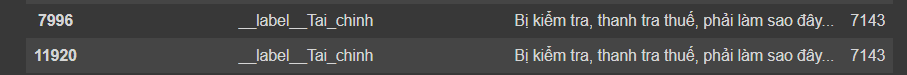
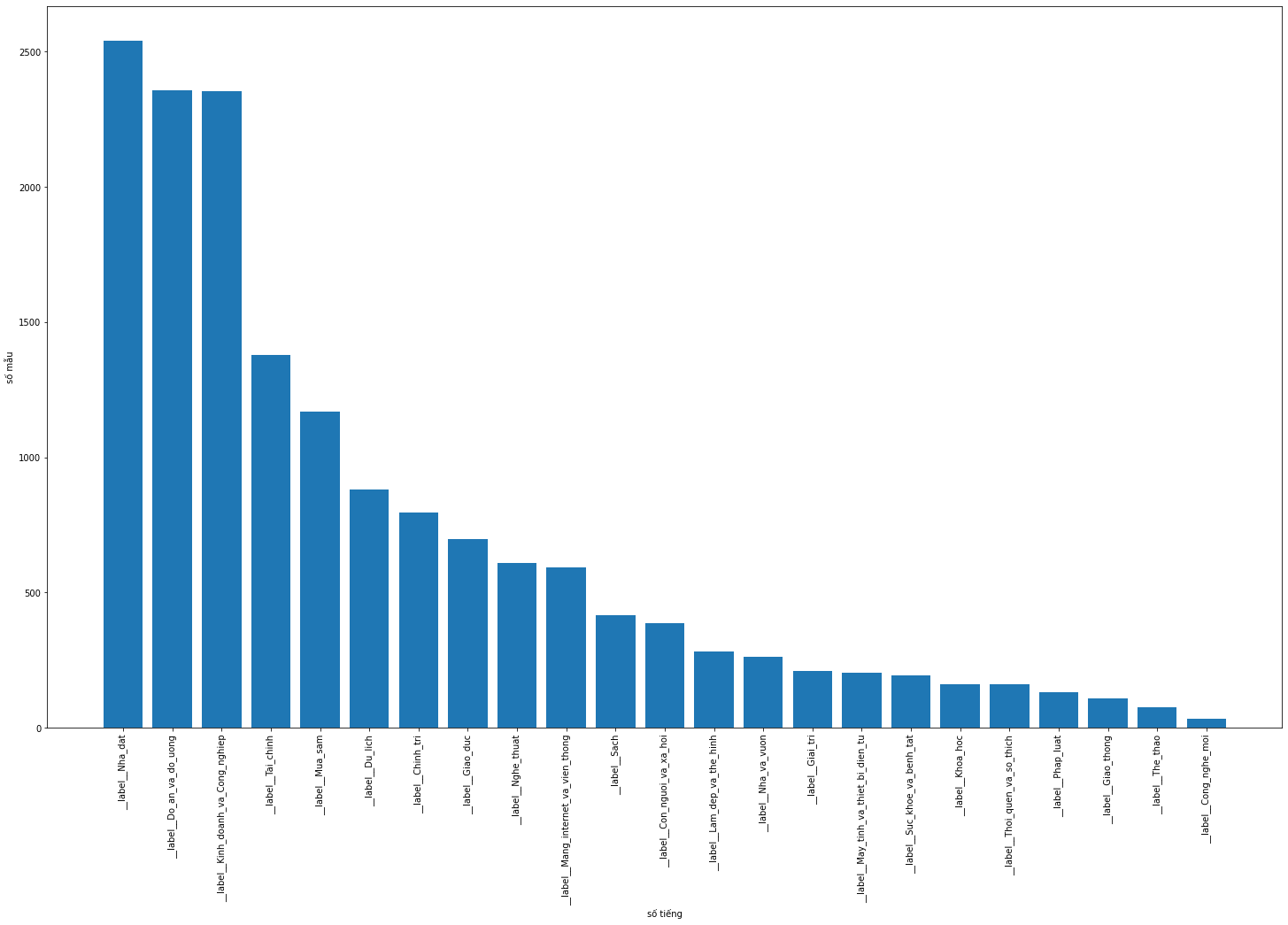
# Phân tích dữ liệu [[src](https://colab.research.google.com/drive/1s26Ztbpv9xWINMrqUnHqCmMIhT7pCMUB?usp=sharing)]

