

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA: CÔNG NGHỆ THÔNG TIN**

-----🙞🕮🙜-----

A yellow and red logo

Description automatically generated

**BÀI TẬP LỚN**

**MÔN: Trí Tuệ Nhân Tạo**

**ĐỀ TÀI:  
Xây dựng Hệ thống phân loại thư rác   
dựa trên giải thuật Naive Bayes**

**Giảng viên hướng dẫn: Trần Thanh Huân**

**Lớp học phần: IT6046**

**Nhóm: 11**

**Sinh viên thực hiện: Nguyễn Đức Trường 2023605862**

**Dương Kim Vũ 2023605127**

**Đào Xuân Thắng 2023605466**

**Nguyễn Cường Thịnh 2023606918**

***Hà Nội, 05/2025***

# LỜI CẢM ƠN

Lời đầu tiên, chúng em xin gửi lời cảm ơn trân thành đến thầy **Trần Thanh Huân**. Trong quá trình học tập và thực hiện đề tài, chúng em đã nhận được sự quan tâm, hướng dẫn tận tình từ thầy. Những gì chúng em nhận được không chỉ dừng lại ở kiến thức môn học mà nhiều hơn thế đó là những lời khuyên, chia sẻ thực tế từ thầy.

Để hoàn thành được đề tài này, chúng em đã nghiên cứu, áp dụng những kiến thức được học trên lớp cùng với các nguồn tài liệu trên Internet và cả những trải nghiệm của bản thân đối với đề tài lần này. Chúng em rất mong sẽ nhận được những lời nhận xét, góp ý từ thầy, cô và các bạn đọc để đề tài của chúng em để có thể hoàn thiện hơn nữa.

Vì thời gian còn hạn chế chúng em chưa có cơ hội cụ thể hóa hết các ý tưởng của mình với đề tài này, em rất mong sẽ được quý thầy cô và các bạn đón nhận cũng như đưa ra những lời góp ý để đề tài của chúng em hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

[LỜI CẢM ƠN 1](#_Toc200571584)

[MỞ ĐẦU 3](#_Toc200571585)

[**1.** **Lí do chọn đề tài:** 4](#_Toc200571586)

[**2.** **Mục đích nghiên cứu:** 5](#_Toc200571587)

[**3.** **Phương pháp nghiên cứu:** 5](#_Toc200571588)

[**4.** **Bố cục của bài báo cáo:** 5](#_Toc200571589)

[Chương I. TỔNG QUAN VỀ KĨ THUẬT PHÂN LOẠI THƯ RÁC 7](#_Toc200571590)

[**1.1** **. Khái niệm thư rác** 7](#_Toc200571591)

[1.2. Tác hại của thư rác 7](#_Toc200571592)

[1.3. Tổng quan về một số kĩ thuật phân loại thư rác: 8](#_Toc200571593)

[1.4. Những thách thức trong việc phân loại thư rác: 9](#_Toc200571594)

[**1.5. Tổng kết chương 1:** 10](#_Toc200571595)

[Chương II : PHÂN TÍCH THUẬT TOÁN NAIVE BAYES TRONG PHÂN LOẠI THƯ RÁC 11](#_Toc200571596)

[2.1. Định lí Bayes: 11](#_Toc200571597)

[2.2. Mô tả thuật toán Naive Bayes: 11](#_Toc200571598)

[2.3.Các bước triển khai thuật toán Naïve Bayes: 12](#_Toc200571599)

[2.3.1. Thu thập và tiền xử lý dữ liệu 12](#_Toc200571600)

[2.3.2. Xây dựng từ điển và tính xác suất 13](#_Toc200571601)

[2.3.3 Tính xác suất điều kiện 14](#_Toc200571602)

[2.3.4 Áp dụng định lý Bayes 15](#_Toc200571603)

[2.3.5 Quyết định phân loại 17](#_Toc200571604)

[2.3.6. Đánh giá và tinh chỉnh mô hình 17](#_Toc200571605)

[2.5. Ưu điểm và hạn chế của thuật toán Naive Bayes 19](#_Toc200571606)

[2.5.1. Ưu điểm 19](#_Toc200571607)

[2.5.2. Hạn chế 19](#_Toc200571608)

[CHƯƠNG II: PHÂN TÍCH THUẬT TOÁN NAIVE BAYES TRONG PHÂN LOẠI THƯ RÁC 20](#_Toc200571609)

[CHƯƠNG 3: THIẾT KẾ , TRIỂN KHAI HỆ THỐNG PHÂN LOẠI THƯ RÁC 3.1. Tải tập dữ liệu: 21](#_Toc200571610)

[3.2. Phân chia dữ liệu huấn luyện và kiểm tra 22](#_Toc200571611)

[3.3. Xử lý tiền dữ liệu 22](#_Toc200571612)

[3.4. Dự đoán hàm 24](#_Toc200571613)

[3.5. Hàm phân loại và áp dụng lên dữ liệu kiểm tra 25](#_Toc200571614)

[3.6. Đánh giá độ chính xác mô hình 27](#_Toc200571615)

[3.7 .TỔNG KẾT CHƯƠNG 3 27](#_Toc200571616)

# 

# MỞ ĐẦU

1. **Lí do chọn đề tài:**

Việc sử dụng Internet ngày càng phổ biến đã mang lại nhiều tiện ích cho người dùng, nhưng đồng thời cũng kéo theo sự gia tăng của các vấn nạn như tin nhắn rác và cuộc gọi rác. Theo thống kê từ Bộ Thông tin và Truyền thông, trong 10 tháng đầu năm 2024, tổng đài phản ánh 156/5656 đã ghi nhận gần 850.000 lượt phản ánh từ người dùng, trong đó có khoảng 185.000 lượt liên quan đến tin nhắn rác, chiếm 22% tổng số phản ánh . Tình trạng này không chỉ gây phiền toái cho người dùng mà còn đặt ra những thách thức lớn về an ninh mạng và quản lý thông tin.

