

DETECTING SPAM: EMAIL CLASSIFICATION MODEL

OBJECTIVE

1. เพื่อศึกษาและวิเคราะห์ความสัมพันธ์ระหว่างคำที่มักปรากฏร่วมกันในอีเมลสแปมโดยใช้เทคนิค **Association Rule Mining**
2. เพื่อจัดทำ **Classification Model** สำหรับจำแนกอีเมลสแปม และอีเมลปกติ

RAW DATA

ounce feather bowl hummingbird opec moment
alabaster valkyrie dyad bread flack desperate iambic
hadron heft quell yoghurt bunkmate divert afterimage

DATA PREPROCESSING

1. Cleaning (การกำกับความสะอาดข้อความ): แปลงข้อความทั้งหมดเป็นตัวพิมพ์เล็ก (lowercase) ลบอักขระพิเศษ เช่น เครื่องหมายวรรคตอน
2. Tokenization (การแยกคำ): แยกข้อความเป็นคำ (tokens) โดยใช้ word_tokenize
3. Stopwords Removal (การลบคำที่ไม่สำคัญ): ดาวน์โหลด และลบ stopwords ภาษาอังกฤษ เช่น "the", "is", "and"
4. Stemming (การลดรูปคำ): ใช้ PorterStemmer เพื่อตัดคำให้เหลือ รูปแบบราก เช่น "running" → "run"
5. Lemmatization (การทำให้อยู่ในรูปแบบพจนานุกรม): ใช้ WordNetLemmatizer เพื่อแปลงคำให้อยู่ในรูปพจนานุกรม เช่น "better" → "good"
6. ลบตัวเลข: ลบตัวเลขทั้งหมดออกจาก lemmatized_tokens
7. ลบคำสั้น ๆ: ลบคำที่มีความยาวน้อยกว่าหรือเท่ากับ 3 ตัวอักษร
8. แปลง String เป็น List: ใช้ ast.literal_eval เพื่อแปลง string ที่เหมือน list ให้เป็น list จริง ๆ
9. ลบค่าเว้นว่าง (" "): ลบคำที่เป็นช่องว่างหรือว่างเปล่าออกจากลิสต์
10. เพิ่ม email_id: เพิ่มคอลัมน์ email_id โดยให้ค่าเป็นลำดับเริ่มจาก

CLEANED DATA

```
['ounce', 'feather', 'bowl', 'hummingbird', 'opec', 'moment',  
 'alabaster', 'valkyrie', 'dyad', 'bread', 'flack', 'desperate', 'iambic',  
 'hadron', 'heft', 'quell', 'yoghurt', 'bunkmate', 'divert', 'afterimage']
```

ASSOCIATION

word 1	word 2	Support	Confidence price → quality	Confidence quality → price	Lift
price	quality	8.16%	32.43%	61.93%	2.46

- คำว่า "price" และ "quality" มักปรากฏร่วมกันในอีเมลสแปมที่มีเนื้อหาเกี่ยวกับข้อเสนอพิเศษหรือโปรโมชั่น
- เมื่อพับคำว่า "quality" ในอีเมล โอกาสที่จะเจอคำว่า "price" ก็สูงถึง 61.93%
- ความสัมพันธ์นี้แข็งแรงพอที่จะใช้เป็นหนึ่งในตัวบ่งชี้ว่าอีเมลนั้นอาจเป็นสแปม

EMAIL CLASSIFICATION MODEL

- คำนวณ Term Frequency (TF): วัดความถี่ของคำในเอกสารโดยการหารจำนวนครั้งที่คำปรากฏในเอกสารด้วยจำนวนคำทั้งหมดในเอกสารนั้น
- คำนวณ Inverse Document Frequency (IDF): วัดความสำคัญของคำในเอกสารทั้งหมดโดยการหารจำนวนเอกสารทั้งหมดด้วยจำนวนเอกสารที่คำนี้ปรากฏ
- คำนวณ TF-IDF: คูณค่า TF และ IDF ของคำแต่ละคำเพื่อหาค่า TF-IDF ซึ่งแสดงถึงความสำคัญของคำในเอกสาร
- ใช้ TfidfVectorizer: ใช้ฟังก์ชัน TfidfVectorizer เพื่อแปลงข้อมูลข้อความเป็นเวกเตอร์โดยเลือกจำนวนคำ 1000 คำที่มีค่า TF-IDF สูงสุด
- แปลงข้อมูล: ใช้ fit_transform() สำหรับชุดข้อมูลฝึก (Training Data) และ transform() สำหรับชุดข้อมูลทดสอบ (Test Data) เพื่อแปลงข้อความเป็นเวกเตอร์ที่ใช้ในโมเดล Machine Learning

