# **Mục lục**

[**Mục lục** 1](#_Toc436945879)

[**Lời nói đầu** 2](#_Toc436945880)

[**1.** **Tài liệu tham khảo** 3](#_Toc436945881)

[**2.** **Giới thiệu** 3](#_Toc436945882)

[2.1 Bài toán thực tế 3](#_Toc436945883)

[2.2 Mục đích bài toán 3](#_Toc436945884)

[2.3 Phạm vi thực hiện 4](#_Toc436945885)

[2.4 Hướng giải quyết 4](#_Toc436945886)

[**3.** **Cơ sở toán học và dữ liệu** 4](#_Toc436945887)

[3.1 Cơ sở toán học 4](#_Toc436945888)

[3.2 Tập dữ liệu 5](#_Toc436945889)

[**4.** **Xây dựng chương trình** 6](#_Toc436945890)

[4.1 Xây dựng tập dữ liệu 6](#_Toc436945891)

[4.2 Xác định text chính của email 8](#_Toc436945892)

[4.3 Phân tích đặc điểm của từ ngữ tiếng Việt với tiếng Anh 8](#_Toc436945893)

[4.4 Tách từ tiếng Anh 9](#_Toc436945894)

[4.5 Tách từ tiếng Việt 9](#_Toc436945895)

[4.6 Loại bỏ từ dừng, nhiễu, lỗi 10](#_Toc436945896)

[4.7 Chuẩn hóa từ về một dạng 10](#_Toc436945897)

[4.8 Xác định tần suất xuất hiện của của các từ đặc trưng 10](#_Toc436945898)

[4.9 Áp dụng thuật toán Naȉve Bayes 11](#_Toc436945899)

[**5.** **Kết quả thực hiện chương trình** 12](#_Toc436945900)

[**6.** **Kết luận** 14](#_Toc436945901)

[**7.** **Phân công công việc** 14](#_Toc436945902)

# **Lời nói đầu**

*Chúng em xin chân thành cám ơn thầy Thân Quang Khoát đã nhiệt tình giảng dạy trong suốt quá trình học để chúng em để có đủ lượng kiến thức có thể hoàn thành cho bài tập nhóm này.*

*Mục đích của để tài để các thành viên trong nhóm hiểu rõ hơn về thuật toán Naȉve Bayes được học trên lớp, các kỹ thuật xử lý để đạt được hiểu quả cao trong việc tính toán kết quả phán đoán. Mỗi thành viên rõ hơn về cách đọc thư điện tử và quản lý hòm thư cá nhân một cách hợp lý. Hơn nữa là để các thành viên trong nhóm hiểu nhau hơn và có một tác phong làm việc nhóm chuyên nghiệp, nhiệt tình, cởi mở hơn nữa.*

*Qua bài tập lớn này, chúng em đã học được cách tự nghiên cứu giải quyết một bài toán thực tiễn đặt ra trong cuộc sống. Bài tập lớn này là thành quả của một quá trình nghiên cứu nghiêm túc, sự phối hợp làm việc nhóm, từng đóng góp của mỗi cá nhân của nhóm chúng em. Đây thực sự là một kinh nghiệm quý báu cho bước đường tự nghiên cứu, học tập của chúng em sau này.*

*Một lần nữa em xin chân thành cảm ơn thầy!*

# **Tài liệu tham khảo**

[1] <https://www.coursera.org/course/nlp>

[2] <http://spamassassin.apache.org/publiccorpus/>

http://archive.ics.uci.edu/ml/

[3] <http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>

[4] <http://tartarus.org/~martin/PorterStemmer/def.txt>

[5] <https://lib.lhu.edu.vn/ViewFile/10652>

(Luận văn thạc sỹ của thầy Lương Quốc Sơn về bộ lọc thư rác)

[6] <http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer>

Thư viện tách từ vnTokenizer của thầy Lê Hồng Phương – Trường Đại Học Quốc Gia Hà Nội.

