



Email Spam Classification Apps

นำเสนอโดย
นายภควันต์ ทุมดี 2010711102021
มหาวิทยาลัยหอการค้าไทย





วัตถุประสงค์

1. ตรวจสอบและกรองสแปม: วัตถุประสงค์หลักของ Email Spam Classification Apps คือการตรวจสอบและกรองอีเมลสแปมออกไปจากกล่องขาเข้าของผู้ใช้ ทำให้ผู้ใช้สามารถมั่นใจได้ว่าอีเมลที่พวกเขาได้รับเป็นเพียงข้อความที่สนใจและควรตอบกลับ
2. ควบคุมความเสี่ยง: การจัดหมวดหมู่สแปมอีเมลช่วยให้ผู้ใช้งานสามารถลดความเสี่ยงในการเปิดอีเมลที่อาจมีลิงก์หรือแนบไฟล์ที่อาจเป็นอันตราย เนื่องจากสแปมมักมีลิงก์หรือแนบไฟล์ที่อาจเป็นส่วนหนึ่งของการโจมตีด้านความปลอดภัย
3. ประหยัดเวลา: การจัดหมวดหมู่อีเมลสแปมช่วยประหยัดเวลาในการค้นหาข้อความสำคัญที่อาจถูกซ่อนอยู่ในอีเมลสแปม นอกจากนี้ยังช่วยลดเวลาในการลบหรือลบอีเมลสแปมที่ไม่ต้องการและสามารถให้ผู้ใช้ทำการอื่นๆ ได้อย่างมีประสิทธิภาพ
4. เพิ่มความน่าเชื่อถือในการใช้อีเมล: การใช้ Email Spam Classification Apps ช่วยเพิ่มความน่าเชื่อถือในการใช้งานอีเมล เนื่องจากผู้ใช้สามารถมั่นใจได้ว่าเฉพาะอีเมลที่มีความสำคัญและควรตอบกลับเท่านั้นที่อยู่ในกล่องขาเข้าของพวกเขา
5. พัฒนาความสามารถในการแยกประเภท: Email Spam Classification Apps ช่วยพัฒนาความสามารถในการแยกประเภทของระบบปฏิบัติการในการจัดหมวดหมู่สแปมอีเมล ซึ่งอาจถูกนำไปใช้ในอย่างอื่นๆ เช่น ส่งข้อความแจ้งเตือนหรือระบบการแจ้งเตือนของอื่นๆ

การรวบรวมข้อมูล

class	message
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
spam	URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
ham	I HAVE A DATE ON SUNDAY WITH WILL!!
spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap.xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
ham	Oh k...i'm watching here:)
ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham	Fine if thatåÕs the way u feel. ThatåÕs the way its gota b
spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/!%1.20 POBOXox36504W45WQ 16+
ham	Is that seriously how you spell his name?
ham	I%0Bm going to try for 2 months ha ha only joking
ham	So ò_ pay first lar... Then when is da stock comin...
ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?

โดยข้อมูลจะมี 2 รูปแบบ

- class ที่แบ่งแยกระหว่าง ham กับ spam โดย ham คือข้อความปกติและ spam มักเป็นอีเมลที่ส่งโฆษณาข้อความขายของ
- message บอกถึง ลักษณะของข้อความ หรือ รูปแบบข้อความ

ประมวลผลข้อมูล



```
In [1]: import pandas as pd
```

```
In [2]: data=pd.read_csv("spam.csv" , encoding="latin-1")
```

```
In [3]: data.head(5)
```

```
Out[3]:
```

	class	message	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

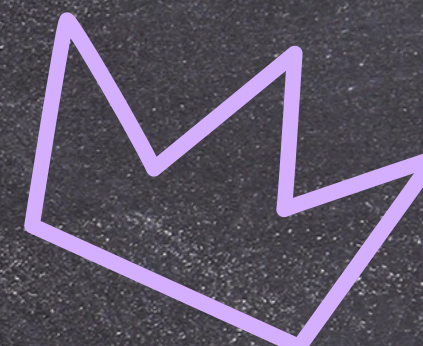
```
In [4]: data.columns
```

```
Out[4]: Index(['class', 'message', 'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], dtype='object')
```

```
In [5]: data.drop(['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'] , axis=1 ,inplace=True)
```

```
In [6]: data.head()
```

