GHARDA FOUNDATION'S
GHARDA INSTITUTE OF TECHNOLOGY GIT

Experiment No. NLP Date :

Assignment No. 1

Q.1 Exercise.

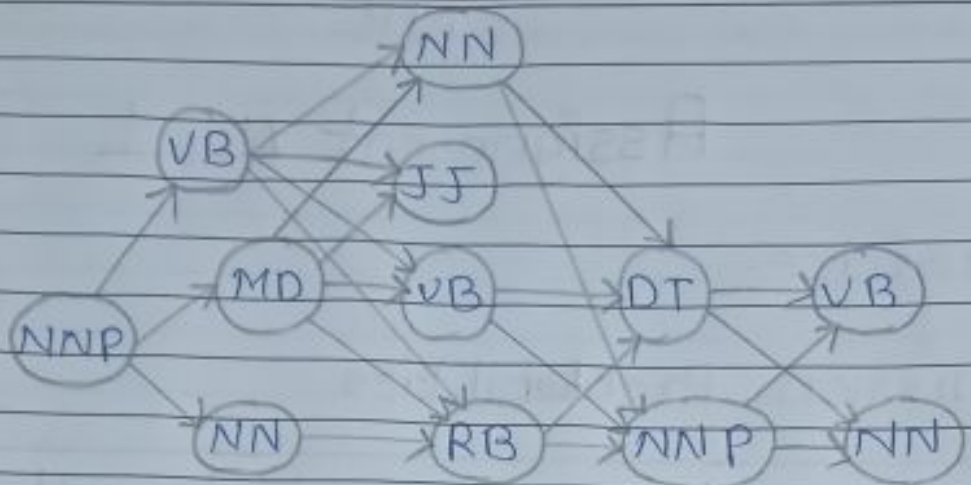
Emission Probabilities.

	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0	0
NN	0	0.000200	0.000223	0	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Transition Probabilities

	NNP	MD	VB	JJ	NN	RB	DT
<S>	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0062	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.6068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.6617

Test sample : Janet will back the bill



Q.2 Estimating Bi-gram Probabilities.

what is the most probable next word. Predicted by the model for the following word sequence?

Given corpus,

<s> I am Henry </s>
<s> I like college </s>
<s> Do Henry like college </s>
<s> Henry I am </s>
<s> Do I like Henry </s>
<s> Do I like college </s>
<s> I do like Henry </s>

word	Frequency
<s>	7
</s>	7
I	6
am	2
Henry	5
College	3
like	5
do	4

1) <s> Do?

Next word Prediction Probability W_{i-1} = do

Next word	Probability Next word
P(</s> / do)	0/4
P(<I> / do)	2/4
P(<am> / do)	0/4
P(<Henry> / do)	1/4
P(<like> / do)	1/4
P(<college> / do)	0/4
P(<do> / do)	0/4

I is more Probable

2) <s> I like Henry?

Next word Prediction Probability W_{i-1} = I like Henry.

Next word	Probability Next word
P(</s> / Henry)	3/5
P(<I> / Henry)	1/5
P(<am> / Henry)	0/5
P(<Henry> / Henry)	0/5
P(<like> / Henry)	1/5
P(<college> / Henry)	0/5
P(<do> / Henry)	0/5

∴ </s> is more Probable

3) <s> Do I like? (use Trigram)

P(<I like>) = 3

Next word Prediction Probability w_{i-2} = I & w_{i-1} = like

Next word	Probability Next word
P(</s> / I like)	0/3
P(<I> / I like)	0/3
P(<am> I like)	0/3
P(<Henry> I like)	1/3
P(<like> I like)	0/3
P(<college> I like)	2/3
P(<do> I like)	0/3

∴ college is probable

4) <s> Do I like college? (use Four-gram)

Next word Prediction Probability w_{i-3} = I w_{i-2} = like w_{i-1} = College

Next word	Probability Next word
P(</s>/I like college)	2/2
P(<I>/I like college)	0/2
P(<am>/I like college)	0/2
P(<Henry>/I like college)	0/2
P(<college>/I like college)	0/2
P(<do>/I like college	0/2

∴ </s> is more Probable

which of the following sentence is better i.e gets a higher probability with this model use Bi-gram.