# **Giới thiệu**

## Bài toán thực tế

Hiện nay với sự phát triển mạnh mẽ và đi kèm với nhiều tiện ích của mạng internet thì việc sử dụng thư điện tử để trao đổi thông tin đã trở nên khá phổ biến. Trao đổi qua thư điện tử có rất nhiều lợi ích như việc truyền tải thông điệp dễ dàng, có thể đính kèm nhiều định dạng dữ liệu, thông tin truyền đi nhanh chóng và dễ dàng trong quản lý… Song song với lợi ích luôn có những vấn đề phát sinh. Điển hình là vieejv phải nhận những email mang nội dung không mong muốn từ những người không quen biết. Spam email là những bức thư chứa nội dung mà ta không chủ động đọc cũng như được gửi tới nhiều lần, chủ yếu liên quan đến nội dung quảng cáo. Nếu như mỗi ngày phải nhận cả chục email như vận sẽ dẫn đến cảm giác khó chịu. Nhưng để giữ cho địa chỉ email không bị rơi vào những đối tượng có mục đích phát tán email spam cũng là khá khó khắn khi ta chỉ là người dung đơn thuần chưa có hiểu biết rõ về bảo mật. Với những vấn đề đặt ra cần có một ứng dụng để chọn lọc, ngăn chặn từ xa với những email spam đến với với người dùng.

## Mục đích bài toán

* Phân loại thư vào hai tập thư rác và thư thường, mục tiêu đối với thư thường đạt tỷ lệ phân loại đúng là trên 90%, đối với thư rác là trên 70%.
* Nhận diện được một văn văn có phải là thư spam hay không dựa trên nội dung của văn bản đó.
* Thời gian xử lý ngắn, tối ưu về bộ nhớ lưu trữ
* Có thể phân loại đối với cả thư tiếng Anh và tiếng Việt
* Kịch bản: thư (dưới dạng text) -> phân loại -> thư thường | thư rác.

## Phạm vi thực hiện

Nhóm đã lựa chọn thuật toán Naïve Bayes để thực hiện phân loại email thành hai loại là email thường và email rác.

## Hướng giải quyết

Xác định nhãn của email sẽ được gắn dựa trên nội dung chính của email, loại bỏ đi nhưng ký tự đặc biệt. Để xác định nội dung của email cần phân tách lấy phần text chính, loại bỏ thành phần không mang ý nghĩa.

Tách ra những từ, cụm từ thường xuyên được sử dụng trong văn bản email và xác định tần suất xuất hiện của chúng từ đó phán đoán nội dung của văn bản dựa trên thuật toán Navie Bayes.

# **Cơ sở toán học và dữ liệu**

## Cơ sở toán học

Nhóm đã tính xác suất số dựa trên công thức Naȉve Bayes là một thuật toán phân tích thông kê thực hiện trên dữ liệu số học.

Thuật toán Naȉve Bayes trong bài toán xử lý nội dung văn bản đó là coi tập email đã được gắn nhãn là biến cố cho trước. Giả thiết này sẽ là một phương pháp đánh giá định lượng để đánh giá chứng cứ có được trong quá trình huấn luyện.

Công thức xác suất Naȉve Bayes:

**P(h|D) =**

Trong đó:

D là tập huấn luyện

*h* là một giả thiết

P(h) là xác suất trước giả thiết của h

P(D) là xác suất trước giả thiết của D

P(h/D) là xác suất sau, là xác suất điều kiện của h sau khi tập huấn luyện được biểu diễn dựa trên D

P(D|h) là xác suất của quan sát D dựa trên h. Xác suất này còn được gọi là khả năng có thể xảy ra.

Phương pháp phân loại email sẽ dựa trên xác suất sau P(h|D), xác suất để phân loại văn bản email vào nhãn thường hay rác sau khi quan sát được dữ liệu D đó là các đặc trưng, từ ngữ xuất hiện trong các văn bản email.

Lúc đó với h1, h2 lần lượt là phân lớp nhãn với email thường và email rác, ta có:

**P(hi|D) =**

Với D là tập từ quan sát được từ những email thu thập được

Tri thức là một tập các từ đặc trưng được trích ra từ cơ sở dữ liệu gọi là một vector n chiều x(x1,x2,...,xn) với n là số lượng từ đặc trưng.

Động cơ suy diễn là phân loại Naive-Bayes, giả sử các thuộc tính trong x là độc lập với nhau và với vị trí của chúng trong văn bản được xét, xét một email mới z với một vector m chiều z(z1,z2,...,zm) với z là tập con của x.

Xác suất để z thuộc nhãn lớp hi (i ϵ [1,2]) là:

**P(z1, …,zm|hi) =**

Phân loại Naive-Bayes tìm nhãn lớp có thể phù hợp nhất với z.

