**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**--------------------**



**ĐỒ ÁN 03**

**XÂY DỰNG HỆ THỐNG LỌC THƯ RÁC**

**SINH VIÊN THỰC HIỆN:** **MSSV:**

**Tôn Thất Tấn 17150086**

**Huỳnh Minh Trí 17110242**

**GIÁO VIÊN HƯỚNG DẪN: Cô Từ Tuyết Hồng**

**TP. HỒ CHÍ MINH - 2020**

Trang Bìa

**LỜI CẢM ƠN**

“Để hoàn thành đồ án môn học này, em xin gửi lời cảm ơn chân thành đến:

Ban giám hiệu trường Đại Học Sư phạm kỹ thuật Thành phố Hồ Chí Minh vì đã tạo điều kiện về cơ sở vật chất với hệ thống thư viện hiện đại, đa dạng các loại sách, tài liệu thuận lợi cho việc tìm kiếm, nghiên cứu thông tin.

Xin cảm ơn giảng viên bộ môn – **Cô Từ Tuyết Hồng** đã giảng dạy tận tình, chi tiết để em có đủ kiến thức và vận dụng chúng vào đồ án môn học này.

Do chưa có nhiều kinh nghiệm làm để tài cũng như những hạn chế về kiến thức, trong bài đồ án chắc chắn sẽ không tránh khỏi những thiếu sót. Rất mong nhận được sự nhận xét, ý kiến đóng góp, phê bình từ phía Cô để bài đồ án môn học được hoàn thiện hơn.

Lời cuối cùng, em xin kính chúc Cô nhiều sức khỏe, thành công và hạnh phúc.”

TP.HCM, ngày 15 tháng 12 năm 2020

Sinh viên thực hiện

**Tôn Thất Tấn 17150086**

**Huỳnh Minh Trí 17110242**

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................

TP.HCM, ngày 15 tháng 12 năm 2020

Giáo viên hướng dẫn

**NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN 1**

................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................

TP.HCM, ngày 15 tháng 12 năm 2020

Giáo viên phản biện 1

**NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN 2**

................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................

TP.HCM, ngày 15 tháng 12 năm 2020

Giáo viên phản biện 2

**MỤC LỤC**

[Phần 1: Mở đầu 12](#_Toc58926978)

[1. Lời mở đầu 12](#_Toc58926979)

[2. Mục đích, phạm vi nghiên cứu 12](#_Toc58926980)

[3. Phân chia công việc 13](#_Toc58926981)

[Phần 2: Nội dung 14](#_Toc58926982)

[Chương 1: Cơ sở lý thuyết 14](#_Toc58926983)

[1.1. Một số khái niệm cơ bản 14](#_Toc58926984)

[1.1.1. Định nghĩa thư rác 14](#_Toc58926985)

[1.1.2. Phân loại thư rác 14](#_Toc58926986)

[1.1.3. Quá trình phát tán thư rác 14](#_Toc58926987)

[1.1.4. Tác hại thư rác 15](#_Toc58926988)

[1.2. Các phương pháp lọc thư rác 17](#_Toc58926989)

[1.2.1. Lọc thư rác bằng luật lệ 17](#_Toc58926990)

[1.2.2. Lọc thư rác dựa trên địa chỉ IP 17](#_Toc58926991)

[1.2.3. Lọc dựa trên chuỗi hỏi/đáp (Challenge/Response filters) 18](#_Toc58926992)

[1.2.4. Phương pháp lọc dựa trên mạng xã hội. 19](#_Toc58926993)

[1.2.5. Phương pháp lọc nội dung 20](#_Toc58926994)

[1.2.5.1. Lọc dựa trên các dấu hiệu nhận biết 20](#_Toc58926995)

[1.2.5.2. Lọc thư rác thông qua danh sách trắng, đen. 21](#_Toc58926996)

[1.2.5.3. Lọc thư rác dựa vào phương pháp heuristic. 21](#_Toc58926997)

[1.2.5.4. Lọc thư rác dựa trên xác suất thống kê và học máy. 22](#_Toc58926998)

[Chương 2: Tổng quan thuật toán 24](#_Toc58926999)

[2.1. Mô hình Bag of Words (BOW) 24](#_Toc58927000)

[2.1.1. Chuẩn hóa dữ liệu 24](#_Toc58927001)

[2.1.1.1. Tách từ 24](#_Toc58927002)

[2.1.1.2. Chuẩn hoá từ 24](#_Toc58927003)

[2.1.1.3. Remove stopwords 24](#_Toc58927004)

[2.1.1.4. Giới hạn số features 24](#_Toc58927005)

[2.1.1.5. Giới hạn độ lớn 1 vector của 1 văn bản 25](#_Toc58927006)

[2.1.2. TF-IDF 25](#_Toc58927007)

[2.1.2.1. TF(Term Frequency) 25](#_Toc58927008)

[2.1.2.2. IDF (Inverse Document Frequency) 25](#_Toc58927009)

[2.1.2.3. TF-IDF 26](#_Toc58927010)

[2.1.3. Count Vectorizer 26](#_Toc58927011)

[2.2. Navie Bayes Classification 26](#_Toc58927012)

[2.2.1. Khái niệm 26](#_Toc58927013)

[2.2.2. Phân loại Naive Bayes 27](#_Toc58927014)

[Chương 3: Triển khai đề tài 28](#_Toc58927015)

[3.1. Tiền xử lí 28](#_Toc58927016)

[3.1.1. Xử lí dữ liệu 28](#_Toc58927017)

[3.1.1.1. Thu thập dữ liệu 28](#_Toc58927018)

[3.1.1.2. Làm sạch dữ liệu 29](#_Toc58927019)

[3.1.1.3. Tách từ 29](#_Toc58927020)

[3.1.1.4. Chuẩn hoá từ 30](#_Toc58927021)

[3.1.1.5. Loại bỏ stop words 30](#_Toc58927022)

[3.1.1.6. Ghép các từ 30](#_Toc58927023)

[3.1.1.7. Chuyển các văn bản sang các vector 31](#_Toc58927024)

[3.1.2. Code hoàn chỉnh 31](#_Toc58927025)

[3.1.2.1. Đọc dữ liệu 31](#_Toc58927026)

[3.1.2.2. In dữ liệu 31](#_Toc58927027)

[3.1.2.3. Đếm số lượng email dựa theo label 31](#_Toc58927028)

[3.1.2.4. Biểu diễn bằng biểu đồ 32](#_Toc58927029)

[3.1.2.5. Đếm chiều dài của email 32](#_Toc58927030)

[3.1.2.6. Chuẩn hóa văn bản 33](#_Toc58927031)

[3.1.2.7. Loại bỏ stopwords dựa theo từ điển 33](#_Toc58927032)

[3.1.2.8. Biểu diễn các từ là spam 34](#_Toc58927033)

[3.1.2.9. Biểu diễn các từ là ham 34](#_Toc58927034)

[3.1.2.10. Đánh nhãn cho email 35](#_Toc58927035)

[3.1.2.11. Chuyển đổi từ thành vector 35](#_Toc58927036)

[3.2. Huấn luyện mô hình 36](#_Toc58927037)

[3.2.1. Chia tập dữ liệu 36](#_Toc58927038)

[3.2.2. Đánh giá mô hình 36](#_Toc58927039)

[3.2.3. Lưu mô hình 36](#_Toc58927040)

[3.2.4. Code hoàn chỉnh 36](#_Toc58927041)

[3.2.4.1. Naïve bayes 36](#_Toc58927042)

[3.2.4.2. Lưu mô hình 37](#_Toc58927043)

[3.3. Xây dựng Flask 38](#_Toc58927044)

[3.3.1. Xây dựng giao diện 38](#_Toc58927045)

[3.3.1.1. Trang chủ 38](#_Toc58927046)

[3.3.1.2. Kết quả là spam 38](#_Toc58927047)

[3.3.1.3. Kết quả là ham 39](#_Toc58927048)

[3.3.1.4. Nhúng mô hình vào website 40](#_Toc58927049)

[3.3.2. Code Hoàn chỉnh 40](#_Toc58927050)

[3.3.2.1. Code giao diện 40](#_Toc58927051)

[3.3.2.2. Code kết quả 40](#_Toc58927052)

[3.3.2.3. Code xử lý mô hình 41](#_Toc58927053)

[3.4. Đưa lên host 41](#_Toc58927054)

[3.4.1. Đăng ký Heroku 41](#_Toc58927055)

[3.4.2. Cấu hình 42](#_Toc58927056)

[3.4.2.1. Tạo requirements 42](#_Toc58927057)

[3.4.2.2. Tạo procfile 42](#_Toc58927058)

[3.4.2.3. Tạo nltk.txt 43](#_Toc58927059)

[3.4.2.4. Tạo buildpack 43](#_Toc58927060)

[3.4.3. Upload 43](#_Toc58927061)

[Chương 4: Kết quả 44](#_Toc58927062)

[4.1. Demo sản phẩm 44](#_Toc58927063)

[4.1.1. Kiểm tra là spam 45](#_Toc58927064)

[4.1.1.1. Kết quả 45](#_Toc58927065)

[4.1.2. Kiểm tra là ham 46](#_Toc58927066)

[4.1.2.1. Kết quả 46](#_Toc58927067)

[4.2. Link demo sản phẩm 46](#_Toc58927068)

[Phần 3: kết luận 47](#_Toc58927069)

[1. Đánh giá 47](#_Toc58927070)

[2. Ưu điểm 47](#_Toc58927071)

[3. Nhược điểm 47](#_Toc58927072)

[Phần 4: Tài liệu tham khảo 47](#_Toc58927073)

**DANH MỤC HÌNH ẢNH**

[Hình 1: Spam email 15](#_Toc58884852)

[Hình 2: Biểu đồ spam email 17](#_Toc58884853)

[Hình 3: Bộ lọc email 19](#_Toc58884854)

[Hình 4: Bộ lọc request 20](#_Toc58884855)

[Hình 5: Bộ lọc dấu hiệu 22](#_Toc58884856)

[Hình 6: File csv 29](#_Toc58884857)

[Hình 7: Làm sạch dữ liệu 30](#_Toc58884858)

[Hình 8: Phân tách từ 30](#_Toc58884859)

[Hình 9: Đọc file 31](#_Toc58884860)

[Hình 10: In file 32](#_Toc58884861)

[Hình 11: Đếm email 32](#_Toc58884862)

[Hình 12: Biểu đồ email 32](#_Toc58884863)

[Hình 13: Đếm chiều dài của email 32](#_Toc58884864)

[Hình 14: Chuẩn hóa 33](#_Toc58884865)

[Hình 15: Loại bỏ stopwords 33](#_Toc58884866)

[Hình 16: Spam words 34](#_Toc58884867)

[Hình 17: Ham words 34](#_Toc58884868)

[Hình 18: Đánh nhãn 35](#_Toc58884869)

[Hình 19: Chuyển đổi 35](#_Toc58884870)

[Hình 20: Chuyển đổi thành vector 36](#_Toc58884871)

[Hình 21: Kết quả huấn luyện 37](#_Toc58884872)

[Hình 22: Lưu mô hình 37](#_Toc58884873)

[Hình 23: Demo trang chủ 38](#_Toc58884874)

[Hình 24: Demo spam 39](#_Toc58884875)

[Hình 25: Demo ham 39](#_Toc58884876)

[Hình 26: Code giao diện 40](#_Toc58884877)

[Hình 27: Code kết quả 40](#_Toc58884878)

[Hình 28: Code xử lí flask 41](#_Toc58884879)

[Hình 29: Tạo app 42](#_Toc58884880)

[Hình 30: Cấu hình requirements 42](#_Toc58884881)

[Hình 31: Tạo procfile 43](#_Toc58884882)

[Hình 32: Tạo nltk.txt 43](#_Toc58884883)

[Hình 33: Add buildpack 43](#_Toc58884884)

[Hình 34: Setup heroku 44](#_Toc58884885)

[Hình 35: Upload thành công 44](#_Toc58884886)

[Hình 36: Demo spam 45](#_Toc58884887)

[Hình 37: Kết quả spam 45](#_Toc58884888)

[Hình 38: Demo ham 46](#_Toc58884889)

[Hình 39: Kết quả ham 46](#_Toc58884890)

# Mở đầu

###### Lời mở đầu

Ngày nay, Internet mở ra nhiều kênh liên lạc, nhiều dịch vụ mới cho người sử dụng, một trong những dịch vụ mà Internet mang lại là dịch vụ thư điện tử (Email), đó là phương tiện giao tiếp rất đơn giản, tiện lợi và hiệu quả đối với cộng đồng người sử dụng dịch vụ này. Chính vì những lợi ích do thư mang lại nên số lượng thư trao đổi trên Internet ngày càng tăng, và một số không nhỏ trong đó là thư rác (Spam).

Trong những năm gần đây, spam hay các thư không mong muốn đã trở thành một vấn nạn và đe dọa khả năng giao tiếp của con người trên kênh liên lạc này, đó là một trong những thách thức lớn mà khách hàng và các nhà cung cấp dịch vụ phải đối phó. Spam đã trở thành một hình thức quảng cáo chuyên nghiệp, phát tán virus, ăn cắp thông tin với nhiều thủ đoạn và mánh khóe cực kỳ tinh vi. Người dùng sẽ phải mất khá nhiều thời gian để xóa những thư “không mời mà đến”, nếu vô ý còn có thể bị nhiễm virus và nặng nề hơn là mất thông tin như thẻ tín dụng, tài khoản ngân hàng qua các thư dạng phishing....

Theo báo cáo tình hình thư rác do Kaspersky Lab vừa công bố, tỷ lệ thư rác trong lưu lượng truy cập thư của quý 3/2014 tăng 1,7 % so với quý trước, đạt trung bình 66,9%. Ba nguồn phát tán thư rác hàng đầu gồm có Mỹ (14%) và Nga (6,1%) và Việt Nam đứng vị trí thứ 3 với 6%.

Để ngăn chặn spam, nhiều tổ chức, cá nhân đã nghiên cứu và phát triển những kỹ thuật phân loại thư thành các nhóm; từ đó xác định, nhận biết giữa thư rác và thư có giá trị. Tuy nhiên, những người tạo nên thư rác luôn tìm mọi cách vượt qua các bộ phân loại này và phát tán chúng.

###### Mục đích, phạm vi nghiên cứu

Nhằm ngăn chặn, hạn chế mất thông tin như thẻ tín dụng, tài khoản ngân hàng, virus… qua các thư từ những gói thư rác khi người dùng vô tình nhấn vào. Nhóm đã đề xuất nghiên cứu xây dựng một chức năng nhằm phân loại tin rác theo hướng lọc theo nội dung với mục đích tìm hiểu, thử nghiệm một số phương pháp tiếp cận cho bài toán phân loại thư, từ đó ngăn chặn thư spam hiệu quả hơn.

Nội dung của báo cáo được trình bày theo 3 chương.

Tổ chức cấu trúc như sau:

**Chương 1** Tổng quan về học máy và thư rác: Chương này giới thiệu tổng quát về học máy và thư rác bao gồm khái niệm, ứng dụng về học máy, các kỹ thuật. Chương cũng giới thiệu khái quát về thư rác, các đặc trưng của thư rác và biểu diễn thư rác dựa trên học máy.

**Chương 2** Phân loại thư rác bằng một số thuật toán: Nội dung chính trong chương này là đi sâu nghiên cứu thuật toán học máy có giám sát là Naïve Bayes và phương pháp BOW(Bag of Words)

**Chương 3** Cài đặt, thử nghiệm và đánh giá thuật toán, bộ dữ liệu thử nghiệm và cài đặt chi tiết hai thuật toán đề cập ở chương 2. Phần cuối của chương trình bày kết quả thu được và đưa ra đánh giá về hai thuật toán được sử dụng trong bài toán lọc thư rác.

