 **TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

KHOA CÔNG NGHỆ THÔNG TIN

**- - - - -o0o- - - - -**



**BÀI TẬP LỚN**

**Môn: Kho dữ liệu và khai phá**

***Đề tài:* Phân loại văn bản bằng thuật toán**

**Naive Bayes**

**Giáo viên hướng dẫn: Ts.Trần Chí Kiên**

**Sinh viên thực hiện:**

* + - **Bùi Công Quỳnh**
    - **Bùi Thị Hoàng Yến**

Hà Nội, tháng 11/2020

# MỤC LỤC

[MỤC LỤC 1](#_Toc58076774)

[LỜI MỞ ĐẦU 2](#_Toc58076775)

[CHƯƠNG 1: GIỚI THIỆU 3](#_Toc58076776)

[1.1. Khái niệm 3](#_Toc58076777)

[1.2. Quy trình 3](#_Toc58076779)

[1.3. Bài toán 3](#_Toc58076781)

[CHƯƠNG 2: TỔNG QUAN 4](#_Toc58076783)

[1. Khai phá dữ liệu là gì? 4](#_Toc58076784)

[2. Đặc trưng của khai phá dữ liệu . 4](#_Toc58076785)

[3. Các phương pháp khai phá 4](#_Toc58076786)

[4. Giới thiệu giải thuật Naive Bayes 5](#_Toc58076787)

[1. Khái niệm: 5](#_Toc58076788)

[2. Định lý: 5](#_Toc58076789)

[3. Phân lớp Naive Bayes 6](#_Toc58076790)

[CHƯƠNG 3: CHƯƠNG TRÌNH 8](#_Toc58076791)

[1. Phầm mềm 8](#_Toc58076792)

[2. Cài đặt chương trình: 8](#_Toc58076793)

[3. Cài đặt mã: 8](#_Toc58076794)

[CHƯƠNG 4: TÀI LIỆU THAM KHẢO 14](#_Toc58076829)

[1.1. Bài viết sử dụng các nguồn: 14](#_Toc58076830)

[1.2. Data: 14](#_Toc58076831)

# LỜI MỞ ĐẦU

Ngày nay công nghê thông tin đang phát triển vượt trội như vũ bão trên mọi lĩnh vực hoạt động của đời sống điều đã xảy ra trong nhiều năm trở lại đây của Việt Nam.

Chính vì vậy mà nhu cầu sử dụng văn bản, chia sẻ tin tức dưới dạng văn bản đang được sử dụng rất nhiều: báo chí, truyền thông, tài liệu, ...

Việc cần thiết được đưa ra là làm sao mà có thể phân loại văn bản đó phù hợp với nội dung mà người đọc hay người tìm kiếm hướng tới.

Qua thời gia tìm hiểu, nhằm đáp ứng nhu cầu đó nhóm em xin đưa ra giải pháp “*phân loại văn bản bằng thuật toán Naive Bayes*” .

Bài viết dưới đây chỉ có thể lột tả phần nào thuật toán. Nếu có sư sai sót mong thầy cô giúp đỡ.

Em xin chân thành cảm ơn!

# CHƯƠNG 1: GIỚI THIỆU

* 1. Khái niệm

Khai phá dữ liệu (*data mining*)

* Là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính.
* Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp. Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD.
  1. Quy trình

Khai phá dữ liệu là một bước của quá trình [khai thác tri thức](https://vi.wikipedia.org/w/index.php?title=Khai_th%C3%A1c_tri_th%E1%BB%A9c&action=edit&redlink=1" \o "Khai thác tri thức (trang chưa được viết)) (*Knowledge Discovery Proces*s), bao gồm:

* Xác định vấn đề và không gian dữ liệu để giải quyết vấn đề (*Problem understanding and data understanding*).
* Chuẩn bị dữ liệu (*Data preparation*), bao gồm các quá trình làm sạch dữ liệu (*data cleaning*), tích hợp dữ liệu (*data integration*), chọn dữ liệu (*data selection*), biến đổi dữ liệu (*data transformation*).
* Khai thác dữ liệu (*Data mining*): xác định *nhiệm vụ khai thác dữ liệu* và lựa chọn *kỹ thuật khai thác dữ liệu*. Kết quả cho ta một *nguồn tri thức thô*.
* Đánh giá (*Evaluation*): dựa trên một số tiêu chí tiến hành *kiểm tra* và *lọc* nguồn tri thức thu được.
* Triển khai (*Deployment*).