**Hnb = arg max P(hi).**

(i ϵ [1,2]) (\*)

Tập các sự kiện là sự xuất hiện của các từ thuộc tập x trong văn bản z.

## 3.2 Tập dữ liệu

* Nguồn:
* Tiếng Anh: Dữ liệu được tải về từ website đối với tập dữ liệu email tiếng Anh : [2]
* Tiếng Việt: Dữ liệu thu thập từ Gmail cá nhân của bạn bè trong phần spam đối với email là thư rác và thư thường được tập hợp từ các thành viên trong nhóm.
* Tổ chức dữ liệu

Tiếng Việt: 117 email thường, 73 email rác

Chia ra làm 4 folder: inputMail và inputMailSpam

Tiếng Anh: gồm 3 tập mail, mỗi tập 1200 mail thường, 1200 mail rác

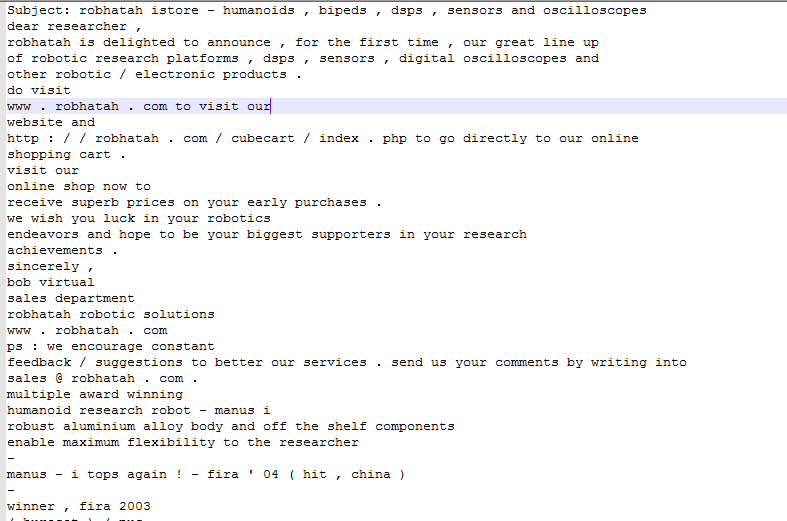
Chia đều mỗi tập mail trên thành 4 tập mail:

* 2 tập mail học gồm 840 mail thường, 840 mail rác
* 2 tập mail test gồm 360 mail thường, 360 mail rác

Được lưu trong hai thư mục test và learn tùy từng ngôn ngữ tương tứng

* Mô tả dữ liệu

Dữ liệu được lưu trữ dưới dạng file text thông thường, có thể bao gồm các thẻ HTML, ký tự đặc biệt, chữ số….



# **Xây dựng chương trình**

## Xây dựng tập dữ liệu

* Đối với dữ liệu là email tiếng Anh được tổng hợp từ nguồn trên internet với bộ dữ liệu khá phong phú đầy đủ => Đem đến kết quả đáng tin cậy hơn
* Đối với dữ liệu là email tiếng Việt được tổng hợp từ nhiều nguồn thông qua những email cá nhân, với những email đã được Gmail gắn thẻ là spam. Thư thường được thu thập tương tự.
* Tập học

## Xác định text chính của email

* Để xác định được nội dung text của văn bản thì tùy với từng tập email thu thập được lưu trữ dưới dạng xml hay html hay text.
* Đối với html ta loại bỏ đi những thẻ tag, link đường dẫn, tệp đính kèm và định dạng khác ngoài text…
* Đối với xml sử dụng Java Dom Parse để đọc tài liệu xml rồi lấy ra text của thẻ chứa nội dung của email ví dụ như lấy nội dung text của thẻ <text\_normal>

## Phân tích đặc điểm của từ ngữ tiếng Việt với tiếng Anh

|  |  |
| --- | --- |
| **Tiếng Việt** | **Tiếng Anh** |
| Từ không biến đổi hình thái, ý nghĩa ngữ pháp nằm ở ngoài từ Ví dụ: Tôi *nhìn*  anh ấy  Anh ấy *nhìn* tôi | Từ có biến đổi hình thái ý nghĩa ngữ pháp nằm trong từ. Ví dụ: I *see* him và He *sees* me |
| Có từ láy như lấp lánh, long lanh… |  |
| Một từ có thể mang nhiều nghĩa tùy từng ngữ cảnh | Từ ngữ thường rõ ràng |

## Tách từ tiếng Anh

* Do đặc trưng của Tiếng Anh là các từ đơn thường là các từ có nghĩa, nhóm đã thực hiện tách riêng từng từ đơn dựa theo các ký tự phân cách như khoảng trắng, dấu xuống dòng, các ký tự ngăn cách câu như [,.!], và các ký tự đặc biệt.

