NLP - by Dr. Ayan Debnath

Day 5, 17th Dec. 2023

Recap: TF-IDF

BOW — weightage ✗

TF-IDF — weightage ✓

	w_1	w_2	w_3 ... w_n
sent1	0.03	0	0.97
sent2			

sent → No } feature set
text → No }

$$TF = \frac{\text{No. of rep of word}}{\text{No. of words sent}}$$

$$IDF = \log\left(\frac{\text{Total no. of sent}}{\text{no. of sent. containing that word}}\right)$$

sent1 : int boy
sent2 : int girl
sent3 : int boy & girl

TF

	sent1	s2	s3
int	1/2	1/2	1/3
boy	1/2	0/2	1/3
girl	0	1/2	1/3

IDF

words	IDF
int	log (3/3) = 0
boy	log (3/2)
girl	log (3/2)

TF * IDF = ✓

words	f_1 int	f_2 boy	f_3 girl
sent1	0	1/2 * log(3/2)	0
sent2	0	0	1/2 * log (3/2)

→ ML

tfidf vectorizer → sklearn

text → Pre-process [clean, lower] → Process (text → No., BOW/ TF-IDF) → ML → predict → performing

← end-to-end project →

TF-IDF Advantage

1. It weights the words → importance to word

I love NLP boy → [0.9 0.3 0.44 | 0] → context, order, semantic information

I NLP love
0.9 0.44 0.3

Disadvantage

1. Context info missing
2. Order " "
3. sensitive to corpus size
4. biased to rare term

corpus

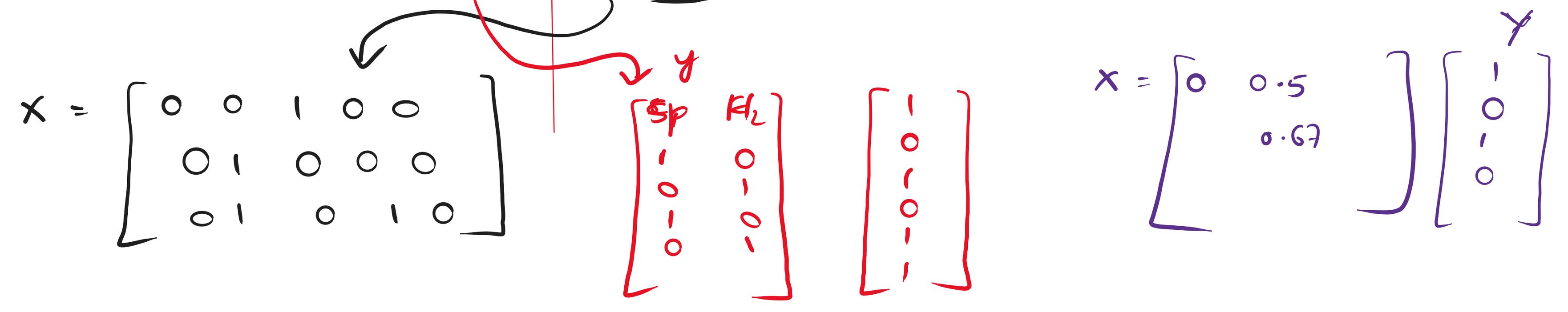
I was born in France.
I can speak French fluently

