MỤC LỤC

[CHƯƠNG 1. BÀI TOÁN PHÂN LOẠI VĂN BẢN 3](#_Toc509578331)

[1.1. Giới thiệu 3](#_Toc509578332)

[1.2. Phát biểu bài toán 3](#_Toc509578333)

[1.3. Mô hình tổng quát 4](#_Toc509578334)

[1.4. Mục đích và phạm vi nghiên cứu 6](#_Toc509578335)

[CHƯƠNG 2. PHÂN LOẠI VĂN BẢN SỬ DỤNG SUPPORT VECTOR MACHINE 7](#_Toc509578336)

[2.1. Giới thiệu về Tf-idf 7](#_Toc509578337)

[Tf- term frequency 7](#_Toc509578338)

[IDF- Inverse Document Frequency 7](#_Toc509578339)

[2.2. Cơ sở lý thuyết 7](#_Toc509578340)

[2.7.1. Tiền xử lý văn bản 7](#_Toc509578341)

[2.7.2. Phương pháp biểu diễn văn bản 8](#_Toc509578342)

[2.7.3. Mô hình không gian vector 8](#_Toc509578343)

[2.7.4. Khái niệm trọng số 8](#_Toc509578344)

[2.3. Thuật toán support vector machine 9](#_Toc509578345)

[2.3.1. Định nghĩa 10](#_Toc509578346)

[2.3.2. Ý tưởng của phương pháp 10](#_Toc509578347)

[2.3.3. Nội dung phương pháp 11](#_Toc509578348)

[2.3.4. Bài toán phân 2 lớp với SVM 11](#_Toc509578349)

[2.3.5. Bài toán nhiều phân lớp với SVM 12](#_Toc509578350)

[2.3.6. Các bước chính của phương pháp SVM 13](#_Toc509578351)

[CHƯƠNG 3. GIAO DIỆN CHƯƠNG TRÌNH DEMO 14](#_Toc509578352)

[3.1. Luồng của chương trình 14](#_Toc509578353)

[3.2. Một số hình ảnh của chương trình 14](#_Toc509578354)

**LỜI MỞ ĐẦU**

Phân loại văn bản là một vấn đề quan trọng trong lĩnh vực xử lý ngôn ngữ. Nhiệm vụ của bài toán này là gán các tài liệu văn bản vào nhóm các chủ đề cho trước. Đây là một bài toán rất thường gặp trong thực tế điển hình như : một nhà chuyên phân tích thị thường chứng khoán, anh ta cần phải tổng hợp rất nhiều tài liệu, bài viết về thị trường chứng khoán để đọc và đưa ra phán đoán của mình. Tuy nhiên, anh ta không thể đọc tất cả các bài viết, bài báo hay các tài liệu để rồi phân loại chúng đâu là tài liệu chứng khoán sau đó anh ta mới đọc kỹ chúng cho mục đích của anh ta. Lý do của vấn đề này là bởi ví số lượng bào viết, bài báo hiện nay rất nhiều, đặc biệt là trên internet, nếu để đọc hết được tất cả tài liệu đó thì sẽ mất rất nhiều thời gian. Một ví dụ khác trong thực tế là việc phân loại spam mail. Khi một mail được gửi đến hộp thư, nếu để người dùng phải đọc tất cả các mail thì sẽ tốn rất nhiều thời gian vì spam mail rất nhiều. Vì vậy, cần có một hệ thống phân loại đâu là spam mail và đâu là mail tốt.

Để giải bài toán này đã có rất nhiều phương pháp được đưa ra như : thuật toán Naïve Bayes, K-NN (K-Nearest-Neighbor), Cây quyết định (Decision Tree), Mạng Neuron nhân tạo (Artificial Neural Network) và SVM (Support Vector Machine). Mỗi phương pháp đều cho kết quả khá tốt cho bài toán này, tuy nhiên phương pháp phân loại văn bản bằng thuật toán Bayes được sử dụng phổ biến hơn cả và dễ dàng cài đặt. Chính vì vậy chúng em lựa chọn đề tài: “**Phân loại văn bản bằng thuật toán Support Vector Machine**“ làm đề tài kết thúc môn học của mình.