## Tách từ tiếng Việt

* Nhóm đã sử dụng thư viện **vnTokenizer** do nhóm tác giả Nguyễn Thị Minh Huyền, Vũ Xuân Lương và Lê Hồng Phương phát triển dựa trên phương pháp so khớp tối đa (Maximum Matching) với tập dữ liệu sử dụng là bản âm tiếng Việt của từ điển từ vựng tiếng Việt. Công cụ được xây dựng bằng ngôn ngữ Java mã nguồn mở.
* vnTokenizer: là bộ thư viện tách từ tự động đối với đoạn văn bản tiếng Việt sang đơn vị từ vựng. Được phát triển trong ngôn ngữ lập trình Java và là nền tảng độc lập. Nó đưa ra kết quả tốt về độ chính xác với tỷ lệ khoảng 96%-98%.

Mô hình thực hiện:

Văn bản

Tách từ

Chuỗi đơn vị từ

Từ điển

* Thư viện vnTokenizer:
* Đầu vào: Một câu, đoạn văn bản tiếng Việt bất kỳ
* Đầu ra: Chuỗi đơn vị từ đã được tách.

## Loại bỏ từ dừng, nhiễu, lỗi

* Đối với cả tiếng Việt và tiếng Anh đều có những từ dừng (stop word). Mặc dù là những từ thường được sử dụng nhiều nhất trong ngôn ngữ tự nhiên nhưng những từ này thường mang mục đích diễn đạt câu văn tốt hơn mà không mang ý nghĩa, nội dung của câu văn.
* Để tránh việc bị loãng khi xác định nội dung văn bản cần loại bỏ đi những từ này.
* Ví dụ trong tiếng Việt như: bởi, có, càng, có thể….
* Ví dụ trong tiếng Anh như: and, then, so …
* Đầu vào của hàm loại bỏ stop word sẽ là một văn bản thông thường và đầu ra không còn các từ stop word trong văn bản nữa.

Bước tiến hành:

Loại bỏ các ký tự đặc biệt như: . ? ! “ “ ( ) @ # …

Loại bỏ các số

Tìm và xóa bỏ từ stop word khỏi văn bản. Để làm việc này cần có sẵn bộ dữ liệu về stop word.

* Đối với tiếng Anh, chúng em lấy tập từ dừng tại địa chỉ *http://www.ranks.nl/stopwords*

## Chuẩn hóa từ về một dạng

* Đối với tiếng Anh thì một từ thường có nhiều dạng khác nhau tùy thuộc vào hoàn cảnh của câu nói ví dụ như organize, organizes, và organizing. Vì vậy để xác định các từ ngữ đặc trưng của văn bản được tốt cần chuẩn hóa từ ngữ tiếng Anh về cùng một dạng chuẩn nguyên dạng.
* Để giải quyết vấn đề này nhóm đã tham khảo thuật toán Stemming and lemmatization được public trên Stanford.edu [3].
* Mục tiêu của thuật toán đó là giảm bớt các hình thức khác nhau của từ về cùng mẫu cơ sở chung.
* Ví dụ: wrote => write
* Reading => read
* car, cars, car's, cars' $\Rightarrow$car
* Các bước của kỹ thuật Stemming: [4]
  + Bước 1 loại bỏ dạng số nhiều -s, đuôi –ed và –ing.
  + Bước 2 chuyển đổi đuôi y-i để cho cuối từ xuất hiện thêm một nguyên âm.
  + Bước 3 loại bỏ các hậu tố có các đuôi ví dụ: -ization, -ational ...
  + Bước 4 giải phóng các hậu tố -full, -ness ...
  + Bước 5 bỏ các hậu tố -ant, -ence ...
  + Bước xóa bỏ ký tự -e cuối cùng.
  + Bước đưa các động từ bất quy tắc về dạng nguyên thể

## Xác định tần suất xuất hiện của của các từ đặc trưng

* + - Đếm số lần xuất hiện của từ, cụm từ trong toàn bộ tập email rác hoặc email thường.
    - Đếm số email rác, thường chứa từ, cụm từ đó.

