Opera File Edit View History Bookmarks Developer Window Help

NER → Named Entity Recognition

Entities	Named Entities
Person	'obama'
locations ←	'Paris'
Artworks ←	Monalisa
Money ←	$100, Rs 100
date ←	Next Wednesday
time ←	
ordinal ←	First president
Cardinal ←	Five laptops

Entity
Intention

'Sachin' → Person
'Delhi' 'Goa' → location

'Spacy' →
'NLTK' →
tokenizer

Opera File Edit View History Bookmarks Developer Window Help

Pipeline

text → | tokenize | pos | stemmy | - - - → doc

↓

Parts of speech

I love icecream → 'I' 'Love' 'icecream'

↓ Pronoun subject

↓ verb

↓ object Noun

→ pos tag — token.pos_dtype

→ Begin Sent → doc

↳ cannot be modified

→ NE (False)_

→ stop word

→ lemma

→ nlp ('I love icecream') → _

Opera File Edit View History Bookmarks Developer Window Help

.exe ←

Recommended

Full project → deployment
→ Streamlit

1. EDA → Examine
→ Visualize
→ clean — Missing Valu
— duplicates
— outliers
— Rename Columns
— drop unwanted rows/cols

→ Feature Engineering
→ Feature Selection -

3 Build Model
↳ hyperparameter tuning ✓

4. Evaluate the model
→ overfitting / underfitting
→ Regularization if reqd

5. Deploy the model
↳ model will be binary
→ UI to use the model
(Streamlit / Flask / Django)

→ DataSet
2. split the data into train & test
1 train_test ✓
2 ↳ stratified
→ shuffle
→ K fold

A new page has been created.

Opera File Edit View History Bookmarks Developer Window Help

(1) train_test_split

Data Validation

x_1	x_2	x_3	y
1			
2			
3			
80			
81			
100			

→ shuffle
→ Randomly Selected

(2) Stratified splitting

— imbalanced data
— No of classes in 'y'

✓ Yes → 50 /
/ No → 50 /

✓
90 → Yes
10 → No
→ 80 → Yes No
→ 20 → Yes No

90 /
10 /

Yes → 90 /
No → 10 /

} accuracy is not reliable

(3) K-Fold Validation — less data

Cross

	Ht	Wt	
1	Tall	Heavy	
2	Tall	Thin	
3	Short	heavy	→
	Short	Thin	→
	Med	Med	→
9	Tall	Thin	
10	Tall	heavy	→

$k = 5, 10$

$\Rightarrow$ split data into k-equal parts

$n = 1000, \quad k = 5$

D1	200
D2	200
D3	200
D4	200
D5	200

logistic →

200

← Training → ← test →

800 ← D1 D2 D3 D4 | D5 → A1 MSE1

→ D1 D2 D3 D5 | D4 → A2 MSE2

D1 D2 D4 D5 | D3 → A3 MSE3

D1 D3 D4 D5 | D2 → A4 MSE4

D2 D3 D4 D5 | D1 → A5 MSE

Avg → MSE

Avg → overall accuracy

— No shuffling

— Guaranteed to test on every part of the data

— No overlap of testing data

Same record will not be tested twice

(3) Leave one out

— leave one record for test

— use the rest for training

1	—
2	—
3	—
4	—
5	—

(1 – 4) → 5th record ✓

1, 2, 3, 5 → 4th record

1, 2, 4, 5 → 3rd record

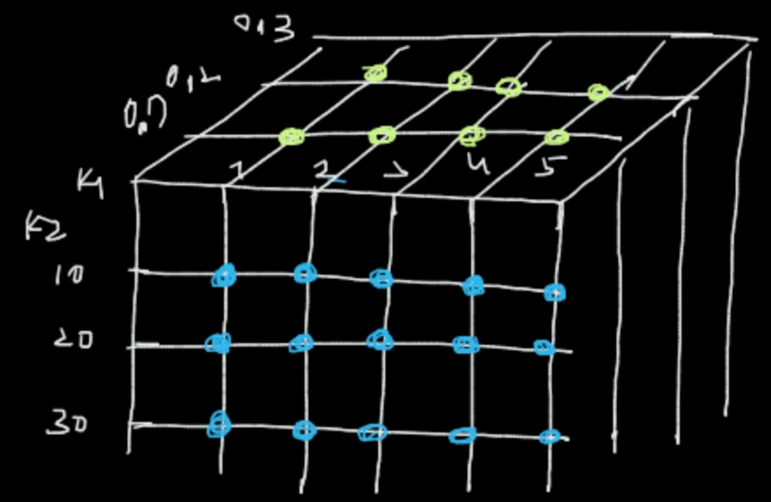
1, 3, 4, 5 → 2nd record

2, 3, 4, 5 → 1st record

Opera File Edit View History Bookmarks Developer Window Help

(5) Shuffle split

- Shuffle & split multiple times (100, 200)
- No guarantee that every part of the data will be tested

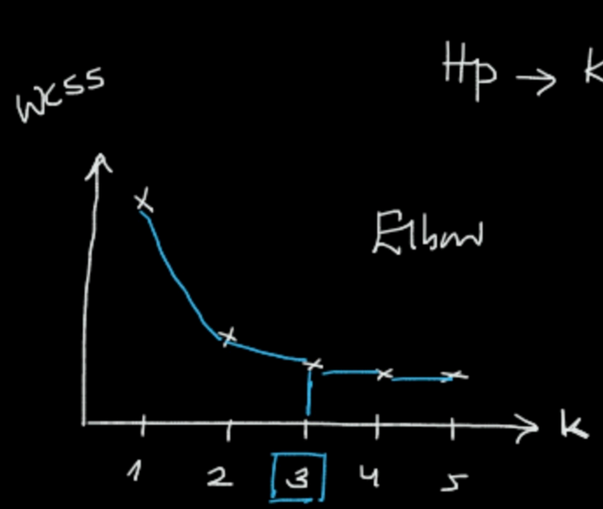


Hyperparameter tuning

one Hp

- Elbow technique

Hp → k



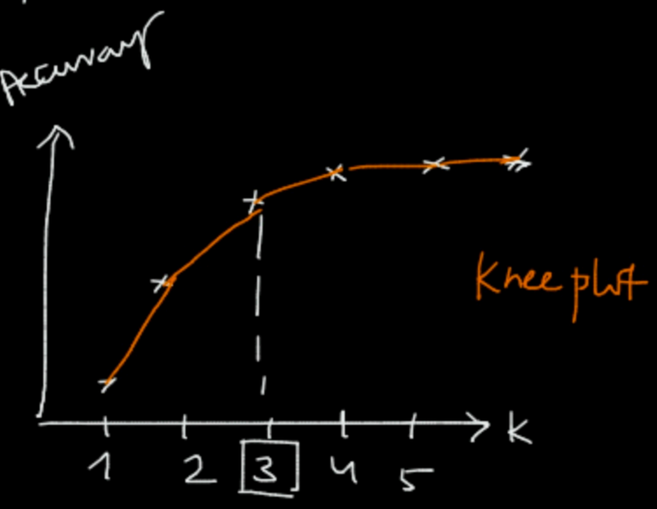
Multiple hyper params

k1 → {1, 2, 3, 4, 5
k2 → 10, 20, 30
k3 → 0.1, 0.2, 0.3 }

45 Models

Grid search

Random search



$k_1 = 2$, $k_2 = 10$, $k_3 = 0.1$
→ Best metric