ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN NGÔN NGỮ HỌC NGỮ LIỆU

BÀI TOÁN PHÂN LOẠI VĂN BẢN

Lớp: **CS321.J21**

Giảng viên hướng dẫn: **TS Nguyễn Thị Quý**

Sinh viên thực hiện:

1. **Nguyễn Phạm Long Duy** **16520299**
2. **Trần Hoàng Phát** **16520918**
3. **Phạm Ngọc Phúc Thuần** **16521206**
4. **Phan Đăng Lâm** **16521710**

Tp. Hồ Chí Minh, tháng 06 năm 2019

MỤC LỤC

LỜI NÓI ĐẦU

Xin cảm ơn khoa Khoa học máy tính và cô Nguyễn Thị Quý đã giúp nhóm có cơ hội thực hiện báo cáo này và hiểu rõ hơn về hiện tại và tương lai của ngành Xử lý ngôn ngữ tự nhiên. Từ những kiến thức nền tảng được truyền tải ở mỗi buổi học và việc tự học hỏi và nghiên cứu những kiến thức mới, nhóm đã hoàn thành báo cáo tìm hiểu về Text classification (tiếng việt: Phân loại văn bản). Trong quá trình thực hiện, những sai sót là điều không thể tránh khỏi. Chính vì vậy, nhóm mong nhận được những ý kiến từ giảng viên để báo cáo và kiến thức được hoàn thiện hơn.

Trong thời đại bùng nổ công nghệ thông tin, mọi thông tin đều được số hoá để lưu trữ trên các thiết bị điện tại hoặc truyền tại trên mạng. Ngày nay, với những ưu điểm như cách lưu trữ gọn nhẹ, thời gian lưu trữ dài, tiện dụng trong trao đổi, dễ dàng sửa đổi, … số lượng văn bản số tăng lên một cách nhanh chóng, đặc biệt là trên nền tảng world-wide-web. Điều này dẫn đến nhu cầu tìm kiếm văn bản cũng tăng theo. Với số lượng văn bản đồ sộ thì việc phân loại văn bản là một nhu cầu bức thiết.

BẢNG ĐÁNH GIÁ CÔNG VIỆC

STT	*MSSV*	*Họ và tên*	*Nội dung phân công*	*Mức độ hoàn thiện*
1	16520299	Nguyễn Phạm Long Duy	- Thu nhập và xử lý dữ liệu - Kiểm tra từ điển stopwords - Huấn luyện model - Chuẩn bị seminar - Báo cáo: chương 4, 5	90%
2	16520918	Trần Hoàng Phát	- Huấn luyện model - Xây dựng mô hình nhận diện chủ đề văn bản - Chuẩn bị seminar - Báo cáo: chương 4	100%
3	16521206	Phạm Ngọc Phúc Thuần	- Quản lý tiến độ của đồ án - Thu nhập và xử lý dữ liệu - Tổng hợp dữ liệu, xây dựng tập dữ liệu train và test - Chuẩn bị seminar - Báo cáo: chương 1, 3	95%
4	16521710	Phan Đăng Lâm	- Thu nhập và xử lý dữ liệu - Kiểm tra từ điển stopwords - Tổng hợp dữ liệu, xây dựng tập dữ liệu train và test - Chuẩn bị seminar - Báo cáo: chương 2	95%

CHƯƠNG 1: TÌM HIỂU VỀ PHÂN LOẠI VĂN BẢN

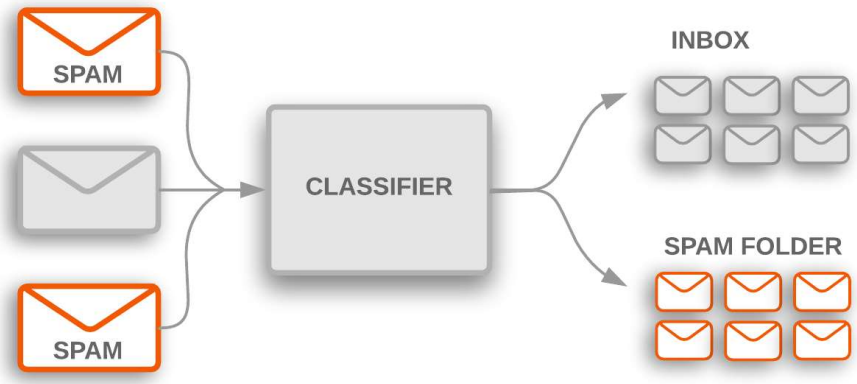
Phân loại văn bản (tiếng anh: Text classification) ra đời và là một trong những bài toán máy học nền tảng thuộc lĩnh vực Natural Language Processing (viết tắt là NLP, tiếng Việt là Xử lý Ngôn ngữ tự nhiên).

1. Thế nào là Phân loại văn bản?

Phân loại văn bản là công việc gán nhãn chủ đề đã được xác định trước cho các văn bản tự do (văn bản chưa được phân loại). Thông qua đó, người sử dụng có thể dễ dàng tìm kiếm, quản lý và sắp xếp thông tin văn bản.

Hình 1.1: Bài toán Phân loại văn bản

Các thuật toán Phân loại văn bản rất quan trọng, được xem là trung tâm của những hệ thống xử lý dữ liệu ở quy mô lớn. Điển hình ở công cụ lọc mail rác, một mail được gửi tới sẽ phải vượt qua "cửa kiểm duyệt" để phân loại vào hộp thư đến hay hộp thư rác.



Hình 1.2: Mô tả đơn giản về công cụ lọc mail rác

2. Cách hoạt động của bài toán Phân loại văn bản:

Hai phương pháp phân loại văn bản chính: thủ công và tự động.

Đối với phương pháp thủ công, người gán nhãn có nhiệm vụ gán nhãn phù hợp cho văn bản dựa theo nội dung. Phương pháp này tuy rằng thường cung cấp kết quả tốt nhưng lại tốn nhiều thời gian và chi phí.