Chúng em xin chân thành cảm ơn các thầy, cô bộ môn đã tận tình giảng dạy em trong suốt thời gian học tập vừa qua. Nhờ có sự chỉ dạy tận tình của cô Nguyễn Thị Thu Hà trực tiếp hướng dẫn giảng dạy, cùng sự đào tạo của các thầy cô bộ môn khác trong thời gian qua giúp chúng em hoàn thành bài báo cáo này. Do đây là lần đầu tiên triển khai một hệ thống có tính thực tiễn cao, nên quá trình triển khai có thể còn nhiều sai sót. Mong các thầy cô đóng góp ý kiến giúp chúng em hiểu rõ hơn về bài toán.

Một lần nữa chúng em xin chân thành cảm ơn các thầy cô!

# BÀI TOÁN PHÂN LOẠI VĂN BẢN

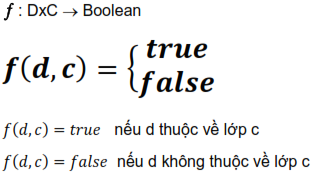
## Giới thiệu

Ngày nay, sự bùng nổ thông tin do bị tác động bởi sự xuất hiện của các siêu phương tiện và World Wide Web (WWW) đã làm cho không gian dữ liệu gia tăng thường xuyên, điều này tạo ra một thách thức cho các hệ thống truy vấn thông tin sao cho có hiệu quả. Một trong những khó khăn mà các hệ thống thông tin thường phải gặp đó là tần suất cập nhật của các thông tin quá lớn .Phương thức sử dụng giấy trong giao dịch đang dần được số hóa, do nhiều tính năng vượt trội mà phương thức này mang lại, như là có thể lưu trữ lâu dài, cập nhật, sửa đổi, tìm kiếm một cách nhanh chóng. Do đó số lượng văn bản số hóa ngày nay đang tăng dần theo cấp số nhân, cùng với sự gia tăng của số lượng văn bản, nhu cầu tìm kiếm văn bản cũng tăng theo, khi đó phân loại văn bản tự động là một yêu cầu cấp thiết được đặt ra. Phân loại văn bản giúp sẽ giúp chúng ta tìm kiếm thông tin một cách nhanh chóng hơn thay vì phải tìm lần lượt trong từng văn bản, hơn nữa khi mà số lượng văn bản đang gia tăng một cách nhanh chóng thì thao tác tìm lần lượt trong từng văn bản sẽ mất rất nhiều thời gian, công sức và là một công việc nhàm chán và không khả thi. Chính vì thế nhu cầu phân loại văn bản tự động là thực sự cần thiết.

Có rất nhiều công trình nghiên cứu về phân loại văn bản và đã có được những kết qủa đáng khích lệ, như là: Support Vector Machine, K–Nearest Neighbor, Linear Least Squares Fit, Neural Network, Naïve Bayes, Centroid Base… 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: đầu tiên 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 qủa phân loại tốt nhất.

## Phát biểu bài toán

Bài toán phân loại văn bản có thể được phát biểu như sau : Cho trước một tập văn bản D={d1,d2,…,dn} và tập chủ đề được định nghĩa C={c1,c2,…,cn}. Nhiệm vụ của bài toán là gán lớp di thuộc về cj đã được định nghĩa. Hay nói cách khác, mục tiêu của bài toán là đi tìm hàm f :



## Mô hình tổng quát

Có rất nhiều hướng tiếp cận bài toán phân loại văn bản đã được nghiên cứu như: tiếp cận bài toán phân loại dựa trên lý thuyết đồ thị, cách tiếp cận sử dụng lý thuyết tập thô, cách tiếp cận thống kê… Tuy nhiên, tất cả các phương pháp trên đều dựa vào các phương pháp chung là máy học đó là : học có giám sát, học không giám sát và học tăng cường.