## Cấu trúc tập dữ liệu

* 1. Lượng dữ liệu
* số lượng bản ghi trong file txt : 16000
* Có trùng lặp dữ liệu

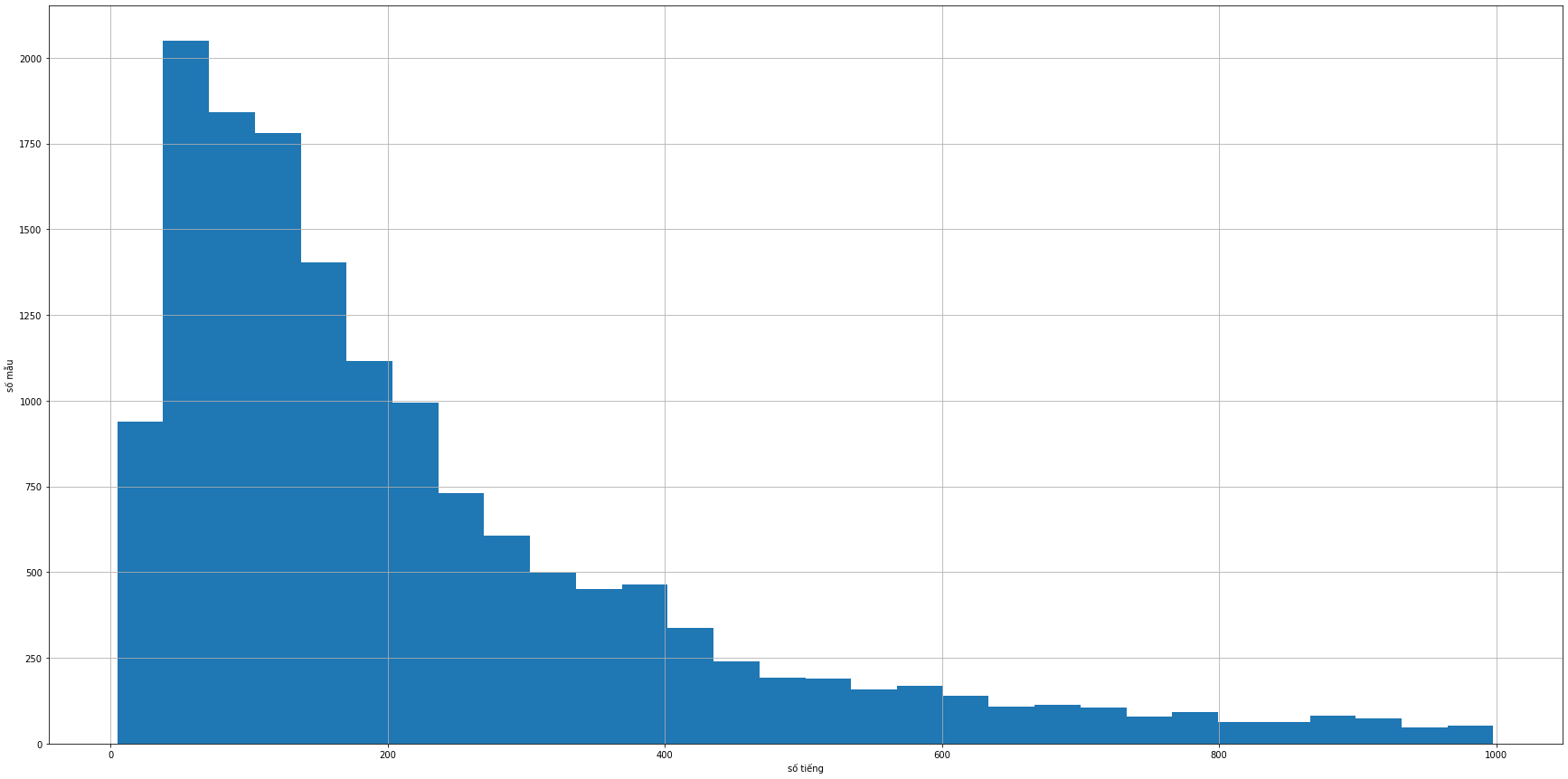


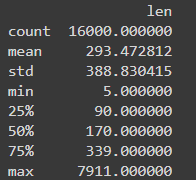
* 1. Phân bổ tập nhãn



* 1. Chiều dài(số lượng tiếng) của bản ghi







## Cấu trúc một mẫu bản ghi

1. mẫu ngắn

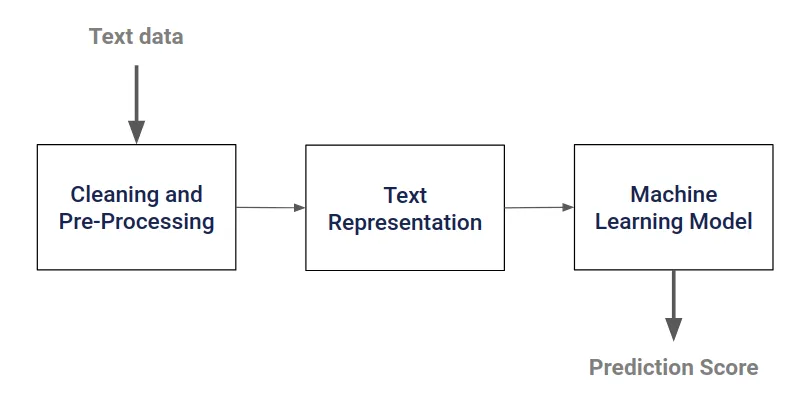
\_\_label\_\_Nha\_dat Cần **thuê** **nhà** **nguyên căn** khu vực đường Láng , Quan Nhân, Trung Kính , Nguyễn Khang

1. mẫu vừa

\_label\_\_Sach #Minigame #**Tặng\_sách**🎁🎁🎁 #Tặng\_COUPON💥 😉 "Tất cả mọi người lớn đều từng là trẻ con. Họ là như thế! Không nên giận họ. Trẻ con phải hết sức rộng lượng đối với người lớn." :) 🌹 HOÀNG TỬ BÉ, Antoine de Saint-Exupéry 🌹 Nguyễn Thành Long dịch, Nguyễn Thành Vũ minh họa 🌀 **Ấn bản** mới tuyệt đẹp của HOÀNG TỬ BÉ đã chính thức phát hành toàn quốc và tại **Hội sách** Chào mừng **Ngày sách Việt Nam**!! Một phiên bản nên thơ với những **minh họa** màu nước bay bổng của họa sĩ trẻ Nguyễn Thành Vũ chắc chắn sẽ làm say lòng bạn đọc yêu HOÀNG TỬ BÉ <3 Và tại sao lại không nhỉ?! 