

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

\_\_\_\_\_ \* \_\_\_\_\_



- BÁO CÁO -

## XỬ LÝ NGÔN NGỮ TỰ NHIÊN

### Đề tài : Phân loại thư rác tiếng Việt với Naive Bayes

- GVHD : Lê Thanh Hương
- Sinh viên thực hiện : Nguyễn Trường Giang - 20173083  
Nguyễn Minh Tuấn – 20173436

*Hà Nội, tháng 01 năm 2022*

## Mục lục

<b>1. MÔ TẢ BÀI TOÁN</b>	<b>3</b>
1.1. Giới thiệu	3
1.2. Mô tả bài toán “Lọc thư rác”	3
<b>2. THUẬT TOÁN PHÂN LOẠI NAÏVE BAYES</b>	<b>4</b>
2.1. Cơ sở lý thuyết:	4
2.1.1 Định lý Bayes	4
2.2. Mô tả thuật toán phân loại Naïve Bayes	4
2.3. Áp dụng thuật toán Naïve Bayes trong phân loại thư rác	5
2.4. Một số tiêu chí đánh giá hiệu năng hệ thống	6
<b>3. CÀI ĐẶT HỆ THỐNG</b>	<b>7</b>
3.1. Tập dữ liệu sử dụng	7
3.2. Tiền xử lý dữ liệu	8
• Loại bỏ ký tự lạ, chuẩn hoá văn bản	8
• Tách biệt các từ vô nghĩa	8
• Loại bỏ toàn bộ các ký tự đứng 1 mình	8
• Loại bỏ hết các số trong văn bản	8
• Ghép các từ tiếng Việt	8
3.3. Tạo Bag of words	8
3.4. Chuyển các văn bản sang các vector	8
3.5. Tính các xác suất cần thiết của Naïve Bayes	8
3.6. Phân loại thư	10
3.7. Kết quả	11
<b>4. KHÓ KHĂN GẶP PHẢI</b>	<b>12</b>
<b>5. HƯỚNG PHÁT TRIỂN TƯƠNG LAI</b>	<b>3</b>
<b>6. TÀI LIỆU THAM KHẢO</b>	<b>4</b>
<b>7. SOURCE CODE</b>	<b>4</b>

## 1. MÔ TẢ BÀI TOÁN

### 1.1. Giới thiệu

Mạng Internet ra đời đã mang lại cho con người những tiện ích hết sức to lớn và quan trọng, một trong những tiện ích đó là dịch vụ thư điện tử. Thư điện tử là phương tiện giao tiếp đơn giản, tiện lợi, rẻ và hiệu quả giữa mọi người trong cộng đồng sử dụng dịch vụ Internet. Tuy nhiên chính vì những lợi ích của dịch vụ thư điện tử mang lại mà số lượng thư trao đổi trên Internet ngày càng tăng và đa số trong số những thư đó là thư rác (spam).

Thư rác (spam mail) là những bức thư điện tử không yêu cầu, không mong muốn và được gửi hàng loạt tới người nhận. Một bức thư nếu gửi không theo yêu cầu có thể đó là thư làm quen hoặc thư được gửi lần đầu tiên, còn nếu thư được gửi hàng loạt thì nó có thể là thư gửi cho khách hàng của các công ty, các nhà cung cấp dịch vụ. Vì thế, một bức thư được coi là thư rác khi nó không được yêu cầu và được gửi hàng loạt. Tuy nhiên, yếu tố quan trọng nhất để phân biệt thư rác với thư thông thường là nội dung thư. Khi một người nhận được thư rác, người đó không thể xác định được thư đó được gửi hàng loạt hay không nhưng có thể xác định được đó là thư rác sau khi đọc nội dung thư. Đặc điểm này chính là cơ sở cho giải pháp phân loại thư rác bằng cách phân tích nội dung thư.