Vấn đề phân loại văn bản theo phương pháp thống kê dựa trên kiểu học có giám sát được đặc tả bao gồm 2 giai đoạn : giai đoạn huấn luyện và giai đoạn phân lớp.

* Giai đoạn huấn luyện

Chúng ta có một tập huấn luyện, mỗi phần tử trong tập huấn luyện được gán vào một hoặc nhiều lớp mà chúng ta sẽ thể hiện chúng bằng một mô hình mã hoá. Thông thường, mỗi phần tử trong tập huấn luyện được thể hiện theo dạng .Trong đó,  là vector biểu diễn cho văn bản trong tập huấn luyện.

Sau đó, chúng ta định nghĩa một lớp mô hình và một thủ tục huấn luyện. Lớp mô hình là họ các tham số của bộ phân loại, thủ tục huấn luyện là một giải thuật (hay thuật toán) để chọn ra một họ các tham số tối ưu cho bộ phân loại.

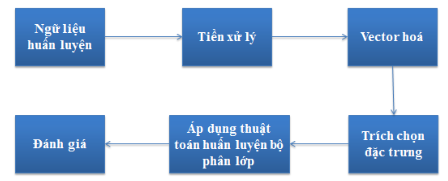


* 1. Mô hình giai đoạn huấn luyện

Đầu vào : ngữ liệu huấn luyện và thuật toán huấn luyện

Đầu ra : mô hình phân lớp (bộ phân lớp – classifier)

Các bước trong giai đoạn huấn luyện:



* 1. Các bước trong giai đoạn huấn luyện

Trong đó :

Ngữ liệu huấn luyện : kho ngữ liệu thu thập từ nhiều nguồn khác nhau.

Tiền xử lý : chuyển đổi tài liệu trong kho ngữ liệu thành một hình thức phù hợp để phân loại.

Vector hoá : mã hoá văn bản bởi một mô hình trọng số

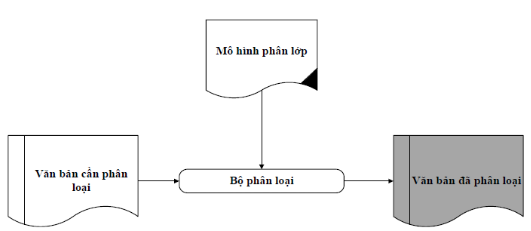
Trích chọn đặc trưng : loại bỏ những từ (đặc trưng) không mang thông tin khỏi tài liệu nhằm nâng cao hiệu suất phân loại và giảm độ phức tạp của thuật toán huấn luyện.

Thuật toán huấn luyện : Thủ tục huấn luyện bộ phân lớp để tìm ra họ các tham số tối ưu.

Đánh giá : bước đánh giá hiệu suất (chất lượng) của bộ phân lớp

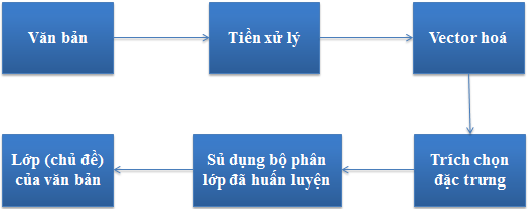
* Giai đoạn phân lớp

Sau khi đã hoàn thành giai đoạn huấn luyện, mô hình phân lớp sẽ được áp dụng cho các văn bản mới cần phân loại.



* 1. Mô hình giai đoạn phân lớp

Các bước trong giai đoạn phân lớp:



* 1. Các bước trong giai đoạn phân lớp

## Mục đích và phạm vi nghiên cứu

Mục đích:

* Tìm hiểu thuật toán Support Vector Machine và phạm vi ứng dụng của nó trong phân loại văn bản.
* Nắm rõ hơn về cơ chế học tập và huấn luyện máy học.
* Xây dựng một chương trình có khả năng phân loại văn bản sau khi tìm hiểu thuật toán.

