1, z2,..., zn | ci).P(ci)

Giả sử trong phương pháp phân loại Naïve Bayes, các thuộc tính là độc lập có điều kiện đối với các lớp thì:

P(z1,z2,...,zn|ci)=

🡪 cMAP = argmax P(ci).

1.3 Các phương pháp xác định giới tính dựa trên các bài biết của người dùng

**Bài toán**

Input: Cho tập văn bản miền ứng dụng D = {dj}, tập đặc trưng được chọn biểu diễn văn bản V = {wi}, ma trân trọng số W = (wi,j) .

Output: Tìm biểu diễn của các văn bản dj ∈D.

Một số mô hình: Mô hình Boolean, Mô hình không gian vector, Mô hình túi các từ (Mô hình xác suất)

*1.3.1 Dự đoán giới tính dựa trên nội dung bình luận trên Youtube*

Một mô hình biểu diễn vector với hàm f cho ra giá trị rời rạc với duy nhất hai giá trị đúng và sai (true và false, hoặc 0 và 1) gọi là mô hình Boolean.

Hàm f tương ứng với từ khóa ti sẽ cho ra giá trị đúng nếu và chỉ nếu từ khóa ti xuất hiện trong văn bản đó.

Mô hình Boolean được xác định như sau:

Giả sử có một tập gồm m văn bản D = {d1, d2, d3, ..., dm}, T là một tập từ vựng gồm n từ khóa T = {t1, t2,......tn}. Gọi w = (wịj) là ma trận trọng số, trong đó wij là trọng số của từ khóa ti trong văn bản dj và được xác định như sau:

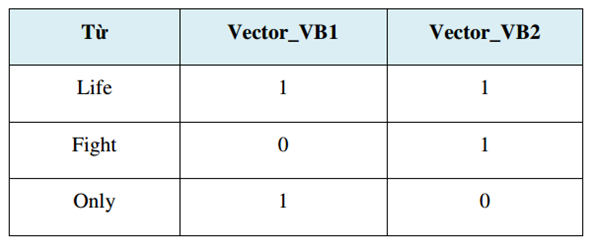
Ta xét 2 văn bản sau:

VB1: Life is not only life

VB2: To life is to fight

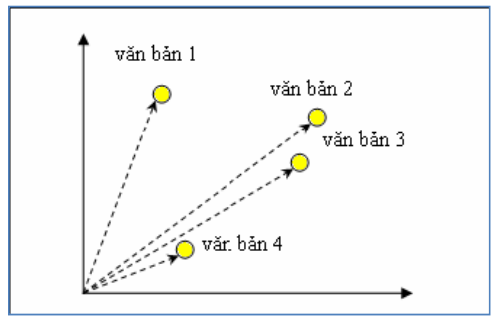
Áp dụng mô hình Boolean sau khi tiền xử lý văn bản ta có biểu diễn sau:

Bảng 1.1 Biểu diễn văn bản mô hình Boolean



Cách biểu diễn văn bản thông dụng nhất là thông qua vector biểu diễn theo mô hình không gian vector (Vector Space Model). Đây là một cách biểu diễn tương đối đơn giản và hiệu quả.

Theo mô hình này, mỗi văn bản được biểu diễn thành một vector. Mỗi thành phần của vector là một từ khóa riêng biệt trong tập văn bản gốc và được gán một giá trị là hàm f chỉ mật độ xuất hiện của từ khóa trong văn bản.



Hình 1.0.2: Biểu diễn vector văn bản trong không gian 2 chiều

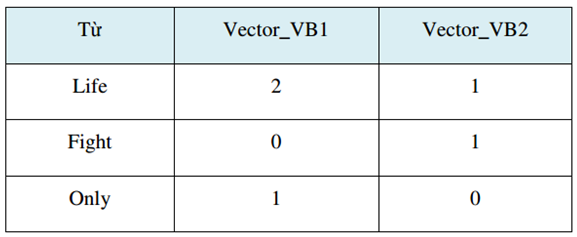
Giả sử ta có một văn bản và nó được biểu diễn bởi vector V(v1, v2,…,vn). Trong đó, vi là số lần xuất hiện của từ khóa thứ i trong văn bản. Ta xét 2 văn bản sau:

VB1: Life is not only life

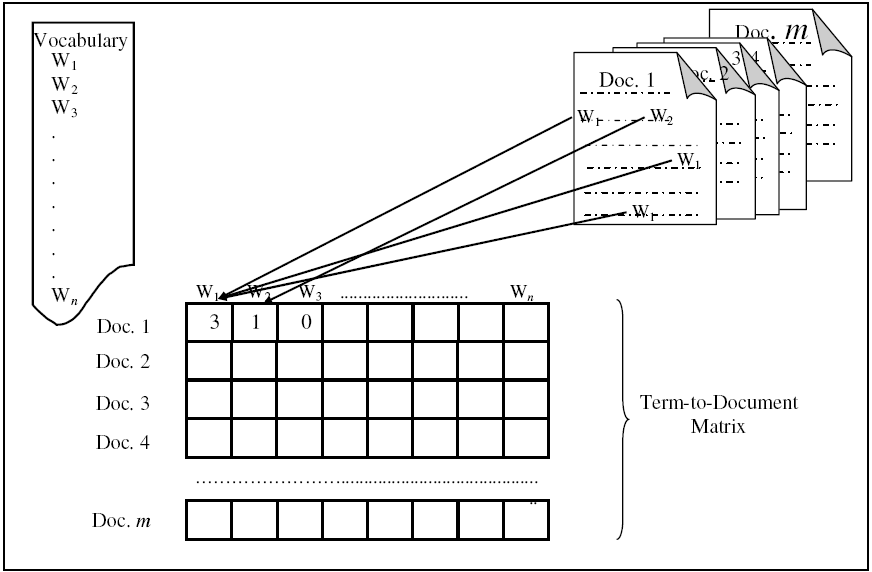
VB2: To life is to fight

Sau khi qua bước tiền xử lý văn bản, ta biểu diễn chúng như sau:

Bảng 2: Biểu diễn văn bản theo mô hình Vector



Trong các cơ sở dữ liệu văn bản, mô hình vector là mô hình biểu diễn văn bản được sử dụng phổ biến nhất hiện nay. Mối quan hệ giữa các trang văn bản được thực hiện thông qua việc tính toán trên các vector biểu diễn vì vậy được thi hành khá hiệu quả. Đặc biệt, nhiều công trình nghiên cứu về mối quan hệ "tương tự nhau" giữa các trang web (một trong những quan hệ điển hình nhất giữa các trang web) dựa trên mô hình biểu diễn vector.



Hình 1.0.3: Biểu diễn văn bản theo mô hình vector

*1.3.2 Dự đoán giới tính sử dụng bài viết từ blog*

Mô hình xác suất là mô hình toán học làm việc với các biến ngẫu nhiên và phân bố xác xuất của nó. Theo thuật ngữ toán

*1.3.3 Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy*

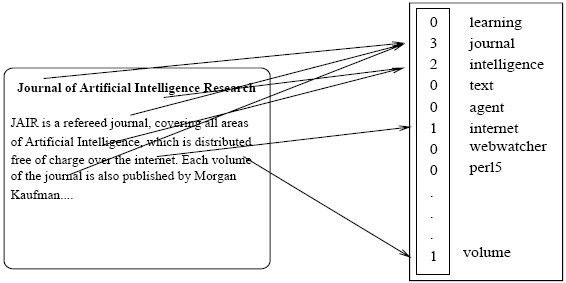
Mô hình xác suất là mô hình toán học làm việc với các biến ngẫu nhiên và phân bố xác xuất của nó. Theo thuật ngữ toán học, một mô hình xác suất có thể coi như một cặp (Y, P), trong đó Y là tập các quan sát (biến ngẫu nhiên) và P là tập các phân bố xác suất trên Y. Khi đó, sử dụng suy diễn xác suất sẽ cho ta kết luận về các phần tử của tập Y.

Văn bản trong mô hình xác suất được coi như một quan sát trong tập Y, trong đó:

* Các từ xuất hiện trong văn bản thể hiện nội dung văn bản
* Sự xuất hiện của các từ là độc lập lẫn nhau và độc lập ngữ cảnh
* Dạng đơn giản: chỉ liệt kê từ
* Dạng chi tiết: liệt kê từ và số lần xuất hiện

*Lưu ý*: Các giả thiết về tính độc lập không hoàn toàn đúng (độc lập lẫn nhau, độc lập ngữ cảnh) song mô hình thi hành hiệu quả trong nhiều trường hợp.

Khi đó văn bản sẽ gồm các từ mà nó chứa trong đó, chính vì vậy phương pháp này được gọi là biểu diễn túi - các - từ (bag-of-word)



Hình 1.0.4: Biểu diên văn bản theo mô hình túi các từ

1.4 Kết luận chương

Chương này tôi đã giới thiệu tổng quan về bài toán phân loại văn bản, một số phương pháp phân loại và các kỹ thuật liên quan đến biểu diễn văn bản. Bên cạnh đó, chương 1 còn đưa ra lý do tại sao lại chọn thuật toán SVM trong luận văn. Ngoài ra cần lưu ý đến 3 yếu tố quan trọng tác động đến kết quả phân loại văn bản:

1) Cần một tập dữ liệu huấn luyện chuẩn và đủ lớn để cho thuật toán học phân loại. Nếu chúng ta có được một tập dữ liệu chuẩn và đủ lớn thì quá trình huấn luyện sẽ tốt và khi đó chúng ta sẽ có kết qủa phân loại tốt sau khi đã được học.