###### Phân chia công việc

|  |  |  |
| --- | --- | --- |
| **Nội dung công việc** | **Người thực hiện** | |
|  | **Tôn Thất Tấn** | **Huỳnh Minh Trí** |
| Thu thập dữ liệu | **x** | **x** |
| Xử lí dữ liệu |  | **x** |
| Tìm hiểu mô hình bag of words | **x** | **x** |
| Triển khai mô hình bag of words | **x** |  |
| Tìm hiểu giải thuật naive bayes | **x** | **x** |
| Triển khai giải thuật naive bayes | **x** | **x** |
| Tìm hiểu flask | **x** | **x** |
| Thực hiện flask | **x** |  |
| Thiết kế giao diện |  | **x** |
| Tìm hiểu heroku | **x** | **x** |
| Triển khai heroku | **x** | **x** |
| Thiết kế báo cáo | **x** |  |
| Kiểm tra, sửa lỗi báo cáo |  | **x** |

# Nội dung

## Cơ sở lý thuyết

### Một số khái niệm cơ bản

#### Định nghĩa thư rác

Hiện nay vẫn chưa có một định nghĩa hoàn chỉnh, chặt chẽ về thư rác. Có quan điểm coi thư rác là những thư quảng cáo không được yêu cầu (Unsolicited Commercial Email-UCE), có quan điểm rộng hơn cho rằng thư rác bao gồm thư quảng cáo, thư quấy rối, và những thư có nội dung không lành mạnh (Unsolicited Bulk Emai -UBE). Sau đây sẽ đưa ra một định nghĩa thông dụng nhất về thư rác và giải thích các đặc điểm của nó để phân biệt thư rác với thư thông thường. 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 bị coi là 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.

#### Phân loại thư rác

Có rất nhiều cách phân loại thư rác:

- Dựa trên kiểu phát tán thư rác: Tính tới thời điểm hiện tại, thư rác có thể bị gửi thông qua thư điện tử, nhóm thảo luận (newsgroups), điện thoại di động (Short Message Service - SMS) và các dịch vụ gửi tin nhắn trên mạng (như Yahoo Messenger, Windows Messenger...)

- Dựa vào quan hệ với người gửi thư rác: bao gồm người lạ mặt, bạn bè, người quen và các dịch vụ quyên góp giúp đỡ…

- Dựa vào nội dung của thư rác: các kiểu nội dung phổ biến như thư về thương mại, thư về chính trị, thư về công nghệ, chuỗi thư (chain e-mail) và các loại khác (như thư phát tán virus...).

- Dựa trên động lực của người gửi: Thông thường, thư rác được gửi đi cho những mục đích quảng bá thông tin. 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. Sự phân loại thư rác rất quan trọng không chỉ trong lĩnh vực tạo những bộ lọc thư rác có hiệu quả cao mà còn giúp cho việc ban hành các bộ luật chống thư rác phù hợp.

#### Quá trình phát tán thư rác

Quá trình phát tán thư gồm 2 bước:

Thu thập các địa chỉ email

Nhận dạng kí tự email

Dụ người dùng điền thông tin địa chỉ email

Mua cơ sở dữ liệu người dùng từ ngân hàng , website…

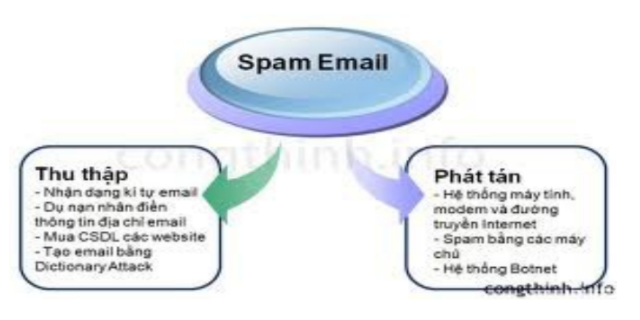
Tạo email bằng dictionary attack

Phát tán thư rác

Sử dụng hệ thống máy tính, modem và đường truyền internet

Spam bằng các máy chủ

Hệ thống botnet

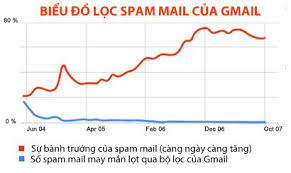


1. Spam email

#### Tác hại thư rác

Theo thống kê thư rác hiện chiếm hơn một nửa số e-mail truyền trên Internet và chính thư rác là nguồn lây lan virus nhanh nhất. Thiệt hại do chúng gây ra rất lớn đối với sự phát triển internet nói chung và người sử dụng thư điện tử nói riêng. Theo thống kê toàn cầu của hãng nghiên cứu Ferris Research ở San Francisco, thư rác gây thiệt hại 50 tỷ USD trong năm 2005. Chỉ tính riêng ở Mỹ, thiệt hại do thư rác gây ra đối với các doanh nghiệp ước tính khoảng 17 tỷ USD/năm. Thư rác chiếm khoảng 80% lưu lượng thư điện tử thế giới trong quý 1/2006, đó là kết luận của nhóm hợp tác chống thư rác gồm các công ty AOL, Bell Canada, Cigular Wireless, EarthLink, France Telecom, Microsoft, Verizon, và Yahoo. Microsoft và AOL cho biết hai hãng này trung bình mỗi ngày chặn gần 5 tỷ thư rác. Ước tính, cứ 9 trong 10 email sử dụng dịch vụ MSN Hotmail của Microsoft là thư rác. Tại Việt Nam, tình hình thư rác cũng đang rất phức tạp. Công ty Điện toán và Truyền số liệu (VDC) - ISP lớn nhất Việt Nam - cho biết, thư rác hiện nay chiếm phần lớn lưu lượng email qua hệ thống máy chủ thư của ISP này. Các thư phàn nàn gửi đến ISP nếu không giải quyết, các khách hàng của ISP đó có thể bị liệt vào danh sách đen, không gửi được email ra địa chỉ nước ngoài. Một số ISP cho biết, cuối năm ngoái, khách hàng của nhiều ISP ở Việt Nam thường xuyên bị tê liệt do bị liệt vào danh sách đen. Mỗi lần thoát ra khỏi danh sách này ISP phải mất khoảng 40 USD. Tại trang web Spamhaus.org (tổ chức theo dõi các nguồn gửi thư rác), có lần vnn.vn đã có trong danh sách top 10 ISP cung cấp nhiều rác nhất. Không chỉ gây thiệt hại về tiền bạc, thư rác còn làm giảm hiệu quả làm việc, gây stress, tiêu tốn thời gian của nhân viên... Những điều này cũng đồng nghĩa với việc, năng suất lao động giảm, ảnh hưởng tới tình hình kinh doanh và doanh thu của công ty. Một số lời khuyên cho người dùng thư điện tử:

* Yêu cầu và đòi hỏi nhà chức trách phải đưa ra những luật lệ nghiêm cấm thư rác và có hình phạt đích đáng cho kẻ cố tình gửi thư rác.
* Mỗi người dùng nên tạo nhiều địa chỉ email, với mục đích khác nhau nên dùng địa chỉ email khác nhau.
* Hạn chế việc đăng kí các dịch vụ vô ích: nên tìm hiểu kĩ thông tin về dịch vụ trước khi cung cấp địa chỉ email của mình.
* Kích hoạt các dịch vụ chống thư rác của ISP.
* Cài đặt một số chương trình xử lý thư trong máy tính cá nhân để xóa thư rác ngay khi chuyển về máy.
* Bảo vệ mật khẩu của mình: chọn mật khẩu lạ, khó đoán chứa chữ cái, xen lẫn chữ số và chữ hoa xen lẫn chữ thường.
* Thường xuyên ghi dự phòng dữ liệu quan trọng. Đồng thời cảnh giác với những thư từ người quen biết nhưng không được báo trước, bởi có thể chúng được gửi đi mà người gửi không biết.
* Số lượng Spam vẫn luôn luôn tăng và ngày càng tinh vi hơn, người ta nhận định rằng việc chống Spam sẽ luôn luôn phải thực hiện, tùy vào ý thức của cư dân Internet và sức mạnh của công nghệ mà việc Spam chỉ được hạn chế phần nào.



1. Biểu đồ spam email

### Các phương pháp lọc thư rác

#### Lọc thư rác bằng luật lệ

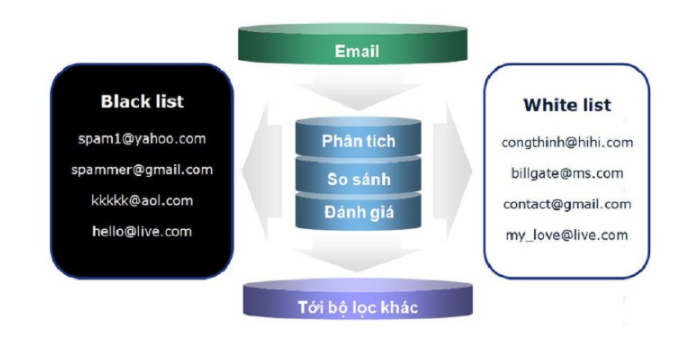
Khi tình trạng thư rác ngày càng tăng trên đường truyền internet gây ra nhiều phiền toái và thiệt hại lớn trên thế giới rất nhiều các quốc gia đã đưa ra các luật để ngăn chặn thư rác. Dưới đây là một số nội dung cơ bản liên quan tới giải pháp ngăn chặn thông qua luật lệ pháp lý được đưa ra trên báo điện tử của bộ viễn thông . Mỹ là một những nước đầu tiên trên thế giới cố gắng ban hành các văn bản pháp luật để giải quyết vấn đề thư điện tử rác tràn ngập. Từ tháng 7 năm 1997, bang Nevada đã dẫn đầu trong việc ban hành các quy phạm pháp luật quy định về hành vi phục vụ và sử dụng thư tín điện tử. Tính đến tháng 3 năm 2003, đã có 26 bang ban hành quy phạm pháp luật quy định về dịch vụ và hành vi sử dụng thư tín điện tử. Đến tháng 11 năm 2003, con số này lên đến 36. Về phía chính quyền liên bang, từ những năm 1990, cả Thượng nghị viện và Hạ nghị viện đều quan tâm đến sự lan rộng của thư tín điện tử quấy rối và thư rác, và đã đưa ra nhiều dự án luật như “Luật bảo vệ hộp thư không bị quấy rối” (1999), “Luật Bảo vệ người sử dụng thư điện tử”, “Luật Khống chế thư điện tử không được phép” (2000), “Luật Khống chế thư rác truyền qua đường điện thoại vô tuyến” (2000) , “Luật Chống thư rác” (2001). Mười năm gần đây, Liên minh Châu Âu cũng đã ban hành một số chỉ lệnh, đưa ra các quy phạm và chỉ dẫn đối với các vấn đề thương mại điện tử, thông tin điện tử, bảo hộ dữ liệu. Trong các chỉ lệnh nói trên, có không ít các qui định có liên quan mật thiết, thậm chí là trực tiếp với phục vụ và sử dụng thư điện tử như “Chỉ lệnh Bảo vệ dữ liệu cá nhân ở Châu Âu”, “Chỉ lệnh về thông tin điện tử và bảo mật dữ liệu” ... Ngày 12 tháng 7 năm 2002, Nghị Viện Liên minh Châu Âu đã thông qua “Chỉ lệnh Bảo mật riêng tư và Thông tin điện tử trong Liên minh Châu Âu”. Chỉ lệnh quy định: Từ 31 tháng 10 năm 2003, trong phạm vi Liên minh Châu Âu, nếu chưa được người nhận đồng ý trước, không được gửi thư điện tử thương mại hay nhằm mục đích tuyên truyền cho cá nhân. Tiếp theo sau khi Liên minh Châu Âu đưa ra các qui định về phục vụ và sử dụng thư điện tử, các nước thành viên Liên minh Châu Âu, như Italia, Anh, Đan Mạch, Tây Ban Nha... đều đã ban hành quy phạm pháp luật trong nước quy định hành vi cung cấp và sử dụng thư điện tử, ngăn chặn sự tràn ngập của thư rác. Tại Việt Nam vấn đề thư rác bắt đầu nhận được sự quan tâm từ phía các cơ quan có trách nhiệm. Bộ Thương mại đang soạn thảo Thông tư quản lý hoạt động quảng cáo thương mại trên các phương tiện điện tử. Trên trang báo điện tử của bộ viễn thông, Bà Lại Việt Anh, Trưởng Phòng chính sách, Vụ Thương mại điện tử, Bộ Thương mại, nhận xét: mục tiêu của Thông tư này trước mắt tập trung quản lý ba hình thức quảng cáo đang bức xúc: thư điện tử, tin nhắn điện thoại di động và quảng cáo trên trang thông tin điện tử.

#### Lọc thư rác dựa trên địa chỉ IP

Phương pháp lọc thư rác thông qua địa chỉ IP là phương pháp đơn giản và được sử dụng sớm nhất trong công cuộc chống thư rác. Dựa vào địa chỉ IP của người gửi để xác định thư đó bị ngăn chặn hoặc cho qua. Có hai cách để thực hiện việc lọc thư: một là duy trì một danh sách các địa chỉ IP bị chặn (còn gọi là danh sách đen blacklist); thứ hai là sử dụng một danh sách các địa chỉ IP cho phép qua (danh sách trắng whitelist).

Danh sách đen (Blacklist) Người ta lập ra một danh sách các địa chỉ gửi thư rác. Các nhà cung cấp dịch vụ thư điện tử (ISP) sẽ dựa trên danh sách này để loại bỏ những thư nằm trong danh sách này. Danh sách này thường xuyên được cập nhật và được chia sẻ giữa các nhà cung cấp dịch vụ. Một số danh sách đen điển hình được lập ra như: SpamCop Blocking List và Composite Block List. Ưu điểm của phương pháp này là các ISP sẽ ngăn chặn được khá nhiều địa chỉ gửi thư rác. Mặc dù danh sách đen này luôn được cập nhật nhưng với sự thay đổi liên tục địa chỉ, sự giả mạo địa chỉ hoặc lợi dụng một mail server hợp pháp để gửi thư rác đã làm số lượng thư rác gửi đi vẫn ngày càng tăng cao. Do đó phương pháp này chỉ ngăn chặn được một nửa số thư rác gửi đi và sẽ mất rất nhiều thư hợp pháp nếu ngăn chặn nhầm.

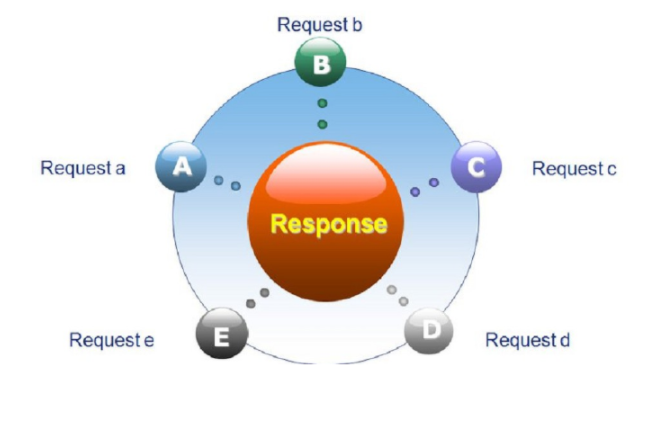
Danh sách trắng (Whitelist) Danh sách các địa chỉ tin cậy (Safe Sender List), danh sách này có thể do một nhà cung cấp dịch vụ nào đó cung cấp. Những địa chỉ thuộc danh sách sẽ được cho qua bộ lọc. Người dùng phải đăng ký với nhà cung cấp danh sách để được nằm trong danh sách. Ưu điểm: số lượng địa chỉ trong danh sách trắng sẽ ít hơn trong danh sách đen vì thế sẽ dễ cập nhật hơn danh sách đen và giải quyết được tình trạng chặn nhầm thư. Tuy nhiên cả hai phương pháp trên đều có nhược điểm là khó cập nhật, nhất là khi ai đó thay đổi địa chỉ IP. Ngoài ra người gửi cũng có thể lợi dụng server mail có trong danh sách trắng để gửi thư rác, khi đó rất khó kiểm soát.