Mới đây, thuật toán máy học (machine learning) đã được áp dụng rộng rãi trong các lĩnh vực khác nhau như xử lý ngôn ngữ tự nhiên, nhận diện hình ảnh, khai thác dữ liệu, phân tích tín hiệu, v.v. Thuật toán máy học là một phương pháp mạnh mẽ trong việc phân loại thư rác vì khả năng tự học từ dữ liệu và cải thiện hiệu suất qua thời gian khi có thêm dữ liệu có sẵn. Những mô hình máy học này có thể phân loại email dựa trên các đặc trưng của email và huấn luyện mô hình trên các tập dữ liệu có sẵn.

Trong số các thuật toán máy học được sử dụng để phân loại thư rác, thuật toán naive bayes (NB) là một trong những thuật toán đơn giản nhất và hiệu quả nhất. Naive Bayes là một thuật toán học máy dựa trên định lý Bayes, sử dụng xác suất có điều kiện để dự đoán khả năng một tin nhắn thuộc vào một lớp cụ thể (ví dụ: spam hoặc ham). Với giả định rằng các đặc trưng (từ ngữ) trong một tin nhắn là độc lập với nhau, thuật toán này cho phép xử lý, phân loại văn bản một cách nhanh chóng.  
 Vì những lý do trên, tôi đã chọn đề tài “Xây dựng hệ thống phân loại thư rác bằng thuật toán Naïve Bayes” để nghiên cứu và thực hiện. Mục tiêu của đề tài là xây dựng một hệ thống phân loại thư rác sử dụng thuật toán NB và đánh giá hiệu quả của hệ thống so với các phương pháp khác. Đề tài này có ý nghĩa thực tiễn, giúp người dùng xử lí dễ dàng các thư rác. Đồng thời, đề tài này cũng có ý nghĩa khoa học, giúp tôi áp dụng kiến thức về máy học vào bài toán thực tế, tìm hiểu về thuật toán NB và vai trò của nó trong phân loại thư rác.

1. **Mục đích nghiên cứu:**

* Nắm vững cơ sở lý thuyết về thuật toán Naive Bayes, bao gồm nguyên lý hoạt động, giả định độc lập và các biến thể phổ biến.
* Phân tích khả năng ứng dụng của Naive Bayes trong lĩnh vực xử lý ngôn ngữ tự nhiên, đặc biệt là trong việc phân loại văn bản.
* Xây dựng một hệ thống có khả năng lọc và phát hiện tin nhắn rác (spam) dựa trên thuật toán Naive Bayes với độ chính xác cao.
* Đánh giá hiệu quả của hệ thống thông qua các chỉ số đo lường như độ chính xác (accuracy), độ nhạy (recall) và độ đặc hiệu (precision).
* So sánh hiệu quả thuật toán Naive Bayes với một số phương pháp phân loại khác (nếu có thời gian mở rộng).

1. **Phương pháp nghiên cứu:**

Phương pháp luận được sử dụng trong đề tài này bao gồm phương pháp nghiên cứu và tổng hợp lý thuyết và phương pháp thực nghiệm. Mục lý thuyết được nghiên cứu và thu thập từ trên mạng,các cuốn sách, bài báo tạp chí chuyên ngành. Sau khi có được nền tảng lý thuyết và mô hình thực nghiệm sẽ tiến hành. Mô hình thực nghiệm được xây dựng, hỗ trợ bởi một số nền tảng lý thuyết. Các kết quả thực nghiệm sẽ được đối chiếu với kết quả lý thuyết để chứng minh sự thành công của thực nghiệm.

1. **Bố cục của bài báo cáo:**

Đề tài thực hiện gồm 3 chương cơ bản như sau:

* Chương 1: Tổng quan về kĩ thuật phân loại thư rác
* Chương 2: Phân Tích thuật toán Naïve Bayes trong phân loại thư rác
* Chương 3:Thiết kế,triển khai hệ thống phân loại thư rác

# Chương I.TỔNG QUAN VỀ KĨ THUẬT PHÂN LOẠI THƯ RÁC

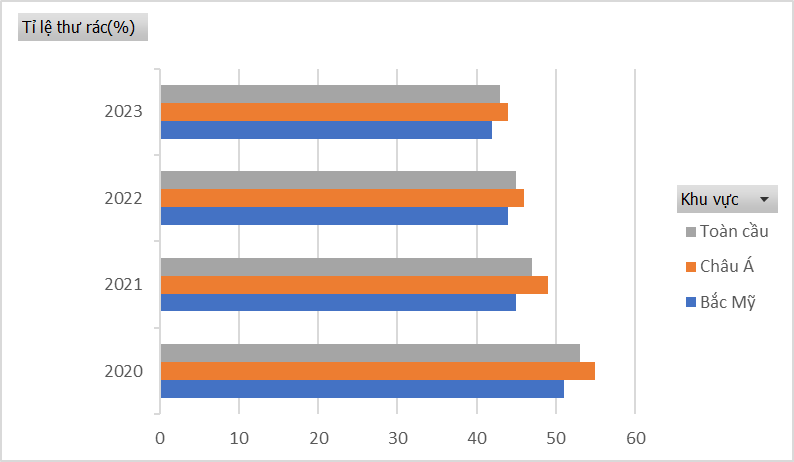
* 1. **. Khái niệm thư rác**

Thư rác (spam mail) là các thông điệp không mong muốn và không được yêu cầu gửi đến người nhận. Thư rác thường chứa các thông tin quảng cáo, thông tin không chính xác, lừa đảo hoặc liên kết độc hại. Nội dung của thư là căn cứ chính quyết định xem bức thư đó có phải là thư rác hay không. Đây cũng là cơ sở cho giải pháp phân loại thư rác bằng nội dung thư.