Thư rác thường được gửi với số lượng rất lớn, không được người dùng mong đợi, thường với mục đích quảng cáo, đính kèm virus, gây phiền toái khó chịu cho người dùng, làm giảm tốc độ truyền internet và tốc độ xử lý của email server, gây thiệt hại rất lớn về kinh tế. Ngoài ra, còn có một số loại thư rác được gửi tới một người nhận xác định nào đó nhằm mục đích phá vỡ và gây cản trở công việc của người nhận hay mạng của nhà cung cấp dịch vụ thư điện tử (ESP) được gọi là “bom thư”. Thư rác còn được cố ý gửi đi nhằm thông báo tin sai lệch, làm xáo trộn công việc và cuộc sống của người nhận.

Vì vậy, ngày nay sự phân loại thư rác là rất quan trọng và cần thiết để mạng lại sự tiện ích và an toàn với người dùng.

### 1.2. Mô tả bài toán “Lọc thư rác”

Bài toán “Lọc thư rác” là bài toán xác định (phân loại) những thư điện tử được gửi đến là thư rác (spam mail) hay thư hợp lệ (ham mail) dựa trên nội dung thư, kết quả phân loại cần đạt độ chính xác cao, đặc biệt giảm thiểu lỗi phân loại sai Ham mail thành Spam mail.

Lọc thư rác theo nội dung là trường hợp riêng của bài toán Phân loại văn bản. Tùy theo nội dung, thư được phân thành hai loại: thư rác và thư hợp lệ. Việc phân loại được tiến hành như sau: Trước tiên, nội dung thư được biểu diễn dưới dạng các đặc trưng hay

các thuộc tính, mỗi đặc trưng thường là một từ hoặc cụm từ xuất hiện trong thư. Tiếp theo, trong giai đoạn huấn luyện, tập thư đã được gán nhãn {rác, bình thường} – gọi là tập dữ liệu huấn luyện, được sử dụng để huấn luyện một bộ phân phân loại. Sau khi huấn luyện xong, bộ phân loại được sử dụng để xác định thư mới (thư cần phân loại) thuộc vào loại nào trong hai loại nói trên. Trong cả giai đoạn huấn luyện và phân loại, thuật toán phân loại chỉ làm việc với nội dung thư đã được biểu diễn dưới dạng đặc trưng.

Có nhiều phương pháp phân loại có thể sử dụng để phân loại thư điện tử, trong đó thông dụng nhất là phân loại dựa trên thuật toán Naïve Bayes và Support Vector Machines (SVM). Trong phạm vi bài tập lớn này, nhóm chúng em lựa chọn Lọc thư rác dựa trên thuật toán phân loại Naïve Bayes.

## 2. THUẬT TOÁN PHÂN LOẠI NAÏVE BAYES

### 2.1. Cơ sở lý thuyết:

#### 2.1.1 Định lý Bayes

Định lý Bayes dựa trên định nghĩa về xác suất có điều kiện – xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Kí hiệu  $P(A|B)$ .

Định lý Bayes có thể phát biểu dưới dạng công thức như sau:

$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

Trong đó:

- $P(h)$ : Xác suất trước (tiên nghiệm) của giả thiết (phân loại)  $h$
- $P(D)$ : Xác suất trước (tiên nghiệm) của việc quan sát được dữ liệu  $D$
- $P(D|h)$ : Xác suất (có điều kiện) của việc quan sát được dữ liệu  $D$ , nếu biết giả thiết (phân loại)  $h$  là đúng
- $P(h|D)$ : Xác suất (có điều kiện) của giả thiết (phân loại)  $h$  là đúng, nếu quan sát được dữ liệu  $D$

### 2.2. Mô tả thuật toán phân loại Naïve Bayes

- Biểu diễn bài toán phân loại (classification problem)
  - Một tập học  $D_{train}$ , trong đó mỗi ví dụ học  $x$  được biểu diễn là một vectơ  $n$  chiều:  $(x_1, x_2, \dots, x_n)$
  - Một tập xác định các nhãn lớp:  $C = \{c_1, c_2, \dots, c_m\}$
  - Với một ví dụ (mới)  $z$ , thì  $z$  sẽ được phân vào lớp nào?
- Mục tiêu: Xác định phân lớp có thể (phù hợp) nhất đối với  $z$

$$c_{MAP} = \arg \max P(c_i | z) \quad (\text{với } c_i \in C)$$

$$c_{MAP} = \arg \max P(c_i | z_1, z_2, \dots, z_n) \quad (\text{với } c_i \in C)$$

$$c_{MAP} = \arg \max \frac{P(z_1, z_2, \dots, z_n | c_i) \cdot P(c_i)}{P(z_1, z_2, \dots, z_n)} \quad (\text{bởi định lý Bayes})$$