อ่านข้อมูลจากไฟล์ CSV และเก็บข้อมูลที่อ่านได้ในรูปแบบของ DataFrame โดยใช้ pandas



```
In [7]: data['class']=data['class'].map({'ham':0 , 'spam':1})

In [8]: data.head()

Out[8]:
```

	class	message
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

ให้พิจารณา DataFrame ที่มีคอลัมน์ชื่อ "class" ซึ่งเก็บข้อมูลเป็นข้อความ (string) ที่บ่งบอกถึงประเภทของอีเมลว่าเป็น "ham" หรือ "spam" ซึ่งเป็นข้อมูลแบบหมวดหมู่ (categorical data) คำสั่งดังกล่าวจะทำการแปลงค่าในคอลัมน์ "class" จากข้อความเป็นตัวเลขดังนี้:

- คำว่า "ham" จะถูกแทนที่ด้วยตัวเลข 0
- คำว่า "spam" จะถูกแทนที่ด้วยตัวเลข 1





```
In [9]: from sklearn.feature_extraction.text import CountVectorizer
```

```
In [10]: cv=CountVectorizer()
```

```
In [11]: x=data['message']|  
         y=data['class']
```

1

2

3

from sklearn.feature_extraction.text import CountVectorizer: คำสั่งนี้เป็นการนำเข้าคลาส CountVectorizer จากโมดูล sklearn.feature_extraction.text ซึ่งเป็นอุปกรณ์ในการแปลงข้อความให้อยู่ในรูปของ Word Count Vectors ที่ใช้ในงานประมวลผลภาษาธรรมชาติ (Natural Language Processing, NLP) และการสร้างแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning Model) ที่ใช้ข้อมูลข้อความ

- ตัวแปร x เพื่อเก็บข้อมูลที่อยู่ในคอลัมน์ "message" ของ
- ตัวแปร y เพื่อเก็บข้อมูลที่อยู่ในคอลัมน์ "class" ของ DataFrame data ซึ่งเป็นข้อมูลเป้าหมาย (target) ที่บ่งบอกถึงประเภทของอีเมลว่าเป็น "ham" หรือ "spam" ซึ่งจะใช้ในกระบวนการฝึกสอนแบบจำลองการเรียนรู้ของเครื่อง



```
In [16]: from sklearn.model_selection import train_test_split
```

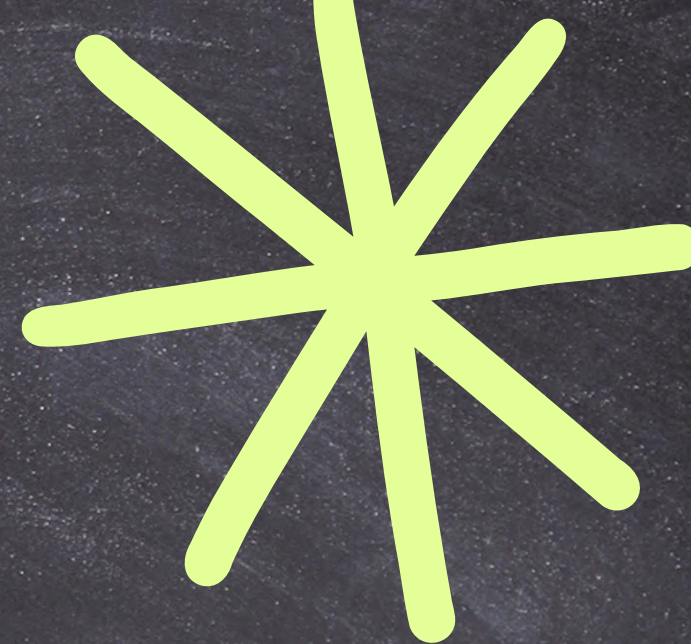
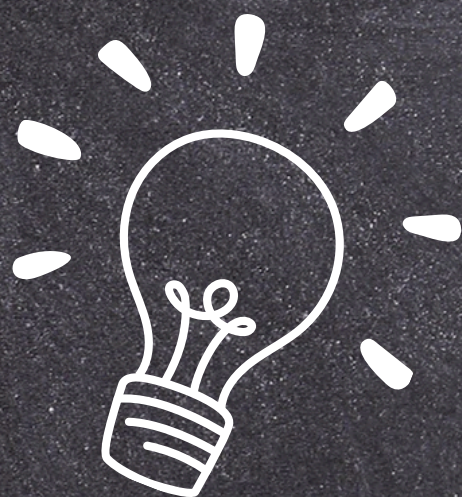
```
In [17]: x_train, x_test, y_train, y_test=train_test_split(x, y, test_size=0.2)
```

```
In [18]: x_train.shape
```



