**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP.HCM**

**KHOA: CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN MÔN HỌC**

**MÔN: KHAI THÁC DỮ LIỆU**

**ĐỀ TÀI : ÁP DỤNG THUẬT TOÁN BAYES, LỌC THƯ RÁC.**

Giảng viên hướng dẫn: Ngô Dương Hà

Lớp: 09DHTH3

# LỜI NÓI ĐẦU

Em xin chân thành cảm ơn thầy Ngô Dương Hà đã tận tâm giúp đỡ chúng em có thêm kiến thức về bộ môn Khai Thác Dữ Liệu. Cảm ơn nhà trường đã tạo điều kiện học tập giúp chúng em có thể thực hành và hiểu sâu hơn về bộ môn. Nay em làm bài báo cáo này nhằm mô tả và tích hợp thông tin mà chúng em hiểu cách để áp dụng thuật toán bayes nhằm khai thác dữ liệu.

**DANH SÁCH NHÓM**

|  |  |  |
| --- | --- | --- |
| **Họ tên** | **MSSV** | **Email** |
| Tô Đình Nhân | 2001181235 | Todinhnhan2004@gmail.com |

MỤC LỤC

[LỜI NÓI ĐẦU 1](#_Toc59085618)

[I. Một số khái niệm cơ bản 5](#_Toc59085619)

[1. Khái niệm về khai thác dữ liệu 5](#_Toc59085620)

[1.1. Khai thác dữ liệu là gì? 5](#_Toc59085621)

[1.2. Tầm quan trọng 5](#_Toc59085622)

[2. Các Phương Pháp Chính Trong Khai Thác Dữ Liệu 6](#_Toc59085624)

[2.1. Phương Pháp 6](#_Toc59085625)

[2.2. Tổng Quan Về Định Lý Bayes 6](#_Toc59085626)

[II. Đặc Điểm Của Định Lý Bayes 7](#_Toc59085630)

[1. Ý tưởng 7](#_Toc59085631)

[2. Ưu và Nhược Điểm 8](#_Toc59085632)

[III. Viết ứng dụng minh họa 10](#_Toc59085634)

[DANH MỤC TÀI LIỆU THAM KHẢO 23](#_Toc59085643)

# I. Một số khái niệm cơ bản

### 1.1. Khái thác dữ liệu là gì ?

**Khai thác dữ liệu**  hay còn được gọi là Data Mining. Data Minng được tách thành 2 từ “Data” và “Mining” để phân tích. “Data” hiểu nôm na là nguồn dữ liệu mà mỗi công ty,tổ chức thu thập được đa dạng ,từ các nguồn khác nhau. “Mining” nghĩa là khai phá , Data Mining chính là quá trình đi sâu vào bộ dữ liệu để phân tích và tìm kiếm các chi tiết,giá trị ẩn bên trong từng dữ liệu, cụ thể là muốn xác định, muốn biết xem dữ liệu cung cấp thông tin gi, thông tin đó có ích không khi chúng ta phải tiến hành chương trình.

### 1.2. Tầm quan trọng :

**Khai thác dữ liệu** chính là trích xuất thông tin từ các bộ dữ liệu khổng lồ , là quá trình khai thác, tiếp thu kiến thức từ dữ liệu . Chính vì thế Khai thác dữ liệu được ứng dụng vào rất nhiều lĩnh vực khác nhau.

Như chúng ta đã biết , Dữ liệu lớn (Big Data) và công nghệ ứng dụng dữ liệu lớn đang chi phối rất nhiều lĩnh vực trong đời sống của chúng ta, không chỉ với riêng các công ty về công nghệ như Google, Facebook, Youtube mà còn lan sang các lĩnh vực quan trọng khác như quốc phòng, y tế, giáo dục, kinh tế, chính trị v…v..Tất cả thông tin chúng ta cung cấp trong mọi lĩnh vực đời song đều được ghi lại và trở thành một phần của dữ liệu lớn, sau đó được các thuật toán sử dụng để đưa ra những dự đoán phù hợp một cách “thần kỳ” đối với mỗi cá nhân. Ví dụ , dựa trên việc kết hợp hàng trăm tỷ từ khóa tìm kiếm , trong thời gian cực kỳ ngắn, Google có thể xác định được sự lây lan của bệnh cúm chính xác như dữ liệu chính thức dựa trên việc bệnh nhân thực sự tới gặp bác sĩ.