- Để tìm được phân lớp có thể nhất đối với  $z \dots$

$$c_{MAP} = \arg \max P(z_1, z_2, \dots, z_n | c_i) \cdot P(c_i) \quad (\text{với } c_i \in C)$$

(vì  $P(z_1, z_2, \dots, z_n)$  là như nhau với các lớp)

- Giả sử trong phương pháp phân loại Naïve Bayes, các thuộc tính là độc lập có điều kiện (conditionally independent) đối với các lớp

$$P(z_1, z_2, \dots, z_n | c_i) \cdot P(c_i) = \prod_{j=1}^n P(z_j | c_i)$$

- Phân loại Naïve Bayes tìm phân lớp có thể nhất đối với  $z$

$$c_{NB} = \arg \max P(c_i) \cdot \prod_{j=1}^n P(z_j | c_i)$$

#### ❖ Giải thuật:

- Giai đoạn học (training phase), sử dụng một tập học: Đối với mỗi phân lớp có thể (mỗi nhãn lớp)  $c_i \in C$ 
  - Tính giá trị xác suất trước:  $P(c_i)$
  - Đối với mỗi giá trị thuộc tính  $x_j$ , tính giá trị xác suất xảy ra của giá trị thuộc tính đó đối với một phân lớp  $c_i$ :  $P(x_j | c_i)$
- Giai đoạn phân lớp (classification phase), đối với một ví dụ mới
  - Đối với mỗi phân lớp  $c_i \in C$ , tính giá trị của biểu thức:

$$P(c_i) \cdot \prod_{j=1}^n P(x_j | c_i)$$

- Xác định phân lớp của  $z$  là lớp có thể nhất  $c^*$

$$c^* = \arg \max P(c_i) \cdot \prod_{j=1}^n P(x_j | c_i) \quad (\text{với } c_i \in C)$$

### 2.3. Áp dụng thuật toán Naïve Bayes trong phân loại thư rác

- **Biểu diễn bài toán “Lọc thư rác”:**
  - Tập học **D\_train**, trong đó mỗi ví dụ học là một biểu diễn mail gắn với một nhãn lớp:  $\mathbf{D} = \{(d_k, c_i)\}$
  - Một tập các nhãn lớp xác định:  $\mathbf{C} = \{c_i\} = \{\text{spam}, \text{ham}\}$
- **Giai đoạn học**

- Từ tập các mail trong **D\_train**, trích ra tập các từ khóa (keywords/terms):  $T = \{t_j\}$  (bag of words)
- Gọi  $D_{c_i} (\subseteq D_{train})$  là tập các mail trong **D\_train** có nhãn lớp  $c_i$
- Đối với mỗi phân lớp  $c_i$ 
  - Tính giá trị xác suất trước của phân lớp  $c_i$ :  $P(c_i)$
  - Đối với mỗi từ khóa  $t_j$ , tính xác suất từ khóa  $t_j$  xuất hiện đối với phân lớp  $c_i$

$$P(t_j | c_i) = \frac{(\sum_{d_k \in D_{c_i}} n(d_k, t_j)) + 1}{(\sum_{d_k \in D_{c_i}} \sum_{t_m \in T} n(d_k, t_m)) + |T|}$$

(Trong đó:  $n(d_k, t_j)$  là số lần xuất hiện của từ khóa  $t_j$  trong mail  $d_k$ )

- **Giai đoạn phân lớp đối với một mail mới  $d$**

- Từ mail  $d$ , trích ra tập  $T_d$  gồm các từ khóa (keywords)  $t_j$  đã được định nghĩa trong tập  $T$  ( $T_d \subseteq T$ )
- Giả sử, xác suất từ khóa  $t_j$  xuất hiện đối với lớp  $c_i$  là độc lập đối với vị trí của từ khóa đó trong mail, nghĩa là:

$$P(t_j \text{ ở vị trí } k | c_i) = P(t_j \text{ ở vị trí } m | c_i) \quad \forall k, m$$