1. Bộ lọc email

#### Lọc dựa trên chuỗi hỏi/đáp (Challenge/Response filters)

Đặc trưng của phương pháp này là khả năng tự động gửi thư hồi đáp cho người gửi để yêu cầu một số hành động chắc chắn về việc gửi thư của họ. Chương trình kiểm tra này được đặt tên là “Turing Test” sau một vài kiểm tra được nghĩ ra bởi nhà toán học người anh tên là Alan Turing. Trong một vài năm gần đây xuất hiện của một vài dịch vụ Internet tự động xử lý hàm Challenge/Response này cho người dùng, chương trình yêu cầu người gửi thư phải vào website của họ và trả lời một số câu hỏi để chắc chắn về e-mail mà người này đã gửi.Việc này chỉ được yêu cầu trong lần gửi thư đầu tiên. Đối với một số người dùng có lượng thư trao đổi thấp, hệ thống đơn lẻ này có thể chấp nhận được như một phương pháp hoàn hảo để loại trừ hoàn toàn thư rác từ hòm thư của họ.



1. Bộ lọc request

#### Phương pháp lọc dựa trên mạng xã hội.

Các nghiên cứu gần đây đã bắt đầu khai thác thông tin từ mạng xã hội cho việc xác định thư rác bằng cách xây dựng một đồ thị (các đỉnh là địa chỉ email, cung được thêm vào giữa 2 node A và B nếu giữa A và B có sự trao đổi thư qua lại). Người ta đã sử dụng một số tính chất đặc trưng của mạng xã hội để xây dựng một công cụ lọc thư rác. Đầu tiên, người ta phân đồ thị thành các thành phần con rồi tính độ phân cụm cho từng thành phần này. Mỗi thành phần con là một đồ thị mạng xã hội của một node, bao gồm tất cả các node xung quanh là “node hàng xóm” (các node có cung liên kết với node này) và những cung liên kết giữa các node hàng xóm này với nhau. Nếu thành phần nào có độ phân cụm thấp thì node tương ứng với thành phần đó là một địa chỉ gửi thư rác. Trong thành phần mạng xã hội của những node gửi thư rác, những node hàng xóm của nó thường là những node rất ngẫu nhiên, không có mối quan hệ (không có sự trao đổi email qua lại với nhau) nên độ phân cụm của mạng xã hội của những node này rất thấp. Ngược lại, mạng xã hội ứng với những người dùng bình thường có độ phân cụm cao hơn. Dựa vào độ phân cụm, người ta tạo được danh sách đen (Blacklist) gồm địa chỉ email tương ứng với những node có độ phân cụm rất thấp, danh sách trắng (Whitelist) ứng với node có độ phân cụm cao, số node còn lại sẽ được đưa vào danh sách cần xem xét (Greylist). Phương pháp này có thể phân loại được 53% tổng số email một cách chính xác là ham hay spam. Nhược điểm của phương pháp là những spammer có thể xây dựng mạng xã hội của chính họ nên khó có thể phát hiện ra.

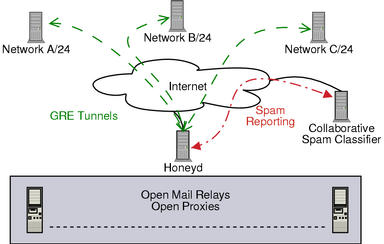
#### Phương pháp lọc nội dung

Phương pháp lọc nội dung để phân loại thư rác đã và đang được quan tâm, nghiên cứu và ứng dụng nhiều nhất. Phương pháp này dựa vào nội dung và chủ đề bức thư để phân biệt thư rác và thư hợp lệ. Phương pháp này có ưu điểm đó là chúng ta có thể dễ dàng thay đổi bộ lọc để nó có thể lọc các loại thư rác cho phù hợp.

Nhược điểm của phương pháp này là: do biết được cách thức lọc nội dung nên các spammer luôn luôn thay đổi hình thức nội dung của thư rác. Phần dưới đây trình bày những nét cơ bản nhất về các phương pháp lọc nội dung thông dụng.

##### Lọc dựa trên các dấu hiệu nhận biết

Trước tiên, tạo ra các địa chỉ email để bẫy thư rác, gọi là honeypots, phương pháp này được nghiên cứu phát triển nhiều vào năm 2003. Honeypots chứa các địa chỉ sao cho không bao giờ thư bình thường có thể gửi đến. Do đó thư gửi đến bẫy địa chỉ này ta có thể coi đó là thư rác. Sau đó hệ thống so sánh thư mới đến với thư đã được bẫy. Sự so sánh dựa trên dấu hiệu nhận biết, nếu chúng có dấu hiệu giống nhau thì có thể kết luận thư mới đến là thư rác. Ưu điểm của phương pháp này là đơn giản, nhanh và không lọc nhầm thư thường thành thư rác. Tuy nhiên spammer có thể dễ dàng vượt qua hệ thống bằng cách sinh ngẫu nhiên các mẩu thư rác sau đó gộp lại nhằm làm cho dấu hiệu của các bức thư rác khác nhau. Bởi vậy tỉ lệ lọc thư rác của hệ thống luôn nhỏ hơn 70%. Do không lọc thư thường thành thư rác nên phương pháp này được triển khai trên server. Một hệ thống lọc thư rác dựa trên honeypots hoạt động rất hiệu quả đó là eTrap. Hệ thống eTrap sử dụng honeypots để thu thập thông tin về spam. Những thông tin về spam được lưu trữ trong cơ sở dữ liệu chia sẻ chung. Hệ thống eTrap lọc thư rác dựa trên những thông tin về spam này.



1. Bộ lọc dấu hiệu

Hình 1.2 : Mô tả tổng quan quá trình hoạt động của honeyd

Trước tiên honeyd bẫy các địa chỉ gửi thư rác, sau đó toàn bộ thông tin về thư rác thu được sẽ được gửi tới Collaborative Spam Classifier để tổng hợp thông tin. Dựa vào những thông tin đó bộ phân loại thư rác sẽ phân tichsm để phân loại thư rác.

##### Lọc thư rác thông qua danh sách trắng, đen.

Hệ thống tìm xem các từ trong danh sách đen/trắng có nằm trong thư mới đến không và đếm số lần xuất hiện của chúng. Nếu số lượng từ thuộc danh sách trắng nhiều hơn rất nhiều số từ thuộc danh sách đen thì bức thư đó là hợp pháp và ngược lại sẽ là thư rác. Đặc trưng của bộ lọc thông qua bỏ phiếu trên danh sách đen/trắng: — Không có biến đổi dữ liệu ban đầu. — Biểu thức chính quy để tách từ ra khỏi thư là: [[:graph:]]+ — Việc chọn đặc trưng đơn giản chỉ là các từ đơn — Cơ sở dữ liệu về đặc trưng chỉ được nạp khi các từ nằm trong danh sách đen hoặc trắng. Nếu nằm trong danh sách đen thì đặt là -1, trong danh sách trắng là +1, các trường hợp còn lại đặt là 0. — Luật tổ hợp là : “Điểm mới = Điểm cũ + trọng số đặc trưng” — Ngưỡng lọc cuối cùng là : Nếu Điểm mới > 0 là thư hợp pháp, nếu < 0 là thư rác. Như vậy bộ lọc thực hiện chấm điểm các từ trong danh sách đen và các từ trong danh sách trắng bằng nhau. Một số cải biên của phương pháp này là đánh trọng số cho các từ trong danh sách đen cao hơn trong danh sách trắng hoặc ngược lại.

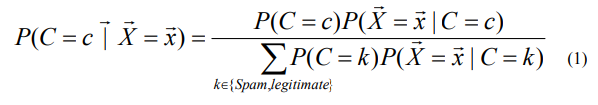
##### Lọc thư rác dựa vào phương pháp heuristic.

Cách thức hoạt động của phương pháp này là dựa trên việc xác định những từ đặc trưng thuộc về thư rác, từ đặc trưng thuộc về thư hợp pháp, sau đó phát hiện những đặc trưng đó trong thư mới nhận để đưa ra kết luận thư đó là thư rác hay thư hợp lệ. Người ta đánh trọng số cho các đặc trưng trên bằng tay hoặc bằng thuật toán và lập một ngưỡng để phân loại thư. Nếu bức thư có trọng số lớn hơn ngưỡng quy định sẽ bị coi là thư rác. Các chương trình lọc thư rác sử dụng phương pháp này có hiệu suất khác nhau. Vì mỗi chương trình sử dụng một luật lọc khác nhau. Một số chương trình lọc theo phương pháp này như hệ thống chấm điểm cho email sử dụng phương pháp hueristic của mail server Mdaemon, SpamAssassin hay SpamGuard của Yahoo. Phương pháp này có ưu điểm là dễ cài đặt và hiệu suất chặn thư rác khá cao khi xây dựng được hệ thống luật tốt. Nhược điểm chính của phương pháp này là tỉ lệ chặn nhầm thư hợp pháp cũng khá lớn 0.5%. Phương pháp này không linh hoạt do các luật được xây dựng luôn chậm hơn sao với sự biến đổi của từ ngữ trong thư rác. Phương pháp này thường được áp dụng cho các bộ lọc thư ở server.

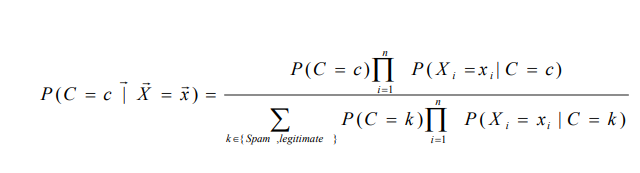
##### Lọc thư rác dựa trên xác suất thống kê và học máy.

Đầu tiên sẽ phân loại các bức thư thành thư rác và thư hợp lệ. Một thuật toán được áp dụng để trích chọn và đánh trọng số cho các đặc trưng của thư rác theo một cách nào đó (thường là áp dụng công thức xác suất). Sau khi trích chọn đặc trưng, hai tập thư rác và thư hợp lệ sẽ được sử dụng để huấn luyện một bộ phân loại tự động. Quá trình huấn luyện dựa trên một phương pháp học máy. Tỉ lệ chặn thư rác của bộ lọc sử dụng phương pháp này rất cao, khoảng 99%. Chương trình SpamProbe có thể đạt tới tỉ lệ lọc thư rác tới 99.9%. Các phương pháp học máy và xác suất thống kê cho phép phân loại cả những thư rác chưa từng xuất hiện trước đó. Phương pháp này còn có tỉ lệ chặn thư hợp pháp rất thấp, thấp hơn nhiều so với phương pháp heuristic. Nhược điểm của phương pháp này là phải có một tập hợp các thư để huấn luyện. Hiệu suất của bộ lọc sẽ phụ thuộc nhiều vào tập huấn luyện này. Tập dữ liệu càng lớn càng chứa nhiều dạng khác nhau thì kết quả phân loại về sau sẽ càng chính xác. Hiện nay phương pháp lọc thư rác theo học máy và xác suất thống kê là một phương pháp có triển vọng với nhiều ứng dụng thương mại như Hotmail, Google, Yahoo. Để có một bộ lọc hoàn hảo dường như không thể thực hiện được, một bộ lọc tốt nhất là bộ lọc kết hợp nhiều bộ lọc. Việc Spam ngày càng được thực hiện tinh vi hơn đòi hỏi các bộ lọc phải có khả năng biến đổi theo sự thay đổi của Spam, sự thay đổi về số lượng, về nội dung và cấu trúc của các thư spam. Vì vậy yêu cầu đặt ra phải có một bộ lọc có khả năng cập nhật để có thể thay đổi, chống lại những thư spam có cấu trúc nội dung mới, bộ lọc học máy lọc dựa trên nội dung Email Classification Using Example(ECUE) đã được chứng minh là có khả năng thực hiện được điều đó. Trong khuôn khổ khóa luận này em xin trình bày hệ thống lọc thư rác ECUE mới do Delany đề xuất và đã xây dựng thử nghiệm thành công.

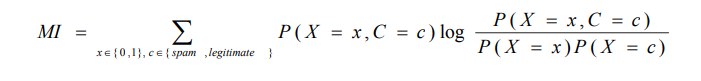
Lọc thư rác dựa trên thuật toán bayes: Coi mỗi email được biểu diễn bởi một vectơ thuộc tính đặc trưng x r = (x1, x2,…,xn). với (x1, x2,…,xn) là các giá trị thuộc tính X1, X2, …, Xn tương ứng trong không gian đặc trưng (space model) . Ta sử dụng giá trị nhị phân 0 và 1 để mô tả email đó có đặc điểm Xi hay không, giả xử nếu email đó có đặc điểm Xi thì ta đặt thuộc tính Xi = 1, còn nếu email đó không có đặc điểm Xi thì ta có thuộc tính Xi = 0. Từ thuyết xác suất của Bayes và xác suất đầy đủ chúng ta có công thức tính xác suất mail với vectơ x r = (x1, x2,…,xn) thuộc vào lớp c như sau:



Để đơn giản khi tính P( X |C r ) ta phải giả sử X1, X2,…,Xn là độc lập. Khi đó biểu thức (1) tương đương với biểu thức sau:



Giá trị được sử dụng rất rộng rãi để đánh hạng cho thuộc tính là giá trị tương hỗ MI(mutual information), ta lấy những thuộc tính có giá trị MI lớn nhất. Ta có thể tính giá trị tương hỗ MI(Mutual information) mà mỗi đại diện của X thuộc về loại C như sau :



Một email được coi là spam nếu:



Giả sử các thuộc tính Xi là độc lập khi đó ta có:



Khi đó (2) tương đương với :



Thuật toán bayes đã được áp dụng vào chương trình lọc thư rác spambayes, và cho kết quả lọc khá hiệu quả.

## Tổng quan thuật toán

### Mô hình Bag of Words (BOW)

Là một mô hình dùng để chuẩn hóa, chuyển đổi văn bản thành dạng số để có thể dạy cho máy tính hiểu được

#### Chuẩn hóa dữ liệu

##### ****Tách từ****

Tiếng Anh: Từ thường được ngăn bởi khoảng trắng (space).

Tiếng Việt: Từ có thể gồm một hoặc nhiều âm tiết, các âm tiết được ngăn với nhau bởi khoẳng trắng.

Tách từ là bước xử lý cơ bản trong xử lý văn bản. Chất lượng của hệ tách từ ảnh hưởng trực tiếp đến các bước sau như gán nhãn từ loại, phân tích cú pháp, phân tích ngữ nghĩa, cũng như các ứng dụng xử lý ngôn ngữ tự nhiên.

Ví dụ

Input: Để đóng vai quần chúng chúng tôi phải mặc quần áo theo lối phong kiến.

Output: Để đóng vai quần\_chúng chúng\_tôi phải mặc quần\_áo theo lối phong\_kiến.

##### Chuẩn hoá từ

Mục đích là đưa văn bản từ các dạng không đồng nhất về cùng một dạng. Dưới góc độ tối ưu bộ nhớ lưu trữ và tính chính xác cũng rất quan trọng.