*Nguồn: https://vi.wikipedia.org/*

* 1. Bài toán

Xét ví dụt:

* Bàn cần xây dựng một mô hình học máy để phân loại các loại văn bản được nạp vào theo các chủ đề : kinh tế, chính trị, văn hóa, thể thao, ... bất kì nào.

# CHƯƠNG 2: TỔNG QUAN

## Khai phá dữ liệu là gì?

**Khai phá dữ liệu** (*data mining*) Là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp. Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD.

## Đặc trưng của khai phá dữ liệu .

* Xác định vấn đề và không gian dữ liệu để giải quyết vấn đề (*Problem understanding and data understanding*).
* Chuẩn bị dữ liệu (*Data preparation*), bao gồm các quá trình làm sạch dữ liệu (*data cleaning*), tích hợp dữ liệu (*data integration*), chọn dữ liệu (*data selection*), biến đổi dữ liệu (*data transformation*).
* Khai thác dữ liệu (*Data mining*): xác định *nhiệm vụ khai thác dữ liệu* và lựa chọn *kỹ thuật khai thác dữ liệu*. Kết quả cho ta một *nguồn tri thức thô*.
* Đánh giá (*Evaluation*): dựa trên một số tiêu chí tiến hành *kiểm tra* và *lọc* nguồn tri thức thu được.
* Triển khai (*Deployment*).

## Các phương pháp khai phá

* 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.

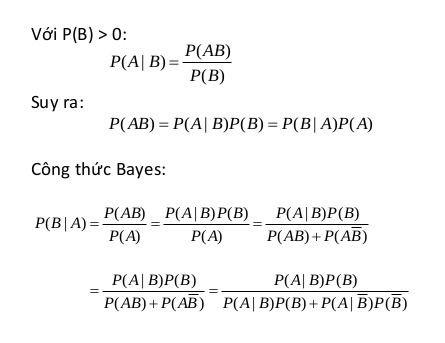
## Giới thiệu giải thuật Naive Bayes

1. Khái niệm:

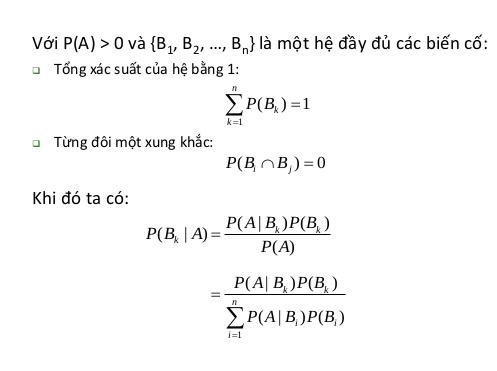
* Là một bộ phận Phân loại (Classification)
* Bộ phân lớp Bayes là một giải thuật thuộc lớp giải thuật thống kê, nó có thể dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp là bao nhiêu. Phân lớp Bayes được dựa trên định lý Bayes (định lý được đặt theo tên tác giả của nó là Thomas Bayes)

1. Định lý:

* Giải thuật Naive Bayes:
* Gọi A, B là 2 biến cố:



* Công thức Naive Bayes tổng quát hóa:



1. **Phân lớp Naive Bayes**

* Bộ phân lớp Naive bayes hay bộ phân lớp Bayes (simple byes classifier) hoạt động như sau:
* Gọi D là tập dữ liệu huấn luyện, trong đó mỗi phần tử dữ liệu X được biểu diễn bằng một vector chứa n giá trị thuộc tính A1, A2,...,An = {x1,x2,...,xn}
* Giả sử có m lớp C1, C2,..,Cm. Cho một phần tử dữ liệu X, bộ phân lớp sẽ gán nhãn cho X là lớp có xác suất hậu nghiệm lớn nhất. Cụ thể, bộ phân lớp Bayes sẽ dự đoán X thuộc vào lớp Ci nếu và chỉ nếu: **P(Ci|X) > P(Cj|X) (1<= i, j <=m, i != j)**  
  Giá trị này sẽ tính dựa trên định lý Bayes.
* Để tìm xác suất lớn nhất, ta nhận thấy các giá trị P(X) là giống nhau với mọi lớp nên không cần tính. Do đó ta chỉ cần tìm giá trị lớn nhất của P(X|Ci) \* P(Ci). Chú ý rằng P(Ci) được ước lượng bằng |Di|/|D|, trong đó Di là tập các phần tử dữ liệu thuộc lớp Ci. Nếu xác suất tiền nghiệm P(Ci) cũng không xác định được thì ta coi chúng bằng nhau P(C1) = P(C2) = ... = P(Cm), khi đó ta chỉ cần tìm giá trị P(X|Ci) lớn nhất.
* Khi số lượng các thuộc tính mô tả dữ liệu là lớn thì chi phí tính toàn P(X|Ci) là rất lớn, dó đó có thể giảm độ phức tạp của thuật toán Naive Bayes giả thiết các thuộc tính độc lập nhau. Khi đó ta có thể tính:  
  **P(X|Ci) = P(x1|Ci)...P(xn|Ci)**