Phương pháp tự động áp dụng thành tựu của Máy học, Xử lý ngôn ngữ tự nhiên và các kĩ thuật khác để tự động phân loại văn bản một cách nhanh hơn và hiệu quả hơn về mặt chi phí. Ba hệ thống phổ biến dùng để phân loại văn bản tự động:

- Rule-based systems (Hệ thống dựa trên quy tắc)
- Machine Learning based systems (Hệ thống sử dụng máy học)
- Hybrid systems (Hệ thống kết hợp)

2.1. Hệ thống dựa trên quy tắc:

Hệ thống dựa trên quy tắc phân loại văn bản theo một bộ quy tắc ngôn ngữ thủ công. Sau khi thực hiện gán nhãn thủ công, người lập trình sẽ suy ra được quy tắc của từng chủ đề văn bản, thường dựa trên yếu tố ngữ nghĩa của nội dung văn bản.

Ví dụ, một văn bản tự do cần được phân loại vào một trong hai chủ đề: thể thao và chính trị. Bộ quy tắc của chủ đề thể thao sẽ bao gồm danh sách các từ liên quan đến lĩnh vực đó như “bóng đá”, “sân cỏ”, “bóng bàn”, “Ronaldo”, … Tương tự, bộ quy tắc

của thể lại chính trị bao gồm "Donal Trump", "Putin", "Kim Jong Un", … Hệ thống sẽ kiểm tra số lượng từ liên quan đến hai chủ đề này. Cuối cùng, chủ đề nào có số lượng từ liên quan nhiều hơn thì văn bản cần phân loại sẽ thuộc chủ đề đó.

Hệ thống dựa trên quy tắc khá đơn giản, người sử dụng dễ dàng hiêu và có thể cải thiện theo thời gian. Nhưng, hệ thống này yêu cầu người sử dụng có kiến thức chuyên sâu về các thẻ tên. Việc tạo quy tắc cho một hệ thống phức tạp có thể gặp nhiều khó khăn, đòi hỏi nhiều sự phân tích và thử nghiệm, dẫn đến tốn nhiều thời gian và chi phí. Đồng thời, việc duy trì chất lượng và kết quả phân loại của bộ quy tắc cũng sẽ bị ảnh hưởng do các quy tắc mới được thêm vào.

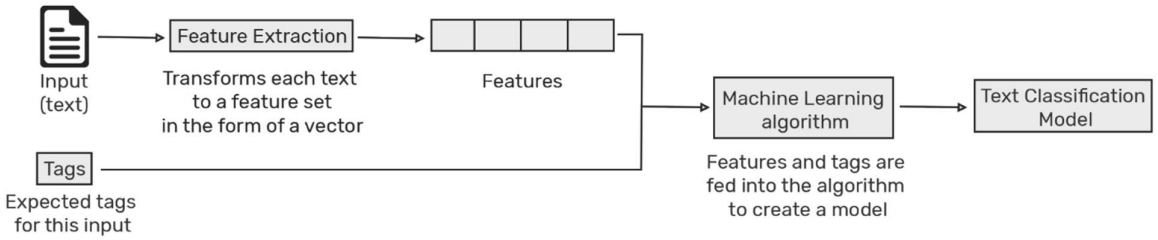
2.2. Hệ thống sử dụng máy học:

Khác với hệ thống dựa trên quy tắc, hệ thống này sử dụng thành tựu của máy học để phân loại văn bản dựa trên quan sát và kinh nghiệm. Tuy nhiên, hệ thống này cũng cần dùng những dữ liệu đã gán nhãn thủ công trước để làm dữ liệu huấn luyện. Thuật toán máy học sẽ học những mối liên quan giữa các chủ đề văn bản, từ đó, thuật toán sẽ đưa ra một output là kết quả phân loại bằng mô hình đã được dạy.

Bước đầu tiên cần làm là khai thác các đặc tính của dữ liệu: phương pháp này được sử dụng để số hoá từng kí tự trong văn bản thành từng biểu diễn số dưới dạng một vector. Một trong những cách tiếp cận phổ biến nhất là *bag of words (túi chứa từ)*, một vector biểu thị tần số của từ trong kho từ ngữ đã được xác định từ trước.

Ví dụ, kho từ vựng ban đầu gồm các từ {Tôi, không, thích, bạn, lắm} và câu được vector hoá là "Tôi thích bạn". Vector biểu diễn cho câu đó sẽ như sau: (1, 0, 1, 1, 0).

Sau đó, dữ liệu huấn luyện bao gồm các cặp đặc tính của dữ liệu (vector cho từng ví dụ văn bản) và các thẻ tên chủ đề (thể thao, chính trị), được cung cấp cho thuật toán máy học để tạo mô hình phân loại:



Sau khi được huấn luyện với đủ mẫu dữ liệu, mô hình máy học có thể bắt đầu đưa ra những dự đoán chính xác. Độ chính xác của hệ thống sử dụng máy học thường cao hơn hệ thống dựa trên luật, đặc biệt là trong các trường hợp phân loại phức tạp. Đồng thời, hệ thống này cũng dễ dàng duy trì và tự học thêm những mẫu dữ liệu mới từ việc phân loại những văn bản đầu vào mới.

Các thuật toán máy học dùng tạo mô hình phân loại văn bản phổ biến:

- Navie Bayes
- Support vector machines
- Deep Learning

2.3. Hệ thống kết hợp:

Hệ thống này là sự kết hợp của hai hệ thống đã nêu trên, nhằm cải thiện độ chính xác của kết quả. Các hệ thống kết hợp có thể dễ dàng tinh chỉnh bằng cách thêm những quy tắc cụ thể cho các thẻ tên mà hệ thống phân loại cơ sở gán nhãn chưa đúng.

CHƯƠNG 2: XÂY DỰNG KHO DỮ LIỆU

Kho dữ liệu dùng để huấn luyện máy (tên gọi là dataset) là một phần không thể thiếu của máy học. Dữ liệu càng tốt giúp máy quan sát và đưa ra những dự đoán càng chính xác. Tuỳ vào nhu cầu và bài toán, việc xây dựng dataset sẽ khác nhau. Sau đây là hướng dẫn xây dựng một dataset cơ bản nhất.