2) Các phương pháp trên hầu hết đều sử dụng mô hình vector để biểu diễn văn bản, do đó phương pháp tách từ trong văn bản đóng vai trò quan trọng quá trình biểu diễn văn bản bằng vector. Yếu tố này rất quan trọng, vì có thể đối với một số ngôn ngữ như tiếng Anh thì thao tác tách từ trong văn bản đơn giản chỉ là dựa vào các khoảng trắng, tuy nhiên trong các ngôn ngữ đa âm tiết như tiếng Việt và một số ngôn ngữ khác thì sử dụng khoảng trắng khi tách từ là không chính xác, do đó phương pháp tách từ là một yếu tố quan trọng.

3) Thuật toán sử dụng để phân loại phải có thời gian xử lý hợp lý, thời gian này bao gồm: thời gian học, thời gian phân loại văn bản, ngoài ra thuật toán này phải có tính tăng cường (incremental function) nghĩa là không phân loại lại toàn bộ tập văn bản khi thêm một số văn bản mới vào tập dữ liệu mà chỉ phân loại các văn bản mới mà thôi, khi đó thuật toán phải có khả năng giảm độ nhiễu (noise) khi phân loại văn bản.

Chương 2 - KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG MÃ HỘI



2.1 Phạm vi áp dụng

* + 1. ***Thu thập địa chỉ trang web URL***

1. Thu thập URL phục vụ cho dữ liệu huấn luyện

Một trong những cách hữu hiệu nhất để thu thập data phục vụ cho việc huấn luyện đó là sử dụng RSS (Really Simple Syndication – định dạng tài liệu). Thông qua RSS chúng ta có thể lấy được những tin chính mới nhất, tiêu đề, tóm tắt và cả đường link để xem toàn bộ tin.

Định dạng RSS cung cấp nội dung web và tóm lược nội dung web cùng với các liên kết đến phiên bản đầy đủ của nội dung tin đó và các [siêu-dữ-liệu](https://vi.wikipedia.org/wiki/Si%C3%AAu_d%E1%BB%AF_li%E1%BB%87u) ([meta-data](https://vi.wikipedia.org/w/index.php?title=Meta_data&action=edit&redlink=1)) khác. Thông tin này được cung cấp dưới dạng một tập tin [XML](https://vi.wikipedia.org/wiki/XML) được gọi là một RSS feed, webfeed, RSS stream hay RSS channel. Cùng với việc hỗ trợ cung cấp chia sẻ thông tin, RSS cho phép những độc giả thường xuyên của một website có thể theo dõi các cập nhật của site đó.

Cụ thể việc lấy tin RSS sẽ được trình bày ở chương 3 của luận văn.

1. Thu thập URL phục vụ cho ứng dụng

Các URL phục vụ cho ứng dụng được thu thập từ các website việc làm không hỗ trợ RSS. Từ các URL đó sẽ được đưa vào ứng dụng để tiến hành phân loại. Cụ thể các website sẽ được trình bày tại chương 3 của luận văn.

* + 1. ***Lọc nội dung chính của trang web***

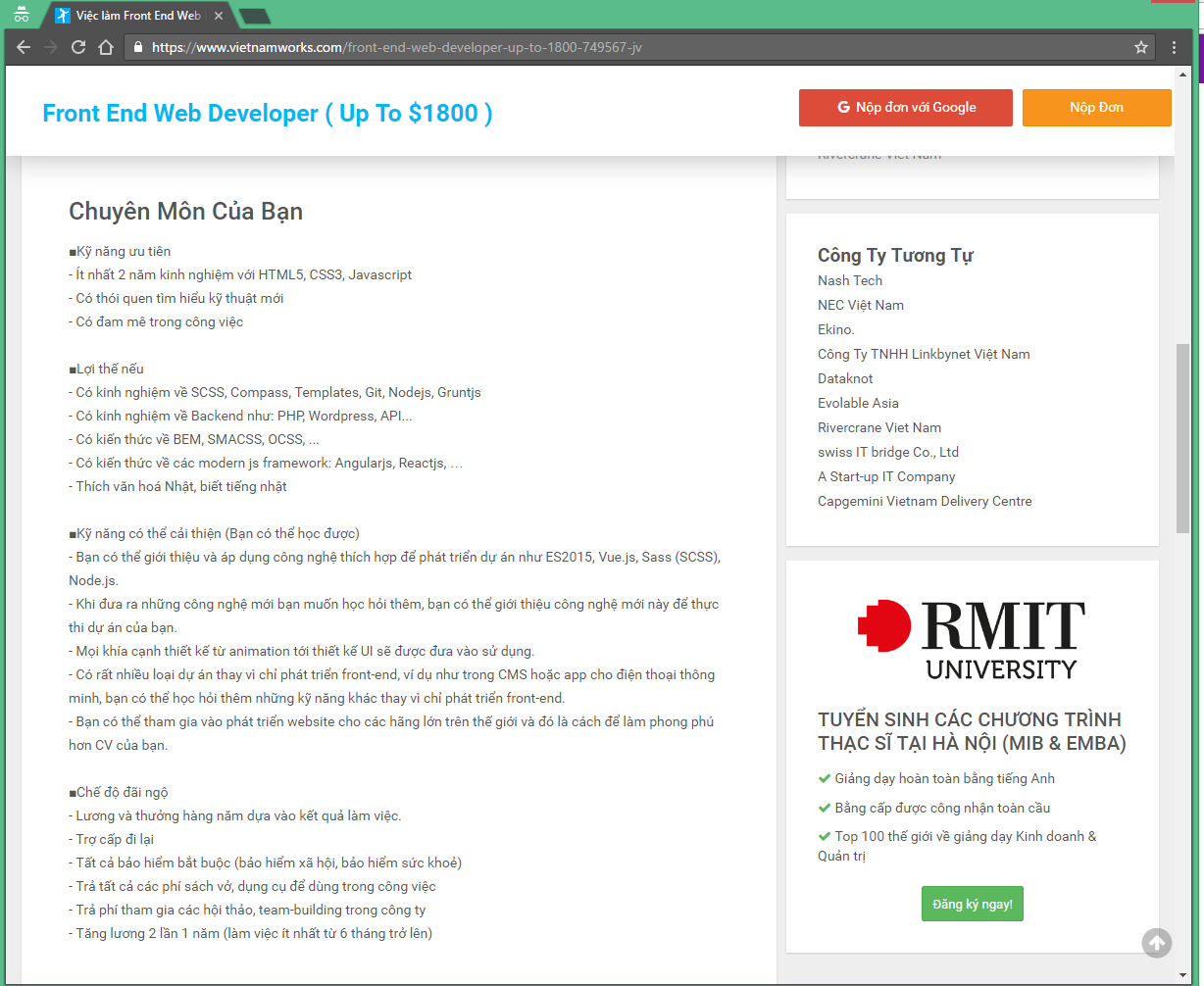
Những trang được tải về bao hàm cả các dấu HTML, văn bản ‘boilerplate’ – các thanh menu, quảng cáo. Đó là các phần tài liệu không hợp lệ, tạm gọi là phần đánh dấu. Trong luận văn, em sử dụng JSOUP (Java HTML Parser) là một thư viện được dùng để phân tích tài liệu HTML. Jsoup cung cấp các API dùng để lấy dữ liệu và thao tác dữ liệu từ URL hoặc từ file HTML. Nó sử dụng các phương thức giống với DOM, CSS , JQuery để lấy dữ liệu và thao tác với dữ liệu. Các trang Web điển hình thường có mục tài liệu mà phần bắt đầu và kết thúc là nhiều ‘boilerplate’ và dấu HTML, còn phần ở giữa là phần nội dung của văn bản trong đó phần đánh dấu tương đối ít, đó là ngôn ngữ và là mục tài liệu cần sử dụng. JSOUP sẽ giúp loại bỏ những phần đánh dấu và chỉ trích rút những văn bản được kết nối.

Phần lớn các trang web được mã hóa ở định dạng HTML (Hyper-Text Mark-up Language). Trong đó, mỗi tệp HTML của trang Web chứa rất nhiều thông tin phụ khác ngoài phần tin chính (main content), như: các pop-up quảng cáo, link liên kết đến các trang, bài khác, hoặc các nhà tài trợ, nhà phát triển, phần ghi quyền tác giả, các thông báo, cảnh báo… Việc làm sạch các văn bản đầu vào ở đây được hiểu chính là quá trình xác định phần nội dung chính của trang web và loại bỏ các thành phần phụ không liên quan. Với các trang web có cấu trúc trang thay đổi thường xuyên, việc bóc tách phần nội dung chính từ các trang web trở nên khó khăn hơn.