Ví dụ: U.S.A = USA

Ví dụ trong từ điển, training data của chúng ta không có U.S.A, chỉ có USA, thì việc convert những từ như U.S.A về USA là điều cần thiết để các bước xử lý sau như text classification, intent detection được chính xác.

Ngoài ra với tiếng Nhật: 猫＝ねこ＝ネコ(full size)＝ﾈｺ(half size)

Có nhiều cách viết, mỗi cách viết khi lưu trữ sẽ tốn lượng memory khác nhau, như half size chỉ tốn 1/2 dung lượng so với full size nên tuỳ theo nhu cầu, tình hình thực tế, chúng ta sẽ đưa văn bản về 1 dạng đồng nhất.

Ngoài ra trong một vài trường hợp, nếu ký tự số không mang lại lợi ích gì thì cũng sẽ tiến hành loại bỏ các ký tự số đó, nếu cứ để nguyên rất có thể các ký tự số sẽ trở thành noise, ảnh hưởng đến tính chính xác của model sau này.

##### ****Remove stopwords****

Chúng ta phải xử lý mới những từ mà xuất hiện nhiều, tuy nhiên lại không mang lại quá nhiều ý nghĩa. Những từ này được gọi là “stop words”. Trong tiếng Anh, đó là những từ như “is”, “this”, “that”. Còn trong tiếng Việt, đó có thể là những từ “để”, “này” , “kia”.

##### ****Giới hạn số features****

Trong dữ liệu của nhóm, chúng ta có khoảng 4.000 bài báo được phân loại, dẫn tới một vocabulary cực kì lớn. Để làm giảm độ lớn của feature vectors, chúng ta có thể chọn một vài từ xuất hiện nhiều nhất. Ví dụ, 5000 từ xuất hiện nhiều nhất ( stop words đã bị loại bỏ ở phần tiền xử lý).

##### ****Giới hạn độ lớn 1 vector của 1 văn bản****

Nếu 1 từ trong vocabulary không xuất hiện trong văn bản thì giá trị feature đó bằng 0. Vì thế ta có thể loại bỏ nó đi. Bằng cách làm nay với tập dữ liệu của nhóm, file train giảm từ 700MB xuống còn 7MB.

#### TF-IDF

Đánh lại trọng số với TF-IDF

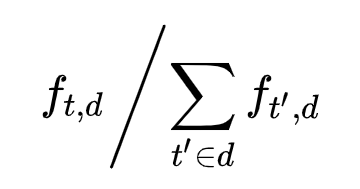
Trong hầu hết các ngôn ngữ, có một số từ có xu hướng xuất hiện thường xuyên như trong tiếng anh có "is", "the"... tương tự tiếng việt có các từ như "là", "của", "cứ"... Chính vì vậy nếu chỉ xét theo tần số xuất hiện của từng từ thì việc phân loại văn bản rất có thể cho kết quả sai dẫn tỷ lệ chính xác sẽ thấp. Vậy giải pháp là gì ?

Phương pháp phổ biến là sử dụng một phương pháp thống kê có tên là **TF-IDF**, 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.

Mà tại sao phương pháp lại có tên là **TF-IDF**.

##### TF(Term Frequency)

là tần số xuất hiện của 1 từ trong 1 văn bản có cách tính như sau:

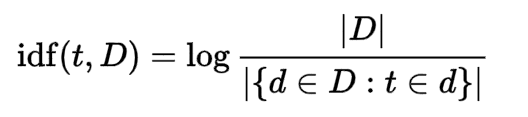


* **f(t,d)** - số lần xuất hiện từ t trong văn bản d.
* mẫu số là tổng số từ trong văn bản d.

##### IDF (Inverse Document Frequency)

Tần số nghịch của 1 từ trong tập văn bản (corpus).

Mục đích của việc tính IDF là giảm giá trị của các từ thường xuyên xuất hiện như "is", "the"... Do các từ này không mang nhiều ý nghĩa trong việc phân loại văn bản.



\* \*\*|D|\*\*: tổng số văn bản trong tập D \* mẫu số là số văn bản có chứa từ t . Nếu từ đó không xuất hiện ở bất cứ 1 văn bản nào trong tập thì mẫu số sẽ bằng 0 => phép chia cho không không hợp lệ, vì thế với trường hợp này thường cộng thêm 1 vào mẫu số.

##### TF-IDF



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 khoá của văn bản đó).

#### Count Vectorizer

Chuyển đổi văn bản thành các vector để biểu diễn cho mô hình học máy có thể hiểu được.

tất cả các văn bản từ giờ sẽ được biểu diễn dưới dạng 1 vector có nhiều thuộc tính, 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.

Ví dụ bag of word của chúng ta chỉ có 3 từ:

Tôi, yêu, Việt\_Nam

Khi đó mọi văn bản đều biểu diễn dưới dạng 1 vector có 3 thuộc tính. Với văn bản

Hoàng Anh là tôi

Ta sẽ có vector biểu diễn cho văn bản trên là

[1, 0, 0]

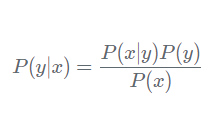
Từ "TÔI" có xuất hiện trong văn bản nên vị trí đầu tiên của vector là 1, từ "YÊU" và từ "VIỆT\_NAM" đều không có nên sẽ là 0.

### Navie Bayes Classification

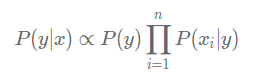
#### Khái niệm

Naive Bayes Classification (NBC) là một thuật toán phân loại dựa trên tính toán xác suất áp dụng định lý Bayes. Thuật toán này thuộc nhóm Supervised Learning (Học có giám sát).

Theo định lý Bayes, ta có công thức tính xác suất như sau:



Do đó ta có:



Công thức chỉ ra xác suất của X xảy ra nếu Y cũng xảy ra, ta viết là **P(X|Y)**. Và nếu ta biết xác suất của Y xảy ra khi biết X, ta viết là **P(Y|X)** cũng như xác suất độc lập của X và Y.

* P(X|Y) là “xác suất của X khi biết Y”
* P(X) là xác suất xảy ra của X
* P(Y|X) là “xác suất của Y khi biết X”
* P(Y) là xác suất xảy ra của Y

Ví dụ, P(lửa) là xác suất có lửa, P(khói) là xác suất ta nhìn thấy khói. Ta sẽ có những trường hợp sau:

P (Lửa | Khói) có nghĩa là tần suất có lửa khi chúng ta nhìn thấy khói. P (Khói | Lửa) có nghĩa là chúng ta thường thấy khói khi có lửa.

Công thức sẽ cho chúng ta biết được điều gì xảy ra tiếp theo nếu ta đã biết một điều.

Ví dụ: Một đám cháy nguy hiểm là có xác suất là 1% nhưng khói lại khá phổ biến là 10% (từ các nhà máy) và 90% đám cháy nguy hiểm tạo ra khói. Vậy ta có:

P (Lửa | Khói) = P (Lửa) P (Khói | Lửa) = 1% x 90% = 9% P (Khói) 10%

Trong trường hợp này, 9% khả năng thấy khói có nghĩa là có một đám cháy nguy hiểm.

#### Phân loại Naive Bayes

Naive Bayes là một thuật toán phân loại cho các vấn đề phân loại nhị phân (hai lớp) và đa lớp. Kỹ thuật này dễ hiểu nhất khi được mô tả bằng các giá trị đầu vào nhị phân hoặc phân loại.

Thuật toán Naive Bayes tính xác suất cho các yếu tố, sau đó chọn kết quả với xác suất cao nhất.

Tuy nhiên, ta cần lưu ý giả định của thuật toán Naive Bayes là các yếu tố đầu vào được cho là độc lập với nhau.

Thuật toán này là một thuật toán mạnh mẽ trong các bài toán:

**Dự đoán với thời gian thực**

**Phân loại Text/ Lọc thư rác**

**Hệ thông Recommendation**

Về mặt toán học, ta có thể viết như sau:

Nếu ta có một Class E và các điểm dữ liệu x1,x2,x3, etc.

Đầu tiên ta sẽ phải tính xác suất P(x1| E) , P(x2 | E) …  (xác suất của x1 thuộc class E xảy ra) và sau đó ta sẽ chọn class có xác suất xảy ra x1 cao nhất.

## Triển khai đề tài

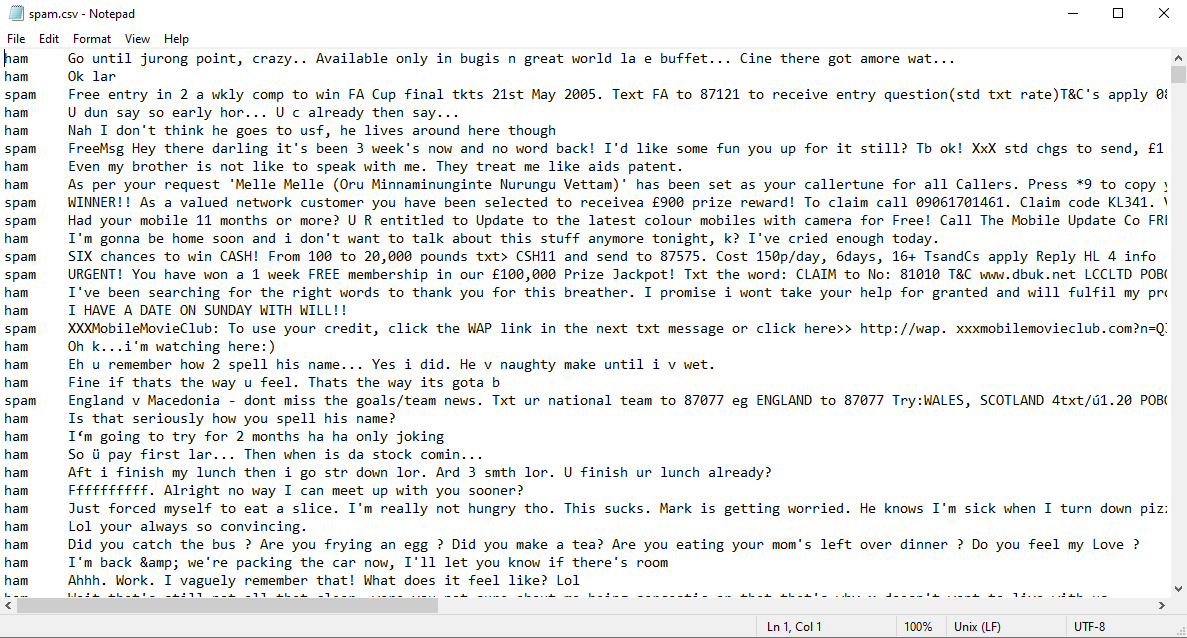
### Tiền xử lí

#### Xử lí dữ liệu

##### Thu thập dữ liệu

Hiện trên mạng, có rất rất nhiều tập dữ liệu cho bài toán lọc thư rác, có thể dễ dàng tìm thấy trên một số trang như Kaggle với keyword "Mail Filter Dataset".Dưới đây là bộ dữ liệu mẫu và được thu thập thêm từ email người dùng và ngôn ngữ của bộ dữ liệu là tiếng anh.

Có tổng cộng 5572 email, trong đó có 4825 email là thư bình thường và 747 email là thư rác và được lưu vào file spam.csv kèm theo nhãn cho từng email (1 là spam , 0 là ham)

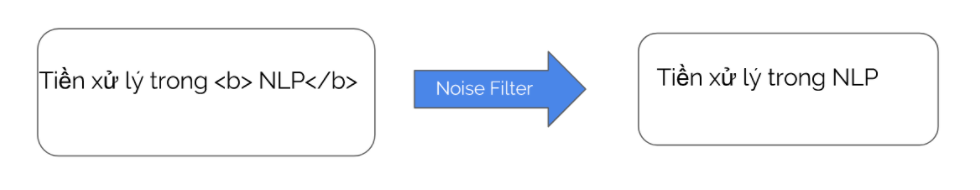


1. File csv

##### Làm sạch dữ liệu

Mục đích bước này là loại bỏ noise trong data của bạn. Đa phần noise là các thẻ HTML, JavaScript, những dấu câu… và đương nhiên nếu cứ để noise để tiến hành xử lý sẽ dẫn đến kết quả xử lý không tốt

Ví dụ:

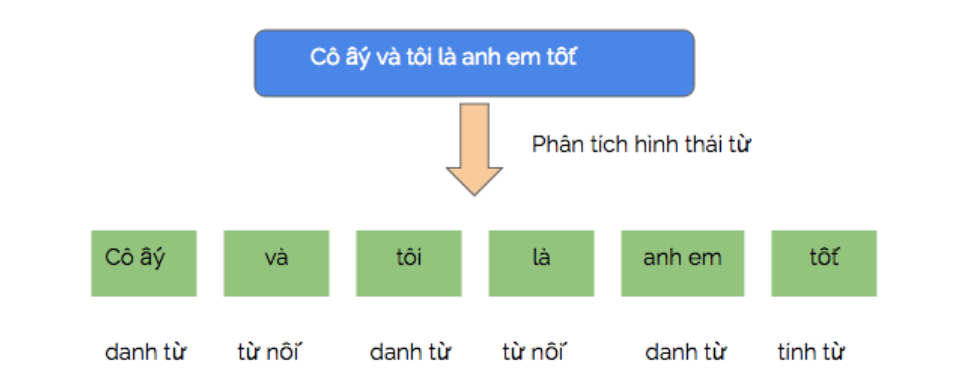


1. Làm sạch dữ liệu

##### Tách từ

Trong tiếng Việt, dấu cách (space) không được sử dụng như 1 kí hiệu phân tách từ, nó chỉ có ý nghĩa phân tách các âm tiết với nhau.Trong tiếng anh cũng vậy, vì thế, để xử lý tiếng anh, công đoạn tách từ (word segmentation) là 1 trong những bài toán cơ bản và quan trọng bậc nhất.

Ví dụ: từ “đất nước” được tạo ra từ 2 âm tiết “đất” và “nước”, cả 2 âm tiết này đều có nghĩa riêng khi đứng độc lập, nhưng khi ghép lại sẽ mang một nghĩa khác. Vì đặc điểm này, bài toán tách từ trở thành 1 bài toán tiền đề cho các ứng dụng xử lý ngôn ngữ tự nhiên khác như phân loại văn bản, tóm tắt văn bản, máy dịch tự động, …



1. Phân tách từ

##### Chuẩn hoá từ

Mục đích là đưa văn bản từ các dạng không đồng nhất về cùng một dạng. Dưới góc độ tối ưu bộ nhớ lưu trữ và tính chính xác cũng rất quan trọng.

Ví dụ: U.S.A = USA

Ví dụ trong từ điển, training data của chúng ta không có U.S.A, chỉ có USA, thì việc convert những từ như U.S.A về USA là điều cần thiết để các bước xử lý sau như text classification, intent detection được chính xác.

Ngoài ra với tiếng Nhật: 猫＝ねこ＝ネコ(full size)＝ﾈｺ(half size)

Có nhiều cách viết, mỗi cách viết khi lưu trữ sẽ tốn lượng memory khác nhau, như half size chỉ tốn 1/2 dung lượng so với full size nên tuỳ theo nhu cầu, tình hình thực tế, chúng ta sẽ đưa văn bản về 1 dạng đồng nhất.

Ngoài ra trong một vài trường hợp, nếu ký tự số không mang lại lợi ích gì thì cũng sẽ tiến hành loại bỏ các ký tự số đó, nếu cứ để nguyên rất có thể các ký tự số sẽ trở thành noise, ảnh hưởng đến tính chính xác của model sau này.

##### Loại bỏ stop words

StopWords là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở tiếng việt StopWords là những từ như: để, này, kia... Tiếng anh là những từ như: is, that, this...