1. Lựa chọn dữ liệu:

Trước hết chúng ta phải chọn dữ liệu phù hợp với bài toán cần giải quyết, ta sẽ quan tâm về các thuộc tính của dữ liệu đó, dữ liệu càng có đầy đủ thuộc tính thì máy học đưa ra dự đoán càng chính xác. Nếu thuộc tính bị thiếu hơn khoảng 25% thì dữ liệu đó không phù hợp dành cho máy học.

Đối với bài toán Phân loại phân loại văn bản, nhóm lựa chọn dữ liệu là các bài báo từ các kênh tin tức điện tử chính thống như Báo Thanh niên (https://thanhnien.vn/), Báo Tuổi trẻ (https://tuoitre.vn/), Báo Người lao động (https://nld.com.vn/) và Báo VNExpress (https://vnexpress.net/).

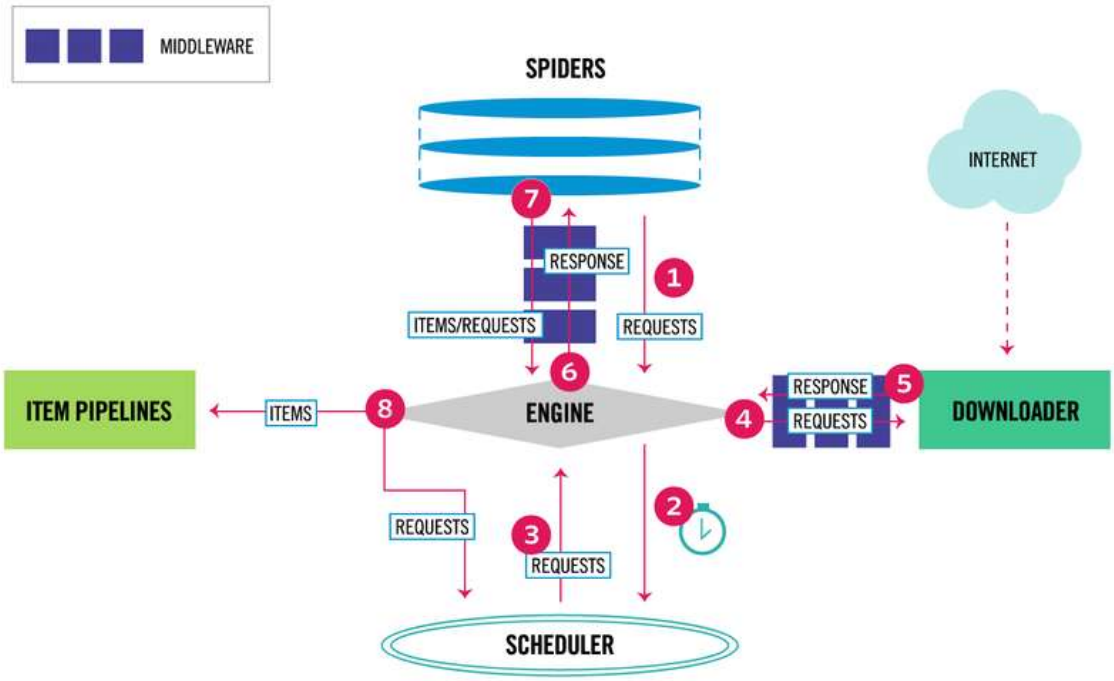
2. Thu nhập dữ liệu:

Sau khi lựa chọn dữ liệu ta phải thu thập chúng sao cho đủ nhiều tùy vào loại bài toán để làm data tranning cho máy.

Một số cách để thu nhập dữ liệu:

- Tải các dữ liệu mở (open data): miễn phí, có nhiều trang web hỗ trợ nhưng không đủ để đáp ứng trong các bài toán phức tạp.
- Tự tạo và gán nhãn cho dữ liệu (data): độ chính xác sẽ cao hơn nhưng tốn nhiều thời gian để thực hiện.
- Scrape data tự động từ các trang web: có thể tạo các crawler tool đủ mạnh để lấy dữ liệu về nhanh chóng, ví dụ như python có thư viện beautifulSoup hỗ trợ parse data html dễ dàng.

Nhóm sử dụng Scrapy Framework – một framework mạnh mẽ về thu nhập dữ liệu.



Hình 2.1: Mô hình hoạt động của một Scrapy

Các thành phần của một Scrapy:

- *Scrapy Engine:* có trách nhiệm kiểm soát luồng dữ liệu giữa tất cả các thành phần của hệ thống và kích hoạt các sự kiện khi một số hành động xảy ra
- *Scheduler:* Giống như một hàng đợi (queue), scheduler sắp xếp thứ tự các URL cần download
- *Dowloader:* Thực hiện tải các trang web xuống và cung cấp cho engine
- Spiders là class được viết bởi người dùng, chúng có trách nhiệm bóc tách dữ liệu cần thiết và tạo các url mới để nạp lại cho scheduler qua engine.
- *Item Pipeline:* Những dữ liệu được bóc tách từ spiders sẽ đưa tới đây, Item pipeline có nhiệm vụ xử lý chúng và lưu vào cơ sở dữ liệu
- *Các Middlewares:* Là các thành phần nằm giữa Engine với các thành phần khác, chúng đều có mục đích là giúp người dùng có thể tùy biến, mở rộng khả năng xử lý cho các thành phần.
    + Spider middlewares: Là thành phần nằm giữa Eninge và Spiders, chúng xử lý các response đầu vào của Spiders và đầu ra (item và các url mới).

