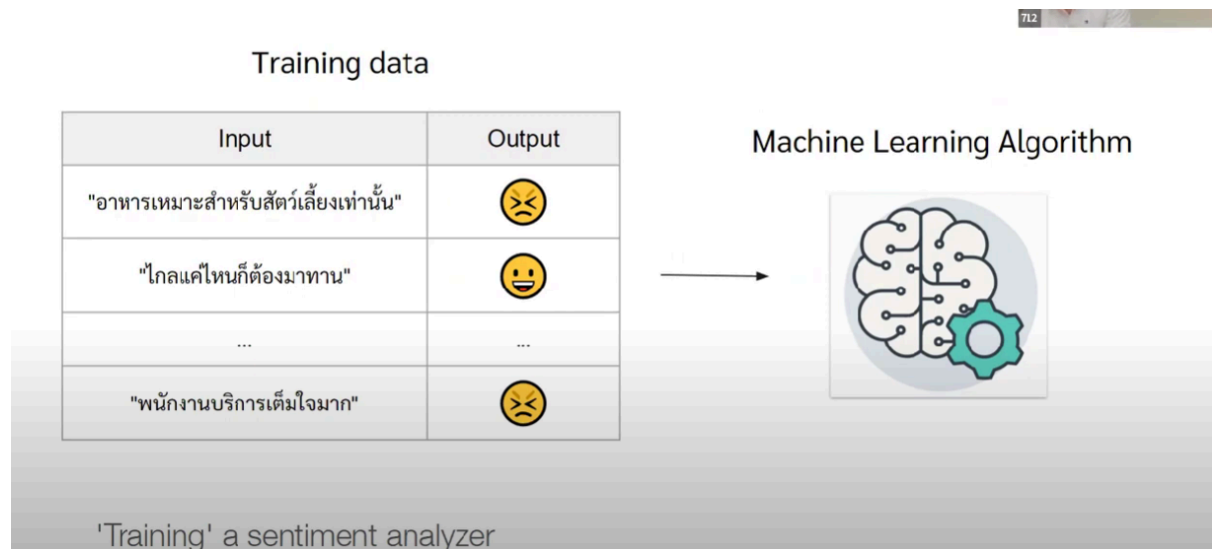


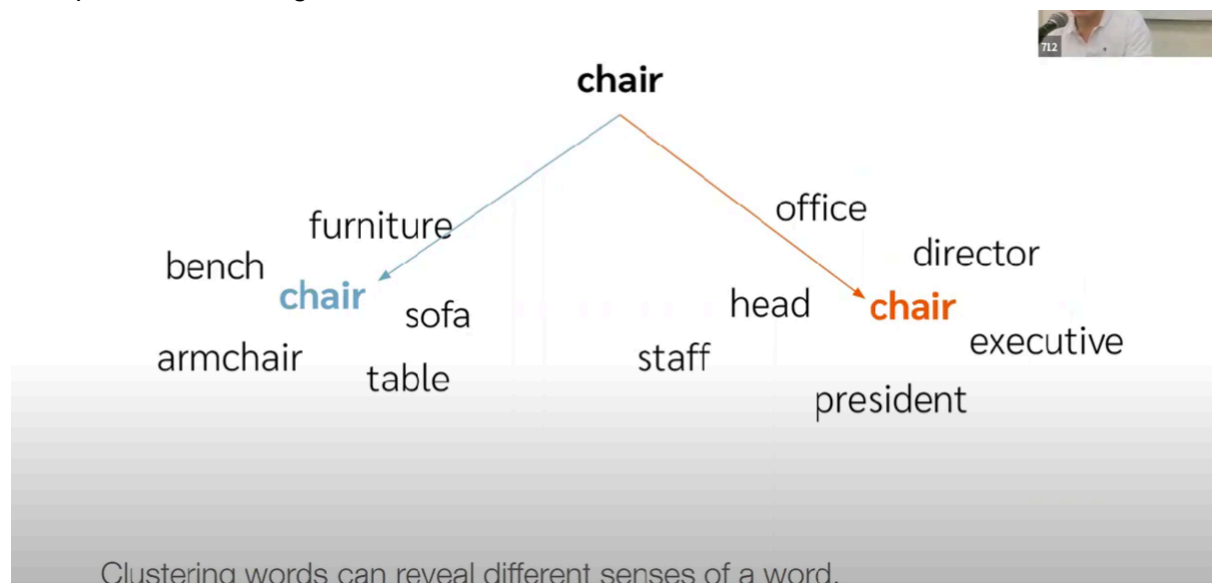
## Day 2 Text Classification with Naive Bayes

