



Bài giảng môn học:

Kỹ nghệ tri thức và học máy (7080510)

CHƯƠNG 3:

HỌC CÓ GIÁM SÁT – PHẦN 3

(Supervised Learning)

Giảng viên: Đặng Văn Nam

Email: dangvannam@hmg.edu.vn

3.3 Naïve Bayes

- a. Giới thiệu Xác suất có điều kiện Và thuật toán Naïve Bayes.
- b. Thực hành xây dựng mô hình phân lớp Comment độc (Toxic) – không độc (No Toxic)
- c. Bài thực hành số 8: Lọc email

1. Thuật toán Naïve Bayes

Thuật toán Naïve Bayes

- Naive Bayes Classifiers (NBC) là một trong những thuật toán tiêu biểu cho bài toán phân lớp dựa trên lý thuyết xác suất áp dụng định lý Bayes.
- Định lý Bayes cho phép chúng ta có thể tính toán một xác suất chưa biết dựa vào các xác suất có điều kiện khác. Với công thức tổng quát tính xác suất của biến cố A với điều kiện biến cố B_k xảy ra trước (hay được gọi là xác suất hậu nghiệm):
- Với $P(A) > 0$ và $\{B_1, B_2, \dots, B_n\}$ là một hệ đầy đủ các biến cố thỏa mãn tổng xác suất của hệ bằng 1 ($\sum_{k=1}^n P(B_k) = 1$) và từng đôi một xung khắc ($P(B_i \cap B_j) = 0$).

Naïve Bayes Classifier

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$



Thomas Bayes
1702 - 1761

Khi đó ta có:

Ví dụ công thức Naïve Bayes

- Một bệnh viện phải làm xét nghiệm với một số lượng lớn bệnh nhân và thấy rằng có 0.1% bị mắc bệnh còn 99.9% là khỏe. Để biết rằng việc xét nghiệm là đúng người ta cho một người khỏe xét nghiệm thì xác suất âm tính 99%. Nếu xét nghiệm trên một người bị bệnh thì xác suất dương tính là 98%. Chọn ngẫu nhiên một người và thấy rằng người này dương tính, tìm xác suất người này bị mắc bệnh?
- $+$ ($-$) là sự kiện người này dương (âm tính). $A(\bar{A})$ là người này khỏe (mắc bệnh).
- Theo dữ kiện, ta có:
 - $P(A) = 0.999$;
 - $P(\bar{A}) = 0.001$;
 - $P(+|A) = 0.01$;
 - $P(-|A) = 0.99$;
 - $P(+|\bar{A}) = 0.98$;
 - $P(-|\bar{A}) = 0.02$

$$P(\bar{A}|+) = \frac{P(+|\bar{A}) \cdot P(\bar{A})}{P(+)} = \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.01 \times 0.999} = 0.0893$$

Thuật toán Naïve Bayes

Trong công thức Naïve Bayes,
Giả thiết các chiều dữ liệu là
độc lập với nhau

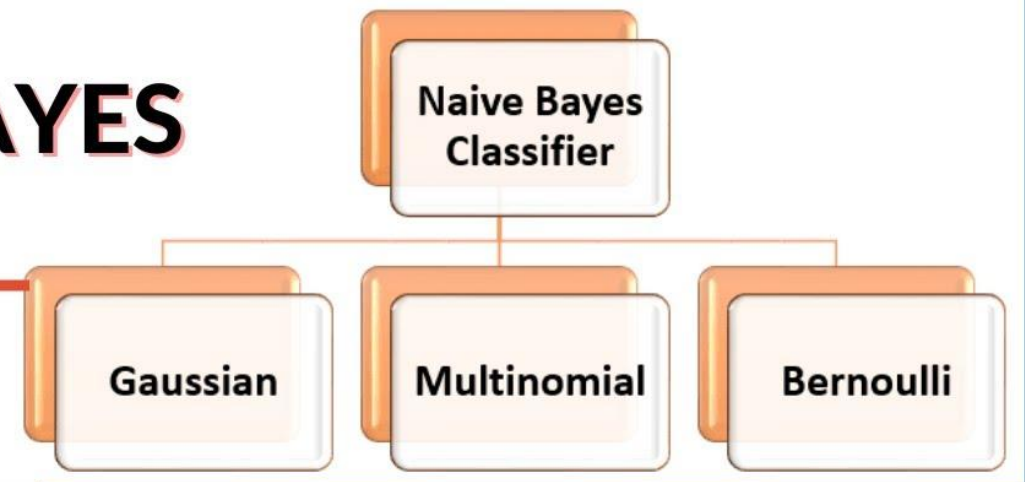
MACHINE LEARNING

NAIVE BAYES ALGORITHM

BIG DATA KNOWLEDGE
HUNT OFFICIAL



Types & Applications



Có 3 loại phân bố xác suất phổ biến là:

- Gaussian naïve bayes
- multinomial naïve bayes
- Bernoulli naïve bayes

Thuật toán Naïve Bayes

Guassian Naive Bayes....



sử dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục. Với mỗi chiều dữ liệu i và một nhãn c , x_i tuân theo một phân phối chuẩn có kỳ vọng μ và phương sai σ

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Thuật toán Naïve Bayes

Multinomial naïve bayes: Mô hình này được sử dụng chủ yếu bài toán phân loại văn bản và vector đặc trưng được xây dựng dựa trên ý tưởng Bag of Words (BoW)

Bernoulli naïve bayes: Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị nhị phân (0|1).

Multinomial Naïve Bayes Classifier

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c) P(c)$$

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c_j) \prod_{x \in X} P(x | c)$$

The Bernoulli distribution

$$p(x) = P[X = x] = \begin{cases} q = 1 - p & x = 0 \\ p & x = 1 \end{cases}$$

2. Ví dụ: Phân lớp comment trên Twitter sử dụng Naïve Bayes

Ứng dụng Naïve Bayes trong phân văn bản

- Xây dựng mô hình học máy phân lớp văn bản thành 2 lớp: 1-Toxic 0-Non Toxic sử dụng thuật toán Naïve Bayes

| | A | |
|----|-------|--|
| 1 | class | tweet |
| 2 | 0 | !!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should |
| 3 | 1 | !!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!! |
| 4 | 1 | !!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confus |
| 5 | 1 | !!!!!! RT @C_G_Anderson: @viva_based she look like a tranny |
| 6 | 1 | !!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told |
| 7 | 1 | !!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking |
| 8 | 1 | !!!!!!"@_BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!" |
| 9 | 1 | !!!!“@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!” |
| 10 | 1 | " & you might not get ya bitch back & thats that " |
| 11 | 1 | " @rhythmixx_ hobbies include: fighting Mariam" |
| 12 | 1 | " Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh |
| 13 | 1 | " Murda Gang bitch its Gang Land " |
| 14 | 1 | " So hoes that smoke are losers ? " yea ... go on IG |
| 15 | 1 | " bad bitches is the only thing that i like " |
| 16 | 1 | " bitch get up off me " |
| 17 | 1 | " bitch nigga miss me with it " |

Thông kê tập dữ liệu:

class 56700

tweet 56700

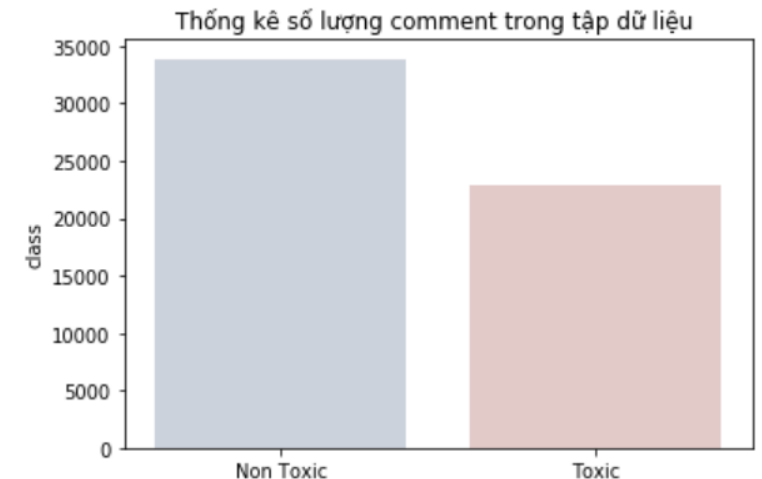
dtype: int64

Thông kê số lượng comment theo lớp:

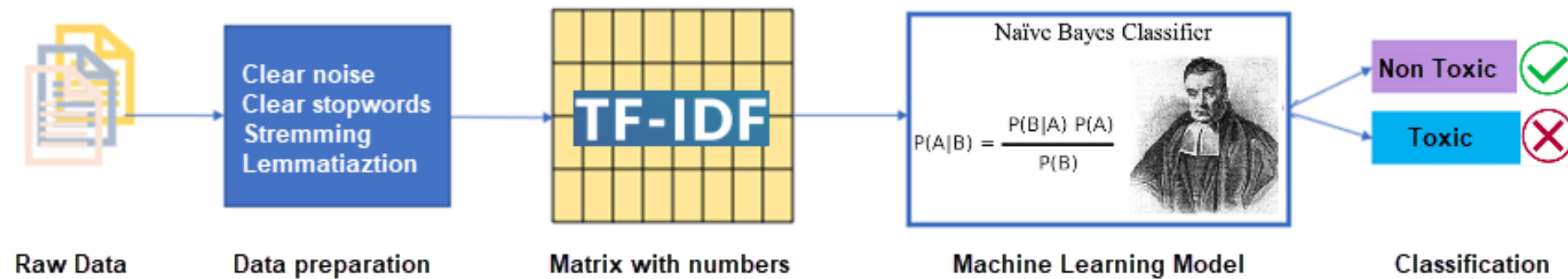
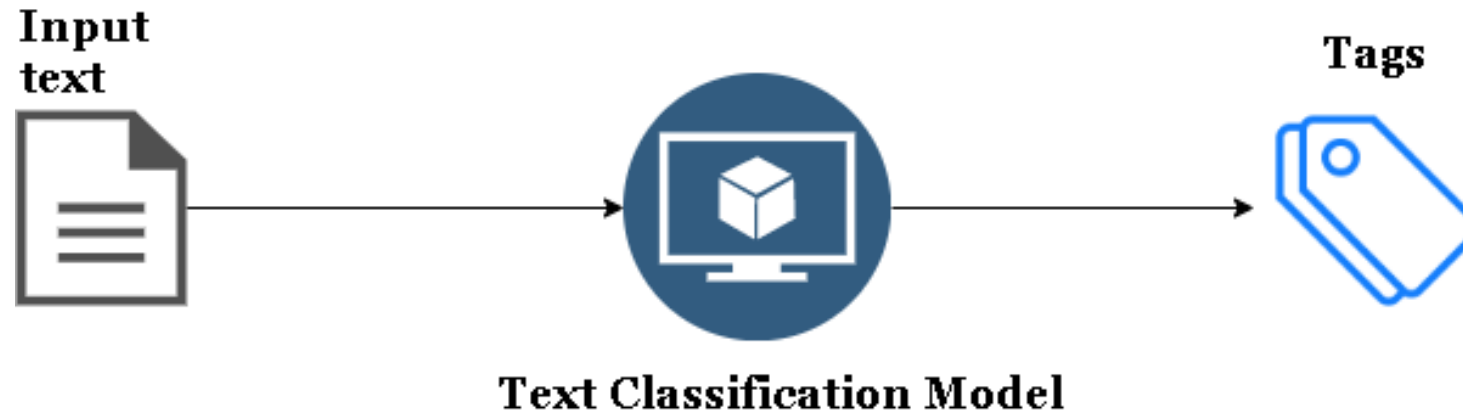
0 33847

1 22853

Name: class, dtype: int64



Ứng dụng Naïve Bayes trong phân văn bản



Ứng dụng Naïve Bayes trong phân văn bản



- 1. Chuẩn bị dữ liệu:

| STT | Comments ban đầu | Comments đã xử lý |
|-----|---|--|
| 1 | !!!!!!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking with hoes! 😂😂😂" | shit blow faithful somebody still fuck hoe |
| 2 | " got ya bitch tip toeing on my hardwood floors " 😂 http://t.co/cOU2WQ5L4q | get ya bitch tip toe hardwood floor |
| 3 | "@Dunderball: I'm an early bird and I'm a night owl, so I'm wise and have worms." | early bird night owl wise worm |
| 4 | "@Tmacc_GFG: “@VoiceOfDStreetz: "@Tmacc_GFG: “@tizzimarie: No slushes 😥”hoes nasty anyway fam"😴”them hoes taste like meds"🍇🍇🍼 | slush hoe nasty anyway fam hoe taste like meds |
| 5 | beÃ°ÄÿÄ'Ä" Ã°ÄÿÄ'Ä"@user @user @user @user @user @user @user @user @user @user | |

Ứng dụng Naïve Bayes trong phân văn bản



- 1. Chuẩn bị dữ liệu:

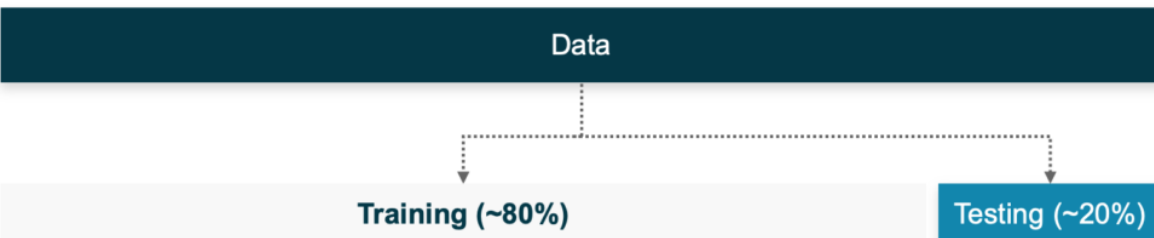
```
1 #Thống kê số row null
2 #trong tập dữ liệu sau xử lý:
3 data_finish.isnull().sum()
```

```
class      0
tweet_ok   43
```

```
1 #Thực hiện xóa tất cả các row có phần tử NaN
2 data_NLP = data_finish.dropna()
3 data_NLP.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 56657 entries, 0 to 56699
Data columns (total 2 columns):
#   Column      Non-Null Count  Dtype
---  ---
0    class      56657 non-null    int64
1    tweet_ok   56657 non-null    object
dtypes: int64(1), object(1)
memory usage: 1.3+ MB
```

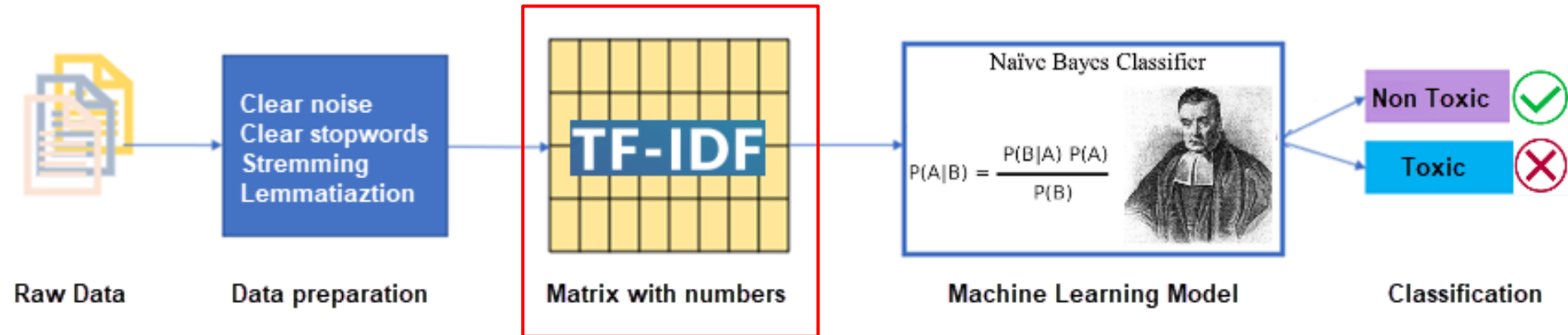
| class | tweet | tweet_ok |
|-------|--|--|
| 0 | !!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should always take the trash out... | woman complain clean house man always take trash |
| 1 | !!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!! | boy dats dwn bad cuffin dat hoe st place |
| 1 | !!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit | dawg ever fuck bitch start cry confuse shit |
| 1 | !!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny | look like tranny |
| 1 | !!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya  | shit hear might true might faker bitch tell ya |
| 1 | !!!!!!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking with hoes! 😂😂😂" | shit blow faithful somebody still fuck hoe |



Ứng dụng Naïve Bayes trong phân văn bản

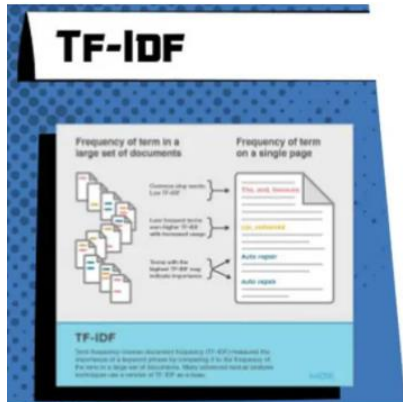


- 2. Trích chọn đặc trưng TF-IDF



Ứng dụng Naïve Bayes trong phân văn bản

- 2. Trích chọn đặc trưng TF-IDF



WHAT IS TF-IDF ?