Với tầm quan trọng của dữ liệu lớn trong mọi mặt của đời sống xã hội như thế, các ngành học liên quan đến kỹ thuật xử lý các dữ liệu này như công nghệ thông tin, khoa học dữ liệu, khoa học máy tính …. Đang ngày càng trở nên những ngành “HOT” trong thời đại công nghệ số , sức hấp dẫn tăng cao do nhu cầu lao động chất lượng cao của xã hội. Một trong các kỹ năng quan trọng mà các chuyên ngành này đều phải học qua đó là Data Mining (Khai thác dữ liệu).

## 2.1. Phương Pháp

* Phân loại (Classification): Là phương pháp dự báo, cho phép phân loại một đối tượng vào một hoặc một số lớp cho trước.
* Hồi qui (Regression):  Khám phá chức năng học dự đoán, ánh xạ một mục dữ liệu thành biến dự đoán giá trị thực.
* Phân nhóm (Clustering):  Một nhiệm vụ mô tả phổ biến trong đó người ta tìm cách xác định một tập hợp hữu hạn các cụm để mô tả dữ liệu.
* Tổng hợp (Summarization): Một nhiệm vụ mô tả bổ sung liên quan đến phương pháp cho việc tìm kiếm một mô tả nhỏ gọn cho một bộ (hoặc tập hợp con) của dữ liệu.
* Mô hình ràng buộc (Dependency modeling): Tìm mô hình cục bộ mô tả các phụ thuộc đáng kể giữa các biến hoặc giữa các giá trị của một tính năng trong tập dữ liệu hoặc trong một phần của tập dữ liệu.
* Dò tìm biến đổi và độ lệch (Change and Deviation Dectection): Khám phá những thay đổi quan trọng nhất trong bộ dữ liệu.

***2.2. Tổng Quan Về Định Lý Bayes***

là phân lớp dựa trên thống kê thực

hiện việc dự đoán xác suất một mẫu sẽ thuộc về lớp nào

dựa trên giá trị của các thuộc tính biết trước.

Hay nói 1 cách khác Định luật bayes là sự liên hệ giữa các xác suất có điều kiện . Gợi ý cho chúng ta rằng có thể tính toán 1 xác suất chưa biết dựa vào các xác suất có điều kiện khác.

Thuật toán Bayes là thuật toán thuộc nhóm Supervised Learning (Học có giám sát).

# II. Đặc điểm của Bayes

## Ý Tưởng

P(Ci|X) =

• Gọi **X** là mẫu dữ liệu chưa biết nhãn.

• Ci là giả thuyết **X** thuộc về phân lớp **Ci**.

• Việc phân lớp là quá trình xác định **P(Ci|X)**, xác suất mà giả thuyết đúng với mẫu dữ liệu **X** cho trước.

• **P(Ci)** là xác suất có thể ước lượng từ dữ liệu huấn luyện.

• **P(X)** là xác suất mẫu dữ liệu được quan sát.

• **P(X|Ci)** là khả năng quan sát mẫu **X** khi cho trước giả thuyết về phân lớp.