Luận văn đã sử dụng Maven cho thư viện này như sau:

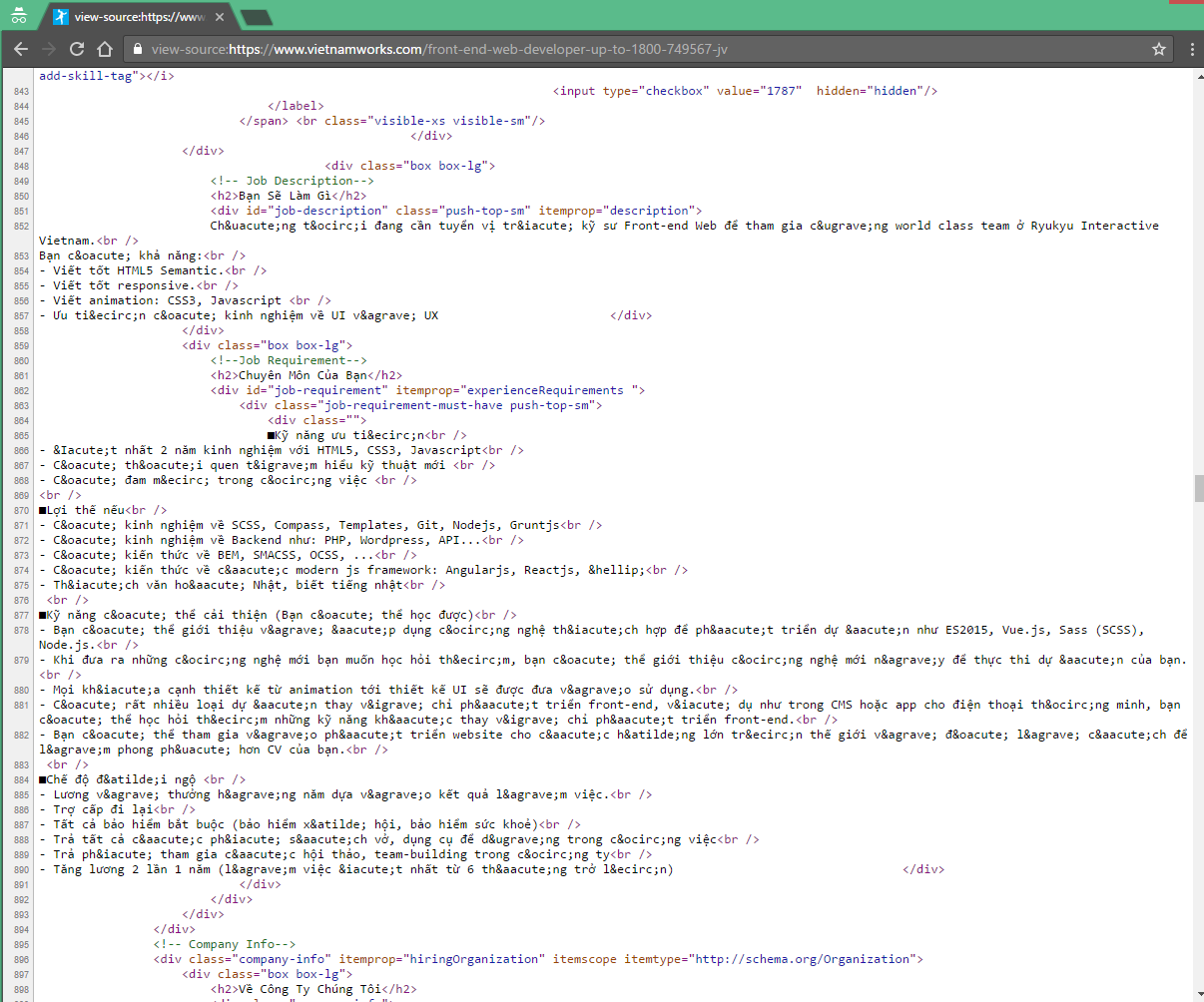


**Hình 2.1: Sử dụng Maven cho thư viên JSOUP**

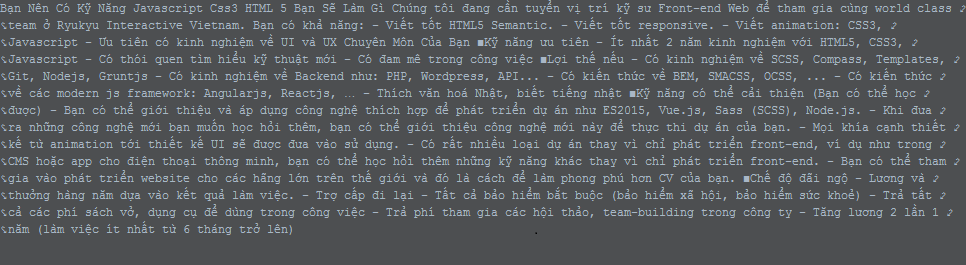


**Hình 2.2: Một website việc làm điển hình**

Hình 2.2 mô tả một bài tuyển dụng việc làm thông thường với phần thân chính của văn bản được đánh dấu. Bên trên cùng là phần header với banner và mục menu; bên dưới là footer, mục liên quan, thông báo bản quyền; bên phải là các liên kết của các nhà quảng cáo và phần không liên quan đến nội dung tuyển dụng.



**Hình 2.3: Nội dung đầu vào là mã html**



**Hình 2.4: Nội dung đầu ra là văn bản đã được trích xuất**

* + 1. ***Tiền xử lý văn bản***

1. Đặc điểm của từ trong tiếng Việt

Tiếng Việt là ngôn ngữ đơn lập. Đặc điểm này bao quát tiếng Việt cả về mặt ngữ âm, ngữ nghĩa, ngữ pháp. Khác với các ngôn ngữ châu Âu, mỗi từ là một nhóm các ký tự có nghĩa được cách nhau bởi một khoảng trắng. Còn tiếng Việt và các ngôn ngữ đơn lập khác thì khoảng trắng không phải là căn cứ để nhận diện từ.

Tiếng:

* Trong tiếng Việt trước hết cần chú ý đến đơn vị xưa nay vẫn quan gọi là tiếng. Về mặt ngữ nghĩa, ngữ âm, ngữ pháp, đều có giá trị quan trọng.
* Sử dụng tiếng để tạo từ có hai trường hợp:
* Trường hợp một tiếng: đây là trường hợp một tiếng được dùng làm một từ, gọi là từ đơn. Tuy nhiên không phải tiếng nào cũng tạo thành một từ.
* Trường hợp hai tiếng trở lên: đây là trường hợp hai hay nhiều tiếng kết hợp với nhau, cả khối kết hợp với nhau gắn bó tương đối chặt chẽ, mới có tư cách ngữ pháp là một từ. Đây là trường hợp từ ghép hay từ phức.

Từ:

Có rất nhiều quan niệm về từ trong tiếng Việt, từ nhiều quan niệm về từ tiếng Việt khác nhau đó chúng ta có thể thấy đặc trưng cơ bản của "từ" là sự hoàn chỉnh về mặt nội dung, từ là đơn vị nhỏ nhất để đặt câu. Người ta dùng "từ" kết hợp thành câu chứ không phải dùng "tiếng", do đó quá trình tách câu thành các "từ" cho kết quả tốt hơn là tách câu bằng “tiếng”.

1. Tách từ

*Bài toán tách từ*

Tách từ là một trong những bước xử lý thông tin đầu tiên trong quá trình xử lý ngôn ngữ tự nhiên, làm đầu vào cơ bản cho các chương trình ứng dụng xử lý ngôn ngữ, tách từ cũng là một nhiệm vụ khó khăn chính trong việc phân loại văn bản, gán nhãn từ loại đối với các ngôn ngữ châu Á như tiếng Hoa, tiếng Thái, tiếng Việt... Bài toán tách từ đặt ra nhằm phân tích một dãy các các từ đưa vào máy tính tách ra thành một tập các từ có nghĩa. Đối với các ngôn ngữ thuộc hệ Ấn – Âu việc tách từ khá đơn giản dựa vào vị trí của các ký tự đặc biệt để chia tách từ (Ví dụ như trong tiếng Anh là dấu cách). Tuy nhiên, với các ngôn ngữ châu Á như tiếng Hoa, tiếng Thái… thì để giải quyết bài toán tách từ lại gặp những vấn đề phức tạp hơn nhiều, sự phức tạp này chủ yếu là do các ngôn ngữ nói trên không có ký tự đặc biệt để phân biệt ranh giới giữa các từ như các ngôn ngữ thuộc hệ Ấn – Âu. Để giải quyết bài toán tách từ cho tiếng Việt cũng gặp phải các vấn đề tương tự nói trên, do đó việc ứng dụng các kết quả nghiên cứu của những nước phương tây cho bài toán tách từ tiếng Việt là hết sức hạn chế và không đem lại hiệu quả.

Trong các năm gần đây các công trình nghiên cứu về bài toán tách từ tiếng Việt thường tập trung vào 2 hướng tiếp cận chính để giải quyết bài toán cho tiếng Việt đó là:

* Hướng tiếp cận giải quyết bài toán tách từ tiếng Việt dựa trên từ (Word - based): Từ điển, thống kê,…
* Hướng tiếp cận giải quyết bài toán tách từ tiếng Việt dựa trên âm tiết (Character-based): Unigram, bigram, trigram, n-gram…

*Một số phương pháp tách từ*

**Phương pháp Maximum Matching: Forward / Backward**

Phương pháp so khớp tối đa (MM-Maximum Matching) hay còn gọi là LRMM - Left Right Maximum Matching. Ở phương pháp này, chúng ta sẽ duyệt một ngữ hoặc câu từ trái sang phải và chọn từ có nhiều âm tiết nhất có mặt trong từ điển và cứ thực hiện lặp lại như vậy cho đến hết câu.

Dạng đơn giản của phương pháp dùng để giải quyết nhập nhằng từ đơn. Giả sử chúng ta có một chuỗi ký tự C1, C2, …, Cn. Chúng ta sẽ áp dụng phương pháp từ đầu chuỗi. Đầu tiên kiểm tra xem C1có phải là từ hay không, sau đó kiểm tra xem C1C2 có phải là từ hay không. Tiếp tục thực hiện như thế cho đến khi tìm được từ dài nhất .