Supervised Learning with Text



Train = การทำให้เครื่องเรียนรู้จากข้อมูล

Unsupervised Learning with text

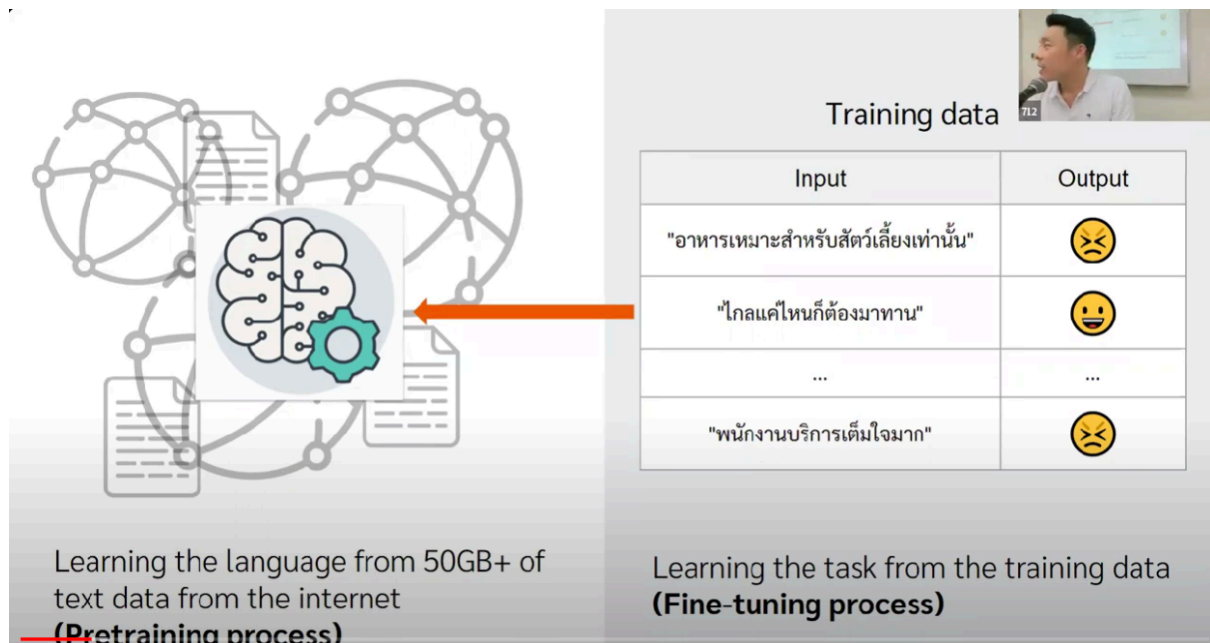


สามารถ Group ตามความหมายได้ เช่น Chair = ประธานก็ได้

เนื้อหาที่เรียนจาก

วิชา Computational Linguistics ปี 2023 สอนโดย รศ. ดร.อรรถพล ธำรงรัตนฤทธิ์ ภาควิชา  
ภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และ Program Director of Data Science True Digital  
Academy

## Transfer Learning + find Tuning



การทำให้โมเดลเรียนรู้ภาษาก่อน เยอะมาก ประมาณ 50 Gb จากนั้น ค่อยมาเทรนให้เฉพาะเจาะจงอีกทีนึง เรียกว่า Find tuning

Such as Chat GPT that learning large data and then they transfer with input by user's question