Dự đoán X thuộc về lớp Ci khi và chỉ khi P(Ci | X) là cao nhất trong số P(Cm | X) của tất cả m lớp

Do P(X) là hằng số cho mọi lớp nên chỉ cần tìm cực đại của P(Ci|X) =

• Naives Bayes giả sử giá trị của mọi thuộc tính là độc lập nên :

**1.2 Làm Trơn Laplace Smooth**

\_Theo công thức như trên thì chúng ta thấy mọi xác suất điều kiện phải khác 0 nhưng trong thực tế thì khó có thể đạt được, Vậy nên ta sử dụng hàm Laplace để khử dữ liệu 0 bằng cách :

Bổ sung mỗi trường hợp (Ci) một dữ liệu ảo

Tổng quát : P(Ci) =

P(xk|Ci) =

Trong đó : m = số lớp ; r = số giá trị của thuộc tính

**Kết Luận**

Vậy ta có thể thấy:

+ Trong mô hình Naive Bayes, có hai giả thiết được đặt ra:

* 1. Các đặc trưng đưa vào mô hình là độc lập với nhau. Tức là sự thay đổi giá trị của một đặc trưng không ảnh hưởng đến các đặc trưng còn lại.
  2. Các đặc trưng đưa vào mô hình có ảnh hưởng ngang nhau đối với đầu ra mục tiêu.

Chính vì hai giả thiết gần như không tồn tại trong thực tế trên, mô hình này mới được gọi là naive (ngây thơ). Tuy nhiên, chính sự đơn giản của nó với việc dự đoán rất nhanh kết quả đầu ra khiến nó được sử dụng rất nhiều trong thực tế trên những bộ dữ liệu lớn, đem lại kết quả khả quan. Một vài ứng dụng của Naive Bayes có thể kể đến như: lọc thư rác, phân loại văn bản, dự đoán sắc thái văn bản, ...

## 2. Ưu và nhược điểm

**+ Ưu Điểm:**

1. Giả định độc lập: hoạt động tốt cho nhiều bài toán/miền sữ liệu và ứng dụng.  
   Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán như phân lớp văn bản, lọc spam,..
2. Cho phép kết hợp tri thức tiền nghiệm (prior knowledge) và dữ liệu quan sát được (obserwed data).  
   Tốt khi có sự chệnh lệch số lượng giữa các lớp phân loại.
3. Huấn luyện mô hình (ước lượng tham số) dễ và nhanh.

**+ Khuyết Điểm:**

1. Giả định độc lập (ưu điểm cũng chính là nhược điểm)  
   hầu hết các trường hợp thực tế trong đó có các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau.
2. Vấn đề zero (đã nêu cách giải quyết ở phía trên)
3. Mô hình không được huẩn luyện bằng phượng pháp tối ưu mạnh và chặt chẽ.  
   Tham số mủa mô hình là các ước lượng xác suất điều kiện đơn lẻ.  
   Không tính đến sự tương tác giữa các ước lượng này.

# III. Viết ứng dụng minh họa

\_Để minh họa thuật toán bayes. Em quyết định xây dựng chương trình lọc thư rác với tập dữ liệu đầu vào là 10 email gốc có nhãn trước đó là Spam và 10 email gốc có nhãn trước đó là nonSpam

\_Chương trình được viết trên trình biên dịch Visual Studio và ngôn ngữ lập trình bằng C#