Dạng phức tạp: Quy tắc của dạng này là phân đoạn từ. Thông thường người ta chọn phân đoạn ba từ có chiều dài tối đa. Thuật toán bắt đầu từ dạng đơn giản, cụ thể là nếu phát hiện ra những cách tách từ gây nhập nhằng, giả sử C1 là từ và C1C2 cũng là một từ, khi đó chúng ta kiểm tra ký tự kế tiếp trong chuỗi C1, C2, ..., Cn để tìm tất cả các đoạn ba từ có bắt đầu với C1 hoặc C1C2.

Ví dụ: Giả sử chúng ta có được các đoạn sau:

- C1 C2 C3 C4

- C1C2 C3C4 C5

- C1C2 C3C4 C5C6

Khi đó chuỗi dài nhất sẽ là chuỗi thứ ba. Do đó từ đầu tiên của chuỗi thứ ba (C1C2) sẽ được chọn. Thực hiện các bước cho đến khi được chuỗi từ hoàn chỉnh.

Nhận xét :

Phương pháp này thực hiện tách từ đơn giản, nhanh và chỉ cần dựa vào từ điển để thực hiện. Tuy nhiên, khuyết điểm của phương pháp này cũng chính là từ điển, nghĩa là độ chính xác khi thực hiện tách từ phụ thuộc hoàn toàn vào tính đủ, tính chính xác của từ điển.

**Phương pháp Transformation – based Learning**

Phương pháp này tiếp cận dựa trên tập ngữ liệu đã đánh dấu. Theo cách tiếp cận này để cho máy tính có thể nhận biết ranh giới giữa các từ để có thể tách từ chính xác, chúng ta sẽ cho máy học các câu mẫu trong tập ngữ liệu đã được đánh dấu ranh giới giữa các từ đúng. Chúng ta thấy phương pháp rất đơn giản, vì chỉ cần cho máy học các tập câu mẫu và sau đó máy sẽ tự rút ra qui luật của ngôn ngữ và để từ đó sẽ áp dụng chính xác khi có những câu đúng theo luật mà máy đã rút ra. Và để tách từ được hoàn toàn chính xác trong mọi trường hợp thì đòi hỏi phải có một tập ngữ liệu tiếng Việt thật đầy đủ và phải được huấn luyện lâu để có thể rút ra các luật đầy đủ.

**Phương pháp tách tách từ tiếng Việt dựa trên thống kê từ Internet và thuật giải di truyền**

Phương pháp tách từ tiếng Việt dựa trên thống kê từ Internet và thuật giải di truyền – IGATEC (Internet and Genetics Algorithm based Text Categorization for Documents in Vietnamese) do H. Nguyễn đề xuất năm 2005 như một hướng tiếp cận mới trong tách từ với mục đích phân loại văn bản mà không cần dùng đến một từ điển hay tập ngữ liệu học nào. Trong hướng tiếp cận này, tác giả kết hợp giữa thuật toán di truyền với dữ liệu thống kê được lấy từ Internet .

*Công cụ vnTokenizer*

Đặt bài toán: Đối với một văn bản tiếng Việt, dấu cách được đặt giữa các âm tiết chứ không phải giữa các từ. Một từ có thể có một, hai hoặc nhiều âm tiết nên có nhiều cách phân chia các âm tiết thành các từ, gây ra nhập nhằng. Việc phân giải nhập nhằng này gọi là bài toán tách từ.

Bài toán: Cho một câu tiếng Việt bất kỳ, hãy tách câu đó thành những đơn vị từ vựng (từ), hoặc chỉ ra những âm tiết nào không có trong từ điển (phát hiện đơn vị từ vựng mới).

Giới thiệu công cụ vnTokenizer: vnTokenizer là công cụ tách từ tiếng Việt được nhóm tác giả Nguyễn Thị Minh Huyền, Vũ Xuân Lương và Lê Hồng Phương phát triển dựa trên phương pháp so khớp tối đa (Maximum Matching) với tập dữ liệu sử dụng là bảng âm tiết tiếng Việt và từ điển từ vựng tiếng Việt. Công cụ được xây dựng bằng ngôn ngữ Java, mã nguồn mở. Có thể dễ dàng sửa đổi nâng cấp và tích hợp vào các hệ thống phân tích văn bản tiếng Việt khác.

Quy trình thực hiện tách từ theo phương pháp so khớp tối đa:



**Hình** 2.5: **Quy trình tách từ**

Đầu vào của công cụ tách từ vnTokenizer là một câu hoặc một văn bản được lưu dưới dạng tệp.

Đầu ra là một chuỗi các đơn vị từ được tách.

Các đơn vị từ bao gồm các từ trong từ điển cũng như các chuỗi số, chuỗi kí từ nước ngoài, các hình vị ràng buộc (gồm các phụ tố), các dấu câu và các chuỗi kí tự hỗn tạp khác trong văn bản (ISO, 2008). Các đơn vị từ không chỉ bao gồm các từ có trong từ điển, mà cả các từ mới hoặc các từ được sinh tự do theo một quy tắc nào đó (như phương thức thêm phụ tố hay phương thức láy) hoặc các chuỗi kí hiệu không được liệt kê trong từ điển.

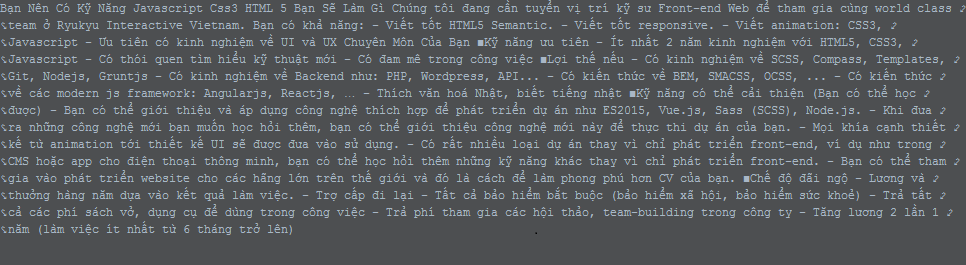
Công cụ sử dụng tập dữ liệu đi kèm là tập từ điển từ vựng tiếng Việt, danh sách các đơn vị từ mới bổ sung, được biểu diễn bằng ôtômat tối tiểu hữu hạn trạng thái, tệp chứa các biểu thức chính quy cho phép lọc các đơn vị từ đặc biệt (xâu dạng số, ngày tháng,…) và các tệp chứa các thống kê unigram và bigram trên kho văn bản tách từ mẫu.

Với các đơn vị từ đã có trong từ điển, khi thực hiện tách từ cũng được xử lý hiện tượng nhập nhằng bằng cách kết hợp với các thống kê unigram và bigram. Chẳng hạn trong tiếng Việt thường gặp các trường hợp nhập nhằng như:

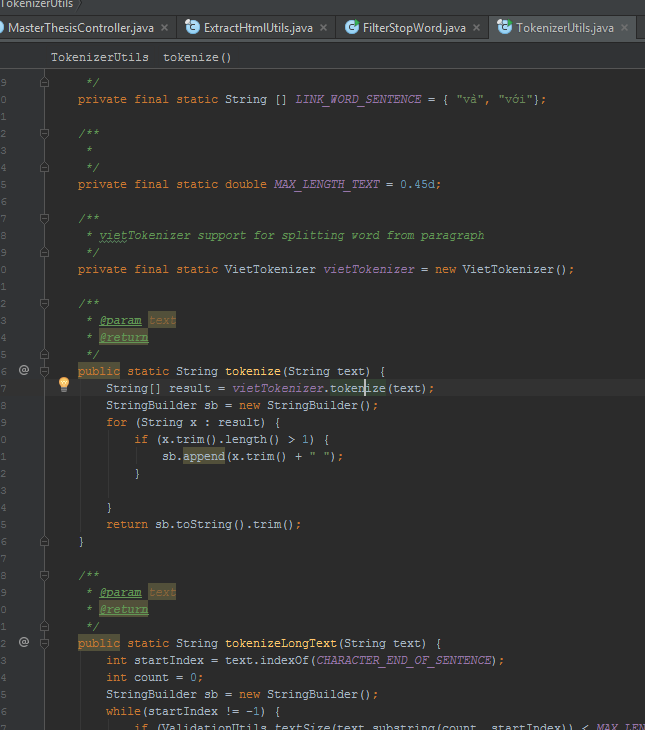
Xâu AB vừa có thể hiểu là 1 đơn vị từ, vừa có thể là chuỗi 2 đơn vị từ A-B.

Xâu ABC có thể tách thành 2 đơn vị AB-C hoặc A-BC.

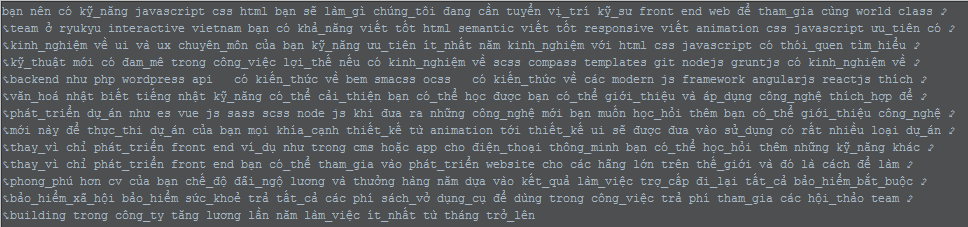
Minh họa sử dụng công cụ vnTokenizer:



**Hình** 2.6**: Đầu vào của thuật toán tách từ**



**Hình** 2.7**: Quá trình xử lý của công cụ vnTokenizer**



**Hình** 2.8**: Đầu ra là văn bản đã được tách từ**

Đánh giá kết quả:

Kết quả đánh giá của công cụ được cho là ổn định đối với nhiều loại văn bản/ văn phong khác nhau. Độ chính xác trung bình đạt được là khoảng 94%.