TF-IDF VECTORIZER

	able	access	according	account	acrobat	across	action	activity	\
0	0.00000	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0
1	0.00000	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0
2	0.09056	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0
3	0.00000	0.0	0.120917	0.0	0.0	0.0	0.0	0.0	0.0
4	0.00000	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0

	actually	added	...	wrote	xescapenumber	yahoo	year	yesterday	\
0	0.0	0.0	...	0.00000	0.0	0.0	0.00000	0.0	0.0
1	0.0	0.0	...	0.00000	0.0	0.0	0.00000	0.0	0.0
2	0.0	0.0	...	0.073743	0.0	0.0	0.00000	0.0	0.0
3	0.0	0.0	...	0.00000	0.0	0.0	0.082863	0.0	0.0
4	0.0	0.0	...	0.00000	0.0	0.0	0.00000	0.0	0.0

	york	youll	young	youre	zero
0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0

1 vectorizer

TfidfVectorizer 
TfidfVectorizer(max_features=1000)

EMAIL CLASSIFICATION MODEL

มีการใช้ 3 โมเดล สำหรับการจำแนกประเภทอีเมล (เช่น สแปมหรือไม่) ได้แก่:

1. Decision Tree Classifier
2. Random Forest Classifier
3. Multinomial Naive Bayes

โดยการประเมินประสิทธิภาพของโมเดลเหล่านี้จะดำเนินการผ่าน 5-fold Cross-Validation ซึ่งเป็นกระบวนการที่แบ่งข้อมูลฝึก (X_train_tfidf และ y_train) ออกเป็น 5 ชุด เพื่อให้แต่ละโมเดลได้รับการทดสอบในแต่ละชุดข้อมูล และคำนวณผลการคำนายนายจากแต่ละชุด หลังจากนั้นผลลัพธ์จะถูกรวบรวมและคำนวณค่าเฉลี่ยสำหรับ KPI ต่างๆ เช่น Accuracy, Precision, Recall, F1 Score, และ ROC-AUC เพื่อช่วยเปรียบเทียบประสิทธิภาพของแต่ละโมเดล

MODEL SELECTION

Model	Mean Accuracy	Mean Precision	Mean Recall	Mean F1 Score	Mean ROC-AUC
Decision Tree	0.9376	0.9376	0.9375	0.9375	0.9375
Naive Bayes	0.9333	0.9337	0.9328	0.9331	0.9329
Random Forest	0.9713	0.9716	0.9709	0.9712	0.9709

Random Forest ถูกเลือกเป็นโมเดลที่ดีที่สุดเนื่องจากมีค่า KPI ที่ดีที่สุดในทุกด้าน โดยมีค่า Mean Accuracy สูงสุดที่ 0.9713, Precision 0.9716, Recall 0.9709, F1 Score 0.9712, และ ROC-AUC 0.9709 ซึ่งแสดงให้เห็นว่า โมเดลนี้สามารถทำนายได้แม่นยำและมีความสามารถในการจับอีเมล์สแปมได้ดี ก็ในแง่ของความแม่นยำ การแยกแยะ และการประเมินผลที่สมดุลระหว่าง Precision และ Recall, จึงทำให้เป็นโมเดลที่เหมาะสมที่สุดในการจำแนกอีเมล์ สแปม

DEPLOYMENT

Congratulations! You've won a free iPhone. Claim now! # Spam

Email 1: Congratulations! You've won a free iPhone. Claim now!

Prediction: Spam 

Meeting reminder: Project discussion at 3 PM tomorrow.# Not Spam

Email 2: Meeting reminder: Project discussion at 3 PM tomorrow.

Prediction: Not Spam 

URGENT: Your account has been compromised! Reset password immediately.

Spam

Email 3: URGENT: Your account has been compromised! Reset password immediately.

Prediction: Not Spam 

DEPLOYMENT

Please find the attached report for your review. # Not Spam

Email 4: Please find the attached report for your review.

Prediction: Not Spam 

Limited-time offer! Get 50% off on all products. # Spam

Email 5: Limited-time offer! Get 50% off on all products.

Prediction: Spam 

MEMBERS

นายจิรภัตร แสงสุกใส	653020203-5
นางสาวเพ็ญนา แก้วมูลเมือง	653020215-8
นายอรัญชัย แสนเทพ	653020219-0
นางสาวรุ่นกรรณ์ ดาษดัน	653020570-8