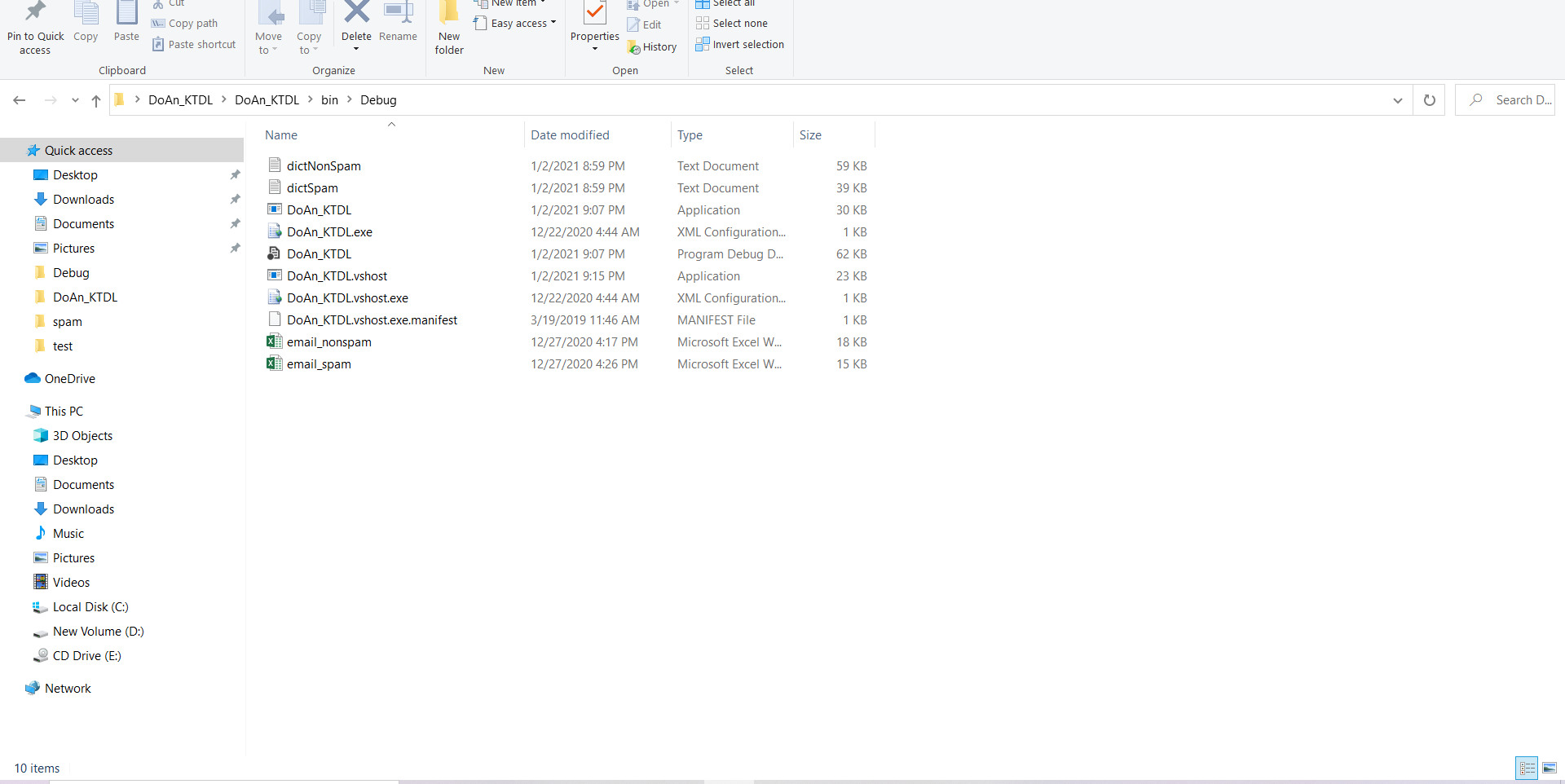
## +Ý Tưởng

\_Xây dựng các thư viện chứa các từ xuất hiện trong cả 2 loại mail kèm với tần số xuất hiện , qua đó dễ dàng phân loại được thư.

\_Tách thư thành các cụm từ ( Từ đây sẽ gọi là các túi từ) và dựa vào thư viện đã xây dựng tính xác suất của từng túi từ.Tích các xác suất lại và ta sẽ có được xác suất của thư cho từng phân lớp.