- Đối với mỗi phân lớp  $c_i$ , tính giá trị likelihood (khả năng có thể) của mail  $d$  đối với lớp  $c_i$

$$P(c_i) \cdot \prod_{t_j \in T_d} P(t_j | c_i)$$

- Phân lớp mail  $d$  thuộc vào lớp  $c^*$

$$c^* = \arg \max_{c_i \in C} P(c_i) \cdot \prod_{t_j \in T_d} P(t_j | c_i)$$

## 2.4. Một số tiêu chí đánh giá hiệu năng hệ thống

- Ma trận nhầm lẫn (Confusion matrix) tính toán các tham số Precision và Recall

Lớp $c_i$		Được phân lớp bởi hệ thống	
		Thuộc	Không thuộc
Phân lớp (thực sự đúng)	Thuộc	$TP_i$	$FN_i$
	Không thuộc	$FP_i$	$TN_i$

Trong đó:

- $TP_i$ : Số lượng các ví dụ thuộc lớp  $c_i$  được phân loại chính xác vào lớp  $c_i$

- $FN_i$ : Số lượng các ví dụ không thuộc lớp  $c_i$  bị phân loại nhầm vào lớp  $c_i$
- $FP_i$ : Số lượng các ví dụ không thuộc lớp  $c_i$  được phân loại (chính xác)
- $TN_i$ : Số lượng các ví dụ thuộc lớp  $c_i$  - bị phân loại nhầm (vào các lớp khác  $c_i$ )
- Precision đối với lớp  $c_i$ : tổng số các ví dụ thuộc lớp  $c_i$  được phân loại chính xác chia cho tổng số các ví dụ được phân loại vào lớp  $c_i$ :

$$\frac{TP_i}{TP_i + FP_i}$$

- Recall đối với lớp  $c_i$ : Tổng số các ví dụ thuộc lớp  $c_i$  được phân loại chính xác chia cho tổng số các ví dụ thuộc lớp  $c_i$ :

$$\frac{TP_i}{TP_i + FN_i}$$

- $F_1$  score: là một trung bình điều hòa của các tiêu chí Precision và Recall

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

### 3. CÀI ĐẶT HỆ THỐNG

#### 3.1. Tập dữ liệu sử dụng

Tập dữ liệu nhóm sử dụng được thu thập từ gmail cá nhân kết hợp thu thập thêm trên internet, sau đó tổng hợp ra một file .xlsx gồm 2 cột (document, label):

	A	B
1	Document	Label
2	<p>🔗 Link thông tin chi tiết và ứng tuyển: <a href="https://forms.gle/M6G2vV3y9AAMttCR7">https://forms.gle/M6G2vV3y9AAMttCR7</a></p> <p>🕒 Thời gian làm việc:</p> <ul style="list-style-type: none"> <li>• Full-time: 08h/ngày, chia 2 ca: 08h00-16h00 / 14h00-22h00 / 06 ngày/tuần.</li> <li>• Parttime: 04h/ca, chia 3 ca: 08h00-12h00/ 13h00 - 17h00/18h00 - 22h00; 06 ngày/tuần</li> </ul>	1
3	<p>👉 Đừng lo 🤖, Careerly đã tổng hợp những tin tuyển dụng cho vị trí Product hoặc ở công ty Product, công nghệ ở các bậc intern, fresher, junior (yêu cầu khoảng 1 năm kinh nghiệm trở xuống) mà Careerly biết được cho độc giả trên Blog của Careerly tại: 🤖  <a href="https://blog.careerly.vn/job-post/">https://blog.careerly.vn/job-post/</a> 🤖🕒</p> <p>Hi vọng với sự xuất hiện của database này, Careerly sẽ không chỉ giúp bạn trang bị kiến thức để apply vào các vị trí Product mà còn</p>	1

