Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2020-2021
M.Tech (Data Science and Engineering)
Mid-Semester Test (Makeup)

Course No. : DSECLZG525

Course Title : Natural Language Processing
Nature of Exam : Open Book
Weightage : 30%

No. of Pages = 3
No. of Questions = *3*

Note: Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1. [3+5=8 Marks]

a) Explain which type of ambiguity exist in following sentences. **[3 marks]**

i. **I saw someone on the hill with a telescope**. (Answer : structural)

ii. **She is walking towards a bank**. (Answer: Lexical)

iii. **The running race was wonderful to watch.** (Answer: Grammatical – race and watch has both noun and verb sense)

b) Given is the following toy corpus. Calculate all the bigram probabilities. **[2 marks]**

<s> I love NLP </s>
<s> NLP is interesting</s>
<s> I am learning NLP </s>

P(I|<s>)=2/3=0.67
P(love|I)=1/3
P(NLP|love) = 1/3
P(</s>|NLP)=2/3

P(NLP|<s>)=1/3
P(is| NLP)=1/3
P(interesting| is)=1/3
P(</s>| interesting)=1/3

P(I|<s>)=2/3=0.67
P(am|I)=1/3
P(learning|am)=1/3
P(NLP|learning)=1/3
P(</s>|NLP)=2/3

c) Calculate the probability of sentence **<s> I am studying NLP</s>** using raw bigram probabilities and using Laplace smoothing. **[1+2=3 marks]**

Without smoothing
P(I|<s>)=2/3=0.67
P(am|I)=1/3
P(studying |am)=0

P(NLP| studying)=0
P(</s>|NLP)=2/3
Unique words=7
With smoothing

Word	Bigram with smoothing
P(I|<s>)	2+1 / 3+7
P(am|I)	1+1 / 3+7
P(studying |am)	0+1 / 3+7
P(NLP| studying)	0+1 / 3+7

Question 2. [6+4 =10 Marks]

a) Let the input sentence be "Bank upon me". Possible Tags are {T1, T2, T3, T4}. Assume all the POS tags are equally likely to be at the starting of the sequence

Table 1: Transition probabilities

	T1	T2	T3	T4
T1	0.18	0.01	0.8	0.01
T2	0.9	0	0.05	0.05
T3	0.4	0.5	0.05	0.05
T4	0.4	0.5	0.05	0.05

Table 2: Emission probabilities

	Bank	Upon	Me
T1	0.1	0.1	0.8
T2	0.8	0.1	0.1
T3	0.2	0.2	0.6
T4	0.8	0.1	0.1

a) Calculate P(x1=Bank, x2=Upon, y1=T1, y2=T2) **[1 Mark]**
b) Which is the most probable POS tag sequence out of these sequences for the given input sentence: I) T4 T1 T3
II) T2 T1 T3
III) T2 T2 T1

IV)T3 T2 T1

[4 Marks]

c) Compute the joint probable sequence of most probable sequence above. **[1 Mark]**

Solution

i. P(x1=Bank, x2=Upon, y1=T1, y2=T2)=P(T1)*P(x1|T1)*P(T2|T1)*P(x2|T2)

=0.25*0.1*0.1*0.01

=0.000025

ii. Here we have to find out the most probable tag sequence

for **I) T4 T1 T3**

P(x1=Bank, x2=Upon, x3=me, y1=T4, y2=T1, y3=T3)

=P(T4) *P(x1|T4)*P(T1|T4)*P(x2|T1)* P(x3|T3)P(T3|T1)

=0.25*0.8*0.4*0.1*0.6*0.4 =0.0019

for II**) T2 T1 T3**

P(x1=Bank, x2=Upon, x3=me, y1=T2, y2=T1, y3=T3)

=P(T2)* P(x1|T2)* P(T1|T2)*P(x2|T1)* P(x3|T3)P(T3|T1)

= 0.25*0.8*0.9*0.1*0.6*0.8 =0.0086

for III**) T2 T2 T1**

P(x1=Bank, x2=Upon, x3=me, y1=T2, y2=T2, y3=T1)=0

For **IV) T3 T2 T1**

P(x1=Bank, x2=Upon, x3=me, y1=T3, y2=T2, y3=T1)

= P(T3)* P(x1|T3)* P(T2|T3)*P(x2|T2)* P(x3|T1)*P(T1|T2)

= 0.25*0.2*0.2*0.1*0.8*0.9=0.0007

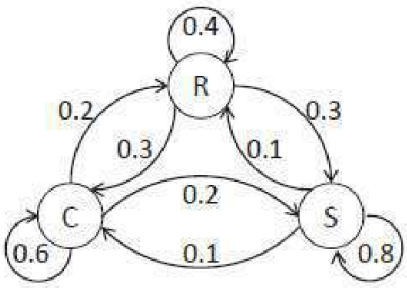
Maximum of all these sequences correspond to T2 T1 T3.