**+Thực hiện**

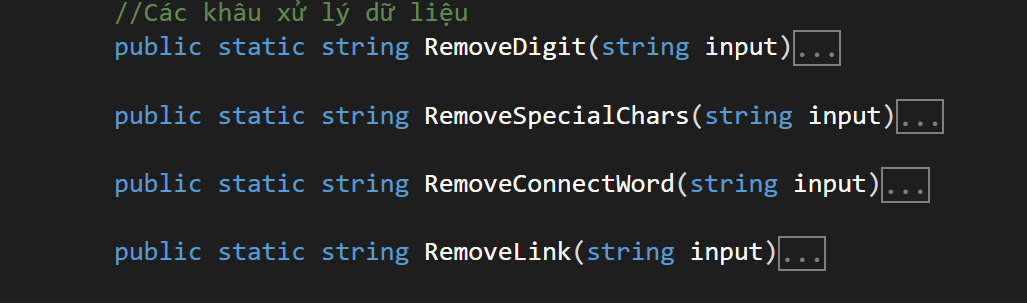
**Dữ liệu đầu vào là 2 file excel gồm thư rác và thư không rác đã được gán nhãn trước**

****

**B1:**

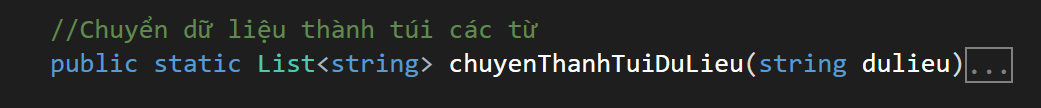
\_Bước đầu giải quyết việc cần thiết nhất là xử lý dữ liệu

\_Em tạo ra các hàm dùng để loại bỏ, chuẩn hóa dữ liệu đầu vào

******

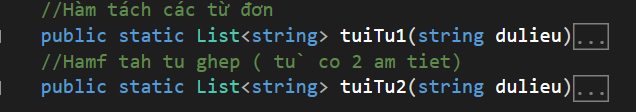
\_Với các hàm xử lý trên đoạn văn bản đầu vào sẽ chỉ còn là các từ có chứa chữ cái .

**B2:**



\_Thực hiện chuyển dữ liệu đã lọc thành một túi từ

**B3:**

****

**\_**Tách các túi từ thành các từ đơn thành từ ghép

Do hạn chế trong kỹ thuật xử lý + chưa có từ điển chuẩn cho việc xử lý ngôn ngữ tiếng việt nên ở đây em chỉ xét tính tần số của các từ đơn và các từ ghép 2 tiếng.

Việc xác định tần số theo công thức sau :

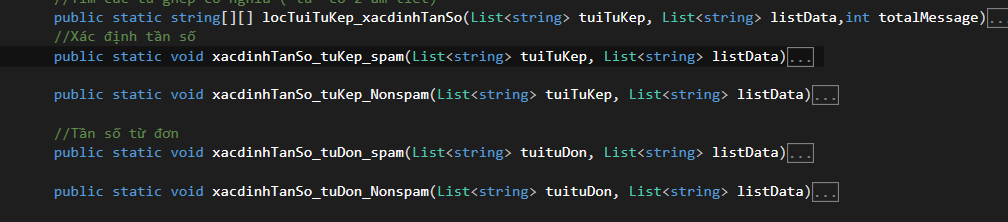
+Đối với từ đơn :

Tần số xuất hiện của từ đơn được tính trên từng bức thư, mỗi lần xuất hiện trong thư được tính là xuất hiện 1 lần, nếu trong thư, từ đó xuất hiện nhiều lần thì cũng tính là 1 lần

+Đối với từ ghép :

Tần số xuất hiện của từ đơn được tính trên từng bức thư, mỗi lần xuất hiện trong thư được tính là xuất hiện 1 lần, nếu trong thư, từ đó xuất hiện k lần thì được tính là k lần

Sau đó thực hiện việc lấy số lần xuất hiện chia cho tổng các loại thư sẽ ra được tần số của từng từ



**B4 :**

\_Sau khi đã thực hiện việc train dữ liệu (Xây dựng các từ điển)