# CHƯƠNG 3: CHƯƠNG TRÌNH

1. Phầm mềm

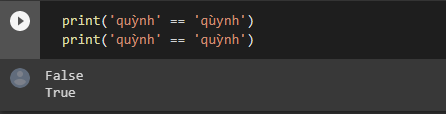
* Sử dụng Machine Learning nên chúng ta yêu cầu một máy chủ có cấu hình mạnh và khả năng tương tác cao: <https://colab.research.google.com/>

1. Cài đặt chương trình:

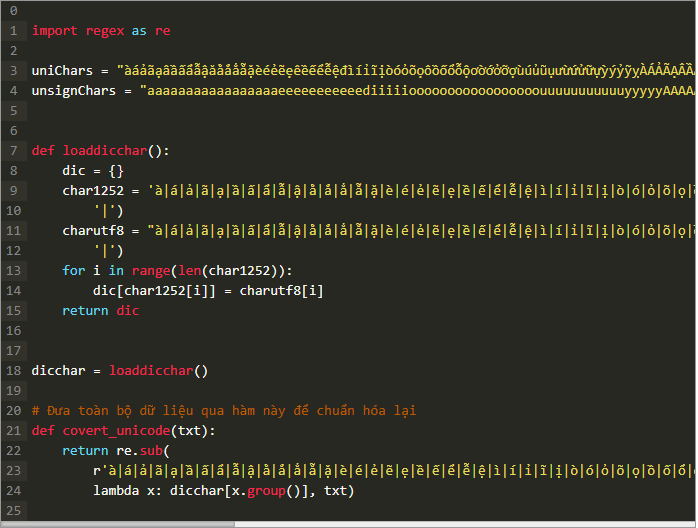
* Cài dặt các chương trình tương quan cho máy chủ: thư viện underthesea, regex, pandas, matplotlip

1. Cài đặt mã:
   1. Tiền xử lý dữ liệu văn bản
      1. Bước tiền xử lý dữ liệu là bước đầu tiên cần làm, nó cần được làm trước bước feature extractor. Việc tiền sử lý dữ liệu là quá trình chuẩn hóa dữ liệu và loại bỏ các thành phần không có ý nghĩa cho việc phân loại văn bản.
      2. Tiền xử lý dữ liệu tiếng Việt cho bài toán phân loại văn bản thường gồm các việc sau:
   * Xóa HTML code (nếu có)
   * Chuẩn hóa bảng mã Unicode (đưa về Unicode tổ hợp dựng sẵn)
   * Chuẩn hóa kiểu gõ dấu tiếng Việt (dùng òa úy thay cho oà uý)
   * Thực hiện tách từ tiếng Việt (sử dụng thư viện tách từ như pyvi, underthesea, vncorenlp,…) đưa về văn bản lower (viết thường)
   * Xóa các ký tự đặc biệt: “.”, “,”, “;”, “)”.
     1. Chuẩn hóa Unicode:

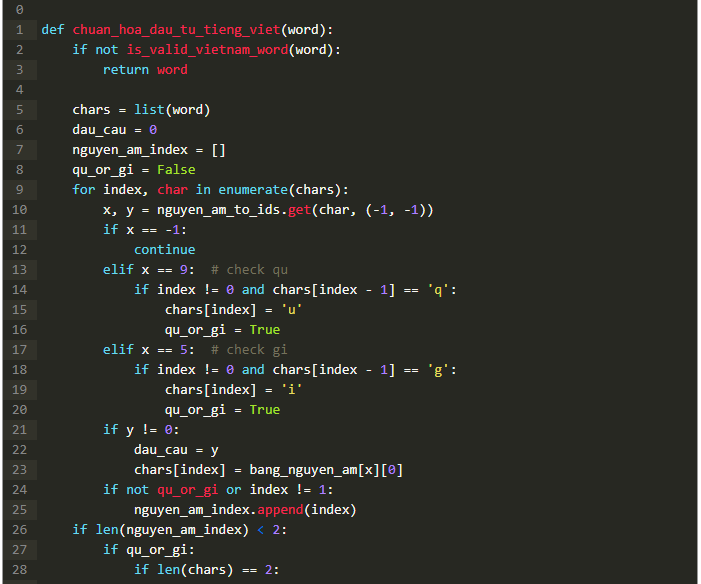
Có 2 loại Unicode: tổ hợp hoặc dựng sãn. Điều đó dẫn tới có 2 loại vấn đề:

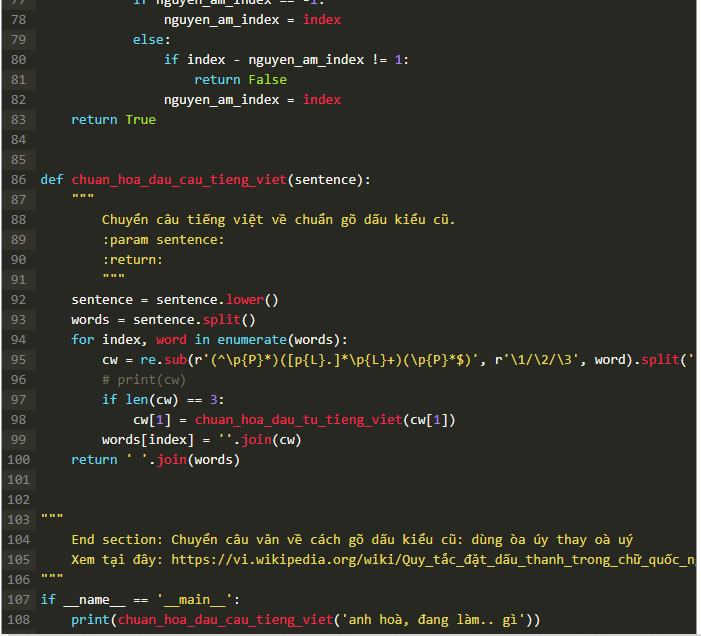


Hướng xử lý: Đưa nó về hàm Unicode dựng sẵn chuẩn:



Chuẩn hóa kiểu gõ: biến đổi từ mà mắt thường nhìn thấy dưới dạng uò = ùo chuẩn





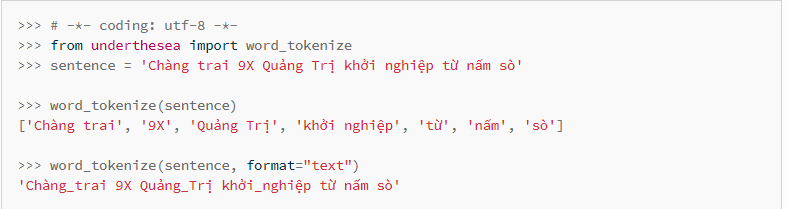
* + 1. Tách từ tiếng việt.

Như các bạn đã biết tiếng việt cấu tạo từ: từ đơn, từ ghép và nhiều bộ phận khác. Nhưng ở đây chúng ta tập trung vào từ đơn và từ ghép. Làm như nào cho máy học biết đâu là từ đơn đâu là từ ghép.

Vd: Học sinh đi học sinh học

Biến đổi: hoc\_sinh di hoc sinh\_hoc

Bài toán sẽ được ta sử dụng thư viện underthesea.word\_tokenize giải quyết:



* + 1. Đưa về dạng Lowercase

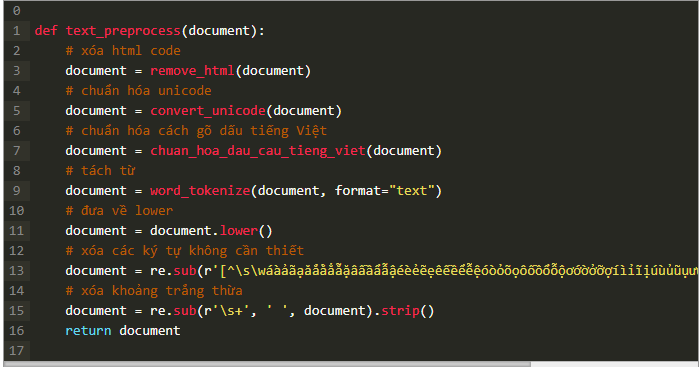
Trong các đoạn văn bản thì sẽ có tên riêng hay tên đâu dong viết hoa nhưng khi đưa vào máy thì ta cần chuẩn hóa nó về dạng viết thường hết giảm số lượng đặc trưng ( do máy tính phân biết giưa in hoa và in thường ).

* + 1. Xóa các kĩ tự không cần thiết.

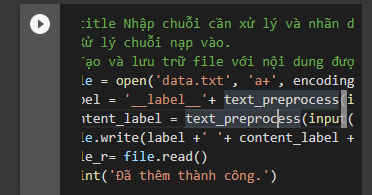
Giảm số lượng chiều đặc trung và tăng tốc độ xử lý.