Hence the most probable sequence is **T2 T1 T3**

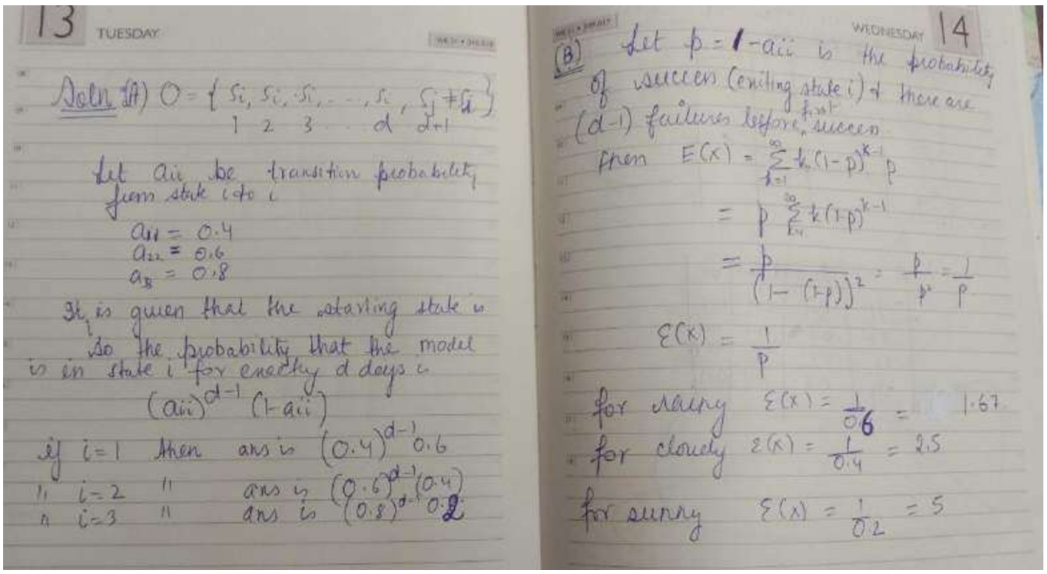
The joint probability for the most probable sequence is 0.0086

b)Once a day, weather is observed as one of the states: [4 marks]

state 1: Rainy (R), state 2:cloudy (C), state 3: Sunny (S)



A) Given that model is in state i, what is the probability that it stays in the state i for exactly d days.
B) What is the expected duration in the state i. (Also conditioned on starting in the state i).



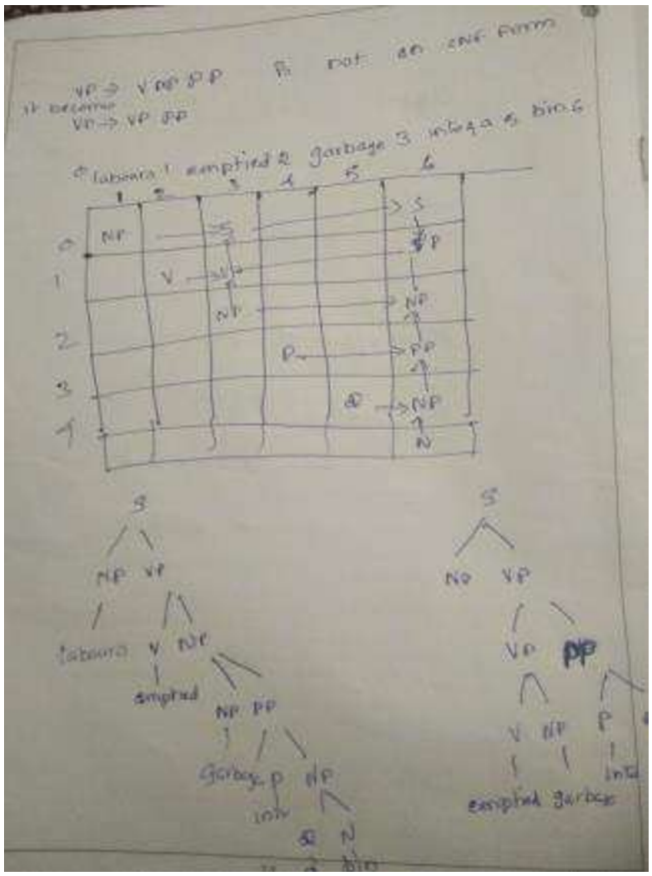
Question 3. [Marks 5+2+5=12 marks]