+ Dowloader middlewares: Nằm giữa Engine và Dowloader, chúng xử lý các request được đẩy vào từ Engine và các response được tạo ra từ Dowloader
+ Scheduler middlewares: Nằm giữa Engine và Scheduler để xử lý những requests giữa hai thành phần

3. Xử lý dữ liệu:

Xử lý dữ liệu bao gồm chọn dữ liệu phù hợp từ bộ dữ liệu có được và xây dựng một bộ tranning set, sẽ có một bài bước say đây:

- *Tổ chức và định dạng dữ liệu*: Dữ liệu ban đầu thu được có thể ở một số file khác nhau, các bảng khác nhau trong database, chúng ta cần định dạng nó thành một bộ nhất định. Ngoài ra có thể các dataset ở các ngôn ngữ khác nhau, chúng ta nên chuyển nó thành một ngôn ngữ chung, thường là tiếng Anh vì đây là ngôn ngữ dễ xử lý và thông dụng.
- *Làm sạch dữ liệu*: Đây là một trong những bước chính trong quá trình xử lý dữ liệu. Bước này xử lý những thuộc tính còn thiếu và loại bỏ những thuộc tính không cần sử dụng. Ví dụ, yêu cầu thuộc tính tuổi nhưng một vài người không ghi tuổi, thì chúng ta có thể xóa thuộc tính này hoặc thay thế nó.
- *Khai thác thuộc tính*: Bước này liên quan đến việc phân tích và tối ưu hóa số lượng thuộc tính. Ta chỉ chọn những thuộc tính quan trọng có ảnh hưởng đến kết quả của bài toán và chọn chúng để tối ưu hóa thuật toán và tiết kiệm tài nguyên hơn.

Đối với các dữ liệu văn bản được lấy từ các trang tin tức, việc xử lý dữ liệu phần lớn là loại bỏ stopwords (từ chức năng) và tổng hợp lại thành một tập tin dạng json theo chủ đề (chi tiết ở phần 4 – phân loại dữ liệu).

4. Phân loại dữ liệu:

Bước này có thể có hoặc không, giúp cho thuật toán tối ưu hơn, nếu ta chuyển những giá trị số vào các nhóm khác nhau.

Ví dụ, ta có một tập dữ liệu khách hàng mua hàng online, ta quan sát và nhận ra rằng tuổi 13 – 14 và 25 – 26 có sự khác biết lớn về thói quen mua hàng. Khi đó, ta có

thể chia nó thành 2 nhóm: trẻ em và người lớn. Thuật toán sẽ tối ưu, ít tốn chi phí và kết quả tốt hơn.

Bài toán phân loại văn bản được phân loại theo chủ đề (hay chủ đề) của nội dung văn bản đó. Tổng cộng 14 chủ đề, bao gồm:

- Health (Sức khoẻ) – 289 văn bản
- Science (Khoa học) – 246 văn bản
- IT (Công nghệ thông tin) – 128 vản bản
- Sports (Thể thao) – 299 văn bản
- Politics and Society (Chính trị và Xã hội) – 133 văn bản
- Business (Kinh tế) – 280 văn bản
- Culture (Văn hoá) – 298 văn bản
- Law (Pháp luật) – 290 văn bản
- Life (Đời sống) – 140 văn bản
- Education (Giáo dục) – 150 văn bản
- Technology (Công nghệ) – 150 văn bản
- Entertainment (Giải trí) – 150 văn bản
- World (Thế giới) – 297 văn bản

CHƯƠNG 3: CƠ SỞ LÝ THUYẾT

1. Tính trọng số TF-IDF:

Trọng số TF-IDDF (Term Frequency – Inverse Document Frequency) được sử dụng để đo tầm quan trọng của từng đặc trưng đối với câu trong kho ngữ liệu. Trọng số TF-IDF càng cao thì dặc trưng càng quan trọng. Trọng số TF-IDF của đặc trưng f trong câu s được tính như sau:

Tính tần số đặc trưng (TF: term frequency):

$$tf_{fs} = \frac{n_{fs}}{\sum n_s}$$

Trong đó, n_{fs} là số lần xuất hiện của đặc trưng f trong câu s và n_s là tổng số lần xuất hiện của các term trong câu s.

Trọng số TF-IDF của đặc trưng f trong câu s được tính như sau:

$$tf - idf_{fs} = \frac{tf\ _{fs} \times \log\frac{S}{sf_f}}{\sqrt{\sum_{r\in F}\left(tf_{rs} \times \log\frac{S}{sf_r}\right)^2}}$$

Trong đó:

S là tổng số câu trong tập ngữ liệu.

sf_s là số câu đặc trưng f

F là tập đặc trưng ($f \in F$)

$\log TF - IDF$: được tính theo công thức sau:

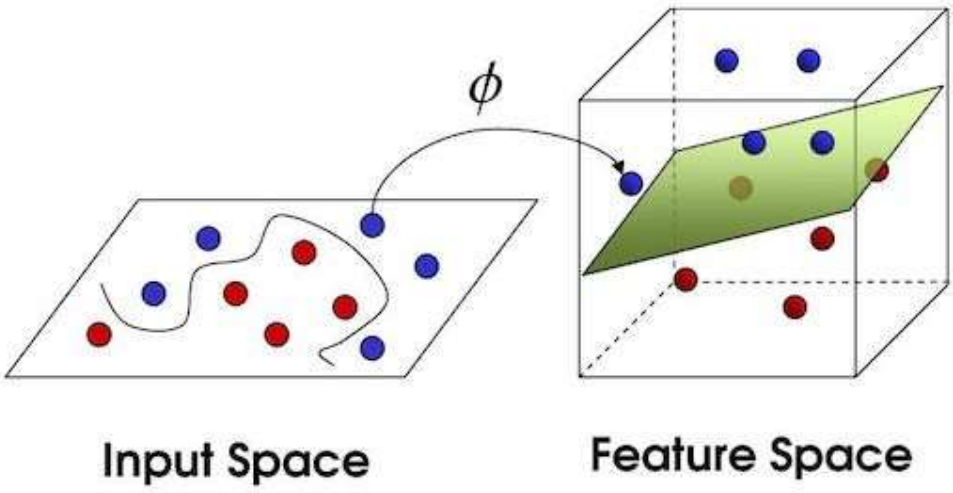
$$tf - idf_{fs} = \frac{l_f^s \times \log\frac{S}{sf_f}}{\sqrt{\sum_{r\in F}\left(l_r^s \times \log\frac{S}{sf_r}\right)^2}}$$

Với:

$$l_f^s = \begin{cases} 0 \\ \log tf_{fs} + 1 \end{cases}$$

2. Thuật toán phân lớp máy học vector hỗ trợ (SVM):

Support Vector Machine (SVM) là một thuật toán thuộc nhóm Supervised Learning (Học có giám sát) dùng để phân chia dữ liệu (Classification) thành các nhóm riêng biệt. SMV có nhiều ưu điểm hơn so với các phương pháp khác như: xử lý với tính ổn định cao trên dữ liệu phức tạp, vector đặc trưng có thể có số chiều lớn và quan trọng hơn cả là xử lý tốt vấn đề thưa.