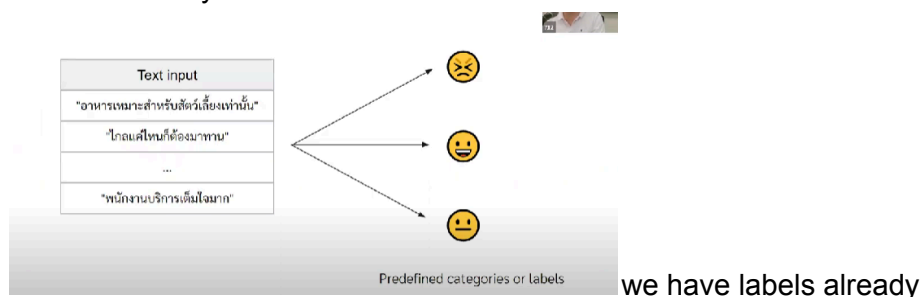
## Text Classification

เรามี ข้อมูล มา เราอยากจะแปะ เลเบล ให้มัน ต้องรู้ว่า เลเบล มีอะไรบ้าง

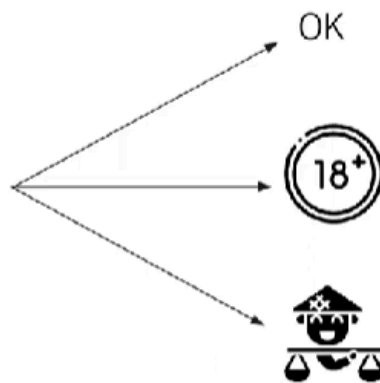
Text classification is the task of assigning predefined categories or labels(output) to given piece of text(input), which can be a sentence, a document, or a set of documents.

การกำหนดหมวดหมู่หรือป้ายกำกับ (เอาต์พุต) ที่กำหนดไว้ล่วงหน้าให้กับที่กำหนดขึ้นส่วนของข้อความ (อินพุต)

ทำได้หลายอย่าง เช่น  
Sentiment analysis

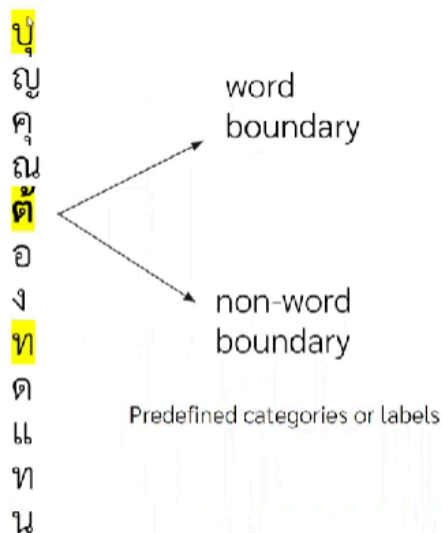


Text input
"ฉันชอบผลงานมานานแล้วค่ะ"
"เงินกู้ตัวนี้ ไม่ต้องคำ ดอกเบี้ยต่ำ คลิกเลย"
"ถ้าจะเดินแค่นี้ กลับบ้านเหอะ เช็ง"
...
"เพลงน่าจะซัดกว่านี้ค่ะ แต่เดินเปะนาก"



Predefined categories or labels

## Spam classification



Word segmentation result

บุญคุณ|ต้อง|ทดแทน

หาจุด หรือ คำนำหน้า

## Other examples

### Automatic Essay Grading

- Input: essay text
- Output: {A, B, C, D}

### Fake news detection

- Input: news text
- Output: fake or not-fake

### Language Identification

- Input: text
- Output: {FR, EN, TH, ZH, ...}

### Intent classification

- Input: utterance
- Output: {apology, request, complaint, ... }

## Type of Classifiers

Rule-based classifier เขียนกฎขึ้นมา ในรูปแบบ keywords ก็ได้ regular expressions ก็ได้  
ML-based classifier (use machine learning)

### Rule-base classifier

A classifier that relies on a set of predefined rules to assign labels to text such as keywords , regular expressions, and lexicons.

A lexicon (or dictionary or word list) is a collection of words and/or phrases optionally along with their meanings or attributes .

ex การทำ spam classifier (easy way)

## Spam Classifier - keyword-based classifier

```
def spam_classify(text):  
    keywords = {'viagra', 'lottery', 'free money'}  
    for word in text:  
        if word in key_words:  
            return 'SPAM'  
    return 'NOT SPAM'
```

must be preprocess ex. tokenize stop word

## Email classifier -- regex-based

```
def email_classify(token):  
    pattern = re.compile('^[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}$')  
    return pattern.match(token)
```

^ return (true false)

หาชื่อเฉพาะ

## Name recognition - Lexicon-based

```
def recognize_name(text):  
    # Where do we get the list of names from?  
    name_list = read_lexicon_from_file('my_name_lexicon.txt')  
    if token in name_list:  
        return 'NAME'  
    return 'NOT NAME'
```

Pros and cons of rule-based classifier

Pros → Simplicity , Transparency, Scalability , Domain specificity  
เขียนง่าย มองเห็นเลยว่าทำอะไร ใช้กับDataใหญ่ยังเร็ว ปรับแต่งได้ง่าย

Cons → Limited performance , Sensitive to noise , Labor intensive  
ชื่ออาจจะสับสน เช่น ชื่อคนกับชื่อถนน อาจเหมือนกัน และ ต้องมานั่งดูกฎเองว่าจะออกแบบอย่างไร

---

## ML - based classifier

An ML-based classifier rely on supervised learning methouds that lean form labeled text data to classify new unseen text data.