`train_test_split` เป็นฟังก์ชันที่ใช้ในการแบ่งชุดข้อมูลเป็นชุดฝึกสอน (training set) และชุดทดสอบ (test set) เพื่อใช้ในระบบการฝึกสอนและทดสอบแบบจำลอง (Machine Learning Model) ในการประมวลผลข้อมูล

- `X`: เป็นชุดข้อมูลฟีเจอร์ (features) ที่ต้องการใช้ในระบบการฝึกสอนและทดสอบ ซึ่งส่วนใหญ่จะเป็นเมตริกซ์หรือ `DataFrame` ซึ่งเก็บข้อมูลที่ใช้ในการฝึกสอนและทดสอบ
- `y`: เป็นชุดข้อมูลเป้าหมาย (target) ที่ต้องการใช้ในระบบการฝึกสอนและทดสอบ ซึ่งส่วนใหญ่จะเป็นเวกเตอร์หรือ `Series` ซึ่งเก็บข้อมูลเป้าหมายที่เกี่ยวข้องกับการฝึกสอนและทดสอบ
- `test_size`: เป็นอัตราส่วนของชุดข้อมูลที่จะถูกแบ่งออกมาเป็นชุดทดสอบ ซึ่งค่าเป็นค่าที่อยู่ในช่วง `[0, 1]` โดยทั่วไปจะกำหนดให้เป็นค่าระหว่าง 0.2 ถึง 0.3 เพื่อให้ชุดทดสอบมีขนาดประมาณ 20% ถึง 30% ของข้อมูลทั้งหมด
- `random_state`: เป็นค่าที่กำหนดเพื่อให้ผลการแบ่งชุดข้อมูลเป็นชุดฝึกสอนและชุดทดสอบมีความน่าเชื่อถือและให้ผลลัพธ์ที่เหมือนกันทุกครั้งที่ยกใช้งาน ค่า `random_state` จะมักใช้เป็นตัวเลขเพื่อกำหนดการสุ่มในระบบการแบ่งข้อมูล

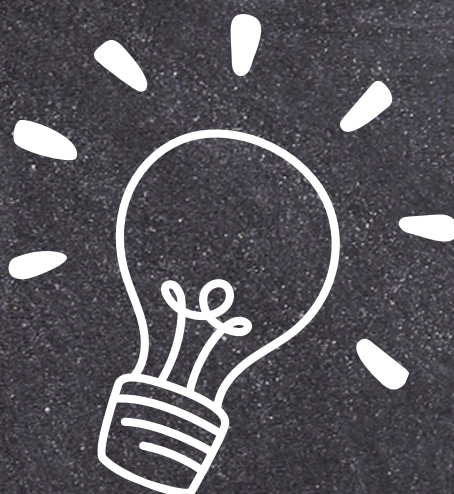


```
In [19]: from sklearn.naive_bayes import MultinomialNB
```

```
In [20]: model=MultinomialNB()
```

```
In [21]: model.fit(x_train, y_train)
```

- ในโค้ดดังกล่าว เราใช้คำสั่ง MultinomialNB จาก scikit-learn เพื่อสร้างแบบจำลอง Multinomial Naive Bayes ซึ่งเป็นอัลกอริทึมสำหรับการจำแนกหมวดหมู่ (classification) ที่ใช้ในงาน NLP (Natural Language Processing) และมีการนำมาใช้ในปัญหาที่เกี่ยวข้องกับข้อความ



```
In [22]: result=model.score (x_test , y_test)
```

```
In [23]: result=result*100
```

```
In [24]: result
```

```
Out[24]: 98.7443946188341
```