📚 Thả tim nếu bạn luôn ủng hộ những thể nghiệm sáng tạo - Dự án minh họa mới các tác phẩm kinh điển trên thế giới nhé <3 XEM NGAY >>> http://bit.ly/Hoang-tu-be-an-ban-minh-hoa-mau ------------------------ 💗💗Với siêu phẩm tuyệt vời thế này, nếu làm quà thì ắc hẳn là rất tuyệt vời nhỉ? 💢 **Mọt sách** Tiki sẽ dành hẳn 10 phần quà may mắn (gồm có 5 **quyển sách** HOÀNG TỬ BÉ và 5 COUPON **mua sách** tại Tiki) dành cho 10 bạn may mắn khi tham gia minigame sau: - B1: LIKE/SHARE ngay post này lên tường FB của bạn (nhớ là công khai mới hợp lệ), kèm hashtag #Mọt\_sách\_Tiki, #Hoàng\_Từ\_Bé - B2: Tag rủ rê thêm 5 người bạn vào post này, nếu có thể kèm "ước muốn" của mình là được tặng sách hoặc coupon nữa là xong rồi đó. 💫 Kết quả minigame sẽ được chọn random kết quả nhé! và được công bố vào ngày 02/05 này. Chúc các bạn thật may mắn <3 <3 <3

===> Phân biệt chủ đề các câu không dựa vào toàn bộ từ ngữ, câu văn,ký hiệu trong mẫu

# Tổng quan quy trình



## Cleaning and pre-processing

Quá trình này sẽ loại bỏ những dữ liệu không cần thiết trong bản ghi[[Nguồn](https://swatimeena989.medium.com/beginners-guide-for-preprocessing-text-data-f3156bec85ca)]

+ Chuẩn hóa:

một số âm tiết viết sai dẫu: oẻ -> ỏe

một số từ viết tắt thông dụng: hssv học sinh sinh viên,...