Có rất nhiều cách để loại bỏ StopWords nhưng có 2 cách chính là:

* Dùng từ điển
* Dựa theo tần suất xuất hiện của từ

##### Ghép các từ

Sau khi thực hiện việc tách các từ trong một câu ra để thực hiện các bước xử lý thì bây giờ sẽ ghép lại để tạo lại một câu hoàn chỉnh.

##### Chuyển các văn bản sang các vector

Ở mô hình Naive Bayes sẽ có xuất hiện của các điều kiện. Ở đây cũng tương tự, nhưng với mỗi từ trong Bag Of Words ở trên, sẽ là 1 thuộc tính. Như vậy, tất cả các văn bản từ giờ sẽ được biểu diễn dưới dạng 1 vector có nhiều thuộc tính, 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.

#### Code hoàn chỉnh

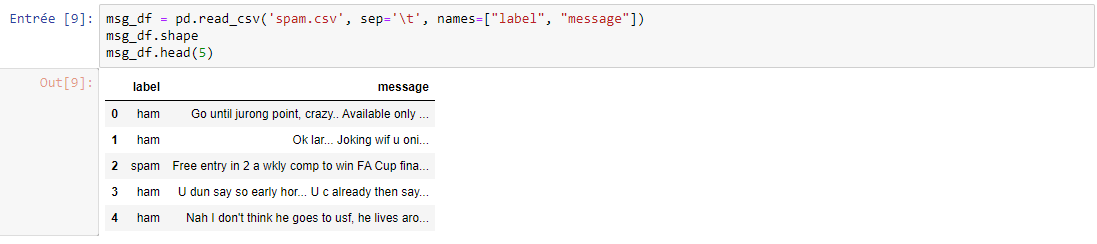
##### Đọc dữ liệu

Import các thư viện cần dùng



1. Đọc file

##### In dữ liệu



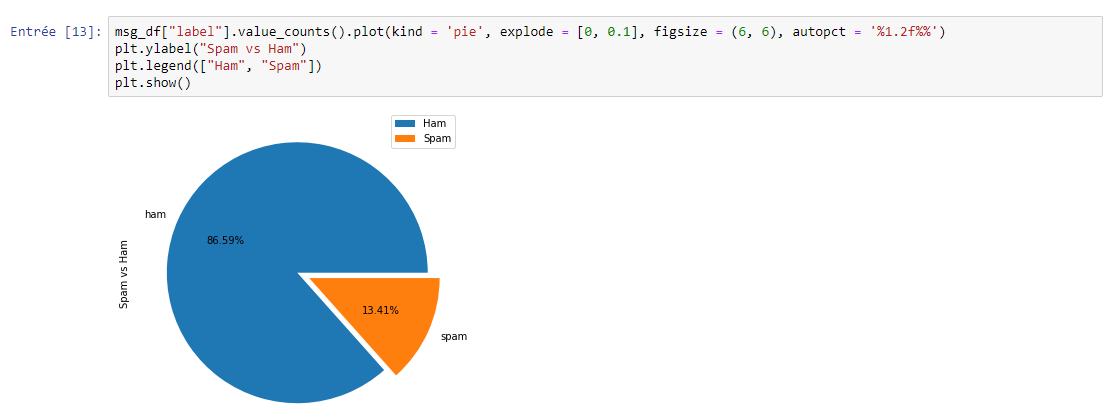
1. In file

##### Đếm số lượng email dựa theo label



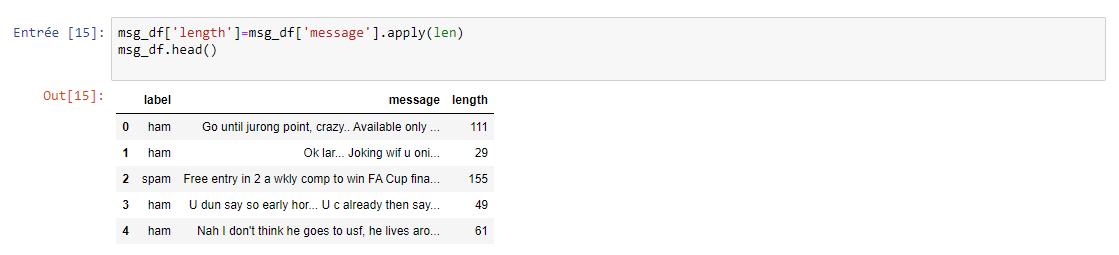
1. Đếm email

##### Biểu diễn bằng biểu đồ



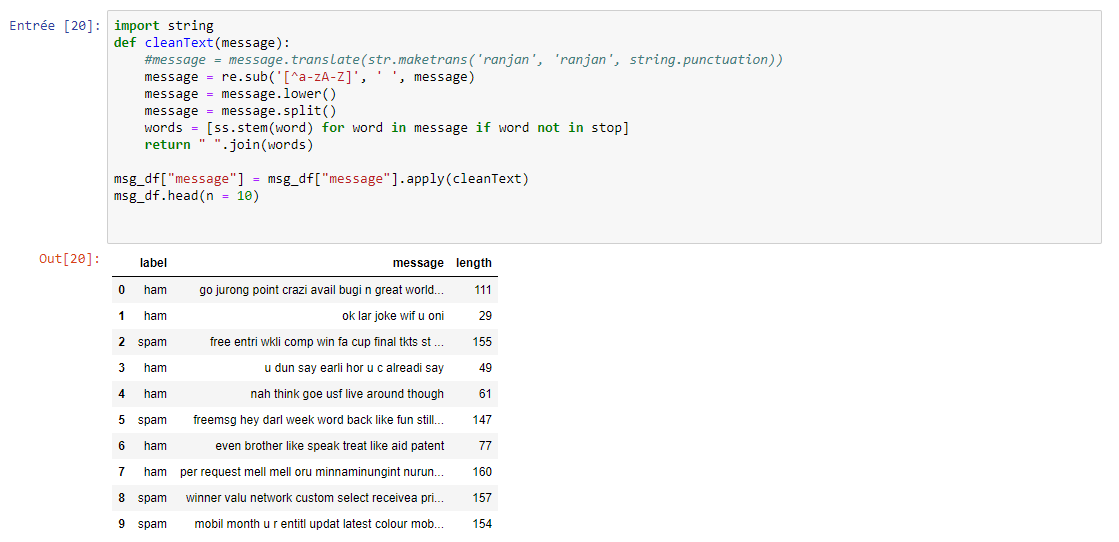
1. Biểu đồ email

##### Đếm chiều dài của email



1. Đếm chiều dài của email

##### Chuẩn hóa văn bản



1. Chuẩn hóa

##### Loại bỏ stopwords dựa theo từ điển



1. Loại bỏ stopwords

##### Biểu diễn các từ là spam



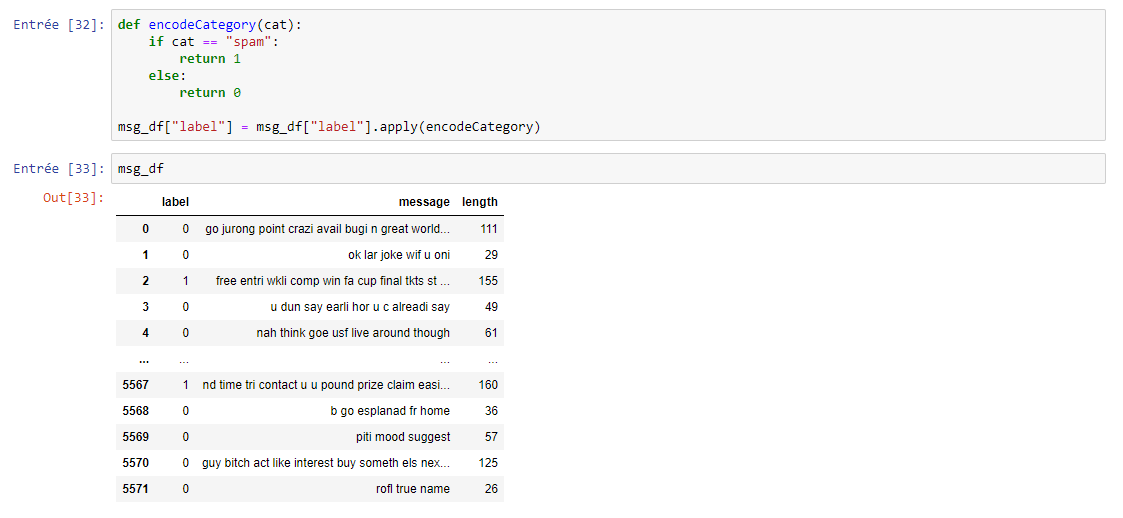
1. Spam words

##### Biểu diễn các từ là ham



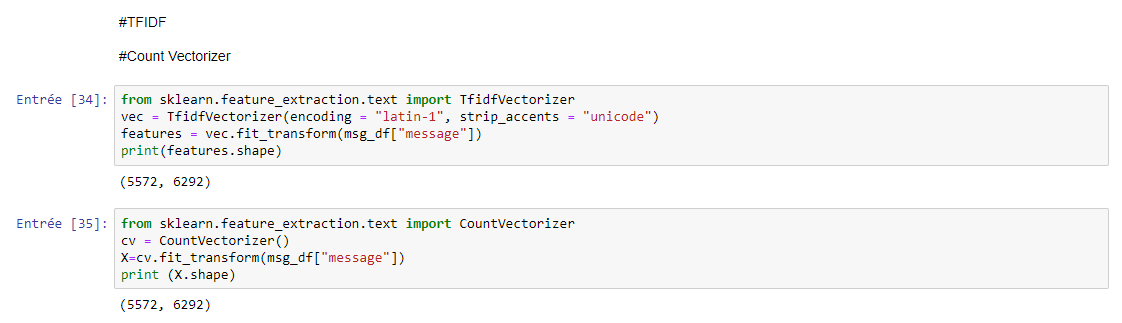
1. Ham words

##### Đánh nhãn cho email

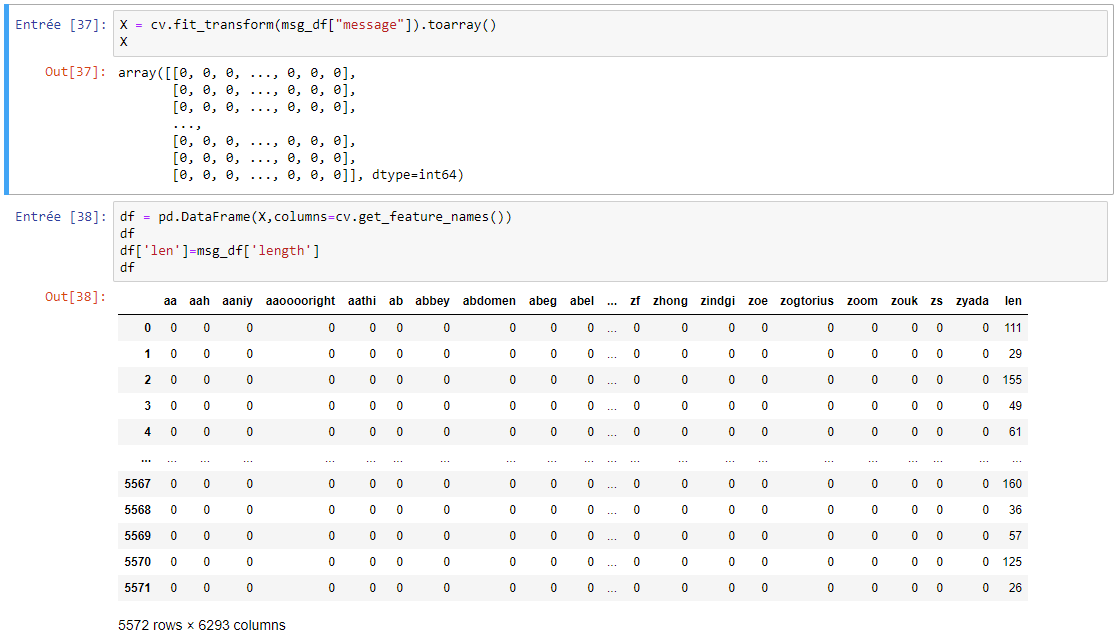


1. Đánh nhãn

##### Chuyển đổi từ thành vector



1. Chuyển đổi



1. Chuyển đổi thành vector

### Huấn luyện mô hình

#### Chia tập dữ liệu

Dữ liệu được chia thành 2 phần là train và test

Train: 80%

Test: 20%

#### Đánh giá mô hình

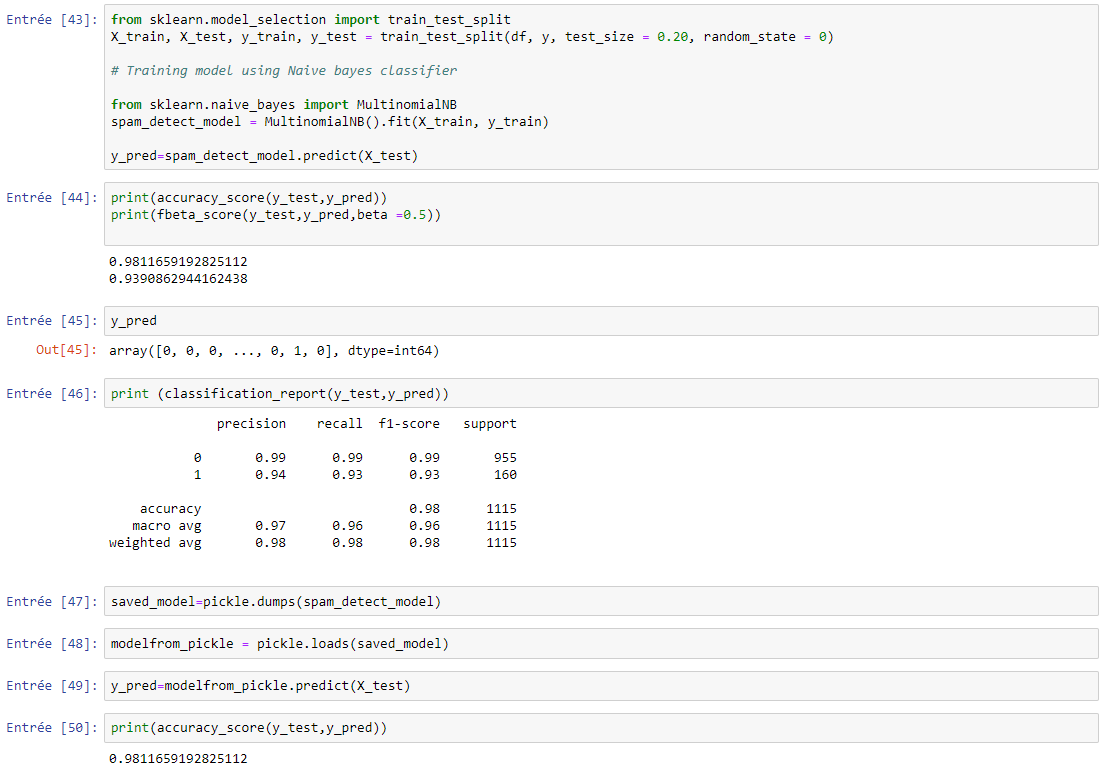
Dùng giải thuật Naive bayes classifier để đánh giá mô hình có đạt yêu cầu hay không.

#### Lưu mô hình

Thực hiện lưu mô hình để hỗ trợ cho việc triển khai lên website.