1) <s> I like college </s>
= P(I/<s>) x P(like/I) x P(college/like) x P(</s>/college)
= 3/7 x 3/6 x 3/5 x 3/3
= 9/70 = 0.13

2) <s> Do I like Henry </s>
= P(Do/<s>) x P(I/Do) x P(like/I) x P(Henry like) x P(</s>/Henry)
= 3/7 x 2/4 x 3/6 x 2/5 x 3/5
= 9/350
= 0.0257

first statement is more probable

Q.3 Emission Probability Matrix

	DT	NN	VB
that	0.40	0.00	0.00
Girl	0.00	0.015	0.0031
Smiles	0.00	0.0004	0.20

State Transition Probability Matrix

	DT	NN	VB
<S>	0.50	0.40	0.1
DT	0.01	0.99	0.00
NN	0.30	0.30	0.40
VB	0.40	0.40	0.20

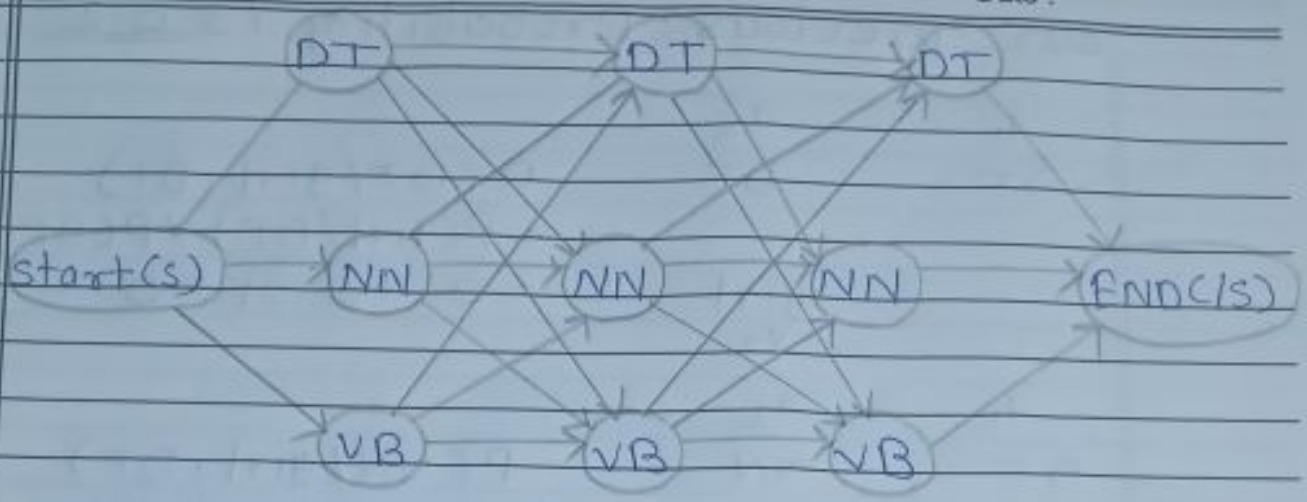
Text Data <s> that girl Smiles </s>
Apply Pos Tagging using HMM.

→

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

that	Girl	Smiles
DT	DT	DT
NN	NN	NN
VB	VB	VB

3 × 3 × 3 = 27 ways



$P(q_2/x_2, q_1) = P(x_2/q_2) \times P(q_2/q_1)$

q_1 q_2

START → DT

$P(DT/that, <S>) =$
$P(that/DT) \times P(DT/<S>)$
$0.40 \times 0.50 = 0.2$

DT → that (x_2)

q_1 q_2

start → NN

$P(NN/that, <S>)$
$P(that/NN) \times P(NN/<S>)$
$0.00 \times 0.40 = 0$

NN → that

q_1 q_2 x_2

start → VB

$P(VB/that, <S>)$
$P(that/VB) \times P(VB \times <S>)$
$0.00 \times 0.1 = 0$

VB → that (x_2)

probability of "that" word is large for "DT" Part of Speech Tag.

our previous Probability is 0.2

q_1	q_2	
DT	DT	$P(DT/girl, DT)$
		$= P(girl/DT) \times P(DT/DT)$
	girl	$= 0 \times 0.01 = 0$
	x_2	

q_1	q_2	
DT	NN	$P(NN/girl, DT)$
		$= P(girl/NN) \times P(NN/DT)$
	girl	$= 0.01458 \times 0.2 = 0.003$
	x_2	