+ Loại bỏ:

* Link : http:....
* emoiji: 💗💗🎁🎁🎁
* số
* kí tự không phải chữ cái, dấu câu ?, !, #,, <, /, - :, …
* từ viết tắt: FB, ...
* stopwords : ắt hẳn, những, chắc chắn,...
* Tokenize

input: hey you want learn natural language processing

output: ['hey', 'you', 'want', 'learn', 'natural', 'language', 'processing']

Với tiếng việt, ta còn muốn tách được ra từ đơn và từ ghép

input: Thịt gà sốt chua ngọt

output: [‘thịt gà’, ‘sốt’, ‘chua’, ‘ngọt’]

## Text Representation

Sau bước 1, một câu của chúng ta được biểu diễn bằng một mảng str.

Để máy tính có thể sử dụng các thuật toán dựa trên kỹ thuật toán học (mathematical methods) như Bayes, SVM, Kmeans, Neutral network,... ta cần số hóa các mảng này

Nói cách khác, mỗi câu sẽ được biểu diễn bằng một véc tơ

input: [‘thịt gà’, ‘sốt’, ‘chua’, ‘ngọt’]

output: [1, 2, 3, 4]

## Machine Learning model

Xây dựng Pipeline để huấn luyện các mô hình học máy trong cùng một điều kiện, tìm ra mô hình có độ chính xác tốt nhất có thể.

# 

# Tiền xử lý dữ liệu

# Toàn bộ quá trình sử dụng code python[[src](https://colab.research.google.com/drive/1zzzRgpQKOaNJTfK_bwNPo6tNkCYfGEP9)]

1. Chuẩn hóa từ:

chuẩn âm tiết : thay thế bằng một dict âm tiết sai liệt kê tay

replace\_list = {

'òa': 'oà', 'óa': 'oá', 'ỏa': 'oả', 'õa': 'oã', 'ọa': 'oạ', 'òe': 'oè', 'óe': 'oé','ỏe': 'oẻ',

'õe': 'oẽ', 'ọe': 'oẹ', 'ùy': 'uỳ', 'úy': 'uý', 'ủy': 'uỷ', 'ũy': 'uỹ','ụy': 'uỵ', 'uả': 'ủa',

'qủa':'quả', 'qùa':'quà', 'qủ':'qu', 'qù':'qu', 'qũ':'qu', 'qú':'qu', 'qụ':'qu',

'ố': 'ố', 'u´': 'ố', 'àk': u' à ','aˋ': 'à', 'iˋ': 'ì', 'ă´': 'ắ','ử': 'ử', 'e˜': 'ẽ', 'y˜': 'ỹ', 'a´': 'á',

}

Chuẩn hóa từ viết tắt:

-> lọc từ viết tắt

-> dùng AI “ chạy bằng cơm” chuẩn hóa ->[file từ tắt](https://drive.google.com/file/d/1ccmaiHA1qYo8GZXbJ3fQSc6mWChvmxuu/view?usp=sharing)

-> loại bỏ chữ viết hoa

1. Loại bỏ:

* Sử dụng thư việt emoji để xóa các emoji
* Sử dụng regular expression và phương thức replace trong python để loại bỏ link, chữ số, dấu câu, từ viết tắt còn lại (các từ viết tắt thường không chứ nguyên âm), và một số từ quá dài (>5 chữ, hoặc quá ngắn <2)

\_\_label\_\_Thoi\_quen\_va\_so\_thich bác nào có thể giúp e về 1 số bài mồi câu lục dc ko em móm thâu 1 tuần rồi ngày nào cũng đi mà chat dc phát nào! 😞😞😞😞😞

=> output

bác nào có thể giúp về số bài mồi câu lục em móm thâu tuần rồi ngày nào cũng đi mà chat phát nào

1. Tokenization

Sử dụng thư viện có sẵn **VnCoreNLP để xử lý việc tokenization tiếng việt**

**input : tôi bảo vệ tổ quốc việt nam thân yêu**

**output: tôi** bảo\_vệ tổ\_quốc việt\_nam thân\_yêu

1. Gia tăng dữ liệu (Data augmentation)

Làm tăng dữ liệu cải thiện quá kết quả khi huấn luyện mô hình bằng việc ta sẽ kết hợp dữ liệu chưa loại bỏ các stopwords và dữ liệu loại bỏ stopwords