- Thuật ngữ **TF-IDF (Term Frequency – Inverse Document Frequency)** là một phương thức thống kê được biết đến rộng rãi để xác định độ quan trọng của một từ trong đoạn văn bản trong một tập nhiều đoạn văn bản khác nhau.

- TF-IDF xác định trọng số của một từ trong văn bản 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. Giá trị TF-IDF cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu.

• 2. Trích chọn đặc trưng TF-IDF

- *TF (Term Frequency)* – Tần suất xuất hiện của từ, là số lần từ xuất hiện trong văn bản. Vì các văn bản có thể có độ dài ngắn khác nhau nên một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn là một văn bản ngắn

$$TF(t, d) = \frac{f(t, d)}{\max \{f(w, d) : w \in d\}}$$

Trong đó:

- $TF(t, d)$ - Tần suất xuất hiện của từ t trong văn bản d .
- $f(t, d)$ - Số lần xuất hiện của từ trong văn bản d .
- $\max\{f(w, d) : w \in d\}$ – Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d .

- *IDF (Inverse Document Frequency)* – Nghịch đảo tần suất của văn bản, giúp đánh giá tầm quan trọng của một từ. Khi tính tần số xuất hiện TF thì các từ đều được coi là quan trọng như nhau. Tuy nhiên có một số từ thường được sử dụng nhiều nhưng không quan trọng để thể hiện ý nghĩa của đoạn văn. Vì vậy ta cần giảm đi mức độ quan trọng của những từ đó bằng cách sử dụng IDF:

$$IDF(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Trong đó:

- $IDF(t, D)$ – Giá trị IDF của từ t trong tập văn bản D .
- $|D|$ - Tổng số văn bản trong tập D .
- $|\{d \in D : t \in d\}|$ – Thể hiện số văn bản trong tập D có chứa từ t .

Ứng dụng Naïve Bayes trong phân văn bản



- 2. Trích chọn đặc trưng TF-IDF

TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

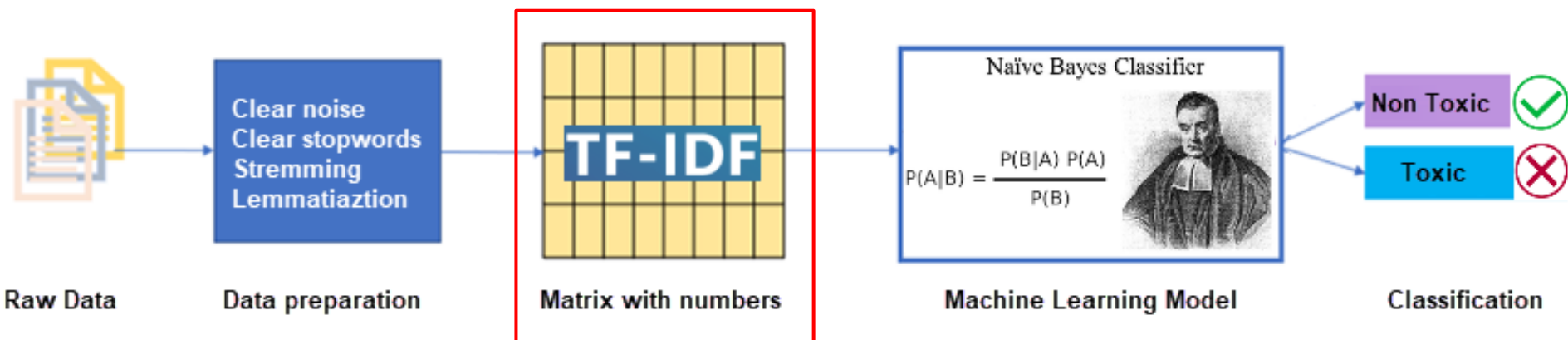
Number of times term t appears in a doc, d

Inverse document frequency

$$\log \frac{1 + \overset{\text{\# of documents}}{n}}{1 + \underset{\text{Document frequency of the term } t}{df(d, t)}} + 1$$