## Áp dụng thuật toán Naȉve Bayes

Thuật toán Naȉve Bayes được áp dụng trong hai giai đoạn:

* Giai đoạn học:

Từ tập các email trong cơ sở dữ liệu, qua quá trình xử lý chúng ta thu được tập từ khóa

**T = {tj}**

Gọi **D\_ci**với **(i ϵ [1,2])** là tập các email trong cơ sở dữ liệu có nhãn lớp **ci** (rác – thường)

Sau đó áp dụng công thức Bayes để tính xác suất **P(tj|hi)**

**P(tj|hi) =**

Vấn đề gặp phải đó là sẽ tồn tại những từ tj mà không xuất hiện trong nhãn lớp hi lúc đó sẽ xảy ra vấn đề P(tj|hi) = 0 cho nên ta sẽ thêm công thức trong trường hợp hợp mà từ **tj**không xuất hiện đó là:

**p =**

Do dữ liệu học là khá nhiều nên tập T cùng tương đối lớn dẫn tới p sẽ tiến tới 0. Do đó để tránh ảnh hưởng tới công thức tính xác suất \* ta sẽ logarit từng thành phần trong công thức tính xác suất \* của nhiều phần tử.

Còn một kỹ thuật khác để có thể tăng độ chính xác khi phán đoán của văn bản đó là có thể đặt một giá trị người nào đó mà có thể loại bỏ đi những từ xuất hiện trong h1 > ngưỡng còn trong h2 < ngưỡng và ngược lại. Khi đó ta đã làm tăng tính đặc trưng của từ ngữ xuất hiện trong email.

* Giai đoạn phân loại
* Từ văn bản d, trích ra tập T\_d gồm các từ khóa tj đã được định nghĩa trong tập T (T\_d T).
* Giả sử xác suất từ khóa tj xuất hiện đối với lớp ci là độc lập đối với vi trí của từ khóa đó trong văn bản P(tj ở vị trí k/ci) = P(tj ở vị trí m/ci) với mọi k, m
* Đối với mỗi phân lớp ci, tính xác suất hậu nghiệm của văn bản đối với ci

**P(ci).**

* Phân lớp văn bản d thuộc vào lớp

=  **P(ci).**

# **Kết quả thực hiện chương trình**

* **Tập Tiếng Anh**

Chúng em có sử dụng một quá trình tối ưu để trích chọn ra tập từ khóa đem đến kết quả cao nhất có thể đối với tập test.

**Thống kê tập học**

|  |  |  |  |
| --- | --- | --- | --- |
| Tập | Định dạng | Mail thường | Mail rác |
| 1 | Text, không có thẻ HTML | 840 | 840 |
| 2 | Text, có thẻ HTML | 840 | 840 |
| 3 | Text, có thẻ HTML | 840 | 360 |

**Thống kê tập test**

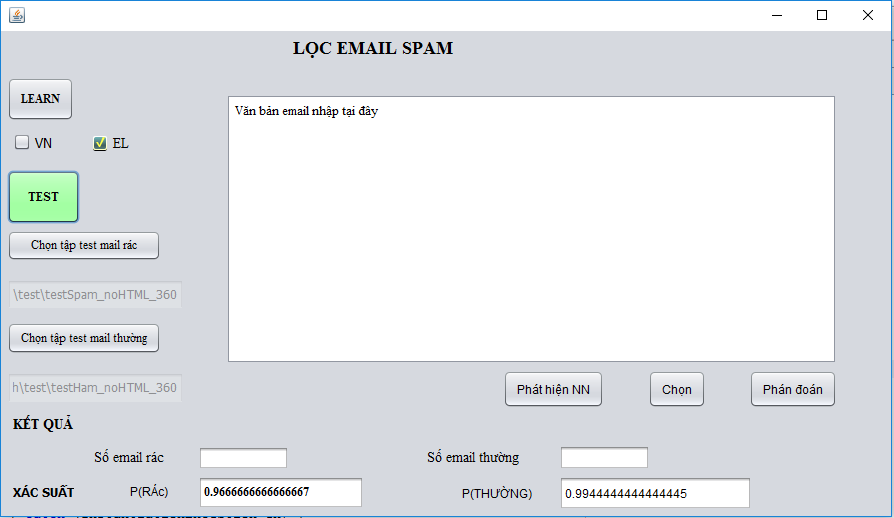
|  |  |  |  |
| --- | --- | --- | --- |
| Tập | Định dạng | Mail thường | Mail rác |
| 1 | Text, không có thẻ HTML | 360 | 360 |
| 2 | Text, có thẻ HTML | 360 | 360 |
| 3 | Text, có thẻ HTML | 360 | 360 |