**input** : bác nào có thể giúp về số bài mồi câu lục em móm thâu tuần rồi ngày nào cũng đi mà chat phát nào

**output sau khi loại bỏ stopwords**

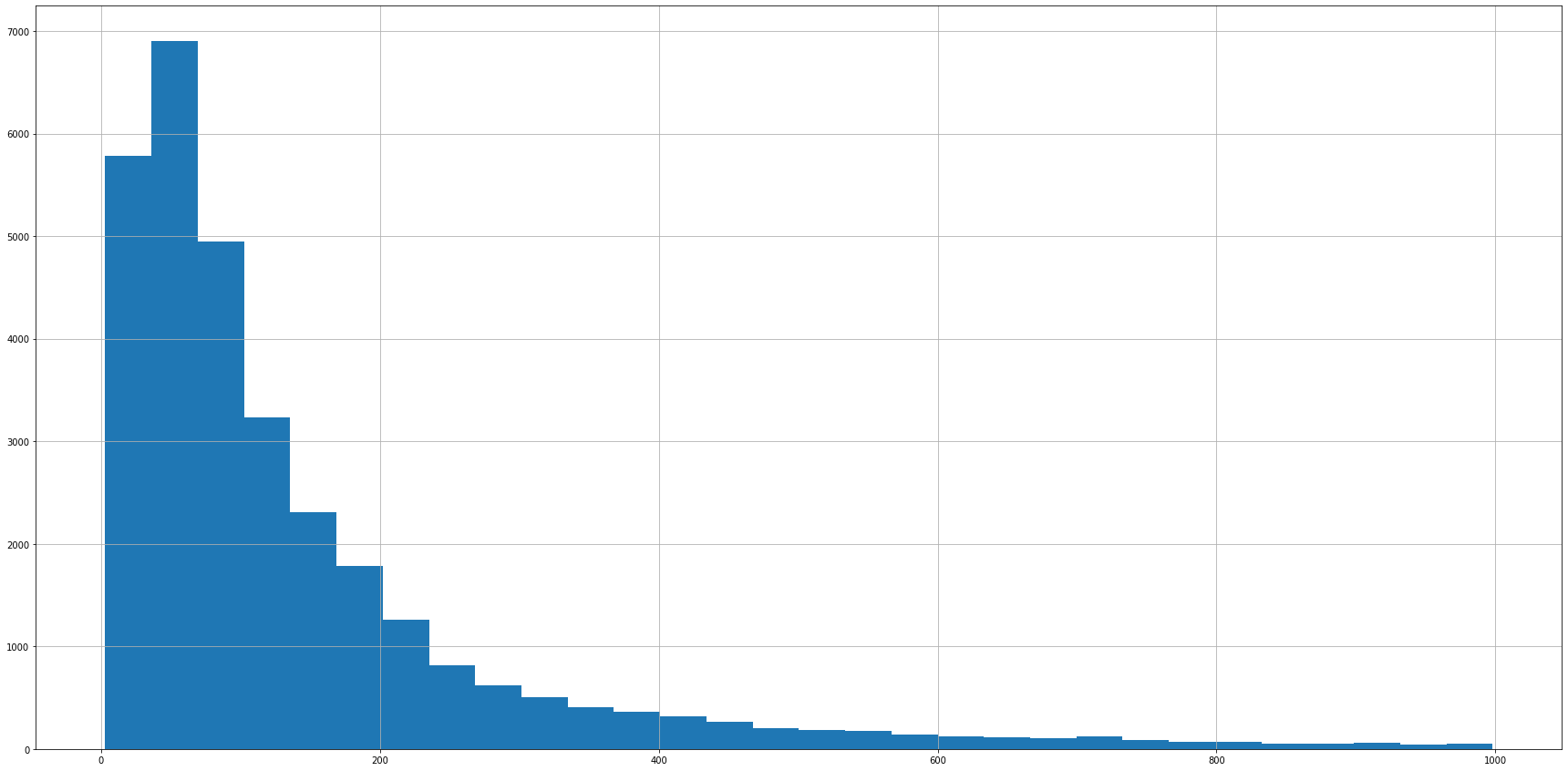
giúp bài mồi Câu lục móm thâu tuần đi chat phát

# Text Representation

Có 2 bước:

1. Cắt bớt câu

Sau tầng tiền xử lý, dữ liệu vẫn gồm tập các mẫu với độ dài(số lượng từ) khác nhau



Thống kê trung bình gồm 158 từ/mẫu -> ta sẽ chọn độ dài cụ thể là 158 từ/mẫu

* Với câu quá dài(>158) sẽ chỉ lấy đúng 158 từ
* Với câu quá ngắn(<158) ta sẽ đệm thêm một số từ rỗng(dạng mã hóa) phía sau.