Phạm vi nghiên cứu: chương trình thực hiện trong một quy mô nghiên cứu nhỏ với một số lượng văn bản không nhiều vào khoảng < 3000 văn bản.

# CHƯƠNG 2. PHÂN LOẠI VĂN BẢN SỬ DỤNG SUPPORT VECTOR MACHINE

## 2.1. Giới thiệu về Tf-idf

### Tf- term frequency

dùng để ước lượng tần xuất xuất hiện của từ trong văn bản. Tuy nhiên với mỗi văn bản thì có độ dài khác nhau, vì thế số lần xuất hiện của từ có thể nhiều hơn . Vì vậy số lần xuất hiện của từ sẽ được chia độ dài của văn bản (tổng số từ trong văn bản đó)

TF(t, d) = ( số lần từ t xuất hiện trong văn bản d) / (tổng số từ trong văn bản d)

### IDF- Inverse Document Frequency

dùng để ước lượng mức độ quan trọng của từ đó như thế nào . Khi tính tần số xuất hiện tf thì các từ đều được coi là quan trọng như nhau. Tuy nhiên có một số từ thường được được sử dụng nhiều nhưng không quan trọng để thể hiện ý nghĩa của đoạn văn , ví dụ :

Từ nối: và, nhưng, tuy nhiên, vì thế, vì vậy, …

Giới từ: ở, trong, trên, …

Từ chỉ định: ấy, đó, nhỉ, …

Vì vậy ta cần giảm đi mức độ quan trọng của những từ đó bằng cách sử dụng IDF :

IDF(t, D) = log\_e( Tổng số văn bản trong tập mẫu D/ Số văn bản có chứa từ t )

## 2.2. Cơ sở lý thuyết

### 2.7.1. Tiền xử lý văn bản

Văn bản trước khi được vector hoá, tức là trước khi sử dụng, cần phải được tiền xử lý. Quá trình tiền xử lý sẽ giúp nâng cao hiệu suất phân loại và giảm độ phức tạp của thuật toán huấn luyện.

Tuỳ vào mục đích bộ phân loại mà chúng ta sẽ có những phương pháp tiền xử lý văn bản khác nhau, như :

Chuyển vẳn bản về chữ thường.

Loại bỏ dấu câu (nếu không thực hiện tách câu).

Loại bỏ các kí tự đặc biệt biệt([ ],[.], [,], [:], [“], [”], [;], [/], [[]], [~], [`], [!], [@], [#], [$],[%],[^],[&],[\*],[(],[)]), các chữ số, phép tính toán số học.

Loại bỏ các stopword (những từ xuất hiện hầu hết trong các văn bản) không có ý nghĩa khi tham gia vào phân loại văn bản.

…

### 2.7.2. Phương pháp biểu diễn văn bản

Một trong những nhiệm vụ đầu tiền trong việc xử lý phân loại văn bản là chọn được một mô hình biểu diễn văn bản thích hợp. Một văn bản ở dạng thô (dạng chuỗi) cần được chuyển sang một mô hình khác để tạo thuận lợi cho việc biểu diễn và tính toán. Tuỳ thuộc vào từng thuật toán phân loại khác nhau mà chúng ta có mô hình biểu diễn riêng. Một trong những mô hình đơn giản và thường được sử dụng trong nhiệm vụ này là mô hình không gian vector. Một văn bản trong nhiệm vụ này được biểu diễn theo dạng , với  là một vector n chiều để đo lường giá trị của phần tử văn bản.

### 2.7.3. Mô hình không gian vector

Mô hình không gian vector là một trong những mô hình được sử dụng rộng rãi nhất cho việc tìm kiếm (truy hồi) thông tin. Nguyên nhân chính là bởi vì sự đơn giản của nó.