**Tỉ lệ phán đoán chính xác**

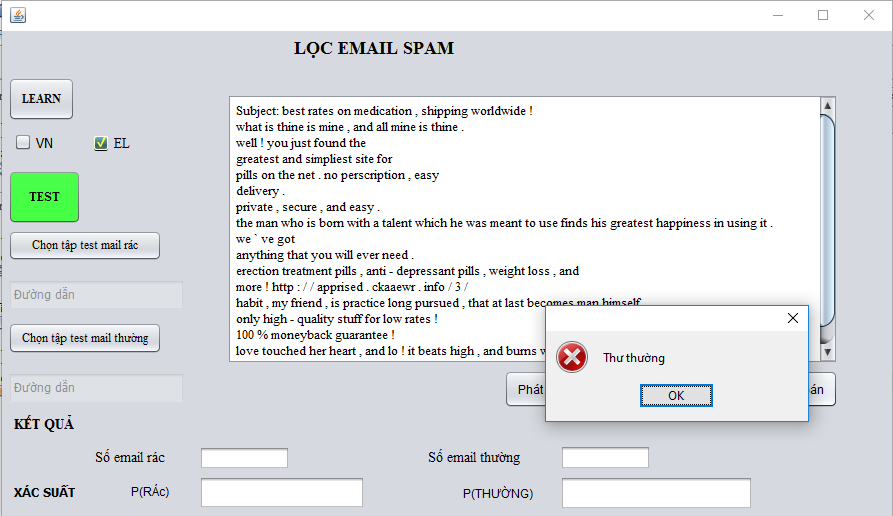
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tập | Định dạng | Số từ khóa | Mail thường | Mail rác |
| 1 | Text, không có thẻ HTML | 12560 | 99.44% | 96.94% |
| 2 | Text, có thẻ HTML | 14739 | 97.82% | 90.83% |
| 3 | Text, có thẻ HTML | 14296 | 98.04% | 89.58% |

* Tập Tiếng Việt

**Giao diện chương trình**



Hình Giao diện sau khi chạy chức năng test



Hình Giao diện sau khi phán đoán một văn bản email là thường hay spam

# **Kết luận**

* Qua kết quả test với 3 tập mail Tiếng Anh đều mang lại kết quả phán đoán chính xác khá cao (>90%), chúng em rút ra nhận xét :
  + Phân loại Naïve Bayes khi được áp dụng vào bài toán phân loại văn bản với đầu vào và tập từ tốt sẽ đem đến kết quả phân loại rất tốt.
  + Phân loại Naïve Bayes có thuật toán không quá phức tạp, chỉ mất nhiều thời gian trong quá trình học, còn trong quá trình phán đoán thì chi phí thời gian thấp.
  + Phân loại Naïve Bayes rất phù hợp cho bài toán phân loại văn bản khi đơn giản, tốc độ thực hiện nhanh.
  + Độ phân bố của tập mail sử dụng có ảnh hưởng lớn đến kết quả phán đoán, tập mail tập trung sẽ cho kết quả cao, tập mail có các đặc trưng phân bố rời rạc không rõ ràng sẽ cho kết quả không thực sự tốt.

# **Phân công công việc**

|  |  |
| --- | --- |
| Phan Công Huân | Xử lý tách từ, đưa ra key word, tần suất của email tiếng Anh  Tinh chỉnh thuật toán |
| Vương Lâm | Xử lý thuật toán ghép tiếng Việt và tiếng Anh |
| Phan Tân | Tách từ văn bản tiếng Việt, đưa ra key word, tần xuất |
| Lê Xuân Duy | Phát hiện ngôn ngữ văn bản, giao diện  Viết báo cáo |