1. Vectơ hóa

Cơ chế: sử dụng từ điển từ vựng(có sẵn ở trên mạng :>)

| vd: |  |
| --- | --- |
| 166 | dự\_án 2123968 |
| 167 | bán 2117131 |
| 168 | giúp 2108545 |
| 169 | một\_số 2087404 |
| 170 | 5 2073774 |
| 171 | Nga 2067041 |
| 172 | dùng 2036691 |
| 173 | xã 2034047 |
| 174 | trẻ 2022703 |
| 175 | vụ 2016205 |
| 176 | đất 2009574 |

. Với từ điển, mỗi từ ứng với một số, trong trường hợp không có từ trong từ điển, ta sẽ gắn cho nó 1 số chưa tồn tại và cập nhật từ mới vào từ điển.

Với ký tự rỗng được đệm vào câu ngắn, thay bằng số 0

vd:input :

theo hành\_trình tour du\_lịch mỹ bờ đông\_du khách của lữ\_hành sẽ đến york giấc mơ được gọi tên của hàng triệu người chinh\_phục du\_khách bằng tượng\_đài nữ\_thần tự\_do duyên\_dáng trên vịnh trung\_tâm trụ\_sở tài\_chính lớn nhất thế\_giới với phố khu và thủ\_đô dạo bước trên mall du\_khách sẽ bị ngợp bởi các công\_trình kỳ\_vĩ nơi đây từ điện tượng\_đài đài tưởng\_niệm cho đến nhà\_trắng mỗi công\_trình đều là một tuyệt\_tác kiến\_trúc được phối\_cảnh hài\_hoà khiến cho từ lâu đã được ghi\_nhận là một trong những thành\_phố đẹp nhất hợp chủng quốc hoa\_kỳ đặc\_biệt hơn khi đi du\_lịch mỹ đến bất\_kỳ thành\_phố nào vào bất\_cứ thời\_gian nào du\_khách cũng có\_thể trải thú\_vui mua\_sắm bất com nuoc ngoai

Câu trên sau tiền sử lý có 162 tiếng sẽ được vector hóa thành

[ 63, 1627, 3066, 388, 9105, 920, 30420, 336, 7, 6798, 38, 30, 1581, 2387, 1947, 3201, 2189, 11, 328, 221, 7, 119, 113, 18, 3252, 902, 121,

8198, 12821, 1344, 6633, 34, 3948, 591, 1671, 699, 103, 67, 198, 15, 835, 316, 6, 1413, 3804, 499, 34, 1387, 9009, 902, 38, 45, 40853, 223,

9, 531, 21987, 189, 97, 39, 462, 8198, 3010, 5599, 13, 30, 52912, 9763, 205, 531, 131, 8, 16, 28807, 2167, 11, 49206, 4557, 122, 13, 39,

700, 14, 11, 1334, 8, 16, 12, 21, 214, 258, 67, 2288, 10631, 3705, 16184, 560, 234, 48, 26, 57, 388, 9105, 30, 1403, 214, 142, 33,

1153, 130, 142, 902, 32, 62, 1289, 12442, 2063, 5408, 9278, 7873, 22899, 60244, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0, 0, 0, 0]