#Popular ML models

Naive Bayes

Logistic Regression

Deep learning models

4 Setp for supervised learning

1,Data preparation - 2.Feature engineering - 3.Model training - 4.Evaluation

### 1.Data preparation

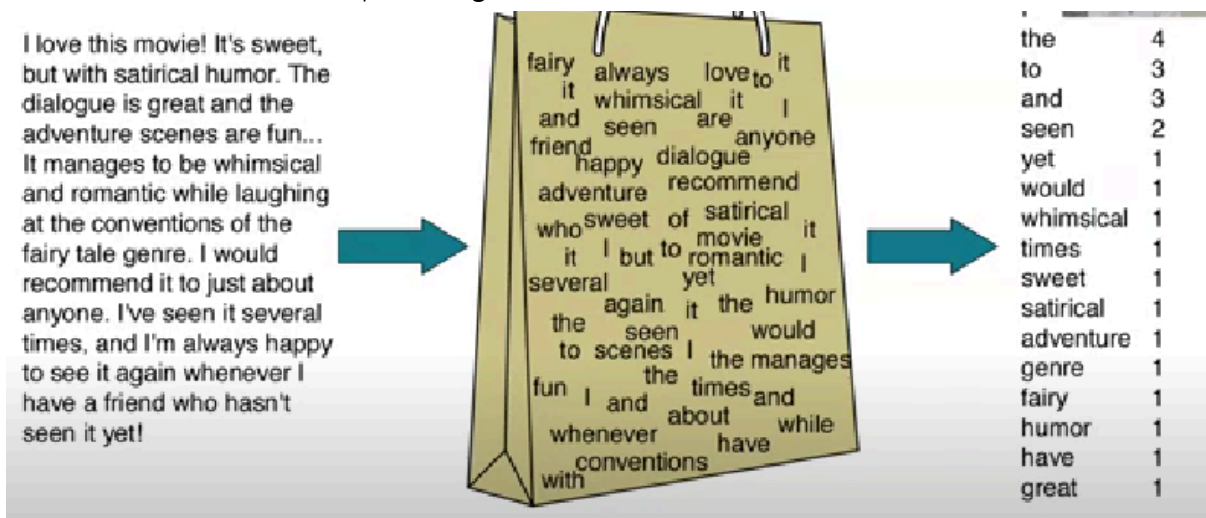
การเลือก sources and **Defining instances** and defining labels  
จะตัดคำ หรือจะใช้ทั้งประโยค

Data Annotation การแปะ labels เอง ตัวอย่าง ในข้อมูล twiter จะไม่มีคำเฉลยให้

## 2.Feature engineering

คือการเปลี่ยน string ให้เป็นตัวเลข

วิธีการที่ง่ายสุดคือ bag of word fature!



that after preprocess such as tokenize stopword !

คำที่เกิดขึ้นบ่อย อาจจะส่งผลกับ labels ของเรา

นับคำอย่างเดียว ว่าประโยคนี มีคำว่า อะไรบ้าง

## 3. Model training

Feeding the training data to the model for the model to learn the relationship between features and labels.