ในโค้ดดังกล่าว เราใช้ `model.score` เพื่อคำนวณคะแนน (score) ของแบบจำลองที่สร้างด้วย Multinomial Naive Bayes โดยใช้ชุดข้อมูลทดสอบ `x_test` และ `y_test` เพื่อทดสอบประสิทธิภาพของแบบจำลองในการทำนายหมวดหมู่ข้อความที่ไม่เคยเห็นมาก่อน





สร้างแอปพลิเคชันเว็บโดยใช้ Streamlit



```
spamDetector.py
1 import pickle
2 import streamlit as st
3
4
5 model=pickle.load(open("spam.pkl", "rb"))
6 cv=pickle.load(open("vectorizer.pkl", "rb"))
7
8
9 def main():
10     st.title("Email Spam Classification Apps")
11     st.subheader("Build With Steamlit & Python")
12     msg=st.text_input("Enter a Text:")
13     if st.button("Predict"):
14         data=[msg]
15         vect=cv.transform(data).toarray()
16         prediction=model.predict(vect)
17         result=prediction[0]
18         if result==1:
19             st.error("This is a spam mail")
20         else:
21             st.success("This is a ham mail")
22
23 main()
24
```

Email Spam Classification Apps


Build With Steamlit & Python

Enter a Text:

Free 10 taka

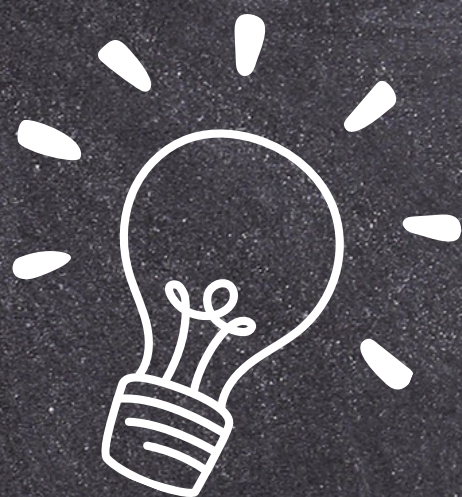
Predict

This is a spam mail





สรุป



มีการใช้ Naive Bayes เป็นอัลกอริทึมที่ใช้ในงานการจำแนกหมวดหมู่ (classification) ในการประมวลผลข้อมูล อัลกอริทึมนี้นิยมใช้ในงานที่เกี่ยวข้องกับปัญหาที่เป็นข้อความหรือข้อมูลที่เป็นเวกเตอร์ (vector data) เช่น การจำแนกอีเมลเป็น "spam" หรือ "ham" (อีเมลที่ไม่ใช่ข้อความขยะและอีเมลที่ไม่ใช่ข้อความขยะ)

ข้อควรทราบเกี่ยวกับ Naive Bayes:

Naive Bayes ถูกเรียก "Naive" เนื่องจากเมื่อนำเสนอแนวคิดของ Bayes' Theorem ในลักษณะของการจำแนกหมวดหมู่ อัลกอริทึมจะถือว่าแต่ละคุณลักษณะ (feature) ของข้อมูล (เช่น คำ, คุณลักษณะของเวกเตอร์) มีความสำคัญแยกตัวอย่างกันเป็นอิสระ นั่นหมายความว่าคุณลักษณะแต่ละตัวจะไม่มีผลกระทบต่อกันคุณลักษณะอื่นๆ ในกระบวนการจำแนก ซึ่งเป็นการประมาณค่าเพื่อให้กระบวนการคำนวณนั้นง่ายและเร็วขึ้น แต่อาจส่งผลให้มีความแม่นยำลดลงในบางกรณี (แต่ก็ยังคงเป็นอัลกอริทึมที่มีประสิทธิภาพและใช้งานได้ดีในหลายสถานการณ์)

THANK

YOU