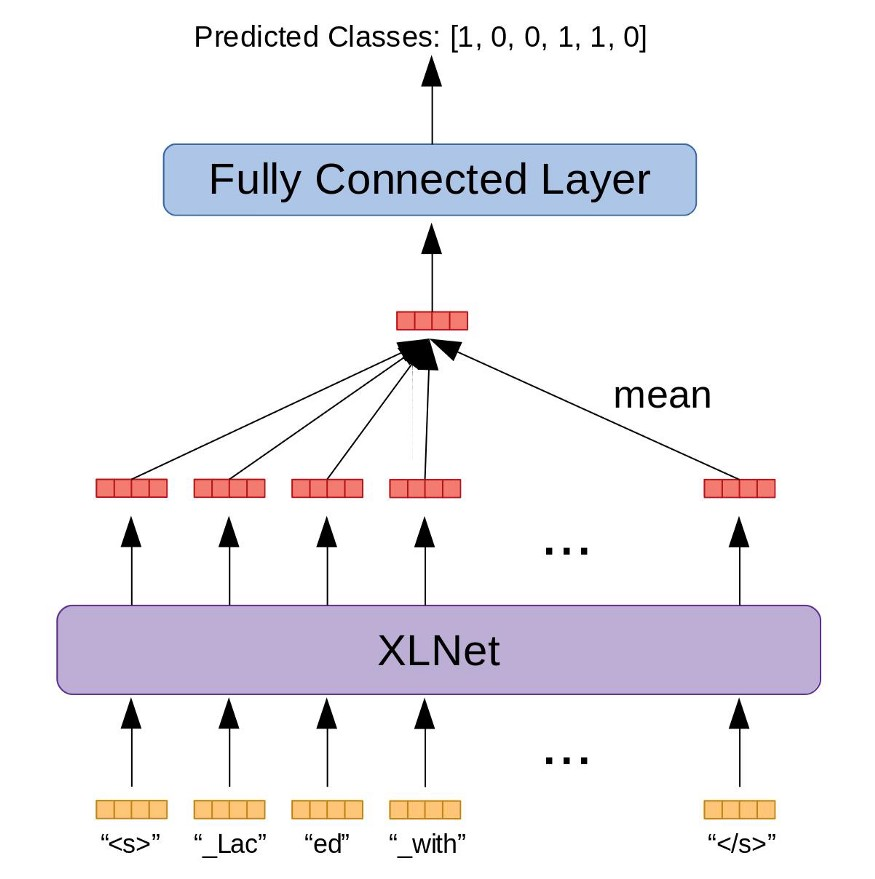
<quá trình này phụ thuộc vào mô hình học máy sử dụng sẽ có chỉnh sửa khác nhau>

# Xây dựng mô hình phân lớp

Các phương án đề xuất:

1. Mô hình Naive Bayes
2. Transfer Learning:
   1. Bert model
   2. PhoBert Model
   3. XLNet Model

Sơ qua:



Với mỗi phương án, sẽ được xây dựng các pipeline phù hợp, toàn bộ code python ở [[đây](https://drive.google.com/drive/folders/1WQ1U9OyuEcDov8bXatbJ-86j0xDo6ND5?usp=sharing)]

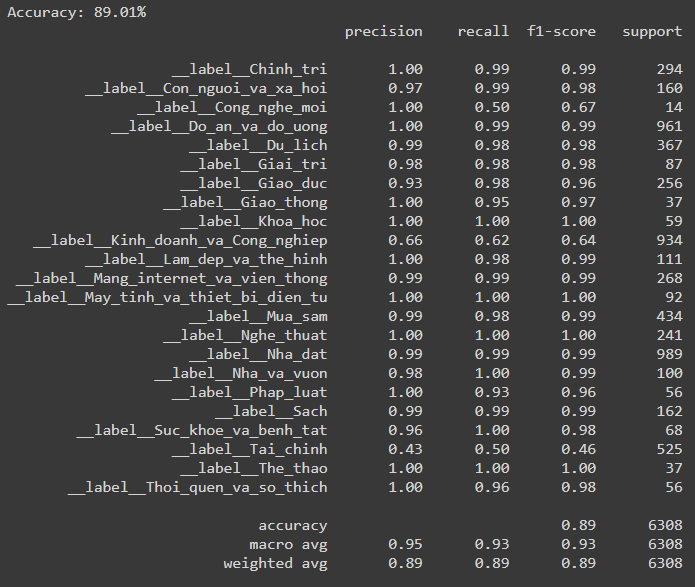
<phần này hơi rộng so với giới hạn bài nên để mọi người tìm hiểu sau :3 >

# Kiểm tra và đánh giá

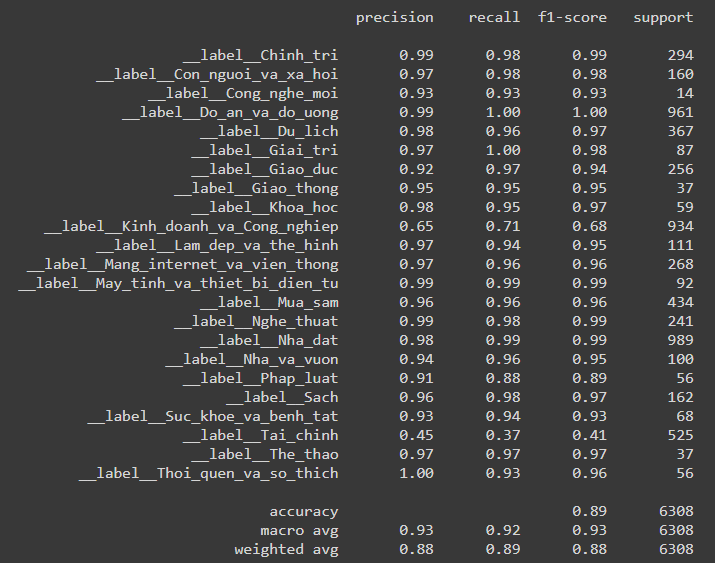
Với tập dữ liệu đã làm sạch và gia tăng dữ liệu (từ 16000-> 31538), ta chia dữ liệu thành hai tập huấn luyện(train dataset) và kiểm thử(valid dataset) theo tỉ lệ 8:2, với mỗi phương án thì các tập vẫn sẽ giống nhau.