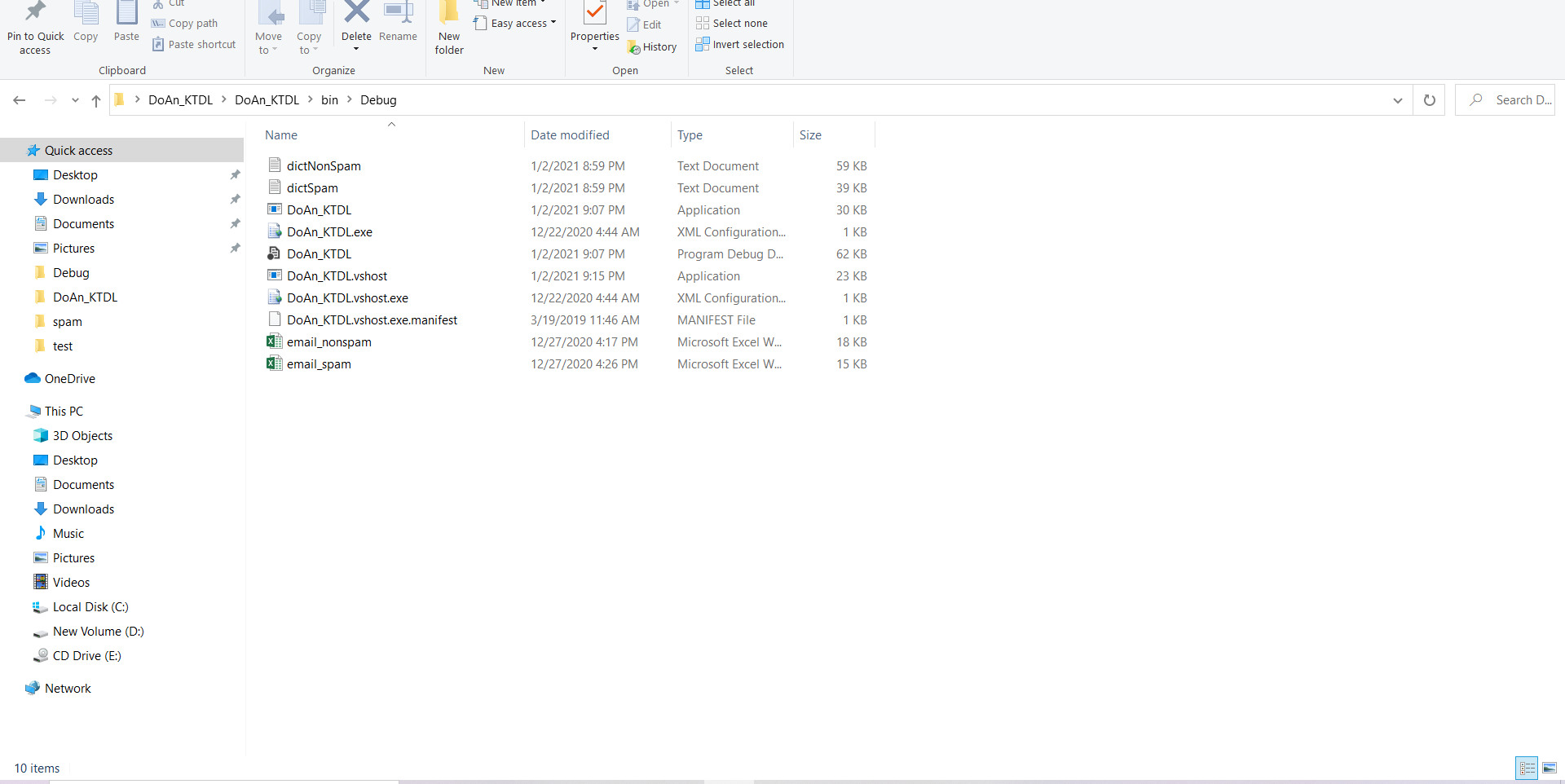
\_Ta tiếp tục tiến hành dự đoán một thư test bằng các cách:

+Lọc dữ liệu

+Tách thành túi từ

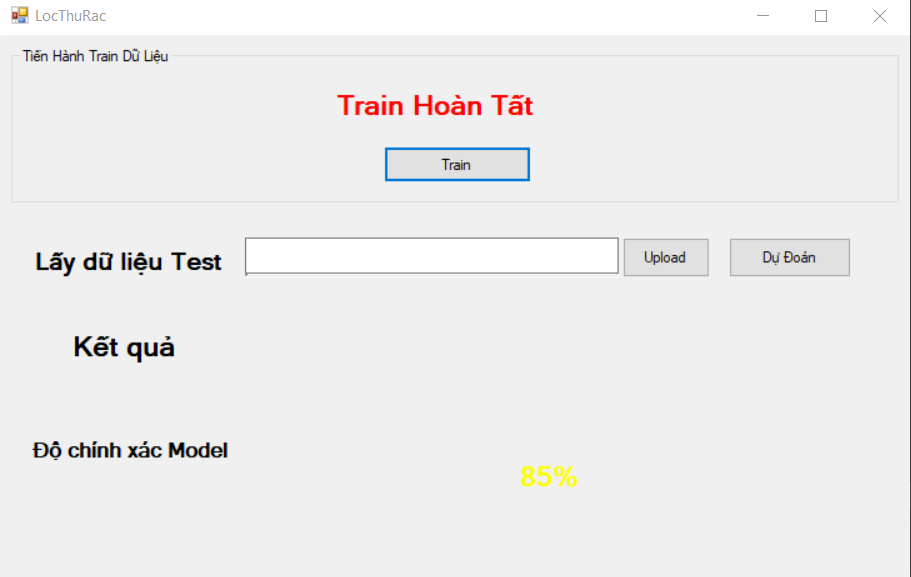
+Sử dụng hàm predict

****

\_Để thuận lợi cho việc chạy chương trình cũng như rút ngắn thời gian train dữ liệu , em tiến hành ghi dữ liệu ra các file txt tương ứng và đọc vào chương trình , nếu như chương trình có thay đổi bộ dữ liệu chuẩn đầu vào , thì ta chỉ cần xóa 2 thư mục ****

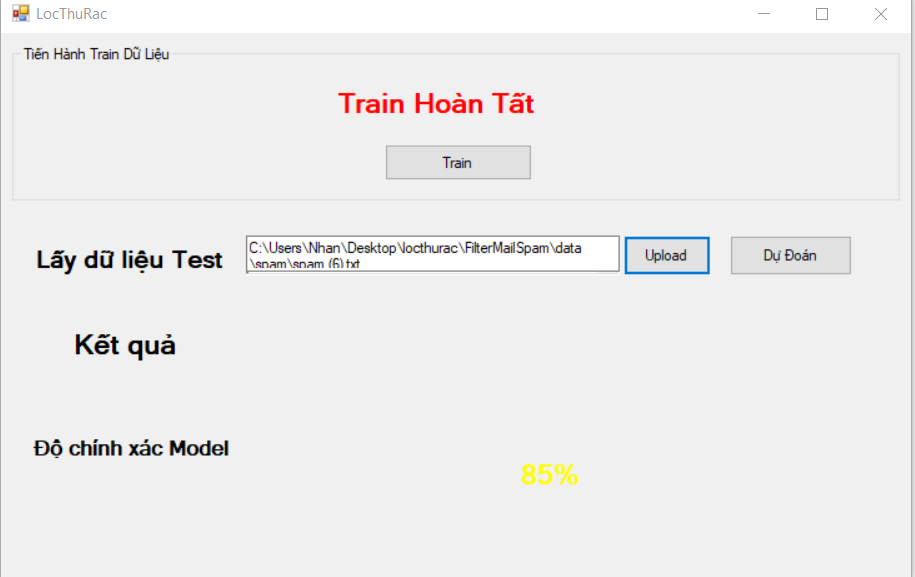
dicNonSpam và dicSpam nằm trog mục debug của chương trình .

