Unit-1 Problems.

3.) D1: ~~Either~~ Sue ~~is~~ rich ~~or~~ ~~she~~ ~~is~~ poor.

D2: Bill eats peanuts ~~and~~ chicken.

D3: Sue eats everything Bill eats.

Remove stop words(articles/pronouns)

Vocabulary = { Sue, rich, , poor, Bill, eats, peanuts, chicken, everything}
(8)

D1: [Sue rich poor]

D2: [Bill eats peanuts chicken]

D3: [sue eats everything Bill eats]

→ use lowercase

	sue	rich	poor	bill	eat	peanuts	chicken	everything
D1	1	1	1	0	0	0	0	0
D2	0	0	0	1	1	1	1	0
D3	1	0	0	1	1	0	0	0 1

$D_1 = [1\ 1\ 1\ 0\ 0\ 0\ 0\ 0]$

$D_2 = [0\ 0\ 0\ 1\ 1\ 1\ 1\ 0]$

$D_3 = [1\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 1]$

Part-B

1) TF-IDF

D1: All vegetarian restaurants serve vegetarian food.
D2: I ate vegetable fried rice from an Indian coffee house.
D3: There is a restaurant near Delhi which serves Italian food.

TF (term frequency) = No of times word appears in a document
IDF (Inverse document frequency) = It measures how rare a word is across collection of document.

$$\boxed{TF = \frac{\text{freq of word in document}}{\text{Total no of words in document}}}$$

$$\boxed{IDF = \log\left(\frac{\text{No of doc}}{\text{words in no of docs}}\right)}$$

Remove stop words

D1 = vegetarian restaurant serve vegetarian food. (5)
D2 = ate vegetable fried rice ~~f~~ indian coffee house (7)
D3 = restaurant near delhi serve italian food (6)

Vocabulary = { vegetarian, restaurant, serve, food, ate, vegetable, fried, rice, indian, coffee, house, near, delhi, italian } (14)

log 3 ÷ log10

word	TF			IDF	TF * IDF		
	D1	D2	D3		D1	D2	D3
vegetarian	2/5	0	0	$\log\left(\frac{3}{2}\right)$ = 0·17	0·06	0	0
restaurant	1/5	0	1/6	log(3/2) = 0·17	0·03	0	0·02
serve	1/5	0	1/6	log(3/2) = 0·17	0·03	0	0·02
food	1/5	0	1/6	log(3/2) = 0·17	0·03	0	0·02
ate	0	1/7	0	log(3/1) = 0·47	0	0·06	0
vegetable	0	1/7	0	log(3/1) = 0·47	0	0·06	0
fried	0	1/7	0	log(3/1) = 0·47	0	0·06	0
rice	0	1/7	0	log(3) = 0·47	0	0·06	0
indian	0	1/7	0	log(3) = 0·47	0	0·06	0
coffee	0	1/7	0	log(3) = 0·47	0	0·06	0
house	0	1/7	0	log(3) = 0·47	0	0·06	0

	TF			IDF	TF * IDF		
	D1	D2	D3		D1	D2	D3
near	0	0	1/6	log(3/1) = 0.47	0	0	0.07
delhi	0	0	1/6	log(3/1) = 0.47	0	0	0.07
italian	0	0	1/6	log(3/1) = 0.47	0	0	0.07
				Total	0.15	0.42	0.27

Avg of $D_1 = \frac{0.15}{14} = 0.0107$

Avg of $D_2 = \frac{0.42}{14} = 0.03$ ✓

Avg of $D_3 = \frac{0.27}{14} = 0.0192$

Unit-2

8.) S1 = John plays in park
NNP VB IN NN

S2 = Park the car
VB DET NN

S3 = John will park the car
NN MD VB DET NN

NN - noun
VB - verb
IN - preposition
DT - determiner
MD - Modal verb

Part-B 1) a.) Identify corresponding POS tag

Jack will park the scooter
NN MD VB DT . NN

b.) HMM model.

Jack will park the scooter
?

S1 = John plays in park.
NN VB IN NN

S2 = Park the car.
VB DT NN

S3 = John will park the car.
NN MD VB DT NN

Lookup table:

	NN	VB	IN	DT	MD
John	2	0	0	0	0
plays	0	1	0	0	0
in	0	0	1	0	0
park	1	2	0	0	0
the	0	0	0	2	0
car	2	0	0	0	0
will	0	0	0	0	1
Total	5	3	1	2	1

Emission probability:

	NN	VB	IN	DT	MD
John	2/5				
plays		1/3			
in			1/1		
park	1/5	2/3			
the				2/2	
car	2/5				
will					1/1

Transition probability:

	NN	VB	IN	DT	MD	<E>	Total
<S>	2/3	1/3	0	0	0	0	3
NN	0	1/5	0	0	1/5	3/5	5
VB	0	0	1/3	2/3	0	0	3
IN	1	0	0	0	0	0	1
DT	2/2	0	0	0	0	0	2
MD	0	1	0	0	0	0	1