- Cột Document: lưu trữ nội dung mail

- Cột Label: lưu giá trị của nhãn dưới dạng số 0 và 1 (0-ham, 1-spam)

### 3.2. Tiền xử lý dữ liệu

- Import module re và thư viện pandas (xử lý dữ liệu), ...
- Đọc dữ liệu từ file
- Loại bỏ tất cả các đường dẫn trong văn bản
- Loại bỏ ký tự lạ, chuẩn hoá văn bản
- Tách biệt các từ vô nghĩa
- Loại bỏ toàn bộ các ký tự đứng 1 mình
- Loại bỏ hết các số trong văn bản
- Ghép các từ tiếng Việt

### 3.3. Tạo Bag of words

Bag of words là 1 túi từ chứa tất cả các từ xuất hiện trong các văn bản spam và ham (ở tập dataset). Tạo Bag of words bằng cách tách các văn bản ra thành các từ, rồi lần lượt "cho hết vào túi", sau đó loại tất cả những từ trùng nhau đi để đảm bảo mỗi từ không xuất hiện quá 1 lần.

Sau đó, tất cả các văn bản từ giờ sẽ được biểu diễn dưới dạng 1 vector có n thuộc tính, với n là số từ có trong Bag of words của văn bản đó

### 3.4. Chuyển các văn bản sang các vector

Với mỗi từ trong Bag of words sẽ là 1 thuộc tính. Như vậy ta sẽ biểu diễn các văn bản dưới dạng 1 vector, nếu 1 từ trong Bag of words có xuất hiện trong văn bản đó, vị trí của nó sẽ là 1, còn lại là 0. Với mỗi một văn bản trong tập dữ liệu, lúc đầu, ta tạo ra 1 vector toàn 0 với số thuộc tính là chiều dài của bag of words. Sau đó, lần lượt kiểm tra xem từng từ của bag of words có nằm trong văn bản đó không, nếu có sẽ gán cho thuộc tính đó bằng 1.

### 3.5. Tính các xác suất cần thiết của Naïve Bayes

Ta sẽ tạo 1 hàm để làm trơn xác suất nhằm tránh việc tích 1 xác suất đang cao, bỗng nhiên gặp 1 xác suất điều kiện bằng 0 kéo theo kết quả cuối bằng 0 với công thức :

$$P(x_i|c_j) = (n_c + 1)/(n + 1)$$

```
def smoothing(a, b):  
    return float((a+1)/(b+1))
```

- P(spam) = số lượng mail spam / tổng số lượng mail



-  $P(\text{non-spam}) = \text{số lượng mail non-spam} / \text{tổng số lượng mail}$

```
spam = 0
non_spam = 0
for l in label:
    if l == 1:
        spam += 1
    else:
        non_spam += 1

spam_coef = services.smoothing(spam, (spam+non_spam))
non_spam_coef = services.smoothing(non_spam, (spam+non_spam))
```

- Lưu lại các giá trị xác suất thành phần của từng từ, và từ lần tiếp theo, việc dự đoán sẽ được tính toán trên nó:

```
bayes_matrix = np.zeros((len(set_words), 4))
```

- Sau đó thực hiện thống kê, đếm việc xuất hiện/ không xuất hiện đồng thời của từng từ khi văn bản là spam hoặc không là spam rồi cập nhật vào ma trận

```
for i, word in enumerate(set_words):

    app_spam = 0
    app_nospam = 0
    nonapp_spam = 0
    nonapp_nospam = 0
    for k, v in enumerate(vectors):
        if v[i] == 1:
            if label[k] == 1:
                app_spam += 1
            else:
                app_nospam += 1
        else:
            if label[k] == 1:
                nonapp_spam += 1
            else:
                nonapp_nospam += 1

    bayes_matrix[i][0] = services.smoothing(app_spam, spam)
    bayes_matrix[i][1] = services.smoothing(app_nospam, non_spam)
    bayes_matrix[i][2] = services.smoothing(nonapp_spam, spam)
    bayes_matrix[i][3] = services.smoothing(nonapp_nospam, non_spam)
```

### 3.6. Phân loại thư

- Với một văn bản mới, trước khi đưa vào dự đoán, chúng ta cũng cần tiền xử lý và biến nó thành vector!