## 1.2. Tác hại của thư rác

Các tác hại mà thư rác gây ra đối với người dùng:

* Phiền toái và quản lý email:
* Thư rác làm cho hòm thư điện tử trở nên quá tải và khó quản lý.
* Tìm kiếm email quan trọng trở nên khó khăn.
* Nguy cơ an ninh thông tin:
* Thư rác có thể chứa các liên kết độc hại hoặc phần mềm độc hại.
* Virus, spyware, ransomware và các mối đe dọa khác có thể được lây lanqua email thư rác.
* Lừa đảo và đánh cắp thông tin:
* Thư rác thường mang tính lừa đảo, cố gắng lừa người nhận để tiết lộ thông tin cá nhân hoặc tài chính.
* Người gửi thư rác có thể cố gắng lừa người nhận thực hiện các giao dịch gian lận.



Hình 1.1. Hình ảnh biểu hiện tỉ lệ thư rác của các khu vực

## 1.3. Tổng quan về một số kĩ thuật phân loại thư rác:

Phân loại thư rác (spam email classification) là một trong những ứng dụng quan trọng của lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và trí tuệ nhân tạo (AI). Nhiều phương pháp đã được nghiên cứu và áp dụng nhằm nâng cao hiệu quả phát hiện và ngăn chặn thư rác. Nhìn chung, các phương pháp này có thể được chia thành bốn nhóm chính: **dựa trên luật định (rule-based), dựa trên thống kê (statistical methods), dựa trên học máy (machine learning),** và **dựa trên học sâu (deep learning).**

* **Phương pháp dựa trên luật định (Rule-based Filtering)**: là cách tiếp cận truyền thống, trong đó các quy tắc được định nghĩa thủ công để phát hiện thư rác.
* **Phương pháp dựa trên thống kê** : sử dụng các mô hình xác suất để ước lượng khả năng một email là thư rác dựa trên tần suất xuất hiện của các từ hoặc cụm từ trong nội dung email. Trong đó, thuật toán **Naive Bayes** là một trong những kỹ thuật phổ biến nhất nhờ tính đơn giản và hiệu quả cao trong nhiều tình huống.
* **Phương pháp dựa trên học máy (Machine Learning):** sử dụng các thuật toán học từ dữ liệu đã gán nhãn để xây dựng mô hình phân loại. Các thuật toán thường được sử dụng gồm **Support Vector Machine (SVM), Decision Tree, Random Forest,** và **K-Nearest Neighbors (KNN).** Những thuật toán này có khả năng học các đặc trưng phức tạp từ dữ liệu, cho độ chính xác cao trong việc nhận diện thư rác.
* **Phương pháp dựa trên học sâu (Deep Learning):** là xu hướng hiện đại trong phân loại thư rác, sử dụng các mô hình mạng nơ-ron sâu như **Convolutional Neural Network (CNN), Recurrent Neural Network (RNN),** hoặc **Long Short-Term Memory (LSTM).** Các mô hình này có khả năng tự động trích xuất đặc trưng từ dữ liệu văn bản và học các mẫu ngữ nghĩa phức tạp mà các mô hình truyền thống khó nắm bắt được.

## 1.4. Những thách thức trong việc phân loại thư rác:

Việc phân loại tin nhắn rác, đặc biệt là trong thời đại thông tin số và mạng xã hội phát triển mạnh, đang đối mặt với nhiều thách thức. Các khó khăn này không chỉ đến từ **sự đa dạng nội dung** mà còn từ **sự thay đổi liên tục của kỹ thuật gửi spam.** Dưới đây là những thách thức tiêu biểu:

* Ngôn ngữ tự nhiên và phức tạp: Tin nhắn rác thường được viết bằng **ngôn ngữ không chuẩn,** sử dụng **viết tắt, ký tự đặc biệt, dấu câu thay thế,** gây khó khăn cho hệ thống phân tích văn bản.
* Spam thay đổi liên tục: Người gửi thư rác thường xuyên **thay đổi nội dung, cách diễn đạt để né tránh bộ lọc, điều này khiến các mô hình phân loại phải liên tục cập nhật để tránh bị lỗi thời.** Ví dụ, một bộ lọc hoạt động tốt hôm nay có thể trở nên vô dụng chỉ sau vài tuần nếu nội dung spam đã thay đổi hoàn toàn về cấu trúc và ngữ nghĩa.
* Thiếu dữ liệu huấn luyện chất lượng cao: Phân loại tin nhắn rác thường sử dụng các mô hình học máy hoặc học sâu, đòi hỏi một lượng lớn dữ liệu đầu vào được gán nhãn chính xác. Tuy nhiên, việc thu thập và tạo lập tập dữ liệu đủ lớn, cân bằng giữa tin nhắn rác và tin nhắn hợp lệ, lại là một công việc tốn thời gian và công sức. Ngoài ra, dữ liệu còn phải được cập nhật liên tục để phản ánh đúng thực tế.