Lưu ý: sau bước tách từ, văn bản sẽ được đưa về dạng chữ thường để loại bỏ stopword.

1. Loại bỏ stop-word

Từ dừng (stop-word) dùng để chỉ các từ xuất hiện quá nhiều trong các văn bản của toàn tập kết quả, thường thì không giúp ích gì trong việc phân biệt nội dung của các tài liệu văn bản. Ví dụ, những từ “và”, “thì”, “là”, “mà” thường xuất hiện hầu hết trong các văn bản thì được gọi là stop-word. Vì đặc điểm của từ dừng nên chúng được loại bỏ mà không ảnh hưởng đến các việc biểu diễn văn bản tiếp theo.

Danh sách các stop-word tiếng Việt sẽ được đính kèm trong phần phụ lục của quyển luận văn.

Dưới đây là một số stop-word tiếng Việt cụ thể:

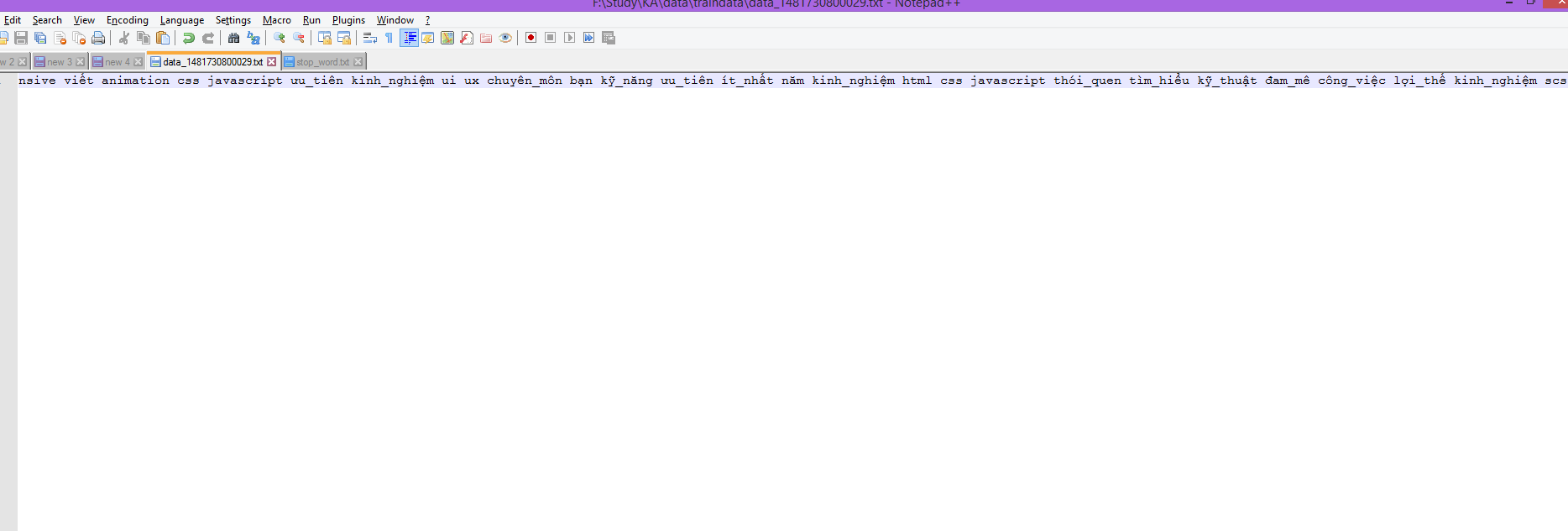
Bảng2**.**1**: Danh sách stopword tiếng Việt**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| a\_ha  a-lô  à\_ơi  á  à | á\_à  ạ  ạ\_ơi  ai  ai\_nấy | chính\_là  chính\_thị  cho\_biết  cho\_hay  cho\_phép | cho\_đến  cho\_đến\_khi  cho\_nên  cho\_tới  cho\_tới\_khi | thế\_à  thế\_là  thế\_mà  thế\_nào  thế\_nên | thế\_ra  thế\_thì  thếch  thì  thi\_thoảng |

1. Làm sạch văn bản

Tiếp theo, ta cần phải “làm sạch” văn bản để tạo ra một tập dữ liệu thiết thực hơn. Ví dụ:

* Loại bỏ dấu câu: “.”, “,”,“?”,“!”,“-”,“(”,“)”,“:”,“[”,“]”
* Loại bỏ các kí tự đặc biệt: [ ],[.], [,], [:], [“], [”], [;], [/], [[]], [~], [`], [!], [@], [#], [$],[%],[^],[&],[\*],[(],[)]
* Các chữ số, phép tính toán số học: các ký tự số cũng thường được loại ra vì không mang lại nhiều ý nghĩa cho văn bản (trừ một số trường hợp đặc biệt, ví dụ thông tin trong lĩnh vực lịch sử, tài chính – ngân hàng).
* Ngoài ra, cần phải loại bỏ các dòng trống, các khoảng trắng có nhiều hơn một khoảng trắng, các ký tự đặc biệt (@, #, $, %, ^, /, \,…)



Hình 2.9: Văn bản sau khi loại stopwords và làm sạch

Sau khi loại bỏ các stopword và làm sạch, văn bản có thể xem như là một tập hợp các đặc trưng, đó là tập hợp các từ “quan trọng” còn lại để biểu diễn văn bản. Việc phân loại văn bản sẽ dựa trên các đặc trưng này. Tuy nhiên, có thể thấy rằng, số đặc trưng của một văn bản là lớn và không gian các đặc trưng (tất cả đặc trưng) của tất cả các văn bản đang xem xét là rất lớn, về nguyên tắc, nó bao gồm tất cả các từ trong một ngôn ngữ. Chính vì vậy, phân loại dựa trên các đặc trưng này cần phải có cách xử lí, lựa chọn đặc trưng nhằm rút ngắn số chiều của không gian đặc trưng. Luận văn đã sử dụng trọng số của một từ được tính theo tần suất xuất hiện của từ trong văn bản (TF) và tần suất nghịch đảo của từ (IDF). Cụ thể phần này đã được trình bày ở chương 1, mục 1.5.

2.2 Các đặc trưng sử dụng

*2.2.1 Đặc trưng text*

Máy vectơ hỗ trợ (SVM - viết tắt tên tiếng Anh support vector machine) là một khái niệm trong [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA) và [khoa học máy tính](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_m%C3%A1y_t%C3%ADnh) cho một tập hợp các phương pháp [học có giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_c%C3%B3_gi%C3%A1m_s%C3%A1t) liên quan đến nhau để [phân loại](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_lo%E1%BA%A1i_%28h%E1%BB%8Dc_m%C3%A1y%29&action=edit&redlink=1) và [phân tích hồi quy](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_h%E1%BB%93i_quy).

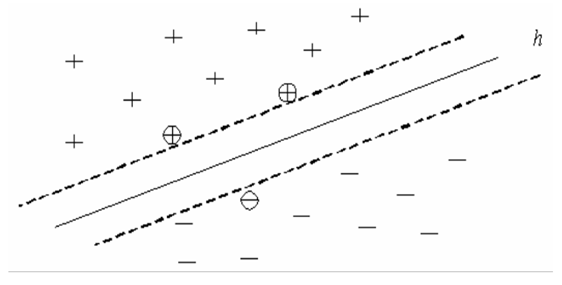
SVM được định nghĩa như sau:

* Là phương pháp dựa trên nền tảng của lý thuyết thống kê nên có một nền tảng toán học chặt chẽ để đảm bảo rằng kết quả tìm được là chính xác.
* Là thuật toán học giám sát (supervied learning) được sử dụng cho phân lớp dữ liệu.
* Là 1 phương pháp thử nghiệm, đưa ra 1 trong những phương pháp mạnh và chính xác nhất trong số các thuật toán nổi tiếng về phân lớp dữ liệu.

*2.2.2 Đặc trưng ký hiệu đặc biệt*

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

Mục đích của phương pháp SVM là tìm được khoảng cách biên lớn nhất để tạo kết quả phân lớp tốt, điều này được minh họa như sau:



**Hình 2.10: Siêu phẳng h phân chia dữ liệu huấn luyện thành 2 lớp + và – với khoảng cách biên lớn nhất. Các điểm gần h nhất là các vector hỗ trợ (Support Vector được khoanh tròn)**

2.3 Mô hình phân loại SVM

Bây giờ xét đến trường hợp phân nhiều lớp K > 2. Chúng ta có thể xây dựng việc phân K-class dựa trên việc kết hợp một số đường phân 2 lớp.