#### Code hoàn chỉnh

##### Naïve bayes



1. Kết quả huấn luyện

##### Lưu mô hình

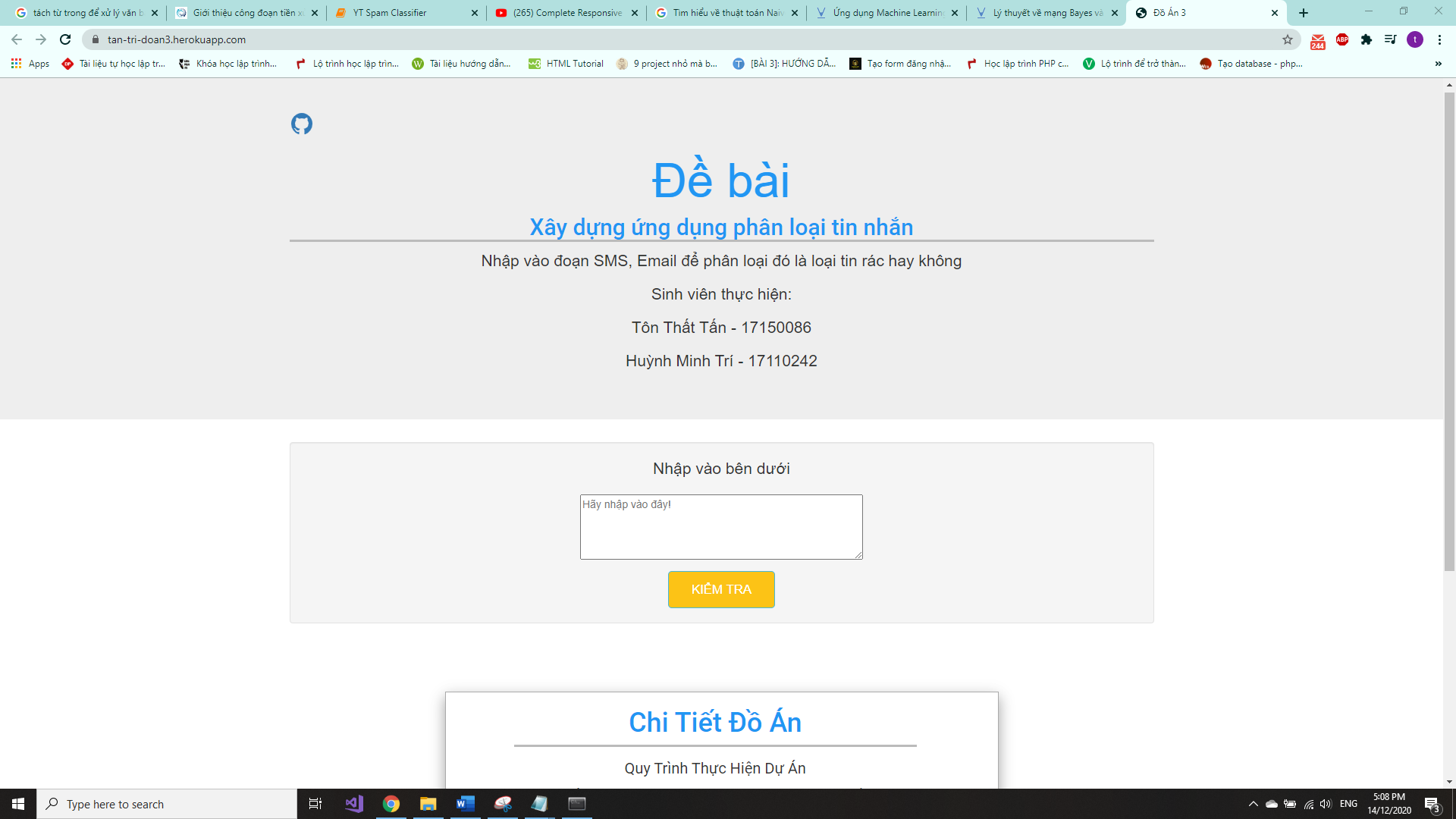


1. Lưu mô hình

### Xây dựng Flask

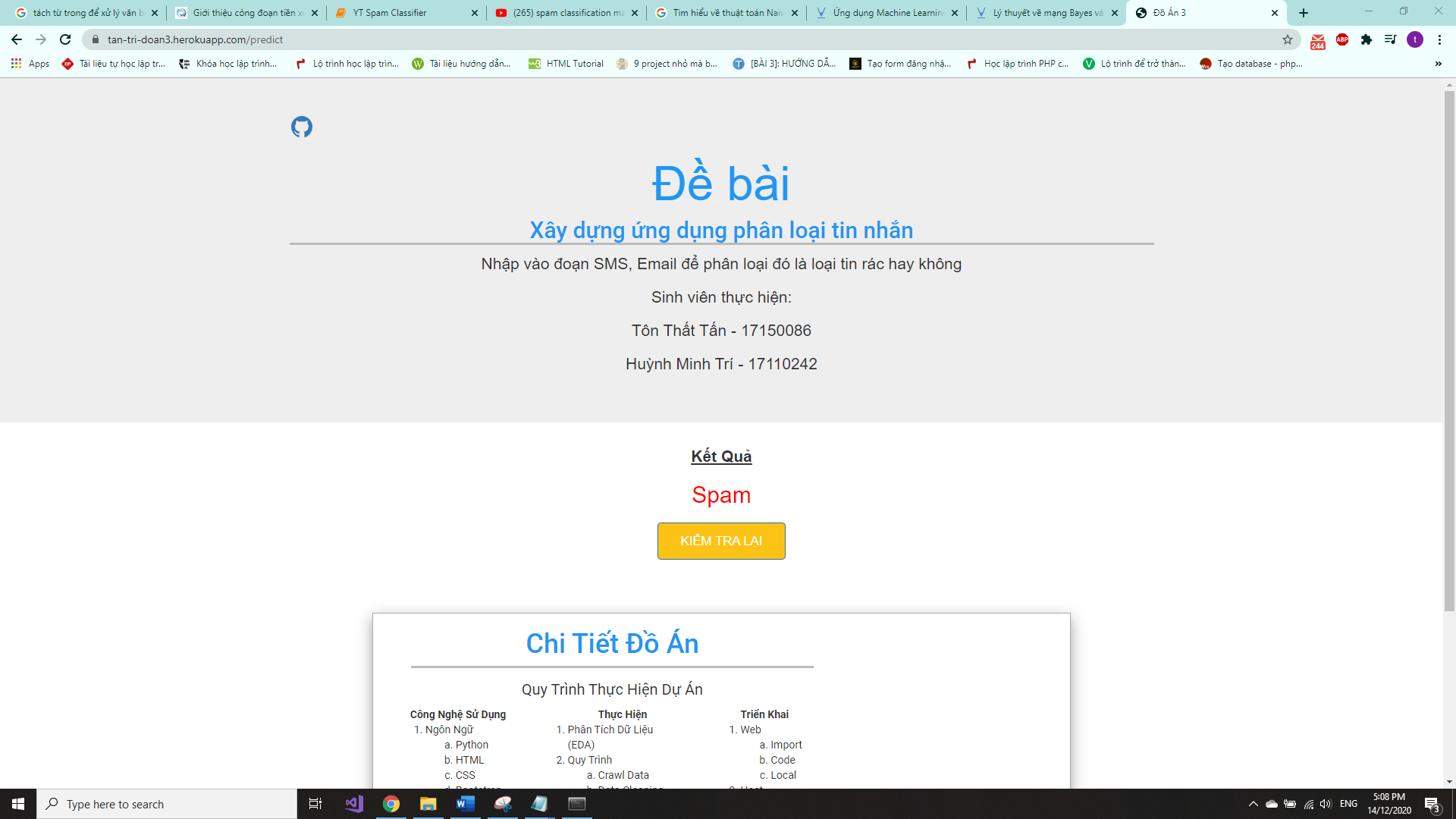
#### Xây dựng giao diện

##### Trang chủ



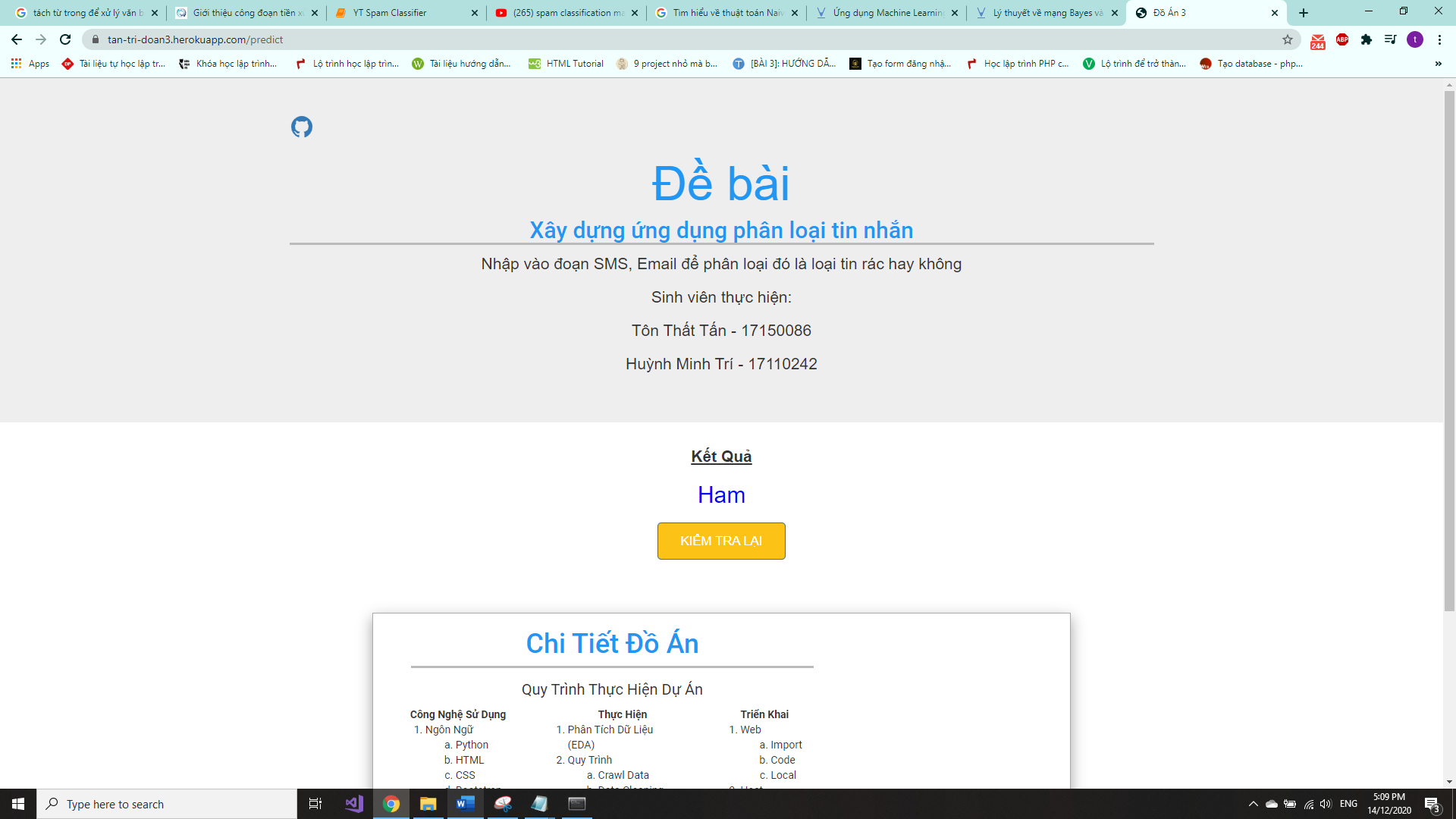
1. Demo trang chủ

##### Kết quả là spam



1. Demo spam

##### Kết quả là ham



1. Demo ham

##### Nhúng mô hình vào website

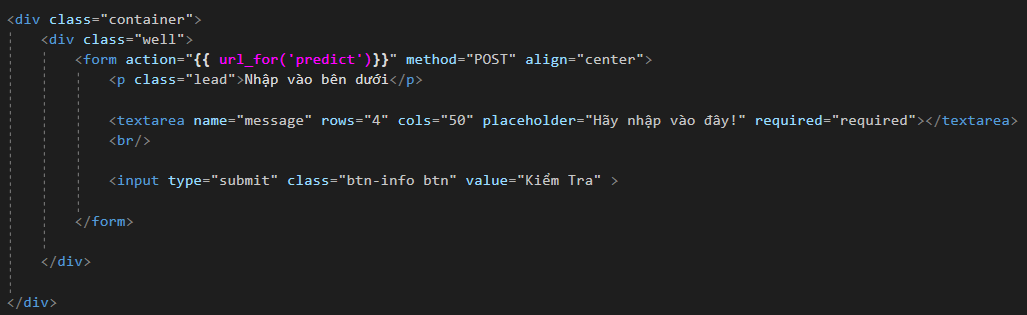
Sử dụng framework Flask để đưa code từ python lên website.

Import các file vào app.py sau khi export từ việc đánh giá mô hình.