**1.5. Tổng kết chương 1:**

Trong Chương I, chúng ta đã tìm hiểu tổng quan về khái niệm thư rác, các tác hại mà thư rác gây ra, cũng như các phương pháp kỹ thuật hiện đang được sử dụng để phân loại và xử lý thư rác. Đây là những nền tảng lý thuyết quan trọng, tạo tiền đề cho các chương tiếp theo đi sâu vào kỹ thuật cụ thể, quy trình triển khai mô hình phân loại cũng như đánh giá hiệu quả của hệ thống.

# Chương II : PHÂN TÍCH THUẬT TOÁN

# NAIVE BAYES TRONG PHÂN LOẠI THƯ RÁC

Trong chương này, chúng ta đã tìm hiểu chi tiết về cách áp dụng thuật toán Naive Bayes trong bài toán phân loại thư rác, bao gồm lý thuyết nền tảng, các bước triển khai và những ưu nhược điểm của phương pháp.

Trước hết, định lý Bayes được giới thiệu như là cơ sở lý thuyết chính, cung cấp công thức tính xác suất hậu nghiệm dựa trên xác suất tiên nghiệm và xác suất có điều kiện. Naive Bayes dựa trên giả định độc lập giữa các đặc trưng, cho phép đơn giản hoá bài toán tính toán xác suất và đưa ra quyết định phân loại dựa trên công thức tối đa hoá xác suất hậu nghiệm.

Tiếp theo, quy trình triển khai Naive Bayes được trình bày chi tiết qua các bước:

* **Tiền xử lý dữ liệu**: chuyển văn bản về chữ thường, loại bỏ dấu câu, tách từ, loại bỏ từ dừng, chuẩn hoá từ (stemming hoặc lemmatization).
* **Xây dựng từ điển và thống kê tần suất**: tạo ra tập từ vựng chung, thống kê số lần xuất hiện của từng từ trong mỗi lớp.
* **Tính toán xác suất điều kiện**: áp dụng công thức có làm trơn Laplace nhằm tránh vấn đề xác suất bằng 0.
* **Áp dụng định lý Bayes để phân loại**: tính toán xác suất hậu nghiệm cho mỗi lớp và lựa chọn lớp có xác suất cao nhất.
* **Đánh giá và tinh chỉnh mô hình**: sử dụng các chỉ số như Accuracy, Precision, Recall, F1-score, Confusion Matrix và các phương pháp như cross-validation, cân bằng dữ liệu, cải thiện tiền xử lý để tối ưu hoá mô hình.

Ngoài ra, chương cũng đã phân loại các biến thể của thuật toán Naive Bayes như Gaussian, Multinomial, Bernoulli, Complement và Categorical Naive Bayes, mỗi loại phù hợp với từng loại dữ liệu khác nhau.

Cuối cùng, các ưu điểm như đơn giản, hiệu quả với dữ liệu nhiều chiều, nhanh chóng và dễ mở rộng, cũng như các hạn chế như giả định độc lập không thực tế và vấn đề zero frequency cũng đã được phân tích rõ ràng.

Việc áp dụng Naive Bayes trong phân loại thư rác đã chứng minh hiệu quả cao trong thực tế nhờ tính đơn giản, khả năng xử lý nhanh trên dữ liệu lớn và vẫn đảm bảo được độ chính xác tốt khi xử lý bài toán phân loại văn bản.

## 2.1. Định lí Bayes:

Định lý Bayes là một quy tắc trong xác suất thống kê, cung cấp phương pháp tính toán xác xuất của một sự kiện dựa vào thông tin trước đó:

P(h | D) = P(D | h).P( h ) / P(D) (1)

* P(h): Xác suất trước của giả thiết (phân loại) h • P(D): Xác suất trước 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 h là đúng
* P(h|D): Xác suất (có điều kiện) của giả thiết h là đúng, nếu quan sát được dữ liệu D

## 2.2. Mô tả thuật toán Naive Bayes:

* + Naive Bayes là một phương pháp học máy dựa trên định lý Bayes, thường dùng trong bài toán phân loại như xử lý văn bản (phân loại email spam, phân tích cảm xúc).
  + Naive Bayes được gọi là "naive" (ngây thơ) vì nó giả định rằng các đặc trưng (features) của dữ liệu là độc lập và có ảnh hưởng riêng biệt lên kết quả dự đoán. Mặc dù giả định này thường không chính xác trong thực tế, Naive Bayes vẫn được sử dụng phổ biến do tính đơn giản và khả năng xử lý hiệu quả với dữ liệu.
* Thuật toán Naive Bayes tính toán xác suất của các lớp (classes) dựa trên xác suất của các đặc trưng. Đầu tiên, dữ liệu huấn luyện được sử dụng để tính toán các xác suất có điều kiện. Sau đó, khi có một dữ liệu mới, thuật toán sử dụng định lý Bayes để tính toán xác suất xảy ra của các lớp và chọn lớp có xác suất cao nhất là kết quả dự đoán.
* Trong huấn luyện, Naive Bayes tính toán các xác suất P(X|C) ở công thức (1) bằng cách giả định rằng các đặc trưng (features) là độc lập nhau. Điều này có nghĩa là xác suất của một đặc trưng trong dữ liệu được tính riêng biệt cho từng lớp, và xác suất tổng hợp của dữ liệu X được tính bằng cách nhân các xác suất này lại với nhau.
* Naive Bayes có thể được áp dụng cho nhiều loại dữ liệu, bao gồm cả dữ liệu dạng văn bản. Trong trường hợp này, các đặc trưng của văn bản có thể là các từ và xác suất P(X|C) được tính dựa trên tần suất xuất hiện của từng từ trong các văn bản huấn luyện thuộc cùng một lớp.