- Thực hiện tính toán với ma trận Bayes để ra các kết quả nhằm xác định nhãn cho văn bản. Dựa vào công thức:

$$y = \operatorname{argmax} P(c_j) \prod P(x_i|c_j)$$

- Ta sẽ nhân tất cả các xác suất điều kiện lại với nhau rồi chọn kết quả lớn hơn => Chính là nhãn dự đoán.

- Đa số các xác suất đều rất nhỏ, nếu cứ nhân liên tiếp với nhau, chắc chắn chúng sẽ bị về 0. Vậy nên mỗi khi nhận thấy xác suất quá nhỏ, chúng ta sẽ nhân xác suất lên để đảm bảo không bị về 0. Tuy nhiên để đảm bảo cuộc "CẠNH TRANH CÔNG BẰNG" giữa 2 xác suất, mỗi lần nhân, chúng ta đều sẽ ghi lại để cuối cùng so sánh xem giá trị nào lớn hơn! Việc so sánh thực hiện một cách khá đơn giản như sau

- Từ các bước trên ta có hàm predict:

```
def predict(mail, set_words, spam_coef, non_spam_coef, bayes_matrix):
    mail = raw_text_preprocess(mail)

    vector = np.zeros(len(set_words))
    for i, word in enumerate(set_words):
        if word in mail:
            vector[i] = 1
    log = np.zeros(2)

    predict_spam = spam_coef
    predict_non_spam = non_spam_coef

    for i, v in enumerate(vector):
        if v == 0:
            predict_spam *= bayes_matrix[i][2]
            predict_non_spam *= bayes_matrix[i][3]
        else:
            predict_spam *= bayes_matrix[i][0]
            predict_non_spam *= bayes_matrix[i][1]

    if predict_spam < 1e-10:
        predict_spam *= 1000
        log[0] += 1

    if predict_non_spam < 1e-10:
```

```

predict_non_spam *= 1000
log[1] += 1

if compare(predict_spam, predict_non_spam, log):
    return 1
return 0

```

### 3.7. Kết quả

- Với tập test 100 email, sau khi gọi hàm predict với mỗi email ta đạt được kết quả phân loại chính xác là 73%

Train

Test

1: Spam  
0: Non-spam

Show 10 entries

Search:

Dữ liệu ban đầu	Dữ liệu sau tiền xử lý	Nhãn
Charles&keith Vincom 191 Bà Triệu ❌❌❌Charles&keith 36 Lê Thái Tổ CẦN TÌM 05 Nam/nữ Nhân Viên Bán Hàng, yêu cầu ngoại hình khá + giao tiếp tốt, đáp ứng được thời gian làm việc theo ca, luân ca, tăng ca. Ưu tiên ứng viên có kinh nghiệm làm việc mảng thời trang cao cấp, không phù hợp với những bạn còn đang đi học nhé. 🌿🌿 Lương + thưởng doanh số + đãi ngộ hấp dẫn. 🙌🙌🙌 Liên hệ Zalo/điện thoại chị Vân để biết thêm chi tiết: 0936052055	charles keith vincom bà triệu charles keith lê thái tổ cần tìm nam_nữ nhân_viên bán hàng yêu_cầu ngoại_hình khá giao_tiếp tốt đáp_ứng được thời_gian làm_việc theo ca luân ca tăng ca ưu_tiên ứng_viên có kinh_nghiệm làm_việc mảng thời_trang cao_cấp không phù_hợp với những bạn còn đang đi học nhé lương thưởng doanh_số đãi_ngộ hấp_dẫn liên_hệ zalo điện_thoại chị vân để biết thêm chi tiết	1
Chiều mộ 2 #LEADERMKT #Marketing và 2 LEADERSALE Cơ hội phát triển tách công ty chi nhánh 🍷 CÓ THỂ TƯ DEAL LƯƠNG CỨNG THEO CON SỐ CÁC BẠN MONG MUỐN 🍷 - Lương cứng 7-10tr (Không có mốc thấp hơn) - Hoa hồng cực cao (2-5%) + THUỞNG NÓNG/TUẦN/TOP + pc ăn 500k+ pc quản lý đội nhóm. - Thu nhập k giới hạn 15-50tr tùy năng lực ✅ YẾU CẦU: - KINH NGHIỆM ADS - Có khả năng lãnh đạo, dẫn dắt đội nhóm 6-7ng. *lb nhận JD.	chiều mộ leadermkt marketing và leadersale cơ_hội phát_triển tách công_ty chi_nhánh Có THỂ Tư DEAL Lương CỨNG THEO CON SỐ CÁC BẠN MONG MUỐN lu ng cu ng tr không có mốc thấp hơn hoa_hồng cực cao thưởng nóng tuần top pc ăn pc quản_lý đội nhóm thu_nhập giới_hạn tr tùy năng_lực ye ca kinh nghiệ ads có khả_năng lãnh_đạo	1