Trong mô hình này, các văn bản được thể hiện trong một không gian có số chiều lớn, trong đó mỗi chiều của không gian tương ứng với một từ trong văn bản. Phương pháp này có thể biểu diễn một cách hình tượng như sau : mỗi văn bản D được biểu diễn dưới dạng  (vector đặc trưng cho văn bản D). Trong đó,  =( x1, x2, … , xn ), và n là số lượng đặc trưng hay số chiều của vector văn bản, xi là trọng số của đặc trưng thứ i (với 1≤ i ≤n).

Như vậy, nếu trong kho ngữ liệu của quá trình huấn luyện nhiều văn bản, ta kí hiệu Dj, là văn bản thứ j trong tập ngữ liệu, và vector j=( x1j , x2j , … , xnj ) là vector đặc trưng cho văn bản Dj, và xij là trọng số thứ i của vector văn bản j.

### 2.7.4. Khái niệm trọng số

Một vấn đề quan trọng nữa trong việc biểu diễn một văn bản đó là tính trọng số cho vector đặc trưng của văn bản. Có nhiều cách khác nhau để tính trọng số này như :

Word frequency weighting

Boolean weighting

tf\*idf weighting

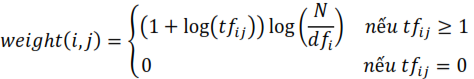
Entropy weighting

Tuy nhiên, để đơn giản cho vấn đề này, chúng ta sẽ chỉ xem xét cách tính Word frequency weighting (trọng số tần suất từ) và tf\*idf, một cách đơn giản đó là đếm số từ đó trong văn bản. Tuy nhiên vẫn có nhiều cách khác nhau để tính trọng số dạng này.

Có ba thông tin được sử dụng trong cách tính trọng số bằng tần suất từ là : term frequency ( tfij số lần suất hiện của từ wi trong văn bản dj ), document frequency (dfi số văn bản có chứa từ wi), collection frequency ( cfi số lần suất hiện của từ wi trong cả tập ngữ liệu). Trong đó, dfi ≤ cfi và ∑j tfij =cfi.

Thông tin được nắm bắt bởi term frequency là sự nổi bật của thông tin (hay từ) trong một văn bản. Term frequency càng cao (số lần xuất hiện càng nhiều trong văn bản) thì đó là từ miêu tả tốt cho nội dung văn bản. Giá trị thứ hai, document frequency, có thể giải thích như là một bộ chỉ định nội dung thông tin. Một từ được tập trung ngữ nghĩa thường xảy ra nhiều lần trong một văn bản nếu nó cũng xuất hiện trong tất cả các văn bản khác. Nhưng từ không được tập trung ngữ nghĩa trải đều đồng nhất trong tất cả các văn bản.

Để thể hiện trọng số phản ánh hết thông tin của từ, thường ta sẽ kết hợp cả hai loại trọng số là tf và df trong một đơn vị chung. Dạng biểu diễn trọng số này được gọi là tf \* idf. Công thức kết hợp hai giá trị trọng số :



Trong đó, N là tổng số văn bản. Biểu thức thứ nhất áp dụng cho các từ có xuất hiện trong văn bản, còn biểu thức thứ hai cho các từ không xuất hiện trong văn bản.

## 2.3. Thuật toán support vector machine

Bài toán phân lớp (*Classification*) và dự đoán (*Prediction*) là hai bài toán cơ bản và có rất nhiều ứng dụng trong tất cả các lĩnh vực như: học máy, nhận dạng, trí tuệ nhân tạo, .v.v . Trong đề tài này, chúng em sẽ đi sâu nghiên cứu phương pháp Support Vector Machines (SVM), một phương pháp rất hiệu quả hiện nay.