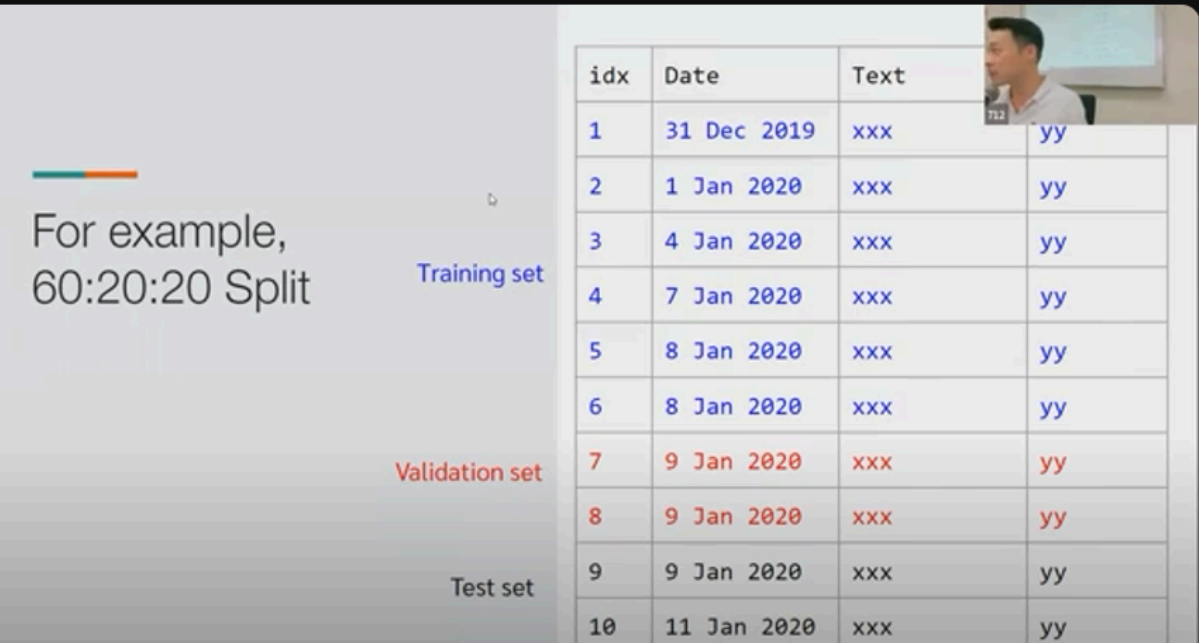
นอกจาก อินพุตแบบนี้ เอาต์พุตแบบนี้

อย่างแรกใน process

Train-validation- test data split

The data must first be split into three subsets:

- Training set – for training models for tech models
- Validation set (holdout set or development set) – for iteratively evaluating models like mini test (to vs another models)
- Test set – for final evaluation of the model



For example,  
60:20:20 Split

idx	Date	Text	
1	31 Dec 2019	xxx	yy
2	1 Jan 2020	xxx	yy
3	4 Jan 2020	xxx	yy
4	7 Jan 2020	xxx	yy
5	8 Jan 2020	xxx	yy
6	8 Jan 2020	xxx	yy
7	9 Jan 2020	xxx	yy
8	9 Jan 2020	xxx	yy
9	9 Jan 2020	xxx	yy
10	11 Jan 2020	xxx	yy

### Shuffling Rows

ถ้าลำดับของแถวมีผลต่อ text หรือ label เราต้องสลับข้อมูลก่อน

### Model training

We train the model on the training set . Each model has its own formula for training the model parameters .

แต่ละโมเดลมีวิธีเทรนที่แตกต่างกัน มีสูตรที่แตกต่างกัน

We evaluate the model on the dev set. Each model has its own way of using the trainer parameters. this process is called '**inference**' (the model infers the labels from the text)

infers == พยายาม predict ก็คือพยายามเดา นั่นแหละ

EXEL > NLP haha

---

# Naive Bayes Model

## Naive Bayes Classifier

A model that uses Bayesian inference from probability theory to infer the label given the text input. This model is outdated now, but it is the simplest ML model, which performs decently well.

model มันล้าสมัยแล้วแต่ก็ง่าย

## Bayesian Inference

Positive or negative? : 'predictable but very fun'

Let Y = label, X = text data

$p(Y = \text{positive} | X = \text{'predictable but very fun'})$

ถ้า text = 'predictable but very fun' แล้วมีความน่าจะเป็นเท่าไรที่ label จะเป็น positive \* คือตีความหมายของด้านบน

$P(Y = \text{negative} | X = \text{'predictable but very fun'})$

ถ้า text = 'predictable but very fun' แล้วมีความน่าจะเป็นเท่าไรที่ label จะเป็น negative

P = probability

## Bayesian Inference (2)

$P(Y = \text{positive} | X = \text{'predictable with no fun'})$

$$= \frac{P(X = \text{'predictable but very fun'} | Y = \text{positive}) P(Y = \text{positive})}{P(X = \text{'predictable but very fun'} | Y = \text{positive}) P(Y = \text{positive}) + P(X = \text{'predictable but very fun'} | Y = \text{negative}) P(Y = \text{negative})}$$

$P(X = \text{'predictable but very fun'} | Y = \text{positive})$  is called likelihood of the data. If you know that the label is positive, what is the chance of seeing 'predictable with no fun'?