## 2.3.Các bước triển khai thuật toán Naïve Bayes:

### 2.3.1. Thu thập và tiền xử lý dữ liệu

Bước đầu tiên là thu thập tập dữ liệu văn bản đã được gán nhãn (ví dụ: “spam” hoặc “ham”). Sau đó, dữ liệu được tiền xử lý bằng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP), bao gồm:

* **Chuyển văn bản về chữ thường**: Giúp đồng nhất từ ngữ, tránh phân biệt giữa “Buy” và “buy”.
* **Loại bỏ dấu câu, ký tự đặc biệt và số không cần thiết**: Giữ lại các từ có ý nghĩa trong phân loại.
* **Tách từ (Tokenization):** Chia văn bản thành các đơn vị từ riêng lẻ.
* **Loại bỏ từ dừng (Stop words):** Loại bỏ những từ phổ biến như “the”, “is”, “at” vì ít mang ý nghĩa phân biệt.
* **Chuẩn hóa từ (Stemming hoặc Lemmatization):**
  + **Stemming**: Cắt bỏ hậu tố để đưa từ về gốc. Ví dụ: "running", "ran" → "run".
  + **Lemmatization**: Dùng từ điển ngôn ngữ để chuyển từ về dạng chuẩn (ví dụ: “better” → “good”).

### 2.3.2. Xây dựng từ điển và tính xác suất

Trong bước này, mục tiêu là tạo ra tập từ vựng (từ điển) và thống kê tần suất xuất hiện của từng từ trong từng lớp (ví dụ: “spam” và “ham”). Đây là nền tảng để tính toán xác suất có điều kiện trong các bước tiếp theo của thuật toán Naive Bayes.

**2.3.2.1. Xây dựng từ điển (Vocabulary)**

* Từ điển được tạo ra bằng cách trích xuất tất cả các từ (token) xuất hiện trong toàn bộ tập dữ liệu huấn luyện sau khi đã qua tiền xử lý (tokenization, loại bỏ stop words, chuẩn hóa).
* Mỗi từ trong từ điển sẽ trở thành một đặc trưng (feature) trong mô hình phân loại.
* Từ điển được xây dựng thống nhất và dùng chung cho tất cả các lớp để đảm bảo tính toàn vẹn và đồng nhất trong huấn luyện và dự đoán.
* Ví dụ với hai câu: "Buy now and get a discount", "Offer valid only today"
* Sau tiền xử lí, từ điển có thể là: → {buy, now, get, discount, offer, valid, today}

**2.3.2.2. Thống kê tần suất xuất hiện**

Sau khi có từ điển, bước tiếp theo là thống kê số lần xuất hiện của từng từ trong từng lớp dữ liệu để phục vụ cho việc tính xác suất có điều kiện.

Các chỉ số chính cần tính:

* count(w, c): Số lần từ www xuất hiện trong các văn bản thuộc lớp ccc (ví dụ: “spam” hoặc “ham”).
* total\_count(c): Tổng số từ (kể cả trùng lặp) xuất hiện trong tất cả văn bản thuộc lớp ccc.

Việc thống kê này giúp đánh giá mức độ liên quan của từ đối với từng lớp và được dùng trong công thức tính xác suất ở bước sau.

Ví dụ từ “discount” có thể:

* Xuất hiện 10 lần trong thư spam
* Xuất hiện 3 lần trong thư ham

### 2.3.3 Tính xác suất điều kiện

Dựa trên từ điển và xác suất từng từ đã tính được, tiến hành tính xác suất có điều kiện cho từng từ trong một thư thuộc về lớp "spam" hoặc "ham". Cụ thể, với mỗi từ trong tập từ vựng, ta xác định xác suất nó xuất hiện trong văn bản khi biết văn bản thuộc một lớp nhất định. Kỹ thuật làm trơn Laplace thường được sử dụng để tránh trường hợp xác suất bằng 0.

Trong thuật toán Naive Bayes, bước quan trọng là tính xác suất có điều kiện của từng từ (hoặc đặc trưng) trong văn bản khi biết trước nhãn lớp (ví dụ: "spam" hoặc "ham"). Công thức cơ như sau:

Trong đó:

count(w, c): Số lần từ w xuất hiện trong các thư thuộc lớp c.

total\_count(c): Tổng số từ (bao gồm cả lặp lại) trong toàn bộ thư thuộc lớp c

|V|: Số lượng từ duy nhất trong toàn bộ tập dữ liệu huấn luyện (kích thước từ điển)

**Khắc phục vấn đề xác suất điều kiện bằng 0**

Nếu trong dữ liệu huấn luyện không có đối tượng X nào có thuộc tính lớp Ck có thuộc tính Fi nhận một giá trị cụ thể vij, xác suất điều kiện P(Fi = xij | Ck) sẽ bằng 0.

Khi phân lớp, nếu có một đối tượng nào mang thuộc tính này thì xác suất phân vào lớp Ck luôn bằng 0.