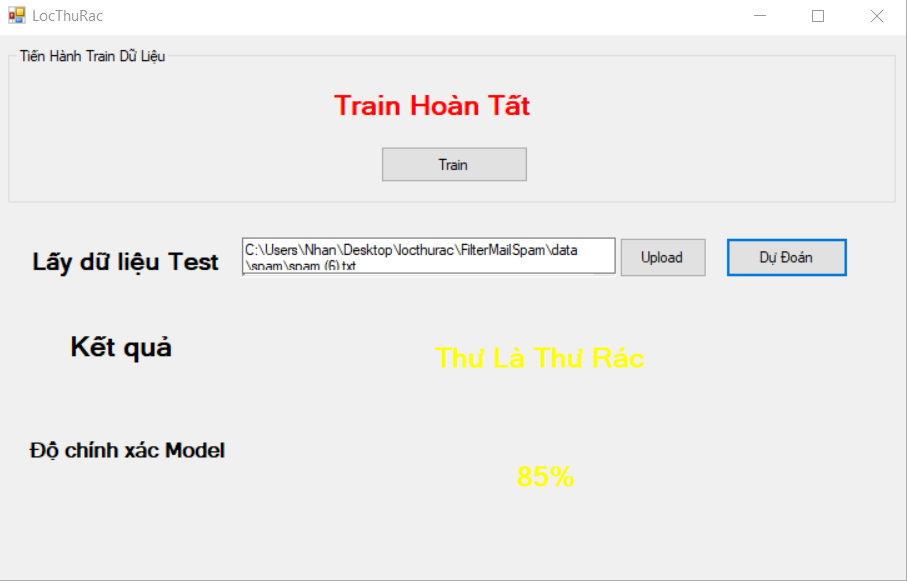
\_Do các hạn chế về kỹ thuật xử lý tiếng việt và tập dữ liệu còn nhỏ , mô hình của em có độ chính xác 85%



\_Tiến hành test file dữ liệu :

\_Chon đường dẫn đến tệp :





**Kết luận :**

**\_** Thư spam tiếng Việt đang trong giai đoạn phát triển, do vậy, vấn đề khó khăn lớn là thu thập tập huấn luyện thư rác và thư bình thường bằng tiếng Việt. Với tập huấn luyện càng lớn thì độ chính xác của việc học từ đơn và từ ghép càng được nâng cao, góp phần rất lớn trong việc tính xác suất theo công thức Naïve Bayes. Do đó chúng tôi sẽ tiếp tục thu thập để có được bộ huấn luyện lớn hơn nhằm nâng cao độ chính xác của phương pháp này

\_Với kiến thức của mình và sự giúp đỡ chỉ bảo tận tình của thầy Ngô Dương Hà , qua chương trình ứng dụng em đã nắm rõ được thuật toán Naives Bayes cũng như viết được chương trình mô phỏng của bài toán.Do việc các kỹ thuật xử lý tiếng việt cũng như tập dữ liệu train chưa thực sự tốt nên độ chính xác của model vẫn còn khá thấp . Em xin chân thành cám ơn thầy đã giúp đỡ em.

# DANH MỤC TÀI LIỆU THAM KHẢO

1. <https://docs.microsoft.com/en-us/archive/msdn-magazine/2019/june/test-run-simplified-naive-bayes-classification-using-csharp>

2. <https://docs.microsoft.com/en-us/archive/msdn-magazine/2013/february/test-run-naive-bayes-classification-with-csharp>

3. https://viblo.asia/p/xay-dung-mo-hinh-loc-thu-rac-bang-naive-bayes-Ljy5Vqxklra

4. https://trituenhantao.io/lap-trinh/phan-2-phan-loai-naive-bayes-coding/