Không ảnh hương tới kết quả mô hình.

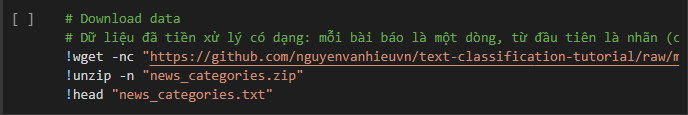
* + 1. Tổng kết các hàm sủ dụng trong tiền xử lý văn bản



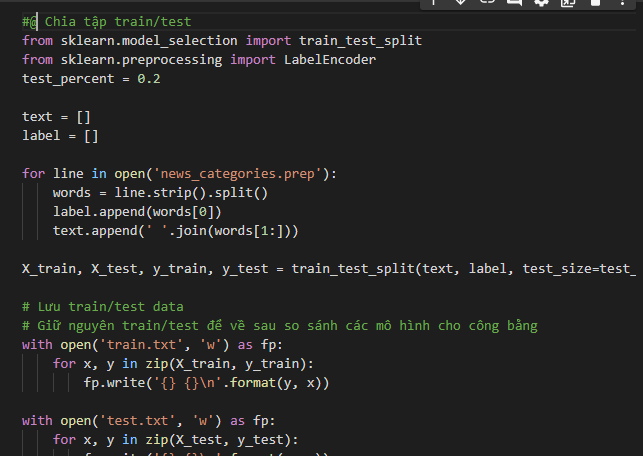
* + 1. Lưu file

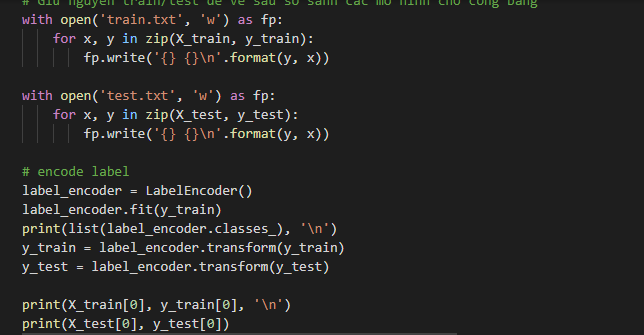


* 1. Naive Bayes trong phân loại văn bản
     1. Download dữ liệu

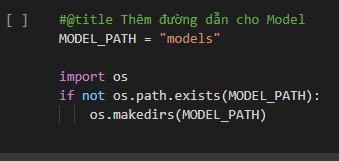


* + 1. Chia tập tin để train / test

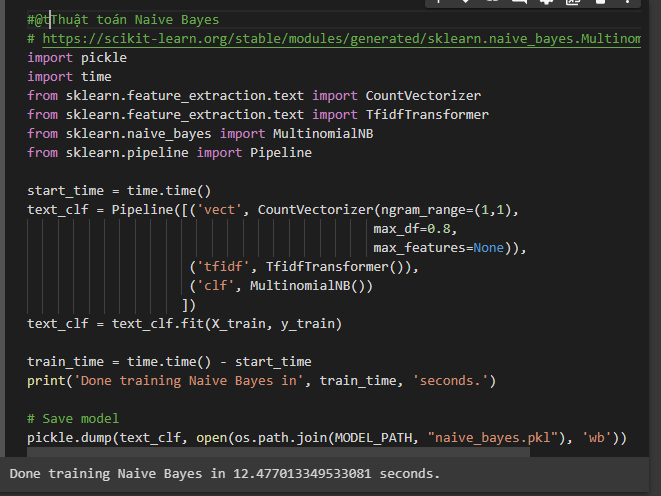




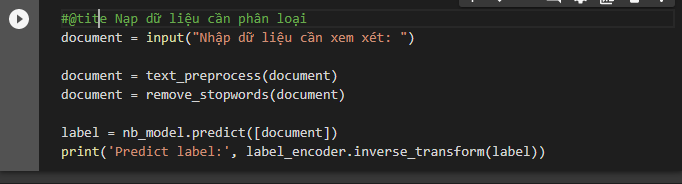
* + 1. Thêm đường dẫn



* + 1. Thuật toán Naive Bayes



* + 1. Đưa dữ liệu cần phân loại vào Model



# CHƯƠNG 4: TÀI LIỆU THAM KHẢO

* 1. Bài viết sử dụng các nguồn:
* <https://vi.wikipedia.org/>
* <https://pypi.org/>
* <https://github.com/>
  1. Data:
* https://github.com/nguyenvanhieuvn/text-classification-tutorial/raw/master/news\_categories.zip