Hình 3.1: Ánh xạ tập dữ liệu từ không gian 2 chiều sang không gian 3 chiều

Cơ sở lý thuyết về SVM như sau:

- Xét bài toán phân loại 2 phân lớp với tập dữ liệu mẫu: $\{(x_i, y_i) | \ i = 1,2, \dots, N, x_i \in R^m\}$

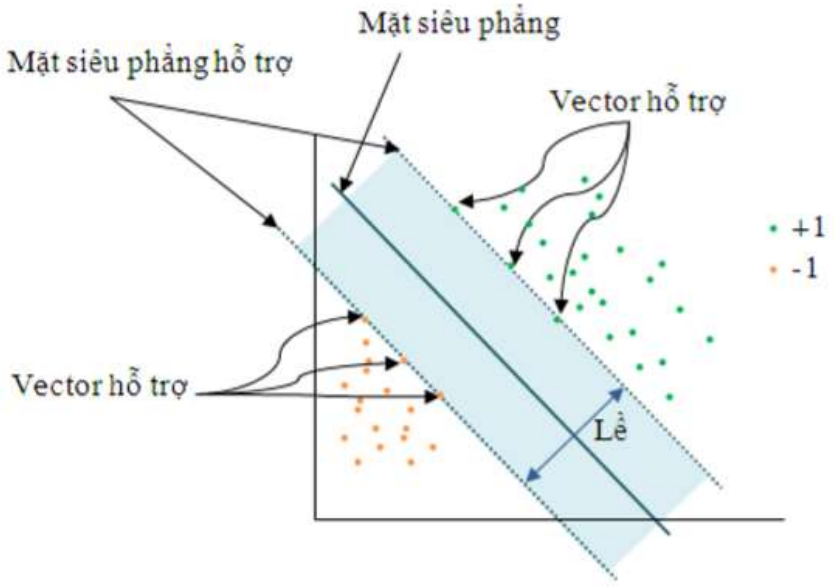
Trong đó, mẫu x_i là các vector đối tợng đợc phân loại thành các mẫu dơng và mẫu âm:

- Các mẫu dương là các mẫu x_i thuộc lĩnh vực quan tâm và được gán nhãn y_i = 1;
- Các mẫu âm là các mẫu x_i không thuộc lĩnh vực quan tâm và đợc gán nhãn $y_i = -1$;

- SVM là bộ phân loại nhị phân, là mặt siêu phẳng phân tách các mẫu dợơng khỏi các mẫu âm với độ chênh lệch (lề - margin) cực đại. Lề được xác định bằng khoảng cách giữa các mẫu dương và các mẫu âm gần mặt siêu phẳng nhất.

Trong đó:

- f(x) = +1 nếu $w^T x + b \geq 0 \rightarrow x$ thuộc lớp dương (lĩnh vực quan tâm)
- f(x) = -1 nếu $w^T x + b < 0 \rightarrow x$ thuộc lớp âm (lĩnh vực không quan tâm)



Hình 3.2: Mặt siêu phẳng tách các mẫu dương khỏi các mẫu âm

Mặt siêu phẳng có phương trình:

$$w^T x + b = 0$$

Trong đó *w* là vector có trọng số, *b* là độ dịch. Khi thay đổi *w* và *b*, hướng và khoảng cách từ gốc toạ độ đến mặt siêu phẳng thay đổi.

Máy học SVM là một họ các mặt siêu phẳng phụ thuộc vào các tham số *w* và *b*.

Mục tiêu của phương pháp SVM là ước lượng *w* và *b* để cực đại hoá lề giữa các lớp dữ liệu dương và âm. Lề càng lớn thì quá trình phân loại dữ liệu càng tối ưu.

Tìm w,b:

(1) Nếu tập dữ liệu huấn luyện *khả tách tuyến tính*:

Ta có các ràng buộc sau:

- $w^T x_i + b \geq +1$ nếu $y_i = +1$
- $w^T x_i + b \leq -1$ nếu $y_i = -1$

Hai mặt siêu phẳng có phương trình là $w^T x_i + b = \pm 1$ được gọi là các mặt siêu phẳng hỗ trợ.

Để xây dựng một mặt siêu phẳng lề tối ưu, ta phải giải bài toán quy hoạch toàn phương sau:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Với: $\alpha_i \geq 0$: Hệ số Lagrange

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

- ❖ <u>*Tính w*</u>: w sẽ được tính từ các nghiệm của bài toán qui hoạch toàn phương trên như sau:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

- ❖ <u>*Tính độ dịch b*</u>: chọn một mẫu x_i sao cho với $\alpha_i > 0$ (vector hỗ trợ), sau đó sử dụng điều kiện Karush-Kuhn-Tucker (KKT) như sau:

$$\alpha_i [y_i (w^T x_i + b) - 1] = 0$$

(2) Nếu tập dữ liệu huấn luyện *không khả tách tuyến tính*: có 2 cách giải quyết.

<u>Cách 1:</u> Sử dụng một mặt siêu phẳng lề mềm, nghĩa là cho phép một số mẫy huấn luyện nằm về phía sai của mặt siêu phẳng phân tác hoặc vẫn ở vị trí đúng nhưng rơi vào vùng giữa mặt siêu phẳng phân tách và mặt siêu phẳng hỗ trợ tương ứng. Trong trường

hợp này, các hệ số Lagrange của bài toán quy hoạch toàn phương có thêm một cận trên *C* dương – tham số do người sử dụng lựa chọn.