Phương pháp SVM được coi là công cụ mạnh cho những bài toán phân lớp phi tuyến tính được các tác giả Vapnik và Chervonenkis phát triển mạnh mẽ năm 1995. Phương pháp này thực hiện phân lớp dựa trên nguyên lý Cực tiểu hóa Rủi ro có Cấu trúc SRM (*Structural Risk Minimization*), được xem là một trong các phương pháp phân lớp giám sát không tham số tinh vi nhất cho đến nay. Các hàm công cụ đa dạng của SVM cho phép tạo không gian chuyên đổi để xây dựng mặt phẳng phân lớp

### 2.3.1. Định nghĩa

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

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

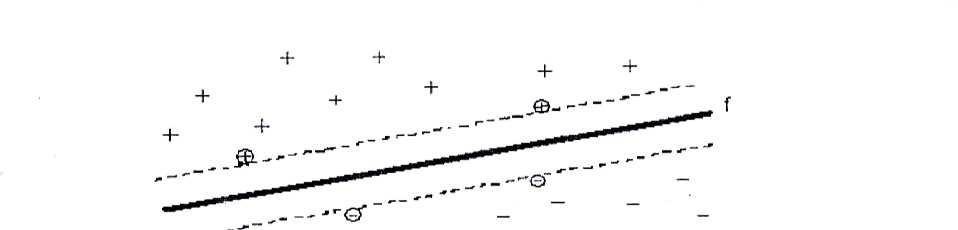
Là 1 phương pháp thử nghiệm, đƣa ra 1 trong những phương pháp mạnh và chính xác nhất trong số các thuật toán nổi tiếng về phân lớp dữ liệu

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

### 2.3.2. Ý tưởng của phương pháp

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

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



Hình 2. 5: Siêu phẳng phân chia dữ liệu học thành 2 lớp + và - với khoảng cách biên lớn nhất. Các điểm gần nhất (điểm được khoanh tròn) là các Support Vector.

### 2.3.3. Nội dung phương pháp

**Cơ sở lý thuyết**

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

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

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

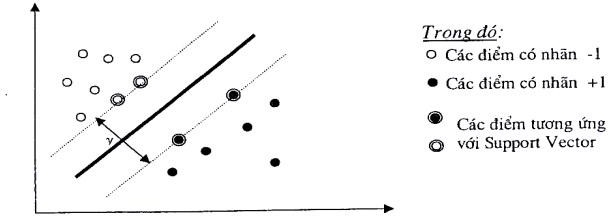
TÓM LẠI: trong trường hợp nhị phân phân tách tuyến tính, việc phân lớp được thực hiện qua hàm quyết định f(x) = sign(<w.x> + b), hàm này thu được bằng việc thay đổi vectơ chuẩn w, đây là vectơ để cực đại hóa viền chức năng

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

### 2.3.4. Bài toán phân 2 lớp với SVM

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

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



Hình 2. 6: Minh họa bài toán 2 phân lớp bằng phương pháp SVM

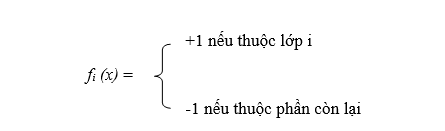
Các điểm mà nằm trên hai siêu phẳng phân tách được gọi là các Support Vector. Các điểm này sẽ quyết định đến hàm phân tách dữ liệu

### 2.3.5. Bài toán nhiều phân lớp với SVM

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

*fi(x)* = wiix + bi

Những phần tử x là support vector sẽ thỏa điều kiện



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

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

### 2.3.6. Các bước chính của phương pháp SVM

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

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

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

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

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

# CHƯƠNG 3. GIAO DIỆN CHƯƠNG TRÌNH DEMO

## 3.1. Luồng của chương trình

B1: Chuẩn bị dữ liệu

Dữ liệu ban đầu khi crawl về sẽ ở định dạng json. Chúng ta cần tách json để lấy được dữ liệu cần thiết. Đồng thời, xử lý trường hợp bị lỗi encoding.