Dữ liệu tập Train

Train

Test

Test

KẾT QUẢ

73 %

1: Spam

0: Non-spam

Show 10 entries

Search:

Dữ liệu ban đầu	Nhãn	Dự đoán
Hiện tại theo em biết trường đang mở đợt tiêm loại vaccine Spunik V Cho em hỏi đây là mũi 1 hay mũi 2 vậy ạ? Vì em đã tiêm mũi 1 từ 9/10 Cho em hỏi thêm về form điển đăng kí, ở đó không có hỏi thêm thông tin gì về sinh viên trừ mục hỏi "Bạn muốn tiêm loại nào", và em đã điển form mà mãi chưa thấy được phản hồi ạ, vậy em phải đăng kí như nào ạ	0	0
hứa thầy cô cho em hỏi: Điểm rèn luyện của em kỳ 20202, mục học tập, bị khác so với điểm GPA hiện trên Ctt của em (đã cập nhật kỳ 20202), thì em có thể liên hệ ở đâu ạ? Em xin cảm ơn ạ	0	1
Kính chào các thầy cô của phòng đào tạo. Em xin phép hỏi một câu ạ: Kỳ 2019.2 em có đăng ký học môn Kỹ thuật điện EE2012 và đã qua môn nhưng hiện tại điểm của em ở trên bảng điểm học phần đang là X và không hiện điểm quá trình, hiện tại em không biết phải làm gì để có thể được hoàn thành điểm môn học ạ?	0	0
Kính chào các thầy cô, Em có một câu hỏi là học phần giải tích 2 em mới được thi bù nên chưa có điểm trên ctt, do đó kết quả xét GPA cho học bổng KKHT kì 20211 của em đang bị thấp. Em hỏi Phòng CTSV thi được bảo là liên hệ PDT. Các thầy cô có thể cho em cách giải quyết không ạ?	0	0
Kính gửi phòng đào tạo, Em có thắc mắc về việc đăng ký tốt nghiệp chương trình cử nhân và chương trình kỹ sư, kính mong phòng đào tạo hoặc anh chị nào biết có thể trả lời cho em ạ. Em có đăng ký các môn kỹ tối là các môn kỹ sư, nhưng chưa vào đợt đăng ký cho kỳ sau, vậy khi vào đợt đăng ký em có thể	0	1

127.0.0.1:5500/front-end/test.html#

Dữ liệu tập Test và kết quả phân loại

## 4. KHÓ KHĂN GẶP PHẢI

Trước hết, vì đề tài này đã được thực hiện khá nhiều trước đó trong thực tế nên nhóm có khá nhiều thuận lợi khi thực hiện bài tập lớn. Thuật toán phân loại Naive-Bayes dựa trên xác suất khá rõ ràng, ứng dụng kiến thức xác suất đã được làm quen trong học phần xác suất thống kê. Các nguồn tài liệu trên mạng khá nhiều và không khó tìm. Bên cạnh đó những thư viện về thuật toán Naïve Bayes và tiền xử lý dữ liệu cũng dễ tiếp cận giúp nhóm tiền hành bài tập lớn được tốt hơn.

Tuy nhiên, trong quá trình thực hiện bài tập lớn nhóm cũng gặp những khó khăn nhất định.