$P(Y = \text{positive})$  is called prior probability of the label. If we don't know the text at all, what's the probability of positive label?

## Bayesian Inference (3)

$P(Y = \text{positive} | X = \text{'predictable with no fun'})$

$$= \frac{P(X = \text{'predictable but very fun'} | Y = \text{positive}) P(Y = \text{positive})}{P(X = \text{'predictable but very fun'} | Y = \text{positive}) P(Y = \text{positive}) + P(X = \text{'predictable but very fun'} | Y = \text{negative}) P(Y = \text{negative})}$$

$P(Y = \text{negative} | X = \text{'predictable with no fun'})$

$$= \frac{P(X = \text{'predictable but very fun'} | Y = \text{negative}) P(Y = \text{negative})}{P(X = \text{'predictable but very fun'} | Y = \text{positive}) P(Y = \text{positive}) + P(X = \text{'predictable but very fun'} | Y = \text{negative}) P(Y = \text{negative})}$$



view in spreads sheet

product คือผล \* รวมกัน

DataSet too large to good

## Likelihood

What's the probability of seeing "predictable" in a + document?

$$= \frac{\text{\# 'predictable' in positive document}}{\text{\# words in positive document}}$$

Text	Label
predictable and boring	negative
very few laughs	negative
short but boring	negative
very powerful	positive
fun and good laughs	positive
predictable but very fun and powerful	?

$P(\text{'predictable'} \mid Y = \text{positive})$

ถ้าเจอค่าเป็น 0 → Smoothing !!

## Smoothing +1

What's the probability of seeing "predictable" in a + document?

$$= \frac{\text{\# 'predictable' in positive document} + 1}{\text{\# words in positive document} + \text{vocab size}}$$

Text	Label
predictable and boring	negative
very few laughs	negative
short but boring	negative
very powerful	positive
fun and good laughs	positive
predictable but very fun and powerful	?

Pros and cons of ML - based classifier

Pros = High accuracy when a good amount of data available

ความแม่นยำสูงเมื่อมีข้อมูลเพียงพอ

Robustness to noise

ความทนทานต่อ noise

Learning very complex relationships with a lot of features

การเรียนรู้ความสัมพันธ์ที่ซับซ้อนพร้อมฟีเจอร์มากมาย

Cons =

Require datasets ต้องมีดาตาเซต

'Black box' : sometimes hard to interpret ดีความ

## 4. Evaluation

### Goal standard

Prediction on dev set

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...	97	98	99	100
Gold standard คำตอบจริง	J	J	J	J	Ad	Ad	Ad	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	OK
Prediction เครื่องเดามา	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	OK

Accuracy = 93 / 100

แต่ทะแม่งๆ

Now compute the accuracy of System A and B

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...	97	98	99	100	
Gold standard คำตอบจริง	J	J	J	J	Ad	Ad	Ad	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	OK	
Prediction เครื่องเดามา	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	OK	
Prediction เครื่องเดามา	J	J	J	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	OK	OK	...	OK	OK	OK	OK

Accuracy of System A = 93 / 100

Accuracy of System B = 93 / 100

จะเห็นว่า system มีความฉลาด เพราะมันพยายาม แยก

## Precision and Recall

**precision** of label A: if classifier predicts 'A', can **we trust it's actually A** ?

# A correctly predicted / # predicted A  
จำนวนที่ทายถูก / จำนวนที่ทาย A

**Recall** of label A: How many 'A' are actually detected?

# A correctly predicted / # true A  
ที่ทายถูก จำนวน A

if recall higher is good that mean nothing out of control

	Precision	Recall
OK email	93/100	93/93
Junk Mail (J)	NA	0/4
Advertisement (Ad)	NA	0/3

## F1 score

Precision and recall are calculated for each label. F1 score is the geometric mean of Precision (p) and Recall(R)

$$F1 \text{ of class A} = 2PR / (P+R)$$

Gold standard คำตอบจริง	J	J	J	J	Ad	Ad	Ad	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	O	
Prediction เครื่องเดมา	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	O	
Prediction เครื่องเดมา	J	J	J	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	OK	OK	...	OK	OK	OK	O