B2: Tiền xử lí dữ liệu

Loại bỏ hết các ký hiệu thừa, loại bỏ các dữ liệu quá ngắn, sử dụng lib tách từ để tách từ tiếng việt

B3: Chuẩn bị dữ liệu phù hợp với input của bài toán – Cấu trúc lại dữ liệu

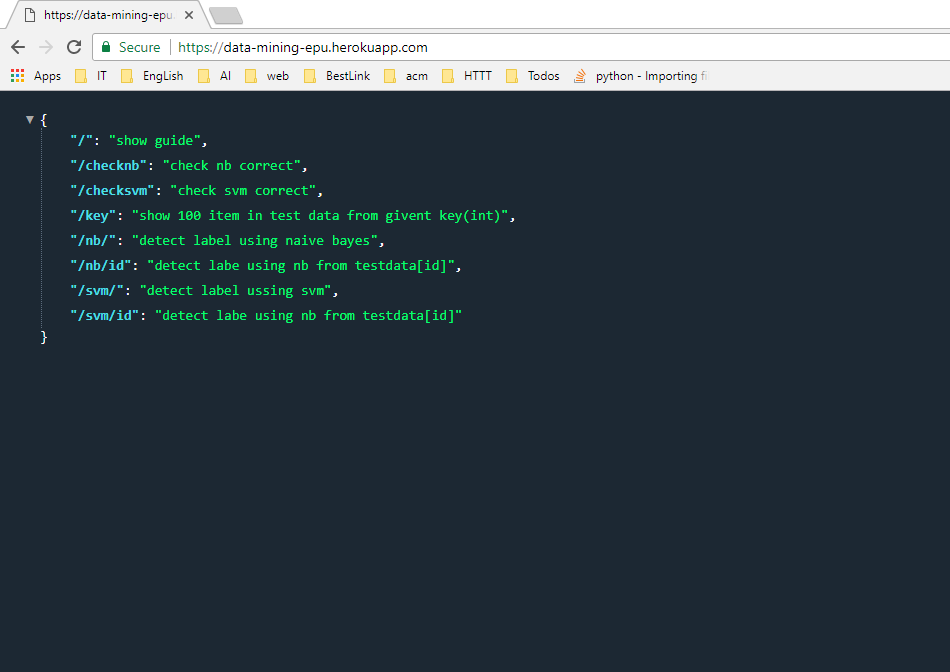
Đưa dữ liệu về một chuẩn có cấu trúc: <content>, <label>. Trộn đều dữ liệu trước khi tách riêng train và test