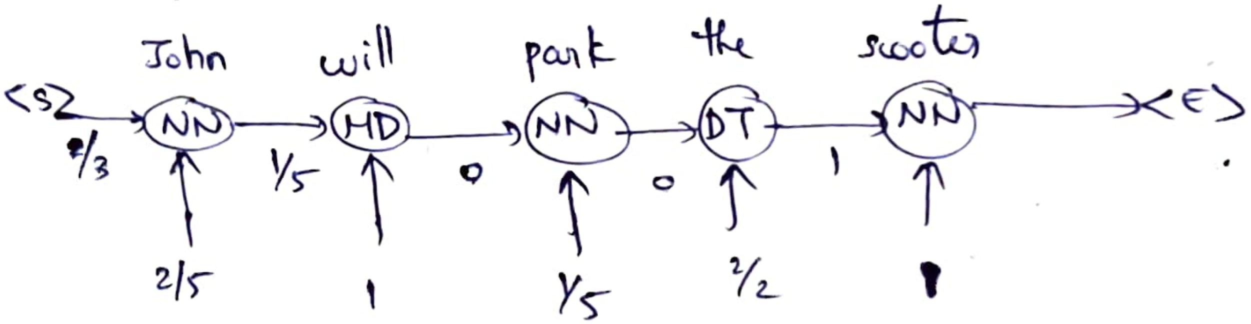
No of hidden states (NN, VB, IN, DT, MD) = 5

No of words (Jack will park the scooter) = 5

⇒ 5^5 possibilities.

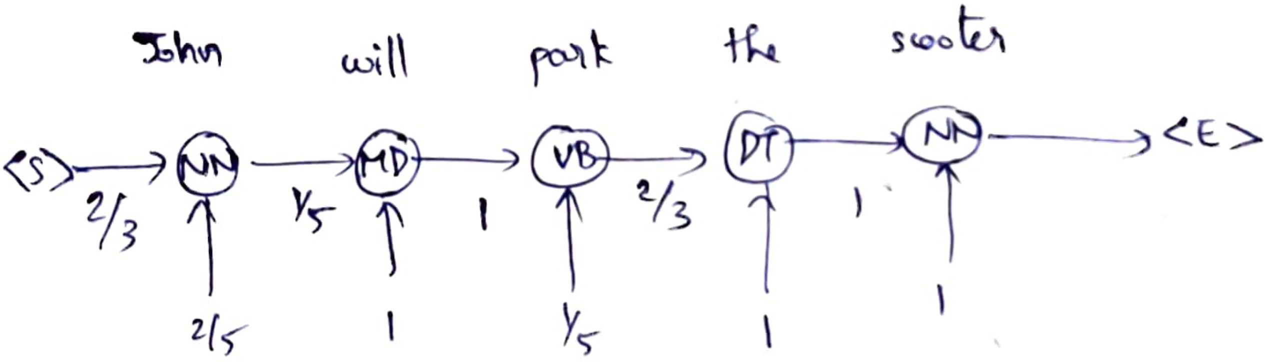
Consider park NN

~~<S> John~~



$$\left(\frac{2}{3}\right)\left(\frac{2}{5}\right)\cdot\left(\frac{1}{5}\right)(1)\cdot 0\left(\frac{1}{5}\right) = 0$$

Consider park VB



$$\frac{2}{3}\cdot\frac{2}{5}\cdot\frac{1}{5}\cdot 1\cdot 1\cdot\frac{1}{5}\cdot\frac{2}{3}\cdot 1\cdot 1 = \underline{\underline{0.01}}$$

Part-B

3) Bi-gram model

I really ______

S1: I really appreciate your help

S2: I am really sorry for not inviting you

S3: I really appreciate your hard work

S4: I really like your watch.

Vocabulary = (I, really, appreciate, your, help, am, sorry, for, not, inviting, you, hard, work, like, watch)

$\boxed{\text{Vocabulary} = 15}$

$$P(w_i / w_{i-1}) = \frac{\text{count}(w_{i-1}\ w_i)}{\text{count}(w_{i-1})}$$

$$P(I / \text{really}) = \frac{\text{count}(\text{really } I)}{c(\text{really})} = \frac{0}{4} = 0$$

$$P(\text{appreciate}/\text{really}) = \frac{\text{count}(\text{really app})}{c(\text{really})} = \frac{2}{4} = \boxed{0.5}$$ highest

$$P(\text{sorry} / \text{really}) = \frac{\text{count}(\text{really sorry})}{c(\text{really})} = \frac{1}{4} = 0.25$$

$$P(\text{like} / \text{really}) = \frac{\text{count}(\text{really like})}{c(\text{really})} = \frac{1}{4} = 0.25$$

$$P(\text{your} / \text{really}) = \frac{\text{count}(\text{really} / \text{your})}{c(\text{really})} = \frac{0}{4} = 0$$

o/p: I really appreciate