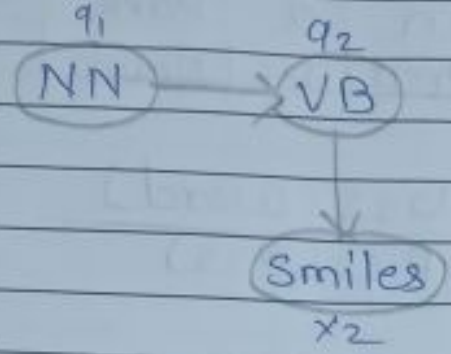
q_1	q_2	
DT	VB	$P(VB/girl, DT)$
		$= P(girl/VB) \times P(VB/DT)$
	girl	$= 0.0031 \times 0. = 0$
	x_2	

Result = Girl is down

our previous Probability is 0.003

q_1	q_2	
NN →	DT	$P(DT/smiles, NN)$
		$= P(smiles/DT) \times P(DT/NN)$
	smiles	$= a \times 0.30 = 0$
	x_2	

q_1	q_2	
NN →	NN	$P(VB/smiles, NN)$
		$= P(smiles/VB) \times P(VB/NN)$
	smile	$= 0.08 \times 0.003 = 0.00024$
	x_2	



$P(VB/smiles, NN)$
$= P(smiles/VB) \times P(VB/NN)$
$= 0.08 \times 0.003 = 0.00024$

Result : smiles is verb

Q.4 Training Corpus

<S>	Book	a	car	</S>			
<S>	Park	the	car	</S>			
<S>	the	book	is	in	the	car	</S>
<S>	The	car	is	in	a	Park	</S>

Text Data <S> the park is a book </S>

Generate state Transition & emission Probability Matrix
Apply Pos Tagging using HMM

(1) Assign concept POS Tag to training corpus

<S>	Book	a	car	</S>			
	verb	Det	Noun				
<S>	Park	the	car	</S>			
	verb	Det	Noun				
<S>	The	book	is	in	the	car	</S>
	Det	Noun	verb	Prep.	Det	Noun	

<s>	The	car	is	in	a	Park	</s>
	Det	Noun	verb	Prep	Det	Noun	

$$P(word/class) = \frac{P(class, word)}{P(class)}$$

(2) Creation of Emission Probability Matrix

	verb (4)	Det (6)	Noun (6)	Prep. (2)
Book	1/4	0	1/6	0
a	0	2/6	0	0
car	0	0	4/6	0
the	0	4/6	0	0
Park	1/4	0	1/6	0
is	2/4	0	0	0
in	0	0	0	2/2
≤	4	6	6	2

(3) Creation of state Transition Probability matrix.

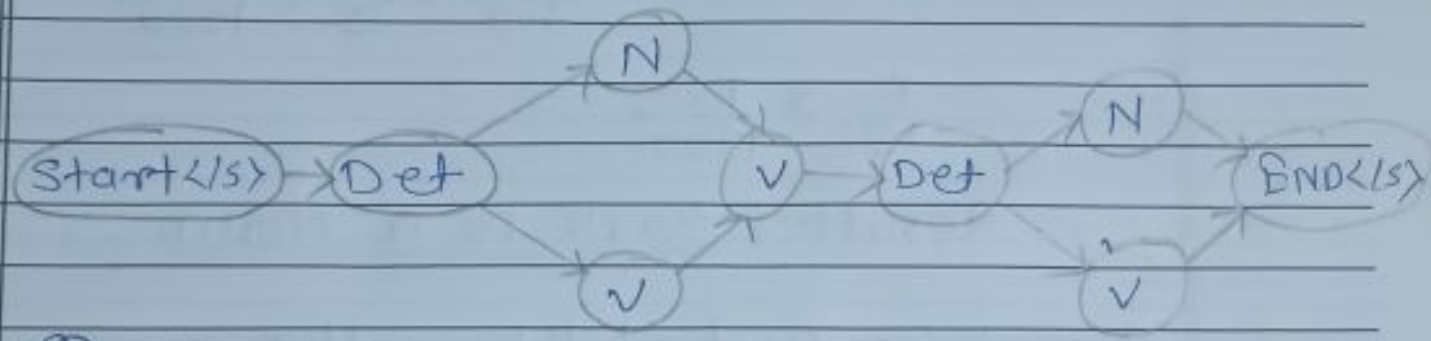
	Noun	verb	Det	Prep	</s>
<s>	0	2/4	2/4	0	0
Noun	0	2/6	0	0	4/6
verb	0	0	2/4	2/4	0
Det	6/6	0	0	0	0
Prep	0	0	2/2	0	0