#### Code Hoàn chỉnh

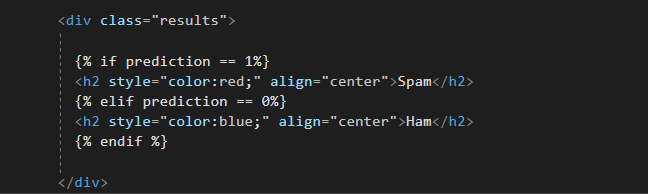
##### Code giao diện

Yêu cầu người dùng nhập



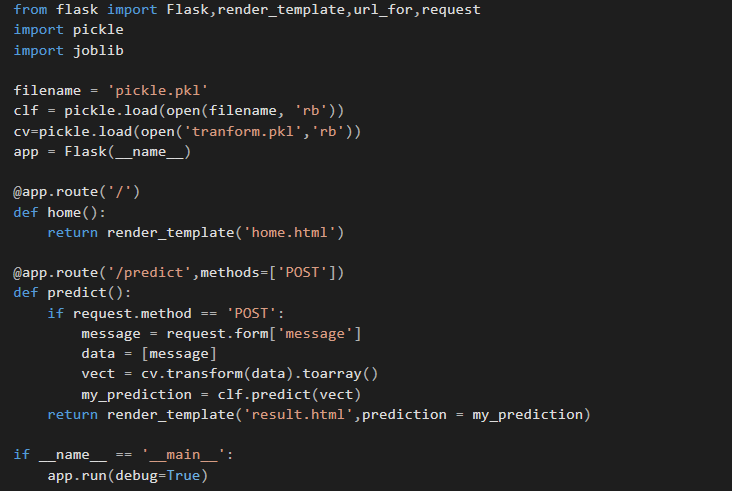
1. Code giao diện

##### Code kết quả



1. Code kết quả

##### Code xử lý mô hình



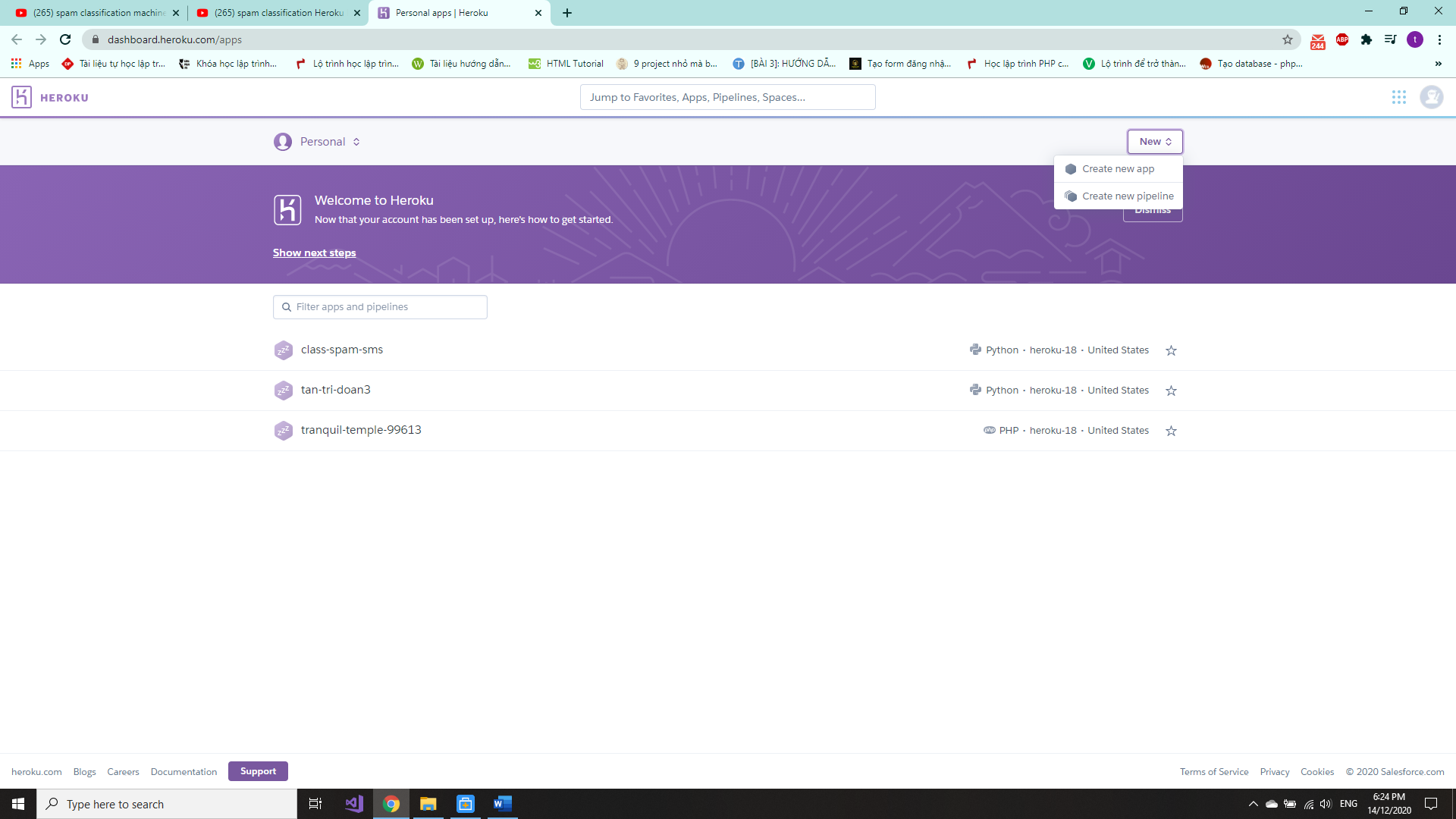
1. Code xử lí flask

### Đưa lên host

#### Đăng ký Heroku

Tạo project trên Heroku

Sử dụng Heroku để làm host cho website

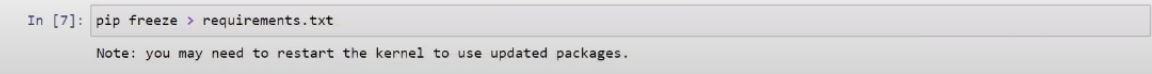


1. Tạo app

#### Cấu hình

##### Tạo requirements

Dùng lệnh pip freeze > requirements.txt để tạo file chứa các thư viện của huấn luyện mô hình.

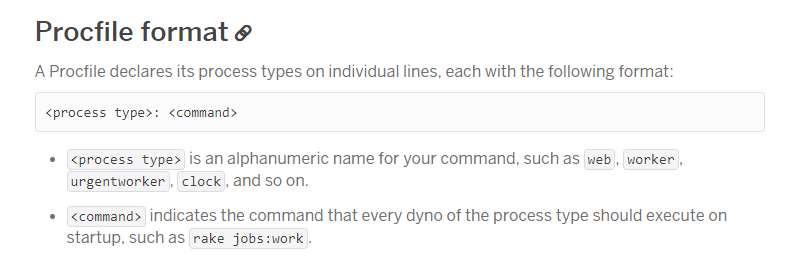


1. Cấu hình requirements

##### Tạo procfile

Tạo file Proc tại Heroku/profile để cấu hình website trên Heroku theo định dạng sau

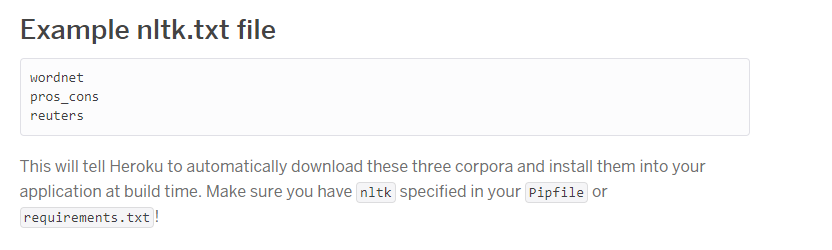
web gunicorn app:app



1. Tạo procfile

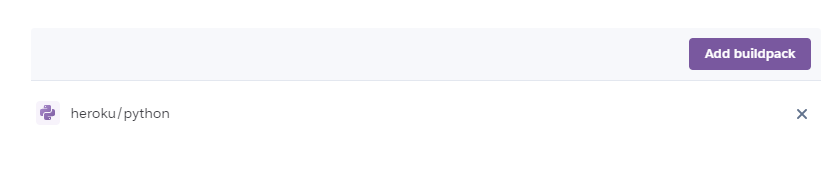
##### Tạo nltk.txt

Tạo file nltk tại Heroku/python-nltk để không xảy ra lỗi khi upload code.



1. Tạo nltk.txt

##### Tạo buildpack

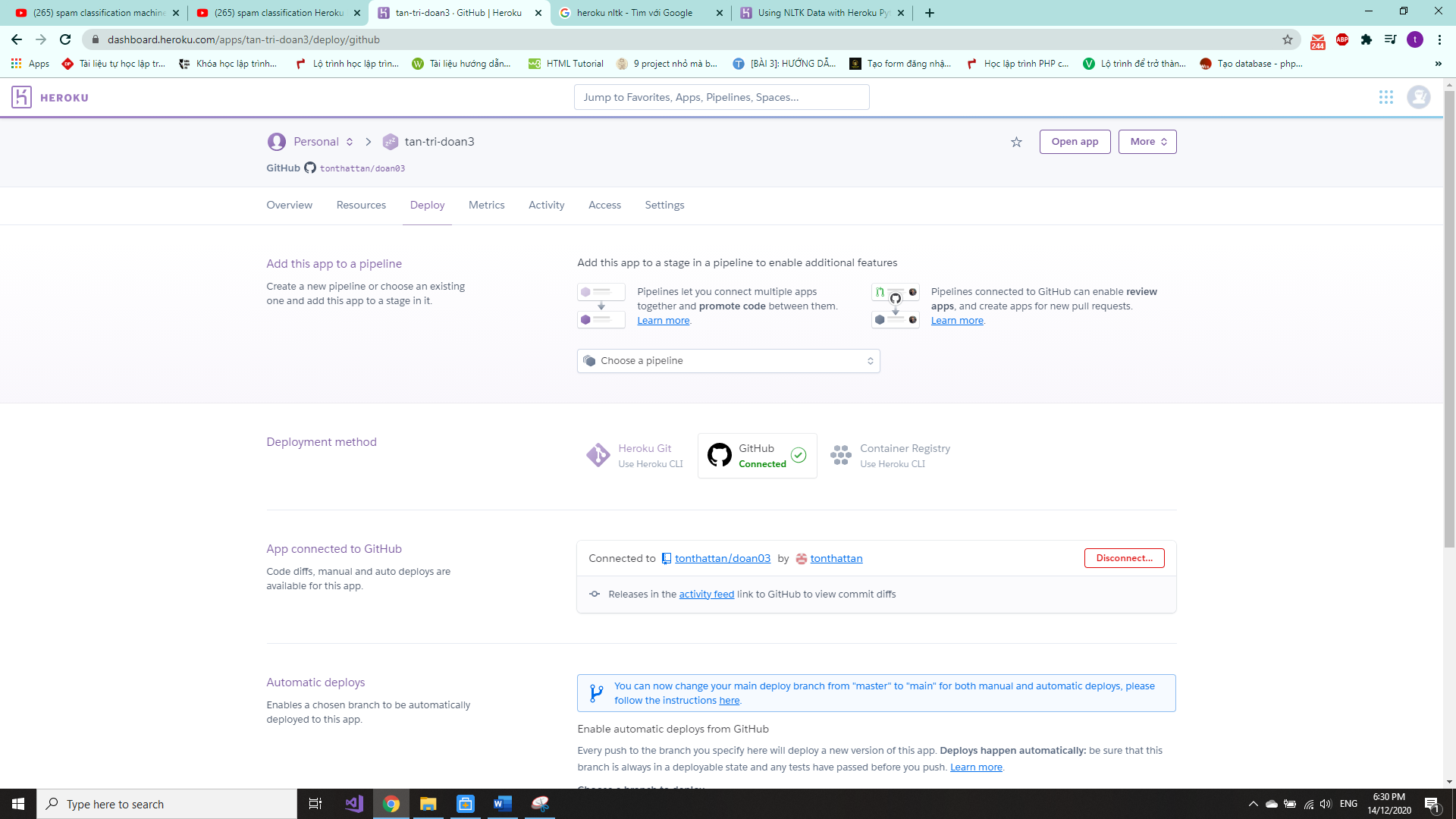


1. Add buildpack

#### Upload

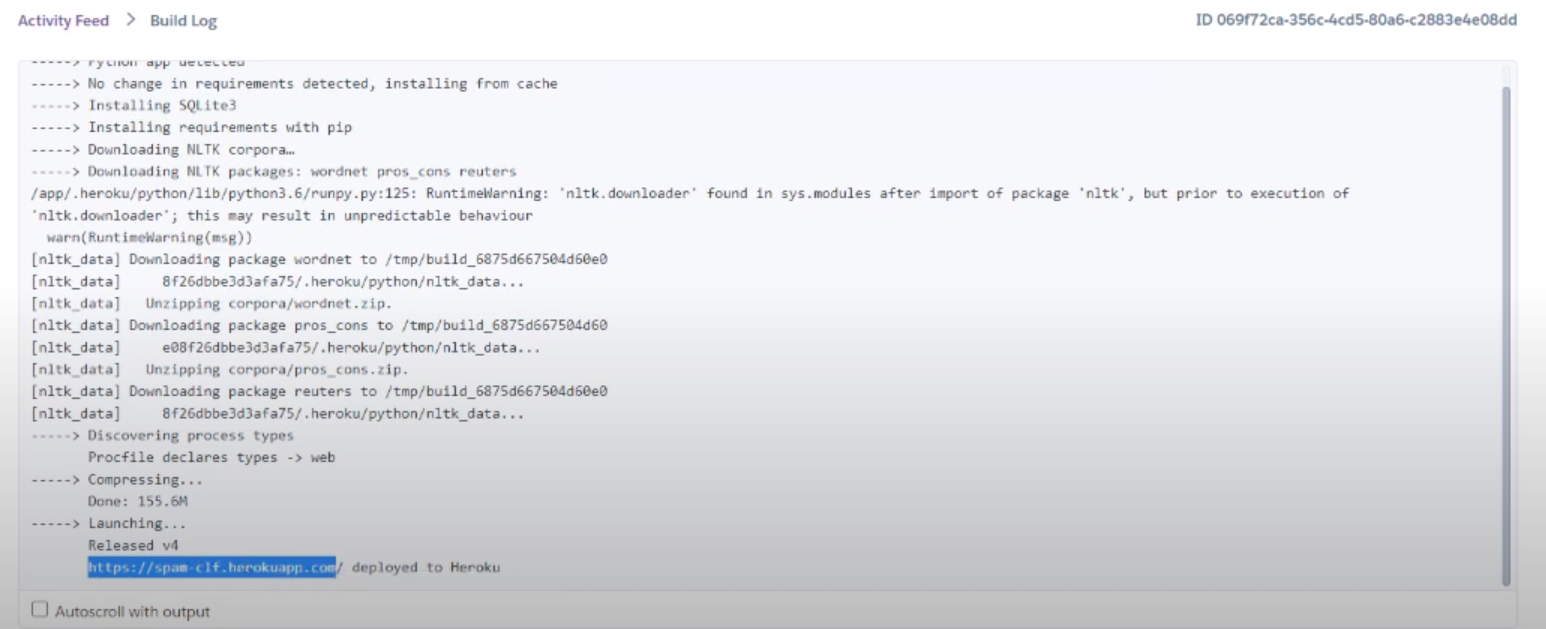
Tải toàn bộ source code của bài lên Github hoặc tải trực tiếp lên Heroku bằng CLI