Ứng dụng Naïve Bayes trong phân văn bản

• 2. Trích chọn đặc trưng TF-IDF



```

1 # Tính TF-IDF cho tập dữ liệu
2 from sklearn.feature_extraction.text import TfidfVectorizer
3 #Convert a collection of raw documents to a matrix of TF-IDF features.
4 vector = TfidfVectorizer(analyzer='word',
5                           max_features=20000,
6                           stop_words = 'english')
7
8 vector.fit(data_NLP['tweet_ok'])
9 xtrain_tfidf = vector.transform(train_x)
10 xtest_tfidf = vector.transform(test_x)
11 print('Kết quả Vector hóa tập Train sang dạng số:')
12 print(xtrain_tfidf.data)
13 print(xtrain_tfidf.shape)
14 print('Kết quả Vector hóa tập Test sang dạng số:')
15 print(xtest_tfidf.data)
16 print(xtest_tfidf.shape)

```

Kết quả Vector hóa tập Train sang dạng số:

[0.54907788 0.53160745 0.64490852 ... 0.33670852 0.13308591 0.26849291]
(45325, 20000)

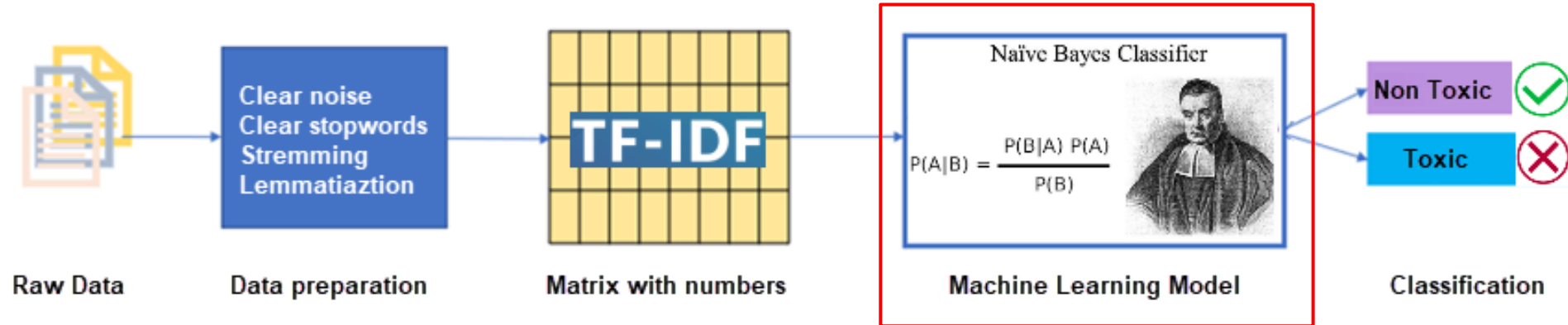
Kết quả Vector hóa tập Test sang dạng số:

[0.73386444 0.51887036 0.43842506 ... 0.58960476 0.39331306 0.52415607]
(11332, 20000)

Ứng dụng Naïve Bayes trong phân văn bản

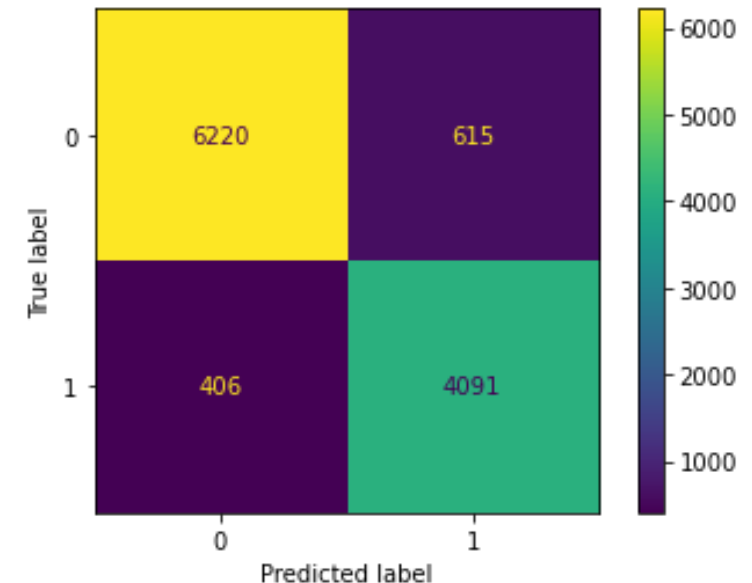


- 3. Naïve bayes text classification



```
1 from sklearn.metrics import accuracy_score
2 #Dự đoán và tính đoán độ chính xác của model trên tập Test:
3 y_pred = MultiNB.predict(X_test_tfidf)
4
5 acc1 = round(accuracy_score(y_test, y_pred)*100, 2)
6 print('1.Độ chính xác của mô hình trên tập Test: ', acc1, '%')
7
8 acc2 = accuracy_score(y_test, y_pred, normalize=False)
9 print('2.Tổng số mẫu dự đoán đúng:', acc2, ' /', len(y_test))
10 print('3.Tổng số mẫu dự đoán sai:', len(y_test) - acc2, ' /', len(y_test))
```