Đầu tiên, nhóm chưa có cơ hội tiếp cận với ngôn ngữ Python nên quá trình học cú pháp và các thư viện gặp khá nhiều khó khăn và tốn thời gian. Nhóm chỉ có 2 thành viên và cả 2 đều chưa tiếp xúc nhiều với Học máy nên chương trình xây dựng được còn đơn giản. Tuy nhiên sau khi làm xong nhóm cũng đã phần nào giải đáp đc câu hỏi “Liệu máy tính học như thế nào?”

Ngoài ra, trên mạng có nhiều tập dữ liệu cho bài toán lọc thư rác bằng Tiếng Anh, tuy nhiên nhóm chúng em lại không tìm thấy bộ dữ liệu nào cho Tiếng Việt cả, vậy nên chỉ

còn các thu thập từ gmail của mình và các nguồn khác trên mạng rồi tổng hợp lại, việc này khá mất thời gian và số lượng email thu thập được cũng không nhiều

## 5. HƯỚNG PHÁT TRIỂN TƯƠNG LAI

Như đã trình bày ở phần lý thuyết thuật toán, Naïve Bayes có giả sử là xác suất từ khóa  $t$  xuất hiện đối với lớp  $c$  là độc lập với vị trí xuất hiện của từ khóa đó trong văn bản. Nhưng rõ ràng trong thực tế thì mức độ ảnh hưởng của một từ đến xác suất email đó có phải là spam hay không là khác nhau nếu so vị trí của từ đó ở tiêu đề với ở trong nội dung mail. Vì vậy ý tưởng phát triển của nhóm là sử dụng kết hợp Bag of words với phương pháp TF-IDF. Với Bags of words, đối với test data mới, ta tiến hành tìm ra số lần từng từ của test data xuất hiện trong "bag", tuy nhiên nó vẫn tồn tại khuyết điểm, nên TF-IDF là phương pháp khắc phục. Cụ thể phương pháp TF-IDF sẽ tính giá trị TF-IDF của một từ là một con số thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản. Mục đích của việc tính IDF là giảm giá trị của các từ thường xuyên xuất hiện như "is", "the"... Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khóa của văn bản đó). Do các từ này không mang nhiều ý nghĩa trong việc phân loại văn bản. Đặc biệt đối với lượng data lớn thì phương pháp này chắc chắn sẽ rất hiệu quả.

Bên cạnh đó nhóm đề xuất sử dụng kết hợp giải thuật SVM (Support Vector Machines), dùng thuật toán này để tối ưu, tăng độ chính xác cho kết quả phân loại sau khi sử dụng thuật toán Naive Bayes. Ý tưởng chính của thuật toán này là 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 mặt phẳng  $h$  quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng lớp 1 và lớp 0. Chất lượng của siêu mặt phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt đồng thời việc phân loại càng chính xác. Mục đích thuật toán SVM tìm ra được khoảng cách biên lớn nhất để tạo kết quả phân lớp tốt.

Ngoài ra, trong tương lai, nhóm mong muốn sẽ có thể phát triển để xây dựng ứng dụng tích hợp vào hệ thống của server cho phép lọc thư rác.

## 6. TÀI LIỆU THAM KHẢO

- Lý thuyết về mạng Bayes và ứng dụng vào bài toán lọc thư rác  
<https://viblo.asia/p/ly-thuyet-ve-mang-bayes-va-ung-dung-vao-bai-toan-loc-thu-rac-07LKXzkeIV4>
- Pandas documentation  
<https://pandas.pydata.org/docs/>
- NLTK – python tutorial  
<https://pythonspot.com/category/nltk/>
- API Reference – scikit-learn 0.23.1 documentation  
<https://scikit-learn.org/stable/modules/classes.html>
- Python tutorial – Matplotlib 3.2.1 documentation  
<https://matplotlib.org/tutorials/introductory/pyplot.html#sphx-glr-tutorials-introductory-pyplot-py>

## 7. SOURCE CODE

<https://github.com/ntgiang3733/NLP-Phan-loai-email-spam>