Cách 2: Sử dụng một ánh xạ phi tuyến Φ để ánh xạ các điểm dữ liệu đầu vào sang một không gian mới có số chiều cao hơn. Trong không gian này, các điểm dữ liệu trở thành khả tách tuyến tính, hoặc có thể phân tách với ít lỗi hơn so với trường hợp sử dụng không gian ban đầu. Khi đó, bài toán quy hoạch toàn phương ban đầu sẽ trở thành:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

Với: $0 \le \alpha_i \le C$

$\sum_{i=1}^{N} \alpha_i y_i = 0$

$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$: Hàm kernal

Dùng hàm kernel, ta không cần biết rõ về ánh xạ Φ

Một số hàm kernel thông dụng:

- Hàm đa thức:

$$k(x_i, x_j) = (x_i^T x_j + 1)^p$$

  - Với *p* là bậc của đa thức, được chỉ định bởi người sử dụng
- Hàm kernel Radial (Radial basis function network)

$$k(x_i, x_j) = \exp\left(-\frac{1}{(\sigma)^2}\left|\left| x_i - x_j\right|\right|^2\right)$$

  - Với σ^2 được chỉ định bởi người sử dụng.
- Mạng neural nhân tạo 2 lớp (two-layer perceptron)

$$k(x_i, x_j) = \tanh(\beta_0 x_i^T x_j + \beta_1)$$

  - Tanh là hàm tang hyperbol
  - Giá trị β_0 và β_1 do người dùng chỉ định.

3. Đánh giá:

Các độ đo được sử dụng để đánh giá hiệu quả của bài toán phân loại văn bản bao gồm: Độ chính xác (P: Precision), độ bao phủ (R: Recall), F_1.

Độ chính xác của chủ đề i được tính theo công tthức:

$$P = \frac{N_1}{N_2}$$

Độ bao phủ của chủ đề i được tính theo công thức:

$$R = \frac{N_1}{N}$$

F_1 của chủ đề i được tính dựa theo độ chính xác và độ bao phủ như sau:

$$F_1 = \frac{2 \times P \times R}{P + R}$$

Trong đó:

N_1: Số câu được nhận diện chủ đề đúng là i

N_2: Số câu được nhận diện chủ đề là i

N: Số câu chủ đề i

Để đánh giá chất lượng cho toàn hệ thống, các độ đo cần được tính ra trung bình trên tất cả các chủ đề văn bản. Giả sử có n chủ đề, trung bình P, R, F_1 được tính theo công thức sau:

$$P = \frac{\sum_{i=1}^{n} P_i}{n}$$

$$R = \frac{\sum_{i=1}^{n} R_i}{n}$$

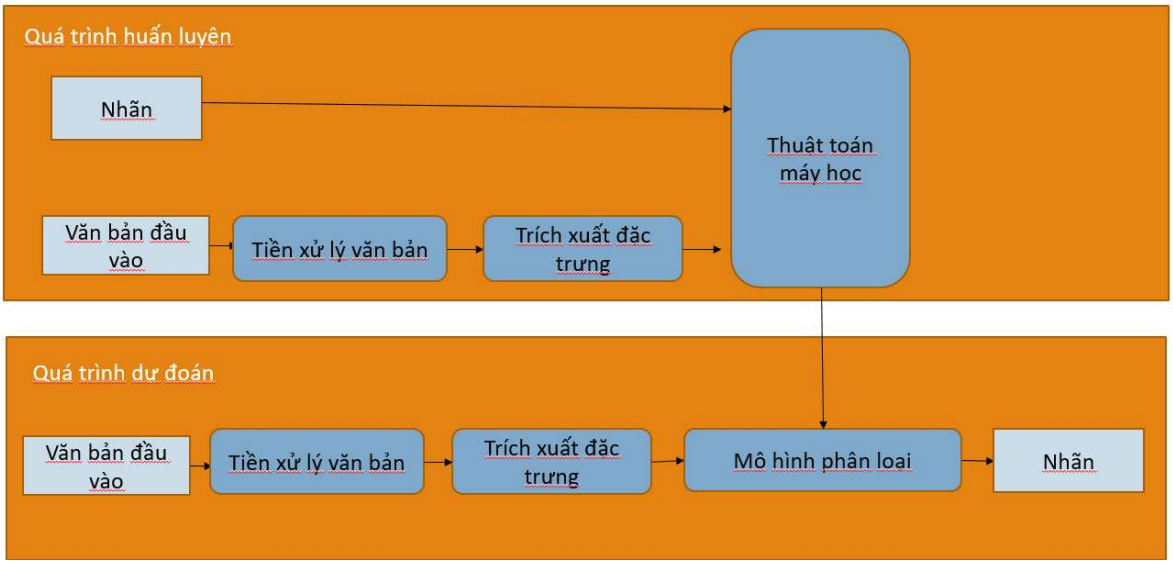
$$F_1 = \frac{\sum_{i=1}^{n} F_{1_i}}{n}$$

CHƯƠNG 4: XÂY DỰNG KHỐI NHẬN DIỆN CHỦ ĐỀ VĂN BẢN

Nhận diện chủ đề văn bản là quá trình nhận biết và gán một chủ đề đã được xác định trước cho văn bản đó. Đối với văn bản tiếng Việt do vẫn còn hạn chế do vẫn chưa có một kho ngữ liệu để train, so khớp đánh giá kết quả. Nguồn dữ liệu của chúng tôi được thu thập từ các trang web: Báo Tuổi Trẻ, Người Lao Động, VN Express được chia ra sẵn thành các tập theo chủ đề. Bài toán phân loại chủ đề văn bản hiện tại có nhiều hướng tiếp cận: Support Vector Machine (SVM), cây quyết định, Naïve Bayes, … và trong đồ án này chúng tôi tập trung vào phương pháp sử dụng thuật toán phân lớp SVM.