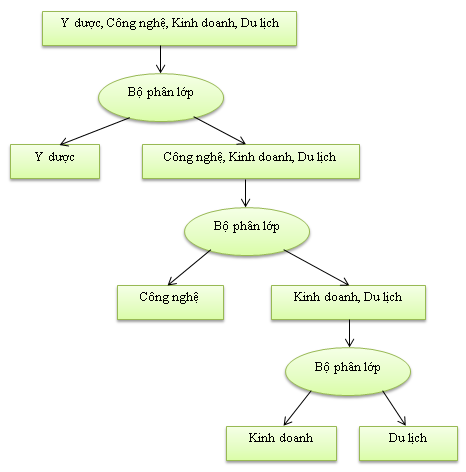
Sau đây ta tìm hiểu các chiến lược áp dụng trong bài toán phân lớp văn bản thuộc nhiều chủ đề khác nhau. Ý tưởng của bài toán phân lớp đa lớp là chuyển về bài toán phân lớp hai lớp bằng cách xây dựng nhiều bộ phân lớp hai lớp để giải quyết. Các chiến lược phân lớp đa lớp phổ biến này là: One-against-One (OAO), và One-against-Rest (OAR).

*Chiến lược One-against-Rest*

Trong chiến lược OAR, ta sẽ dùng K-1 bộ phân lớp nhị phân để xây dựng K-class. Bài toán phân lớp K lớp được chuyển thành K-1 bài toán phân lớp hai lớp. Trong đó bộ phân lớp hai lớp thứ I được xây dựng trên lớp thứ i và tất cả các lớp còn lại. Hàm quyết định thứ I dùng để phân lớp thứ i và những lớp còn lại có dạng:

Siêu phẳng yi(x) = 0 hình thành siêu phẳng phân chia tối ưu, các support vector thuộc lớp i thỏa yi(x) = 1 và các support vector thuộc lớp còn lại thỏa yi(x) = -1. Nếu vector dữ liệu x thỏa mãn điều kiện yi(x)>0 đối với duy nhất một I, x sẽ được phân vào lớp thứ i.

*2.3.1 Giới thiệu kỹ thuật học máy SVM*



**Hình** 2.12**: Ví dụ phân lớp văn bản thuộc các chủ đề chiến lược OAR**

*2.3.2 Áp dụng kỹ thuật học máy SVM vào dự đoán giới tính*

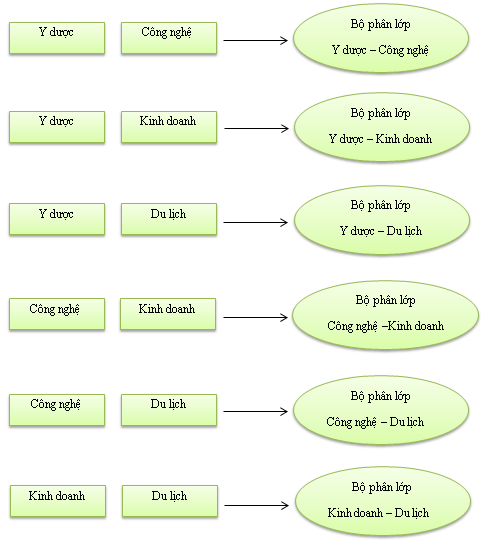
*Chiến lược One-against-One*

Chiến lược OAO, dùng K(K-1)/2 bộ phân lớp nhị phân được xây dựng bằng cách bắt cặp hai lớp một nên chiến lược này còn được gọi là pairwise và sử dụng theo phương pháp đa số kết lợp các bộ phận lớp này để xác định được kết quả phân lớp cuối cùng. Số lượng các bộ phân lớp không bao giờ vượt quá K(K-1)/2.

So với chiến lược OAR thì chiến lược này ngoài ưu điểm giảm bớt vùng không thể phân lớp mà còn làm tăng độ chính xác của việc phân lớp. Trong chiến lược OAR ta phải xây dựng một siêu phẳng để tách một lớp ra khỏi các lớp còn lại, việc này đòi hỏi sự phức tạp và có thể không chính xác. Tuy nhiên trong chiến lược OAO ta chỉ cần tách một lớp ra khỏi một lớp khác mà thôi.

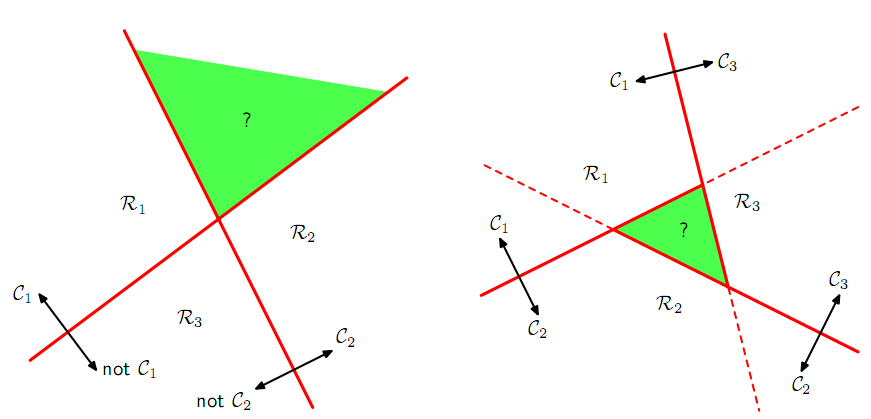
Chiến lược OAR chỉ cần K-1 bộ phân lớp cho K lớp. Trong khi đó chiến lược OAO lại cần đến K(K-1)/2 bộ phân lớp. Nhưng số mẫu huấn luyện cho từng bộ phân lớp trong OAO lại ít hơn và việc phân lớp cũng đơn giản hơn. Vì vậy chiến lược OAO có độ chính xác cao hơn nhưng chi phí để xây dựng lại tương đương với chiến lược OAR.

Hàm quyết định phân lớp của lớp i đối với lớp j trong chiến lược OAO là:



**Hình** 2.13**: Ví dụ phân lớp văn bản thuộc các chủ đề chiến lược OAO**

Tuy nhiên cả hai chiến lược trên đều dẫn đến vùng mập mờ trong phân lớp (như hình 2.14).



**Hình 2.14: Vùng mập mờ trong phân lớp**

Ta có thể tránh được vấn đề này bằng cách xây dựng K-Class dựa trên K hàm tuyến tính có dạng:

Và một điểm x được gán vào lớp Ck khi với mọi .

2.4 Kết luận chương

Chương này đã giới thiệu chi tiết thuật toán Máy vector hỗ trợ SVM và trình bày cách thu thập, trích rút và xử lý một cách tự động các tin tức tuyển dụng. Sau đó, luận văn sẽ sử dụng SVM để thực hiện việc huấn luyện bộ phân loại cho bài toán phân loại thông tin tuyển dụng trên hệ thống website tuyển dụng.

Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ

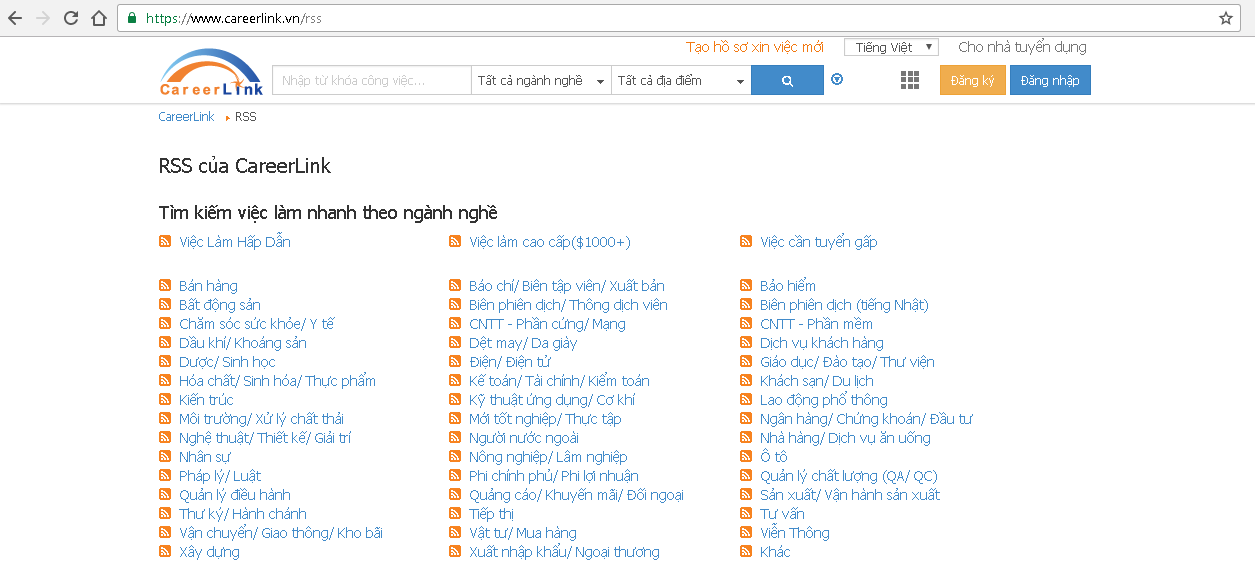
3.1 Thu thập và mô tả dữ liệu đầu vào

Xây dựng bộ dữ liệu và tiền xử lý văn bản

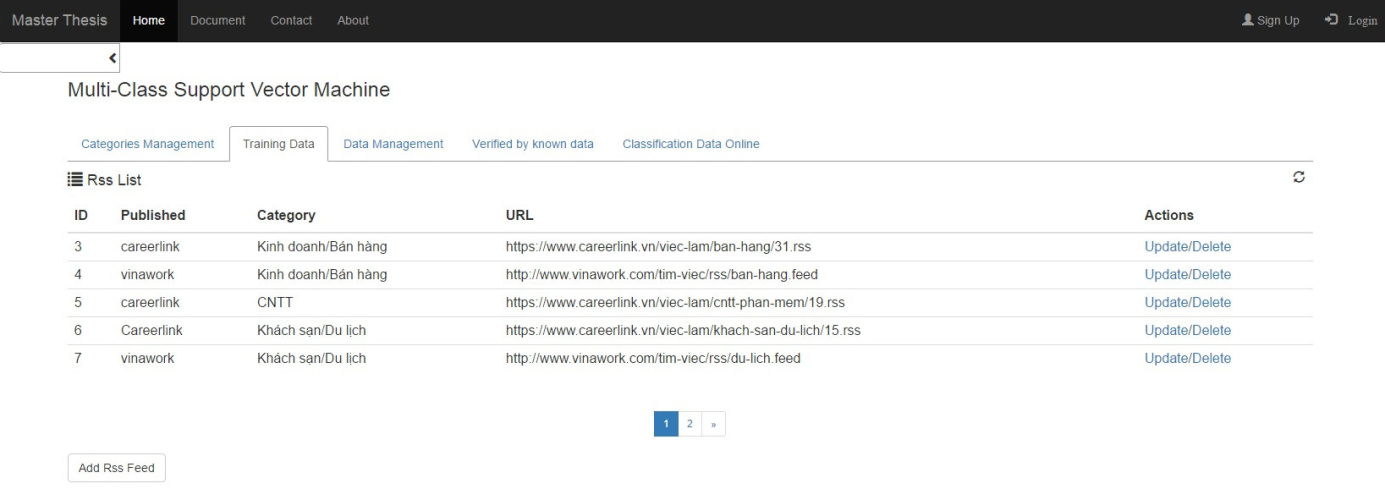
RSS (viết tắt từ Really Simple Syndication) là một tiêu chuẩn định dạng tài liệu dựa trên XML nhằm giúp người sử dụng dễ dàng cập nhật và tra cứu thông tin một cách nhanh chóng và thuận tiện nhất bằng cách tóm lược thông tin vào trong một đoạn dữ liệu ngắn gọn, hợp chuẩn.

Công nghệ của RSS cho phép người dùng Internet có thể đặt mua thông tin từ các websites có cung cấp khả năng RSS (RSS feeds); chúng thường là các site có nội dung thay đổi và được thêm vào thường xuyên. Để có thể dùng công nghệ này, các người quản trị site đó tạo ra hay quản lí một phần mềm chuyên dụng (như là một hệ thống quản lí nội dung - content management system - CMS) mà với định dạng XML máy có thể đọc được, có thể biểu diễn các bài tin mới thành một danh sách, với một hoặc hai dòng cho mỗi bài tin và một liên kết đến bài tin đầy đủ đó. Khác với việc mua nhiều ấn bản của các tờ báo hay tạp chí in giấy, hầu hết việc mua RSS là miễn phí.

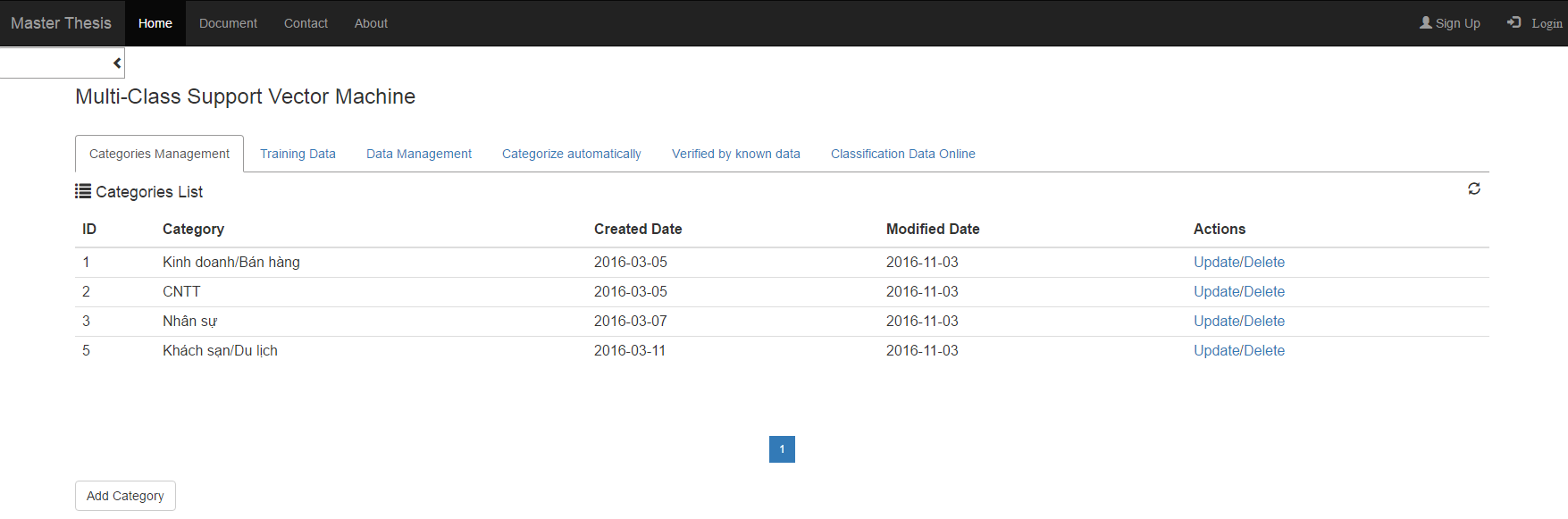
Trong phạm vi luận văn, em đã sử dụng lấy tin tự động từ RSS để cập nhật các việc làm hằng ngày nhằm xây dựng bộ dữ liệu huấn luyện. Dữ liệu được lấy từ 2 trang web: <https://www.careerlink.vn/> và <http://www.vinawork.com/>



Hình 3.1: Thông tin RSS trên website <https://www.careerlink.vn/>



Hình 3.2: Danh sách các RSS được thu thập phục vụ giai đoạn huấn luyện



Hình 3.3: Các category được sử dụng

Các bài báo việc làm được thu thập được thông qua RSS từ tháng 8/2016. Các bài báo này được lấy về sau đó tiến hành tiền xử lý, cuối cùng được lưu dưới dạng file phục vụ cho huấn luyện và kiểm tra. Dữ liệu thu thập được gồm có 1409 bài báo, trong đó chọn 800 bài dùng cho huấn luyện và 609 bài dùng cho việc kiểm tra. Cụ thể số liệu như sau:

Bảng 3.1: Số lượng tập dữ liệu huấn luyện và kiểm tra

|  |  |  |
| --- | --- | --- |
| Chủ đề | Tập dữ liệu huấn luyện | Tập dữ liệu kiểm tra |
| Kinh doanh/Bán hàng | **200** | **233** |
| Công nghệ thông tin (CNTT) | **200** | **179** |
| Nhân sự | **200** | **117** |
| Khách sạn/Du lịch | **200** | **80** |

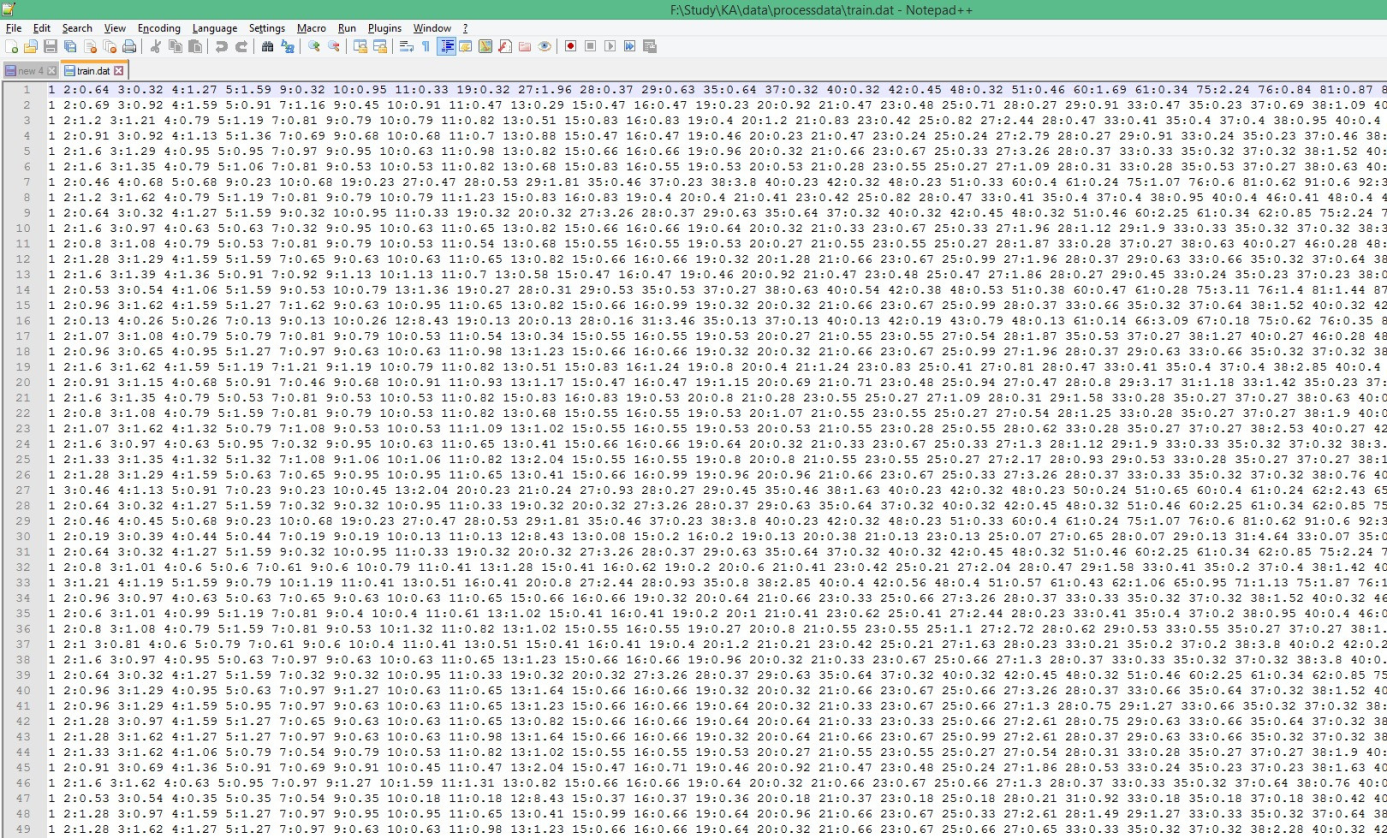
Sau các bước tiền xử lý, văn bản được biểu diễn dưới dạng:

