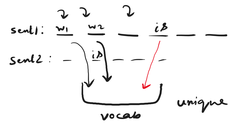
NLP - by mentor Dr Ayon Debnath

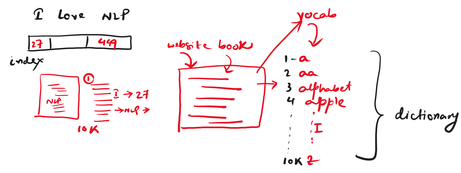
Day 4, 16/12/23

Recap

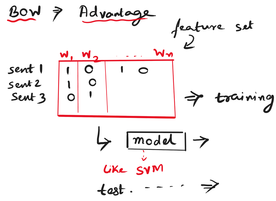
① Text → neumerical representation
- of BOW
- of coding implementation

Topic → BOW, TF-IDF, Coding

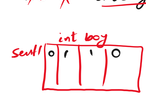




BOW → Advantage



~~An~~ ~~a~~ intelligent boy



① Easiest
② Small dataset, BOW works well
if " is huge, BOW " X

Disadvantages
① Big dataset, X
② sparsity
③ semantic information X
④ context X
⑤ Order X
* ⑥ Word weightage is not give

	w_1	(w_2) intelligent	(w_3) boy	w_4
sent1	0	1 → 0.1	1 → 0.9	0

imp
TF - IDF (Term Frequency - Inverse Document Frequency)

sent1: intelligent boy
sent2: intelligent girl
sent3. intelligent boy girl

$$\text{Term Frequency (TF)} = \frac{\text{No. of rep. of words in sentence}}{\text{Total no. of words in sentence}}$$

$$\text{IDF} = \log\left(\frac{\text{Total no. of sentences}}{\text{No. of sentences containing words}}\right)$$

TF * IDF = ✓

histogram

words	Frequency
intelligent	3
boy	2
girl	2

Vector ⇒

TF

words	sent1	sent2	sent3
intelligent	1/2	1/2	1/3
boy	1/2	0	1/3
girl	0	1/2	1/3

IDF

Words	IDF
intelligent	log(3/3) = 0
boy	log(3/2) =
girl	log(3/2) =

	f_1 intelligent	f_2 boy	f_3 girl
sent1	0	1/2 * log(3/2)	0
sent2	0	0	1/2 * log(3/2)
sent3	0		

⇒ unique words (feature)

⇝ feature set

↳ model ⇒ training data