1. Xây dựng khối nhận diện văn bản:

1.1. Mô hình:



Hình 4.1: Mô hình bài toán phân loại văn bản

1.2. Một số kỹ thuật trong bài toán nhận diện chủ đề văn bản:

1.2.1. Xoá các từ chức năng (stop words):

Các từ chức năng (stop words) là những từ xuất hiện nhiều trong ngôn ngữ , tuy nhiên lại không mang nhiều ý nghĩa về mặt ngữ nghĩa, từ vựng, mà chỉ có ý nghĩa về mặt ngữ pháp. Ở tiếng Việt các từ chức năng là những từ như: ở, ủa, đâu, ơi, nhưng,…Có rất nhiều cách để xóa các từ chức năng nhưng hiện nay có 2 cách chính là:

- Dùng từ điển: Đây là cách đơn giản nhất, ta sẽ lọc lại văn bản, loại bỏ những từ xuất hiện trong từ điển

```
1797  đâu đây
1798  đâu đó
1799  đây
1800  đây này
1801  đây rồi
1802  đây đó
1803  đã
1804  đã hay
1805  đã không
1806  đã là
1807  đã lâu
1808  đã thế
1809  đã vậy
1810  đã đủ
1811  đó
1812  đó đây
1813  đúng
1814  đúng ngày
1815  đúng ra
1816  đúng tuổi
1817  đúng với
1818  đơn vị
1819  đua
1820  đua cho
1821  đua chuyện
1822  đua em
1823  đua ra
1824  đua tay
1825  đua tin
1826  đua tới
1827  đua vào
1828  đua về
1829  đua xuống
1830  đua đến
1831  được
1832  được cái
1833  được lời
1834  được nước
1835  được tin
1836  đại loại
1837  đại nhân
1838  đại phàm
1839  đại để
1840  đạt
```

Hình 4.2: Từ điển các từ chức năng (stopwords)

- Dựa theo tần suất xuất hiện của từ: Trong cách này, ta sẽ tiếng hành lọc và loại bỏ khỏi văn bản những từ xuất hiện nhiều lần (hoặc ít lần). Thường những từ rơi vào 2 trường hợp này sẽ là những từ không mang nhiều ý nghĩa.

Đối với đồ án này, nhóm chúng tôi sử dụng phương pháp dùng từ điển, với 1,942 từ chức năng, được tham khảo và lấy từ tác giả Phạm Văn Toàn[4].

1.2.2. Tách từ:

Trong tiếng Việt, dấu cách không được sử dụng như một kí hiệu để nhận biết sự phân tách giữa các từ như các văn bản tiếng Anh, mà nó chỉ có ý nghĩa phân tách các âm tiết với nhau. Việc tách từ chính xác là rất quan trọng, vì nếu tách không chính xác sẽ làm lệch lạc đi ý nghĩa của câu, văn bản, từ đó giảm tính chính xác của chương trình. Vì vậy khi xử lý văn bản tiếng Việt, tách từ (word segmentation) là một công đoạn cơ bản và quan trọng bậc nhất.

1.3. Chuyển đổi văn bản về dạng vector :

Văn bản sẽ được chuyển về và lưu trữ dưới dạng vector. Ta sẽ vector hóa từng câu thay vì cả văn bản để giữ độ tính chính xác cao cho chương trình. Thành phần của vector được trình bày từ được sắp xếp theo thứ tự từ điển và trọng số TF-IDF của chúng được tính nhờ vào tool scikit-learn.

1.4. Xây dựng mô hình phân lớp:

Sau các quá trình tiền xử lý và trích xuất đặt trưng xong, các vector của dữ liệu đầu vào sẽ được đưa vào thuật toán SVM (sử dụng tool scikit-learn) để xây dựng nên mô hình phân lớp.

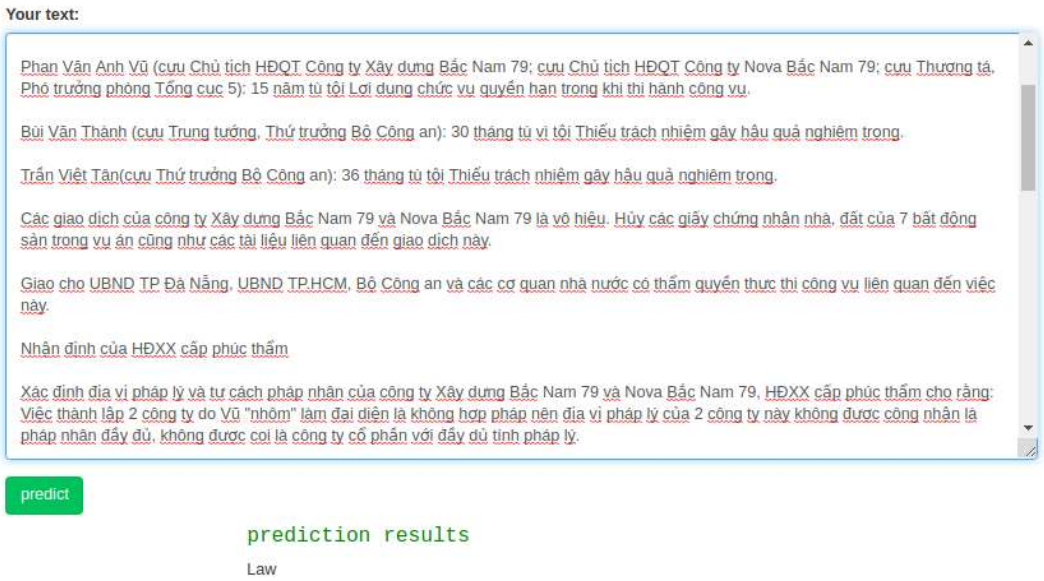
2. Kết quả thực nghiệm:

2.1. Ngữ liệu chưa được loại bỏ từ chức năng (stop words):

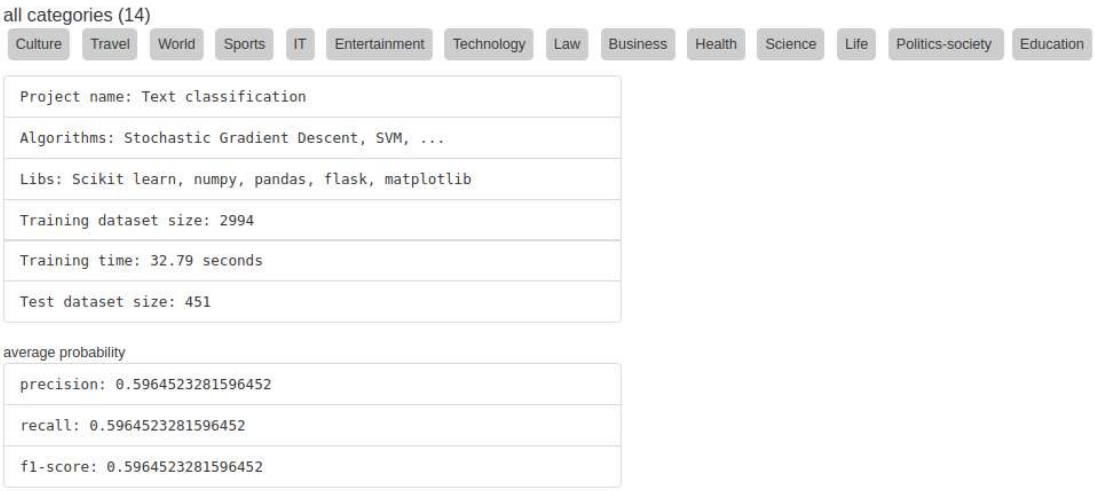
Hình 4.3: Thông số độ kết quả khi dùng ngữ liệu thô

Thực nghiệm huấn luyện chương trình bằng ngữ liệu thô không qua loại bỏ các từ chức năng, với 2,994 dataset huấn luyện và 451 dataset thử nghiệm, chương trình thu về được kết quả chỉ số trung bình F_1 đạt xấp xỉ 60.98%.

2.2. Ngữ liệu đã được loại bỏ từ chức năng (stop words):



Hình 4.4: Thử nghiệm nhận diện chủ đề với một bài báo ngẫu nhiên



Hình 4.5: Thông số huấn luyện, thử nghiệm và kết quả nhận diện chủ đề

Sau quá trình huấn luyện chương trình với 2,994 dataset và thử nghiệm với 451 dataset, kết quả thu về được vẫn còn thấp, chỉ số trung bình của F_1 chỉ đạt xấp xỉ 59.96%.

2.3. So sánh:

So sánh kết quả F1 của 2 lần thực nghiệm ta thấy được kết quả thu được từ nguồn ngữ liệu thô (~60.98%) cao hơn từ ngữ liệu đã loại bỏ từ chức năng (~59.96%). Sự khác biệt này là lượng dataset để huấn luyện vẫn còn nhỏ, phân bố dataset giữa các chủ đề vẫn chưa đồng đều nên sinh ra sự chênh lệch không đáng kể này.

3. Kết luận:

Ta có thể thấy kết quả F1 chương trình qua tiền xử lý loại bỏ các từ chức năng vẫn còn thấp (~59.65%) dù kết quả khi thử nghiệm các bài báo ngẫu nhiên vẫn chính xác. Vấn đề cho độ chính xác thấp có thể do dataset để train còn ít nên với những tên riêng bị trùng giữa các văn bản thuộc các chủ để khác nhau đã gây ra việc giảm độ chính xác của chương trình.

CHƯƠNG 5: ĐÁNH GIÁ VÀ HƯỚNG PHÁT TRIỂN

1. Đánh giá:

Trong thời gian ngắn thực hiện đồ án, nhóm đã học được cách xây dựng và sử dụng kho ngữ liệu. Đồng thời, nhóm cũng đã xây dựng và huấn luyện một mô hình phân loại văn bản sử dụng mô hình SVM đơn giản. Tuy nhiên, độ chính xác sản phẩm vẫn chưa hoàn thiện. Do số lượng dữ liệu đối với từng chủ đề còn ít và chưa đồng đều.

2. Hướng phát triển:

Ngày nay, ứng dụng của bài toán phân loại văn bản không chỉ để nhận diện thể loại và chủ đề (topic labeling) mà còn được áp dụng vào nhiều lĩnh vực khác như:

- Sentiment analysis (phân tích cảm xúc)
- Language detection (nhận diện ngôn ngữ)
- Intent detection (nhận diện ý muốn)

Để cải thiện độ chính xác của mô hình hiện tại, nhóm sẽ tập trung cải thiện thuật toán và kho ngữ liệu ban đầu. Các thành viên trong nhóm có thể cài đặt mô hình phân lớp SVM vào các đồ án hoặc dự án trong tương lai.

TÀI LIỆU THAM KHẢO

1. Luận văn Nhận diện chủ đề của văn bản và ứng dụng trong dịch máy thống kê Anh-Việt-Anh – Tiến sĩ Nguyễn Thị Quý

2. Tổng quan về bài toán Phân loại văn bản: https://developers.google.com/machine-learning/guides/text-classification/

3. Luận văn Tìm hiểu các hướng tìm hiểu các hướng tiếp cận bài toàn phân loại văn bản và xây dựng phần mềm phân loại tin tức bảo điện tử - Cử nhân Nguyễn Trần Thiên Thanh và Cử nhân Trần Khải Hoàng: http://doan.edu.vn/do-an/luan-van-tim-hieu-cac-huong-tiep-can-bai-toan-phan-loai-van-ban-va-xay-dung-phan-mem-phan-loai-tin-tuc-bao-dien-tu-21366/

4. Bài viết Phân loại văn bản tiếng Việt tự động – tác giả Phạm Văn Toàn: https://viblo.asia/p/phan-loai-van-ban-tieng-viet-tu-dong-phan-1-yMnKM3bal7P?fbclid=IwAR383Kcc277EbkHeEjGN6kH_L1CUkbGN9tqhS_RJ7_BT2AfmRE86R5y8xa4#_source-code-10

5. Bài viết Giới thiệu tiền xử lý trong xử lý ngôn ngữ tự nhiên – tác giả T.T.T: https://kipalog.com/posts/Gioi-thieu-tien-xu-ly-trong-xu-ly-ngon-ngu-tu-nhien

6. Bài toán Phân loại văn bản – MonkeyLearn: https://monkeylearn.com/text-classification/

7. Các phương pháp đánh giá một hệ thống phân lớp - Machine Learning cơ bản: https://machinelearningcoban.com/2017/08/31/evaluation/