<classi> <label1>:<value1> <label2>:<value2> ... <labeln>:<valuen>

Trong đó:

* classi là nhãn phân loại của mỗi chủ đề.
* labeli là chỉ số của từ đặc trưng thứ i trong không gian đặc trưng có xuất hiện ở phản hồi đó.
* valuei là trọng số của indexi được tính bởi công thức TF.IDF, nếu valuei = 0 thì không cần phải ghi đặc trưng đó.

Định dạng này tuân theo định dạng dữ liệu đầu vào của chương trình SVMMulticlass.



Hình 3.4: Biểu diễn văn bản dưới dạng vector

*Đánh giá độ chính xác của thuật toán:*

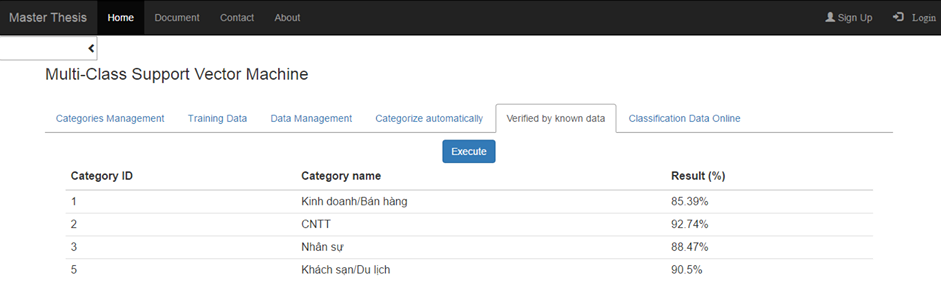
Chúng ta không thể khẳng định một phương pháp phân lớp văn bản cụ thể nào là chính xác hoàn toàn. Vì vậy việc đưa ra độ đo để đánh giá hiệu quả của thuật toán phân lớp giúp chúng ta có thể xác định được độ chính xác của thuật toán, từ đó áp dụng thuật toán đó vào việc phân lớp.

Độ chính xác có thể được tính theo các công thức:

Công thức đánh giá số 1:

Công thức đánh giá số 2:

Sau đây là kết quả thực nghiệm trên bộ test gồm 609 văn bản:



Hình 3.5: Kết quả độ chính xác của việc phân loại

3.2 Các tiêu chuẩn dùng để đánh giá

* + 1. ***URL phục vụ cho ứng dụng***

Mục đích của ứng dụng tìm kiếm thông tin việc làm đó là lấy những thông tin việc làm từ các website đưa vào hệ thống, sau đó sẽ tiến hành phân loại các việc làm đó theo chủ đề đã có. Danh sách các URL em sử dụng để phục vụ cho ứng dụng như sau:

[http://worklink.vn](http://worklink.vn/?gclid=CjwKEAiAjvrBBRDxm_nRusW3q1QSJAAzRI1tTwJW1a4LIhZQT60REsSbsTqh9dG44LdT6XBYtO2n7hoC5T_w_wcB)

<http://careerbuilder.vn/tim-viec-lam.html>

<https://www.timviecnhanh.com/>

<https://www.vietnamworks.com/tim-viec-lam>

<https://mywork.com.vn/>

3.3 Phương pháp thực nghiệm

Mục đích của ứng dụng là xây dựng một danh sách các việc làm được lấy từ các website được liệt kê trong phần 3.2.1 và tiến hành phân loại chúng vào 4 category có sẵn là Kinh doanh/Bán hàng, Công nghệ thông tin, Nhân sự và Khách sạn/Du lịch

3.4 Kết quả thực nghiệm

**Bước 1**: Input URL từ các danh sách các website việc làm không support RSS

**Bước 2**: Trích xuất văn bản từ các URL

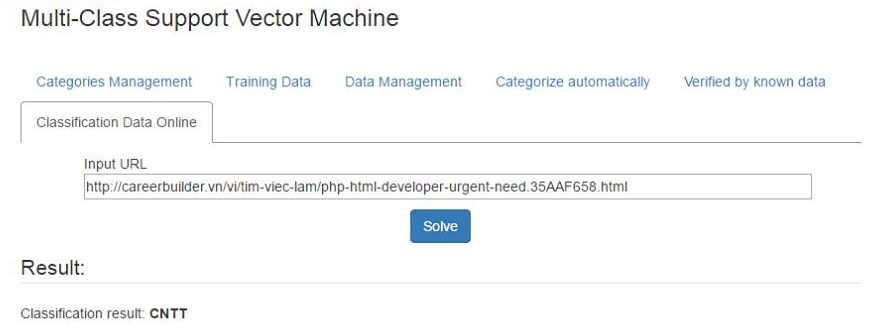
**Bước 3**: Tiền xử lý văn bản

2.5 So sách với một số phương pháp khác

* Tách từ
* Loại bỏ từ dừng
* Làm sạch văn bản
* Vector hóa văn bản theo định dạng đầu vào của thuật toán SVM.

**Bước 4**: Thực hiện phân loại theo bốn chủ đề

2.6 Độ phức tạp và thời gian thực hiện phương pháp



Hình 3.6: Phân loại URL được input vào ứng dụng

3.7 Kết luận chương

Chương này đã trình bày cụ thể về việc xây dựng bộ dữ liệu huấn luyện và thử nghiệm, từ đó tính toán được độ chính xác của thuật toán; và đưa ra ứng dụng về việc phân loại URLs của các website việc làm vào nhãn tương ứng.

KẾT LUẬN VÀ KIẾN NGHỊ

1 Kết quả đạt được

Luận văn tiến hành nghiên cứu giải quyết bài toán phân loại văn bản tiếng Việt dựa vào đặc trưng. Bài toán là nền tảng cho nhiều ứng dụng quan trọng thực tế như lọc thư rác, hệ thống khuyến cáo/quan điểm người dùng,…

Những kết quả chính mà đồ án đạt được:

* Nghiên cứu và tìm hiểu về bài toán phân lớp, trình bày một số thuật toán phân lớp tiêu biểu
* Tìm hiểu một số đặc điểm của văn bản tiếng Việt phục vụ cho quá trình tiền xử lý
* Tìm hiểu và áp dụng các công cụ tiền xử lý dữ liệu đầu vào
* Nghiên cứu và tìm hiểu về thuật toán Support Vector Machine trên hai lớp và nhiều lớp
* Xây dựng chương trình lấy tin RSS hàng ngày
* Xây dựng chương trình huấn luyện và kiểm thử với bộ dữ liệu lấy từ chương trình lấy tin RSS.
* Xây dựng chương trình phân loại việc làm vào các category có sẵn từ các website không hỗ trợ RSS

2 Hạn chế

* Một văn bản đầu vào sau khi phân loại chỉ thuộc về một loại nhãn.
* Hạn chế số lượng và chất lượng của kho dữ liệu tin tức ảnh hưởng đến chất lượng phân loại của hệ thống.
* Cần xác định giá trị chuẩn để một văn bản thuộc vào 1 hoặc nhiều thể loại, hoặc không thuộc thể loại nào.

3 Hướng phát triển

* Xây dựng bộ dữ liệu lớn hoàn chỉnh, phong phú về các lĩnh vực việc làm và chia các lĩnh vực càng nhỏ càng tốt
* Cần xác định giá trị chuẩn để một văn bản thuộc vào 1 hoặc nhiều thể loại, hoặc không thuộc thể loại nào
* Cải thiện hiệu xuất, tăng tốc độ xử lý dữ liệu
* Kiểm soát được thông tin lấy về theo ngày tháng năm mong muốn đảm bảo tính cập nhật của việc làm

DANH MỤC TÀI LIỆU THAM KHẢO

1. Do Viet Phuong and Tu Minh Phuong. “*Gender Prediction Using Browsing History”.* [KSE (1) 2013](http://dblp.uni-trier.de/db/conf/kse/kse2013-1.html#PhuongP13): 271-283.
2. Argamon, S., M. Koppel, J. Fine & A. R. Shimoni (2003). Gender, genre, and writing style in formal written texts. Text, 23(3).
3. Popescu, A. & G. Grefenstette (2010). Mining user home location and gender from Flickr tags. In Proc. of ICWSM-10, pp. 1873–1876.
4. Katja Filippova. User Demographics and Language in an Implicit Social Network
5. Claudia Peersman, Walter Daelemans, Leona Van Vaerenbergh. Predicting Age and Gender in Online Social Networks

PHỤ LỤC