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

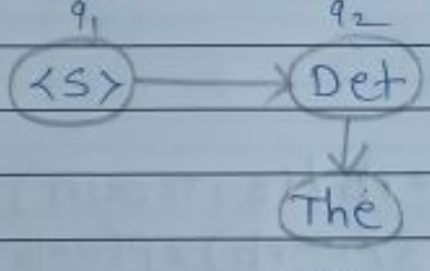
Text data <S> The Park is a book </S>

The	Park	is	a	book
Det	Noun	Verb	Det	Noun
	Verb			Verb

1 x 2 x 1 x 1 x 2 = 4 ways



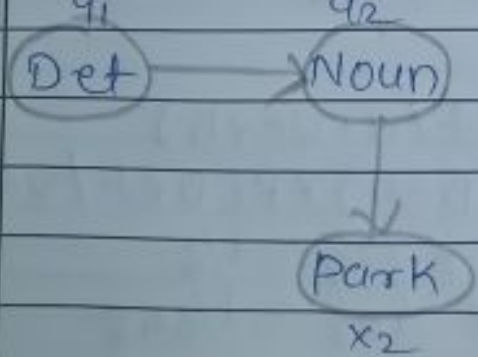
(i) The



P(Det / The, <S>)
= P(The / Det) x P(Det / <S>)
= 4/6 x 2/4 = 1/3

Previous Probability is 1/3

Park as Noun



P(Noun / Park, Det) =
P(Park / Noun) x P(Noun / Det)
P(Noun / Park, Det) = 1/6 x 6/6 = 1/6
P(Noun / Park, Det) = 1/6 x 1/3 = 1/18

Park as verb

q_1 q_2

Det verb $P(\text{verb}/\text{park}, \text{Det}) = P(\text{Park}/\text{verb}) \times P(\text{verb}/\text{Det})$

$= \frac{1}{4} \times \frac{0}{1} = 0$

Park

x2

$\frac{1}{4} \times \frac{1}{100} = \frac{1}{400}$

$\frac{1}{400} \times \frac{1}{3} = \frac{1}{1200}$

$\frac{1}{18} > \frac{1}{1200}$

∴ Result : Park is a Noun.

Previous Probability = 1/18

(3) is a verb

q_1 q_2

Noun → verb → is (x2)

$P(\text{verb}/\text{is}, \text{Noun}) = P(\text{Is}/\text{verb}) \times P(\text{veb}/\text{Noun})$

$= \frac{2}{4} \times \frac{2}{6} = \frac{1}{6} = \frac{1}{6} \times \frac{1}{18} = \frac{1}{108}$

Previous Probability = 1/108

(4) a as Det

q_1 q_2

verb → Det → a (x2)

$D(\text{Det}/\text{a}, \text{verb}) = P(\text{a}/\text{Det}) \times P(\text{Det}/\text{verb})$

$= 2/6 \times 2/4 = 1/6$

$= 1/6 \times 1/108 = 1/648$

Page No. :

Previous Probability is 1/648

Det (5) book as a verb

q_1 Det → q_2 verb → book x_2

P(verb/book, Det)
= P(book/verb) × P(verb/Det)
= 1/4 × 0 = 0 = 0
$= \frac{1}{4} \times \frac{1}{100} = \frac{1}{400}$

book as a Noun

q_1 Det → q_2 Noun → book x_2

P(Noun/book, Det)
= P(book/Noun) × P(Noun/Det)
= 1/6 × 6/6 = 1/6
= 1/6 × 1/648 = $\frac{1}{3888}$

$\frac{1}{400} \times \frac{1}{648} = \frac{1}{259200}$

$\frac{1}{3888} > \frac{1}{259200}$

∴ Result : book is Noun.