A math equations and formulas

AI-generated content may be incorrect.Khắc phục bằng cách ước lượng theo công thức sau:

## 2.3.4 Áp dụng định lý Bayes

**Cơ sở lý thuyết xác suất**

Gọi A, B là hai biến cố

A white sheet with black text

AI-generated content may be incorrect.

Công thức Bayes tổng quát:

A white paper with black text

AI-generated content may be incorrect.  
 Trong đó ta gọi A là một chứng cứ (evidence) (trong bài toán phân lớp A sẽ là một phần tử dữ liệu), B là một giả thiết nào để cho A thuộc về một lớp C nào đó. Trong bài toán phân lớp chúng ta muốn xác định giá trị P(B/A) là xác suất để giả thiết B là đúng với chứng cứ A thuộc vào lớp C với điều kiện ra đã biết các thông tin mô tả A. P(B|A) là một xác suất hậu nghiệm (posterior probability hay posteriori probability) của B với điều kiện A.

**Áp dụng vào hệ thống phân loại thư rác**

Trong mô hình Naive Bayes, ta muốn tính xác suất để một thư điện tử với nội dung cụ thể thuộc về một lớp nhất định (ví dụ: **thư rác** hoặc **thư hợp lệ**). Theo định lý Bayes, ta có:

Trong đó:

Ck: là một lớp (ví dụ: spam)

w1,w2,…wn​: là các từ trong thư

P(Ck): là xác suất tiên nghiệm của lớp Ck

P(w1,...,wn∣Ck):là xác suất các từ xuất hiện đồng thời khi biết thư thuộc lớp Ck

## 2.3.5 Quyết định phân loại

Sau khi tính toán được xác suất hậu nghiệm P(C∣w1,w2,...,wn) cho từng lớp Ck​ (ví dụ: “spam” hoặc “ham”) dựa trên danh sách các từ w1,w2,...,wn ​ xuất hiện trong thư điện tử, mô hình Naive Bayes sẽ thực hiện bước quyết định phân loại.

Cụ thể, hệ thống sẽ so sánh xác suất giữa các lớp và lựa chọn lớp có giá trị xác suất hậu nghiệm lớn nhất. Việc lựa chọn này đảm bảo rằng thư sẽ được gán vào lớp có khả năng xảy ra cao nhất theo mô hình thống kê đã học từ dữ liệu huấn luyện.

Công thức quyết định phân loại như sau:

Ĉ = arg max₍Cₖ₎ [ log P(Cₖ) + ∑₍ᵢ₌₁ ₙ₎ log P(wᵢ | Cₖ) ]

Trong đó:

Ĉ: lớp dự đoán cuối cùng cho thư điện tử cần phân loại.

P(Ck): xác suất tiên nghiệm của lớp Ck

P(wi∣Ck): xác suất có điều kiện của từ wi trong lớp Ck

## 2.3.6. Đánh giá và tinh chỉnh mô hình

Sau khi huấn luyện mô hình Naïve Bayes, bước đánh giá và tinh chỉnh là cần thiết để đảm bảo hệ thống phân loại thư rác hoạt động chính xác và ổn định trên dữ liệu thực tế.

**2.3.6.1. Đánh giá hiệu suất mô hình**

Mô hình thường được đánh giá trên tập kiểm tra độc lập, sử dụng các chỉ số phổ biến trong học máy:

* Accuracy (Độ chính xác): Tỷ lệ dự đoán đúng trên tổng số mẫu.
* Precision (Độ chính xác theo lớp): Tỷ lệ thư được phân loại là “spam” thực sự là spam.
* Recall (Độ nhạy): Tỷ lệ thư “spam” được nhận diện đúng trên tổng số thư spam thực tế.
* F1-score: Trung bình điều hòa giữa Precision và Recall, phản ánh sự cân bằng giữa độ chính xác và độ bao phủ.
* Confusion Matrix (Ma trận nhầm lẫn): Biểu diễn trực quan số lượng mẫu được phân loại đúng và sai theo từng lớp.

Ngoài ra, kỹ thuật đánh giá chéo (k-fold cross-validation) cũng thường được sử dụng để kiểm tra độ ổn định và khả năng khái quát của mô hình, tránh hiện tượng quá khớp (overfitting) vào một tập dữ liệu cụ thể.

**2.3.6.2. Tinh chỉnh mô hình**

Để nâng cao hiệu quả mô hình, có thể áp dụng một số chiến lược tinh chỉnh sau:

* Cải thiện bước tiền xử lý: Loại bỏ từ dừng, chuẩn hóa văn bản, áp dụng stemming hoặc lemmatization giúp giảm nhiễu và giữ lại thông tin quan trọng.
* Chọn đặc trưng nâng cao: Thay vì chỉ sử dụng tần suất từ đơn giản, có thể áp dụng TF-IDF để tăng trọng số cho các từ mang tính phân biệt cao.
* Tối ưu kỹ thuật làm trơn: Có thể điều chỉnh hệ số Laplace hoặc áp dụng các phương pháp khác như m-estimate để cải thiện xử lý với các từ hiếm.
* Cân bằng dữ liệu: Nếu dữ liệu bị mất cân bằng (ví dụ: thư "ham" nhiều hơn thư "spam"), có thể sử dụng các kỹ thuật như oversampling (tăng mẫu lớp ít) hoặc undersampling (giảm mẫu lớp nhiều) để mô hình học được đầy đủ hơn và tránh thiên lệch.

Việc đánh giá và tinh chỉnh nên được thực hiện liên tục trong quá trình phát triển để đảm bảo mô hình đạt được hiệu suất tối ưu khi triển khai thực tế.