1.Độ chính xác của mô hình trên tập Test: 90.99 %
2.Tổng số mẫu dự đoán đúng: 10311 / 11332
3.Tổng số mẫu dự đoán sai: 1021 / 11332



Ứng dụng Naïve Bayes trong phân văn bản



- Tham khảo các bước chi tiết trong file jupyter notebook!

Chuong3_Phan03_NaiveBayes - Jupyter Notebook

localhost:8888/notebooks/Documents/7080510_MachineLearning/Chuong3/Chuong3_Phan03_NaiveBayes/Chuong3_Phan03_NaiveBayes.ipynb

jupyter Chuong3_Phan03_NaiveBayes Last Checkpoint: 17 minutes ago (autosaved) Logout


File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

A| Phân loại (Classification)

THUẬT TOÁN NAIVE BAYES

- Naive Bayes Classifiers (NBC) là một trong những thuật toán tiêu biểu cho bài toán phân lớp dựa trên lý thuyết xác suất áp dụng định lý Bayes.
- Định lý Bayes cho phép chúng ta có thể tính toán một xác suất chưa biết dựa vào các xác suất có điều kiện khác. Với công thức tổng quát tính xác suất của biến cố A với điều kiện biến cố Bk xảy ra trước (hay được gọi là xác suất hậu nghiệm):

Naïve Bayes Classifier

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$


Thomas Bayes
1702 - 1761

Trong công thức Naïve Bayes, Giả thiết các chiều dữ liệu là độc lập với nhau

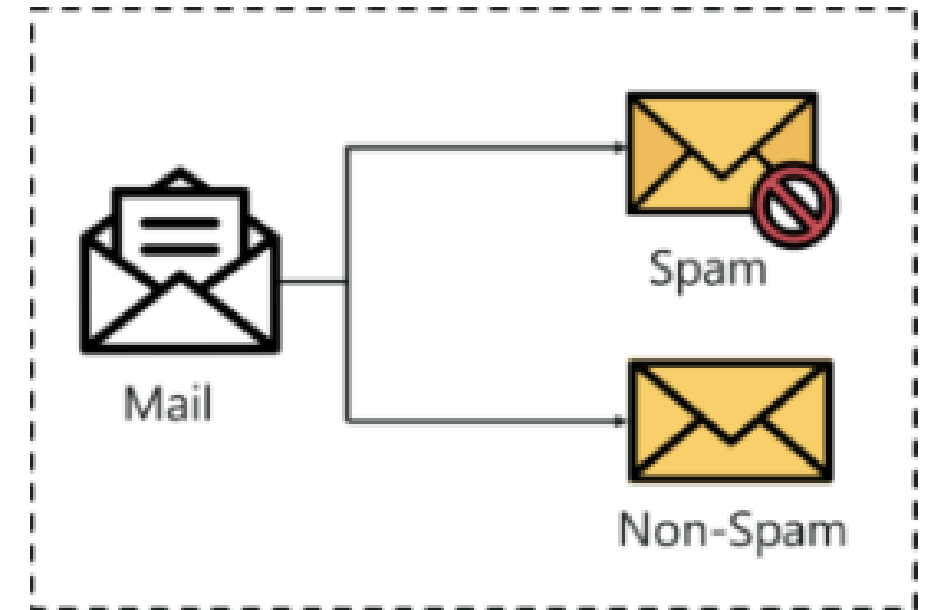
Có 3 loại phân bố xác suất phổ biến là:

THỰC HÀNH 8

Ví dụ: Phân loại Email

- Sinh viên tìm hiểu và thực hiện theo hướng dẫn của bài toán lọc Email mục 3.3 bài 32 trang học máy cơ bản của Vũ Hữu Tiệp

<https://machinelearningcoban.com/2017/08/08/nbc/>



- Sinh viên làm trên Jupyter Notebook



Thank you!