	Precision	Recall	F1	F1 of class OK = $2PR / (P+R)$ = $2 * (0.93) * (1.0) / (0.93 + 1.0)$
OK email	0.93	1.0	0.96	
Junk Mail (J)	NA	0/4	0	
Advertisement (Ad)	NA	0/3	0	

## machine learning = system B

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...					
Gold standard คำตอบจริง	J	J	J	J	Ad	Ad	Ad	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	OK	
Prediction เครื่องเดามา	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	OK	
Prediction เครื่องเดามา	J	J	J	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	OK	OK	...	OK	OK	OK	OK

		Precision		Recall	F1
OK email	86/86	1.0	86/93	0.9247	0.9609
Junk Mail (J)	3/3	1.0	3/4	0.75	0.8571
Advertisement (Ad)	3/10	0.3	3/3	1	0.4615



	Precision	Recall	F1
OK email	0.93	1.0	0.96
Junk Mail (J)	NA	0/4	0
Advertisement (Ad)	NA	0/3	0

System A

	Precision	Recall	F1
OK email	1.0	0.9247	0.9609
Junk Mail (J)	1.0	0.75	0.8571
Advertisement (Ad)	0.3	1	0.4615

System B

## Macro-average precision, recall, and F1

	Precision	Recall	F1
OK email	1.0	0.9247	0.9609
Junk Mail (J)	1.0	0.75	0.8571
Advertisement (Ad)	0.3	1	0.4615
Macro-average	0.76	0.89	0.75

## Confusion Matrix ความ งง ??

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...	97	98	99	100	
Gold standard คำตอบจริง	J	J	J	J	Ad	Ad	Ad	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	OK	
Prediction เครื่องเดามา	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	...	OK	OK	OK	OK	
Prediction เครื่องเดามา	J	J	J	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	Ad	OK	OK	...	OK	OK	OK	OK

		Actual (ความเป็นจริง)		
		Junk	Ad	OK
Predicted (เครื่องบอกมา)	Junk	3	0	0
	Ad	1	3	6
	OK	0	0	87

ตีความ แถวแรก เครื่องบอกว่าเป็น Junk และ เป็น Junk 3 ครั้ง  
บอกว่าเป็น Junk แต่เป็น Ad 0 ครั้ง เช่นเดียวกับ Ok แปลว่า ถูกหมด

แถวสองมีผิดตรง junk and ok

การคำนวณ Recall จาก confusion matrix

		Actual (ความเป็นจริง)		
		Junk	Ad	OK
Predicted (เครื่องบอกมา)	Junk	3	0	0
	Ad	1	3	6
	OK	0	0	87

		Junk	Ad	OK
Recall		3 /	3 /	87 /
		(3 + 1 + 0)	(3 + 0 + 0)	(0 + 6 + 87)

		Junk	Ad	OK		
Predicted (เครื่อง บอกมา)	Junk	3	0	0	Precision	Junk
	Ad	1	3	6		Ad
	OK	0	0	87		OK
					Junk	$3 / (3 + 0 + 0)$
					Ad	$3 / (1 + 3 + 6)$
					OK	$87 / (0 + 0 + 87)$

Accuracy

$$(3 + 3 + 87) / (3 + 3 + 87 + 1 + 6)$$

หาแนว ทะแยง / จำนวนทั้งหมด

Actual (ความเป็นจริง)

		Junk	Ad	OK
Predicted (เครื่องบอกมา)	Junk	3	0	0
	Ad	1	3	6
	OK	0	0	87

ดูว่า อ่อนตรงไหน

## Conclusion

1. We learn how to use rule-based models and supervised learning models for text classification, which can solve many NLP tasks.

เราเรียนรู้ วิธีการทำ ผ่าน rule-based และ machine learning

2. Supervised learning tasks involve tasks formulation, feature engineering, model training , and evaluation

supervised ประกอบด้วย tasks formulation, feature engineering, model training , and evaluation

3. Machine learning models differ in how they learn and infer.

แต่ละโมเดล จะมีวิธีที่ เรียนรู้ และ การทำนาย ที่แตกต่างกัน

4. Evaluation is very important in understanding the performance of various models

การประเมินมีความสำคัญมากในการทำความเข้าใจประสิทธิภาพของโมเดลต่างๆ