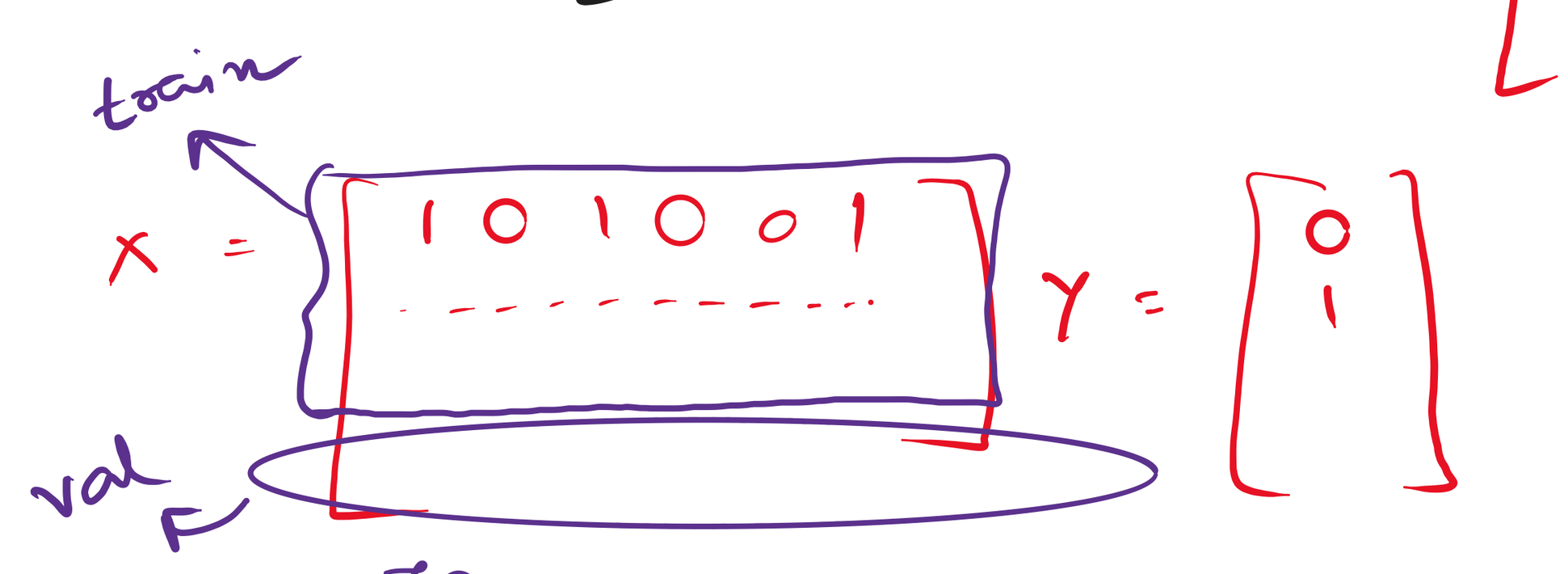
w_1	w_2	w_3 ...
✓	✓	✓

Data

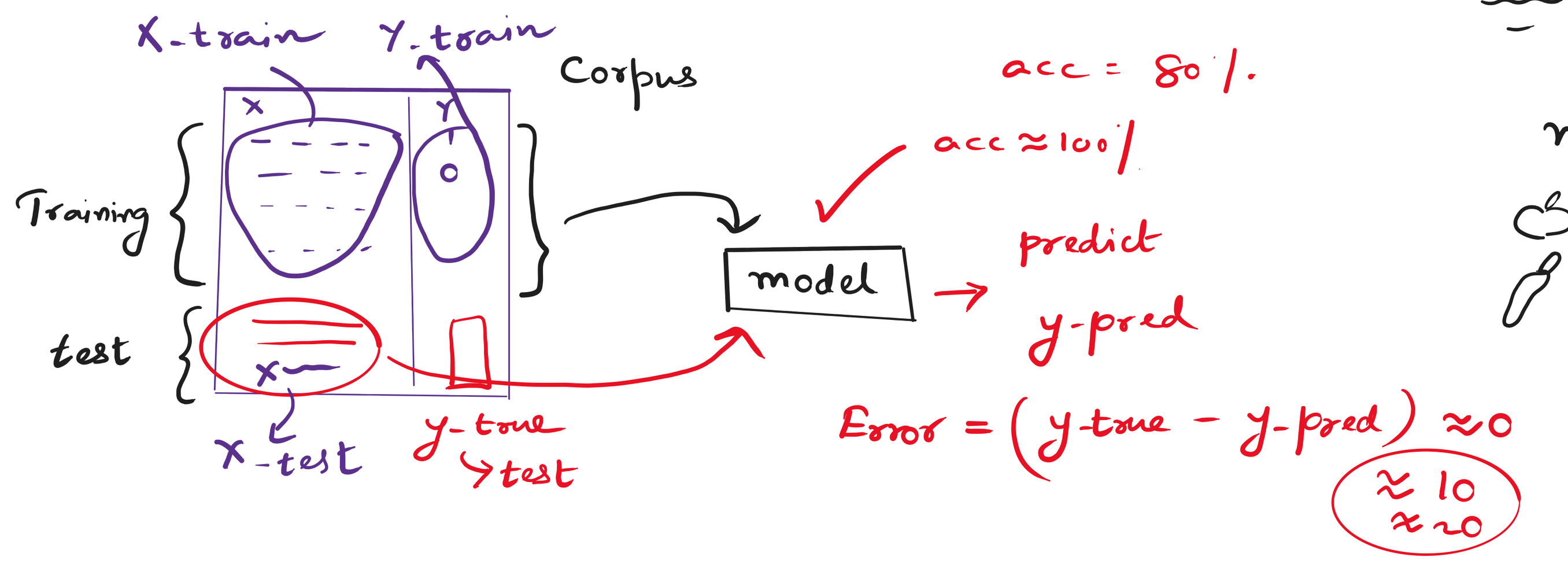
sent1 spam
sent2 ham / not spam
⋮
10K

y label	X msg
1 — spam	—
0 ham	—





70



acc = 80%.
acc ≈ 100% ✓
predict
y-pred

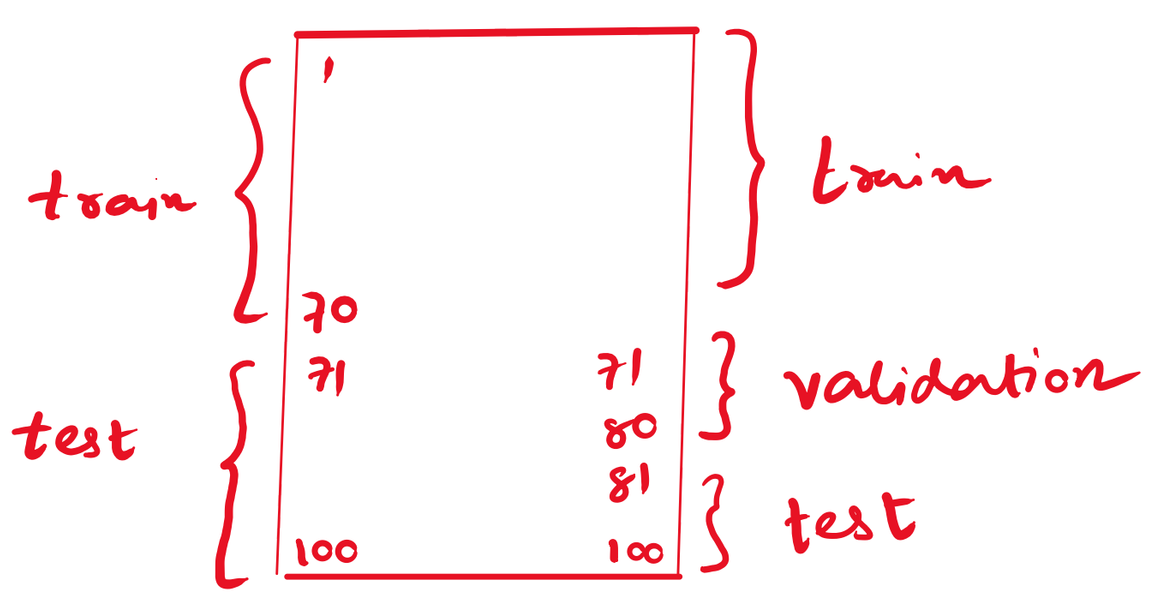
Error = (y-true − y-pred) ≈ 0 (≈ 10, ≈ 20)

X Y
— 1
— 0

model/ robot/child
apple — 0
banana — 1

X test → model → y-pred → apple

Train Validation Test



100 → 70, → 30

test	y-true	y-pred
X		
1 —	—	1
—	—	0
30 —	—	1

test1 Error$_1$ = y-true − y-pred
= 0 − 0 = 0 ✓
test2 error$_2$ = 1 − 1 = 0 ✓
= 1 − 0 = 1 → misclassification

✓ accuracy = ✓

70	10	20
train	Validation → model B	test model B
✓	✓✓	✓
prepare	mock	final

98% 97%
acc BOW > acc TF IDF

model BOW / model TF-IDF → model

test → model → spam/ham

[train, test]