Kết quả:

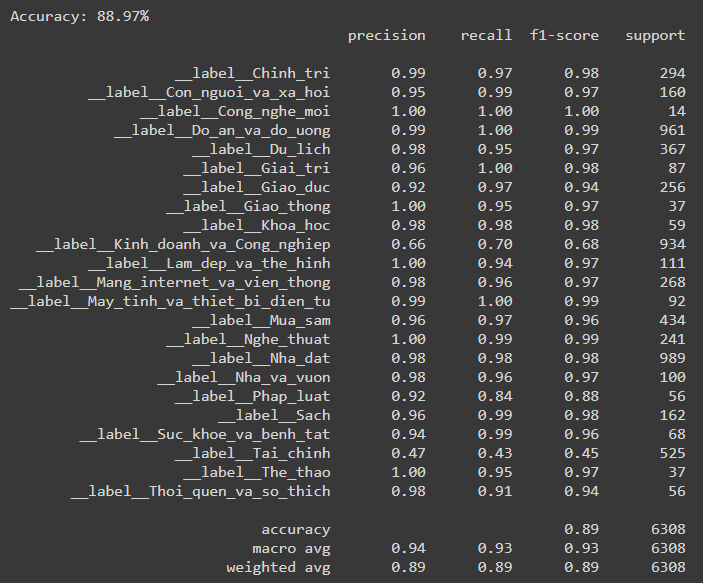
***Mô hình Mô hình Naive Bayes***

******

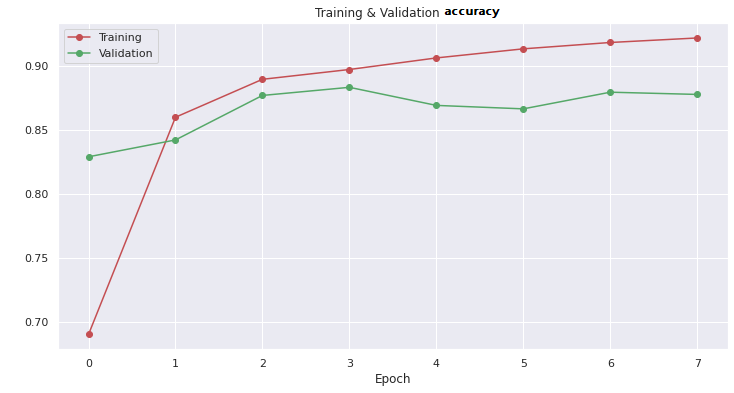
**Bert model với chiều dài một câu là 128 - Accuracy: 88.62%**

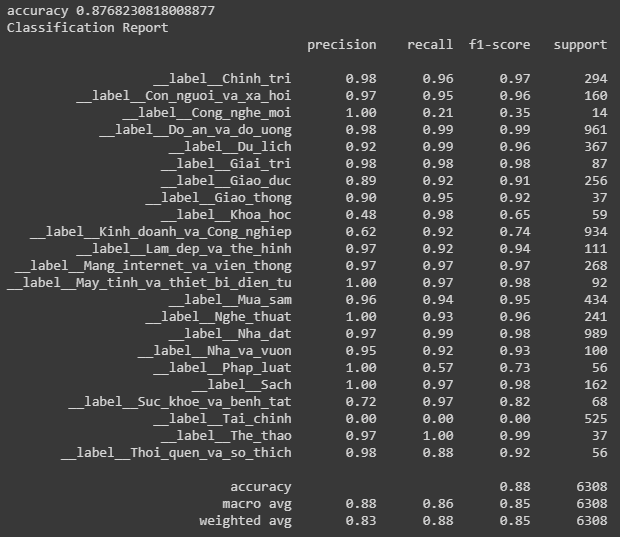
****

**Bert model với chiều dài một câu là 158**

****

**PhoBert model với chiều dài một câu là 128 (overfitting)**

****

****