## 2.5. Ưu điểm và hạn chế của thuật toán Naive Bayes

### 2.5.1. Ưu điểm

* **Đơn giản và dễ triển khai:** So với nhiều thuật toán phân loại khác, Naive Bayes được xem là đơn giản hơn do quá trình ước lượng tham số dễ dàng.
* **Tốc độ xử lý nhanh, dễ mở rộng:** So với hồi quy logistic, Naive Bayes có tốc độ xử lý nhanh, hoạt động hiệu quả khi giả định độc lập có điều kiện (conditional independence) được đảm bảo. Ngoài ra, nó yêu cầu ít dung lượng lưu trữ.
* **Xử lý tốt dữ liệu nhiều chiều:** Trong các bài toán như phân loại tài liệu, dữ liệu thường có số chiều rất lớn – điều mà nhiều thuật toán khác gặp khó khăn, nhưng Naive Bayes vẫn xử lý tốt.

Hiệu quả trong các bài toán có số lượng đặc trưng lớn.

Vẫn hoạt động tốt ngay cả khi dữ liệu huấn luyện bị giới hạn.

Xử lý hiệu quả với các đặc trưng dạng phân loại (categorical features).

Với các đặc trưng dạng số (numerical features), thuật toán giả định dữ liệu tuân theo phân phối chuẩn (normal distribution).

### 2.5.2. Hạn chế

**- Vấn đề tần suất bằng 0 (Zero Frequency):** Xảy ra khi một giá trị hạng mục không tồn tại trong tập huấn luyện.

# CHƯƠNG 3: THIẾT KẾ , TRIỂN KHAI HỆ THỐNG PHÂN LOẠI THƯ RÁC

**3.1. Tải tập dữ liệu:**



Nhập thư viện pandas để xử lý dữ liệu.

**pd.read\_csv(file\_url, sep ='\t', header =None, names = [“label”, “sms”]):** Đây là lệnh gọi hàm đọc tệp dữ liệu Bộ sưu tập thư rác SMS và lưu trữ dữ liệu trong đối tượng DataFrame có tên data.

**pd.read\_csv():** Đây là chức năng tích hợp từ thư viện Pandas để đọc tệp CSV vào DataFrame.

Đọc tập tin SMSSpamCollection với dấu tab (\t) phân cách.

pd.read\_csv(file\_url, sep =’\t’, header =None, names = [“label”,

Không có tiêu đề nên header=None.

Đặt tên cột là Label (spam/ham) và SMS (nội dung tin nhắn).

## 3.2. Phân chia dữ liệu huấn luyện và kiểm tra

Sau khi tải dữ liệu, tiến hành xáo trộn ngẫu nhiên và chia thành 2 tập: 80% huấn luyện và 20% kiểm tra.

**Xáo trộn toàn bộ tập dữ liệu** (frac=1), dùng random\_state=1 để kết quả tái lập được:



**Tính chỉ số phân chia** tại 80% số dòng (để chia train/test):



Lấy **80% đầu tiên** làm tập **huấn luyện** và reset lại chỉ số dòng:



Lấy **20% còn lại** làm tập **kiểm tra** và reset chỉ số dòng:



## 3.3. Xử lý tiền dữ liệu

Xử lý tiền dữ liệu đóng vai trò rất quan trọng trong bài toán phân loại thư rác. Trước hết, nó giúp chuẩn hóa dữ liệu đầu vào bằng cách loại bỏ các ký tự đặc biệt, dấu câu và chuyển toàn bộ văn bản về chữ thường, từ đó giảm thiểu sự trùng lặp do khác biệt chữ hoa và chữ thường. Quá trình tách từ (tokenization) biến mỗi tin nhắn thành danh sách các từ riêng biệt, giúp mô hình dễ dàng đếm tần suất xuất hiện của từng từ và tạo ra đặc trưng đầu vào dạng bảng đếm từ.

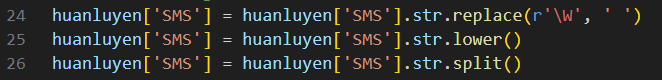
**3.3.1. Làm sạch thư:**

Việc làm sạch thư có một số tác dụng như sau:

**Loại bỏ các ký tự không cần thiết** như dấu câu, ký hiệu đặc biệt, số… giúp dữ liệu trở nên nhất quán và dễ xử lý hơn, tránh gây nhầm lẫn khi phân tích từ ngữ.

**Chuẩn hóa văn bản** bằng cách chuyển toàn bộ chữ về dạng chữ thường giúp giảm thiểu sự phân tách không cần thiết giữa các từ có cùng nghĩa nhưng khác dạng chữ.

**Giảm nhiễu và dữ liệu thừa,** giúp mô hình tập trung vào các từ khóa quan trọng, từ đó cải thiện khả năng nhận diện các mẫu đặc trưng của thư rác và thư thường.



**Tăng tính chính xác của việc tách từ và đếm tần suất,** tạo nền tảng vững chắc cho việc xây dựng từ vựng và tính toán xác suất trong thuật toán Naive Bayes.

split(): được dùng để tách câu thành danh sách các từ.

lower(): đươc dùng để chuyển văn bản thành chữ thường

replace(r'\W', ' '): được dùng để xóa dấu câu, giữ lại chữ và số

**3.3.2. Xây dựng từ vựng:**

Xây dựng từ vựng trong bài toán phân loại văn bản có nhiệm vụ thu thập tất cả các từ duy nhất (unique words) xuất hiện trong tập dữ liệu huấn luyện. Quá trình này bao gồm việc duyệt qua từng tin nhắn, tách tin nhắn thành danh sách từ (tokens), sau đó thêm từng từ vào danh sách từ vựng và loại bỏ các từ trùng lặp. Kết quả thu được là một tập hợp các từ không trùng nhau đại diện cho toàn bộ nội dung trong tập huấn luyện.