Sau đó connect link github tới App vừa mới tạo



1. Setup heroku

Quá trình upload thành công



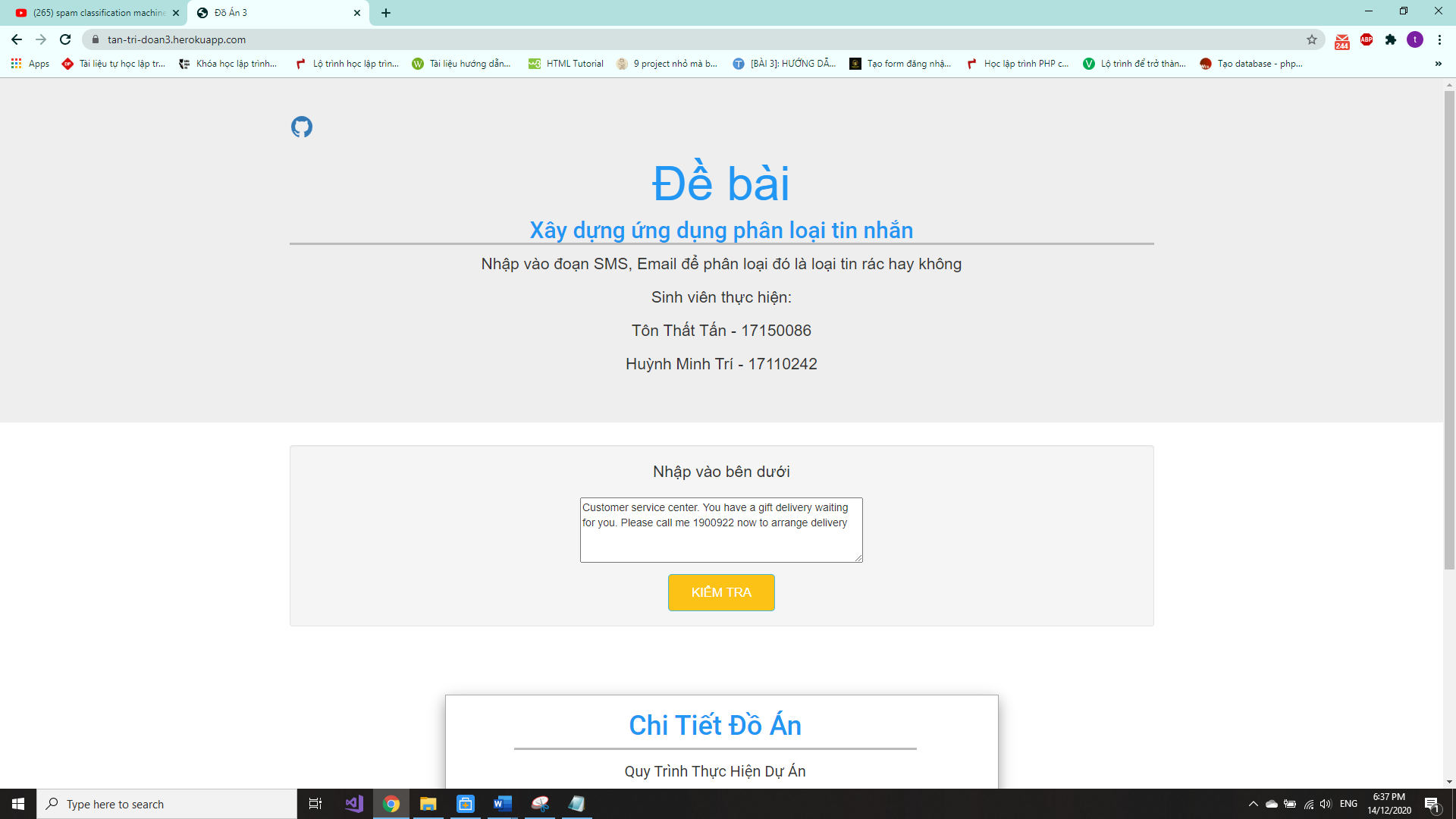
1. Upload thành công

## Kết quả

### Demo sản phẩm

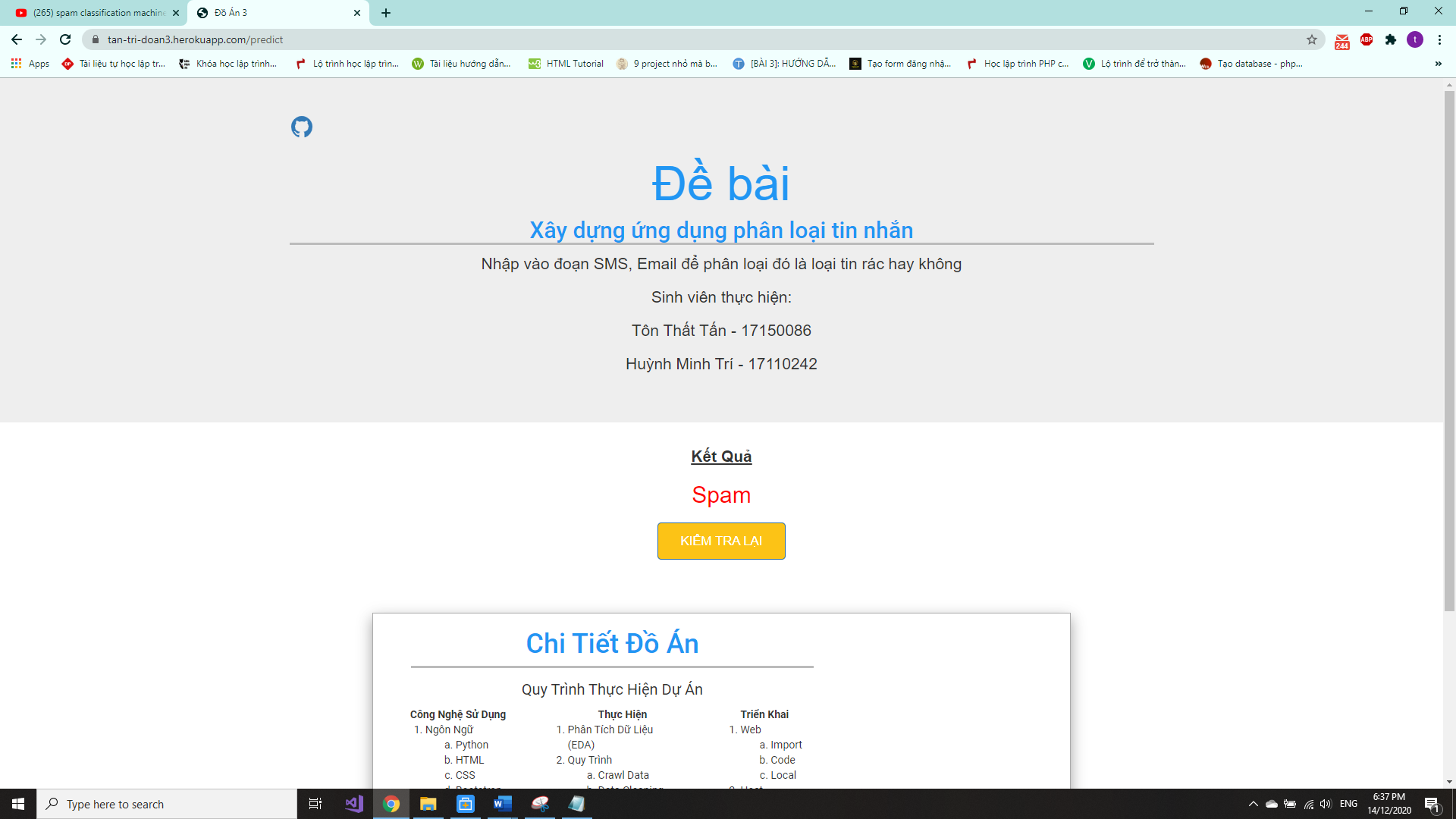
Sau khi upload hoàn tất , khởi động app lên và kiểm tra sản phẩm.

#### Kiểm tra là spam



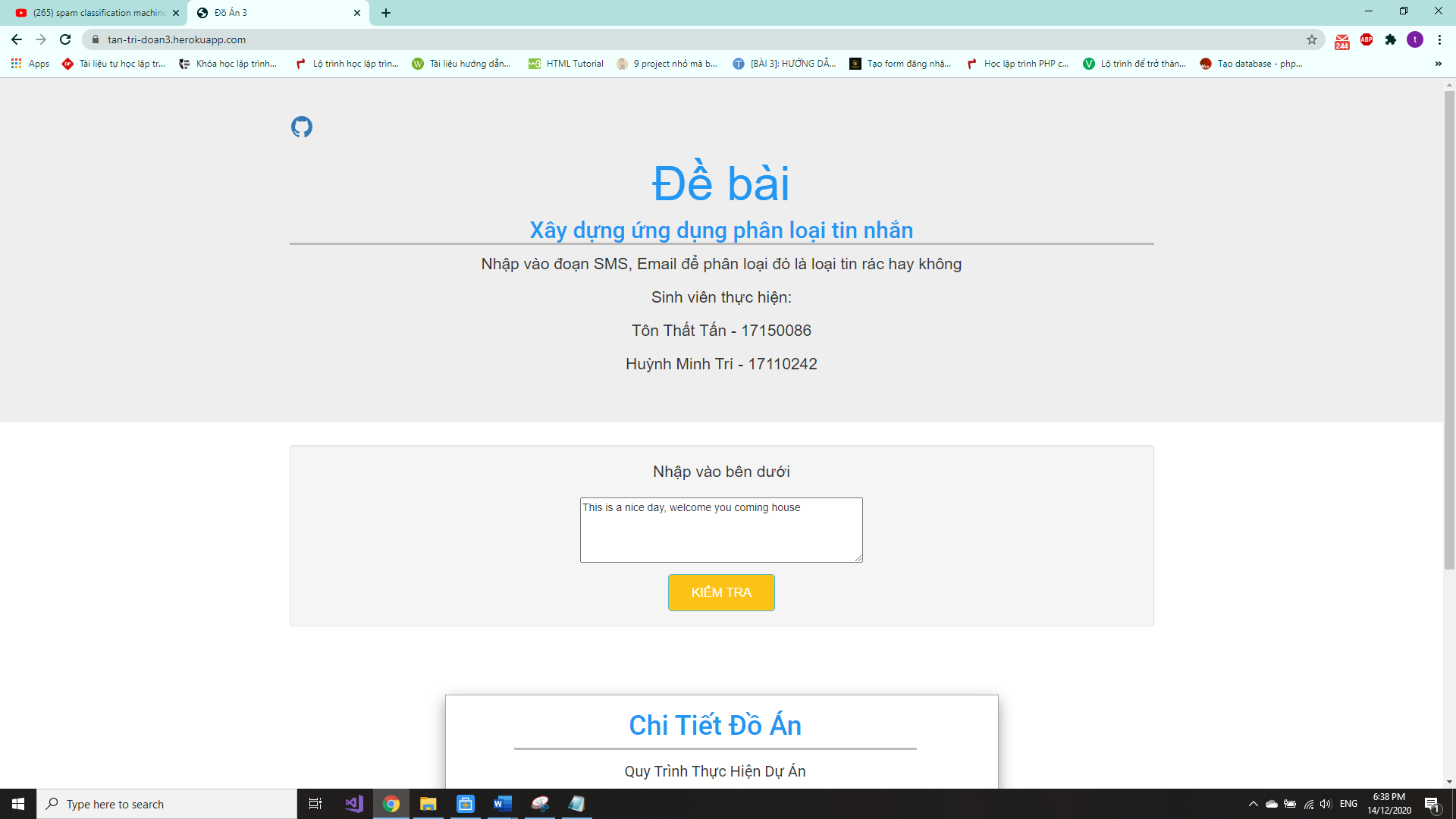
1. Demo spam

##### Kết quả



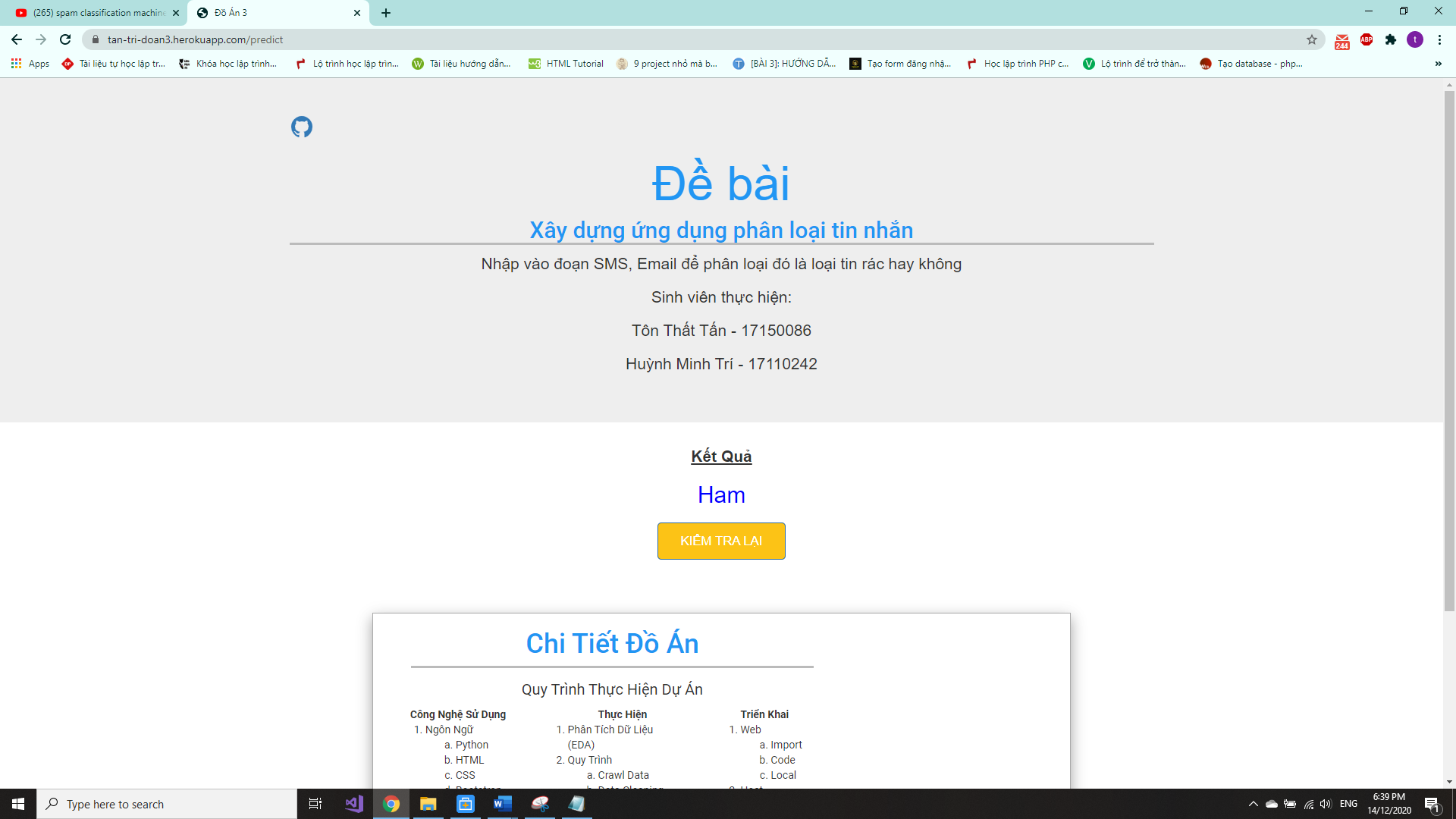
1. Kết quả spam

#### Kiểm tra là ham



1. Demo ham

##### Kết quả



1. Kết quả ham

### Link demo sản phẩm

<https://tan-tri-doan3.herokuapp.com/>

# kết luận

1. Đánh giá

Sau khi trải qua các bước làm đồ án, nhóm chúng em đã thu được kết quả như mong đợi

Từ việc thu thập được nhiều dữ liệu từ nhiều nguồn khác nhau, đảm bảo tính khách quan, đa dạng về dữ liệu mẫu.

Tiếp theo là việc xử lý dữ liệu sau khi thu thập, làm cho dữ liệu trở nên sạch, không còn bị nhiễu, triển khai được mô hình Bag of Words theo yêu cầu đề ra.

Kế đến là việc dùng giải thuật để đánh giá mô hình, cho thấy việc dùng giải thuật Naïve Nayes là đúng, đạt được hiệu quả cao trong vấn đề phân loại.

Cuối cùng là việc triển khai mô hình lên website, tạo ra được ứng dụng đi vào thực tế, đáp ứng được nhu cầu người dùng.

1. Ưu điểm

Tạo ra được sản phẩm đáp ứng được nhu cầu sử dụng hiện nay khi có quá nhiều thư rác làm gây nhiễu, gây hại cho người dùng internet.

Sản phẩm được đưa lên host, giúp bổ sung thêm kiến thức cho nhóm em về việc triển khai lên server.

1. Nhược điểm

Phải phụ thuộc nhiều vào dữ liệu mẫu.

Tốn thời gian, chi phí cho việc thu gom dữ liệu.

Chỉ mới áp dụng được việc phân loại cho email là tiếng anh, chưa có làm được cho email là tiếng việt.

Chưa sử dụng được nhiều thuật toán, mô hình khác nhau để đưa ra so sánh cái nào là tốt nhất.

# Source code và demo

1. Lưu trữ source code

<https://github.com/tonthattan/tan-tri-doan03>

1. Link demo sản phẩm

<https://tan-tri-doan3.herokuapp.com/>

# Tài liệu tham khảo

[1] <https://viblo.asia/p/ly-thuyet-ve-mang-bayes-va-ung-dung-vao-bai-toan-loc-thu-rac-07LKXzkelV4>

[2] <https://www.kaggle.com/balakishan77/spam-or-ham-email-classification/output>

[3] <https://www.youtube.com/channel/UCT1Me5Pe2a-5a0GVSEeb04g>

[4] <https://www.youtube.com/watch?v=inp6h0HDN74>

[5] <https://devcenter.heroku.com/>

[6] <https://kipalog.com/posts/Machine-Learning---NLP--Text-Classification-su-dung-scikit-learn----python>

[7] <https://viblo.asia/p/ung-dung-machine-learning-de-phan-loai-spam-sms-07LKXwg85V4>

[8] <https://codetudau.com/bag-of-words-tf-idf-xu-ly-ngon-ngu-tu-nhien/index.html>