B3: Tiến hành huấn luyện và lưu lại model

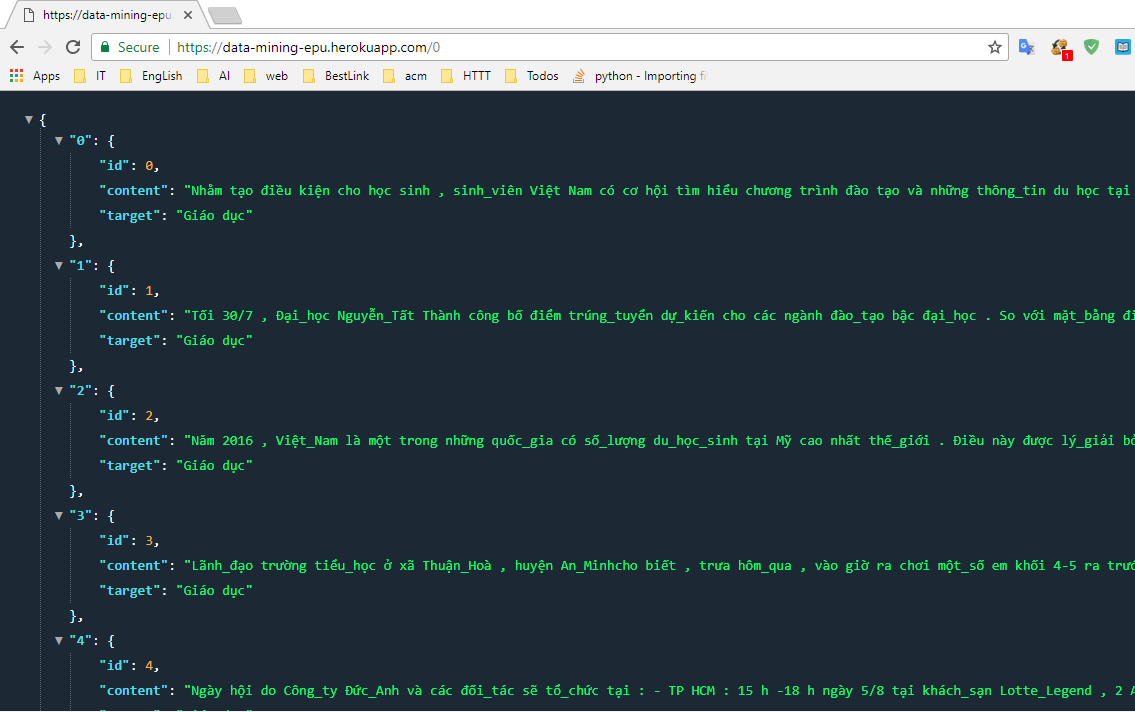
B5: Đọc model đã lưu và tiến hành kiểm tra

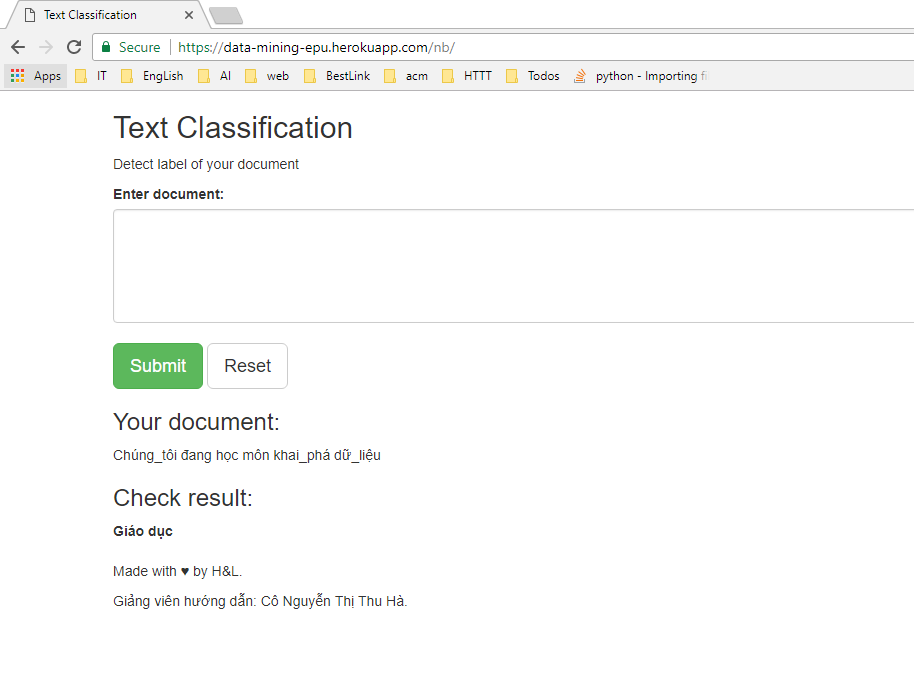
## 3.2. Một số hình ảnh của chương trình

**Giao diện hướng dẫn sử dụng**



**Giao diện xem dữ liệu của tập test**



**Giao diện kiểm tra nhãn của text nhập từ bàn phím dung thuật toán navie bayes.**

**Giao diện hiện thị độ chính xác của thuật toán svm**