A screen shot of a computer program

AI-generated content may be incorrect.

Từ vựng này sẽ được dùng để tạo ma trận đặc trưng đầu vào, trong đó mỗi từ tương ứng với một cột thể hiện số lần từ đó xuất hiện trong từng tin nhắn. Đây là bước quan trọng giúp mô hình Naive Bayes tính toán xác suất xuất hiện của từng từ trong các lớp spam và ham, từ đó phân loại tin nhắn một cách chính xác.

vocab = [] : Tạo một **danh sách rỗng** để chứa tất cả các từ (không loại trùng).

vocab = list(set(vocab)) : chuyển danh sách sang tập hợp (set) → **loại bỏ từ trùng lặp.**

Đếm số lần từ xuất hiện

Tạo một từ điển, với mỗi từ là key, giá trị là danh sách đếm số lần xuất hiện trong từng tin nhắn:



Duyệt qua từng tin nhắn (theo chỉ số):

Với mỗi từ trong tin nhắn, tăng số đếm tại dòng tương ứng:

Chuyển từ điển number\_appear thành DataFrame – mỗi từ là một cột:

A screen shot of a computer program

AI-generated content may be incorrect.

**Gộp** bảng huấn luyện với bảng đếm từ – tạo DataFrame đầy đủ để huấn luyện mô hình Naive Bayes:



## 3.4. Dự đoán hàm

Tính xác suất cơ bản

Tách dữ liệu thành 2 bảng: tin nhắn spam và ham:

A black background with white text

AI-generated content may be incorrect.

Tính **xác suất tiên nghiệm** (prior): tỷ lệ spam và ham:

A black background with white text

AI-generated content may be incorrect.

Tính tổng số từ (tokens) trong tin nhắn spam và ham:

A black screen with white text

AI-generated content may be incorrect.

Kích thước từ vựng.

alpha = 1: giá trị **Laplace smoothing** (tránh xác suất bằng 0).



Tính xác suất có điều kiện cho từng từ

Khởi tạo từ điển chứa xác suất P(word|spam) và P(word|ham):

A black background with white text

AI-generated content may be incorrect.

Tính xác suất có điều kiện: mỗi từ word xuất hiện trong spam và ham.

Áp dụng Laplace smoothing.

A screen shot of a computer code

AI-generated content may be incorrect.

## 3.5. Hàm phân loại và áp dụng lên dữ liệu kiểm tra

**Hàm classify()** nhận đầu vào là một chuỗi tin nhắn, xử lý giống bước tiền xử lý, sau đó tính tích xác suất của từng từ theo công thức Naive Bayes để đưa ra nhãn dự đoán.

A computer screen with text

AI-generated content may be incorrect.

**Hàm classify\_test()** được dùng để áp dụng lên từng dòng trong tập kiểm tra test. Mỗi tin nhắn sẽ được đưa vào hàm phân loại và trả về nhãn dự đoán tương ứng.

A computer screen shot of text

AI-generated content may be incorrect.

## 3.6. Đánh giá độ chính xác mô hình

-Áp dụng hàm classify\_test\_set() cho toàn bộ dữ liệu trong tập kiểm tra.

-So sánh nhãn dự đoán với nhãn thực tế (Label) để đếm số lượng dự đoán đúng và tính độ chính xác.

A screen shot of a computer code

AI-generated content may be incorrect.

## 3.7 .TỔNG KẾT CHƯƠNG 3

Trong chương 3, hệ thống phân loại thư rác dựa trên thuật toán Naive Bayes đã được thiết kế và triển khai với các bước cụ thể và tuần tự. Đầu tiên, dữ liệu được tải về và chia thành hai tập: huấn luyện và kiểm tra. Sau đó, dữ liệu được xử lý tiền xử lý để chuẩn hóa văn bản và loại bỏ các yếu tố gây nhiễu.

Tiếp theo, quá trình xây dựng từ vựng và thống kê tần suất từ được thực hiện để tạo nền tảng cho việc tính toán xác suất có điều kiện. Bằng cách áp dụng kỹ thuật làm trơn Laplace, hệ thống đảm bảo xử lý tốt các từ mới không xuất hiện trong tập huấn luyện, khắc phục vấn đề xác suất bằng 0.

Hàm phân loại được xây dựng dựa trên công thức Naive Bayes, thực hiện việc dự đoán nhãn cho các tin nhắn mới. Cuối cùng, mô hình được đánh giá thông qua việc so sánh nhãn dự đoán với nhãn thực tế trên tập kiểm tra, cho phép xác định độ chính xác của hệ thống.

Việc triển khai mô hình Naive Bayes trong bài toán phân loại thư rác đã minh chứng được sự hiệu quả của thuật toán với ưu điểm đơn giản, dễ thực hiện nhưng mang lại kết quả chính xác cao. Đây cũng là bước chuẩn bị quan trọng để có thể tiếp tục nghiên cứu, tối ưu và mở rộng mô hình cho các bài toán phân loại phức tạp hơn trong thực tế.