a) Find the following the context free grammar is in Chomsky normal form. Justify your answer **[1 Marks]**

b) Create a CKY table for parsing the sentence "**labours emptied garbage into a bin** "with the grammar G and make all possible parse trees. **[4 Marks]**

```
S  -> NP VP
PP -> P NP
VP -> V NP PP
VP -> V NP
NP -> D N
NP -> NP PP
NP -> NP CNP
N  -> A N
```

```
CNP -> C NP
NP -> " labours " | "sacks"
    | "garbage" | "junk"
N -> "worker" | "bin" | "sack"
V -> "dumped" | "emptied"
P -> "of" | "into"
D -> "a" | "the"
C -> "and"
A -> "big" | "small"
```

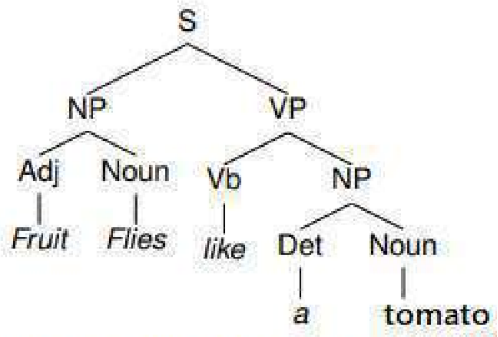


c) Find the probability of the sentence “Fruits flies like a tomato” using PCFG parsing method

[2 Marks]

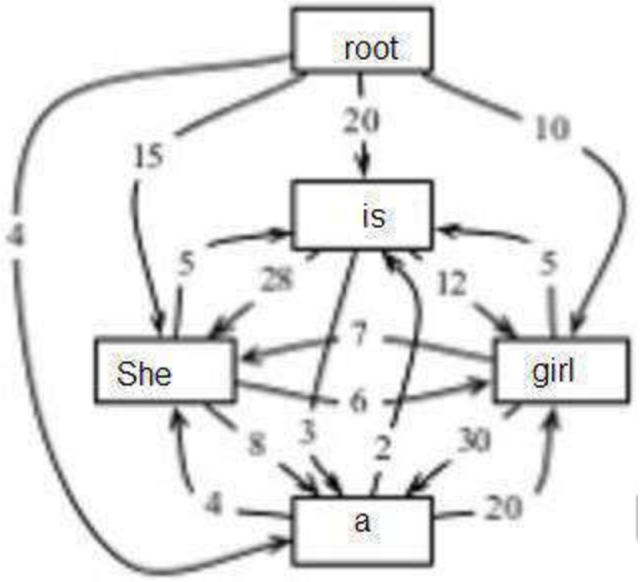
1.0 S → NP VP
0.3 NP → Adj Noun
0.7 NP → Det Noun
1.0 VP → Vb NP
-
0.2 Adj → fruit
0.2 Noun → flies
1.0 Vb → like
1.0 Det → a
0.4 Noun → banana
0.4 Noun → tomato
0.8 Adj → angry

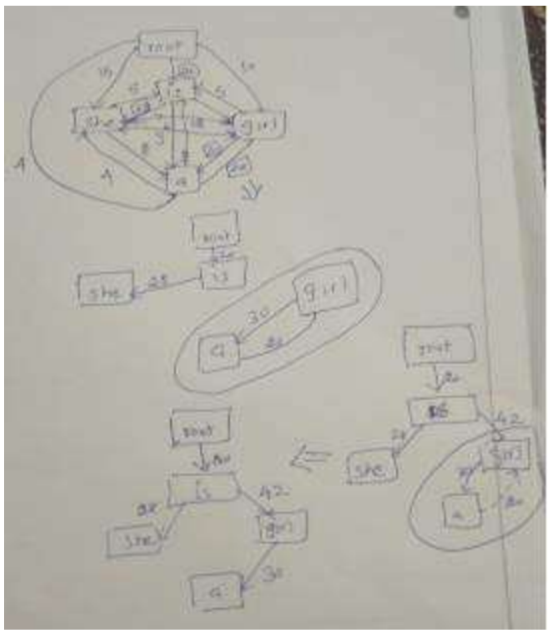
Solution:



$1 * 0.3 * 0.2 * 0.7 * 1.0 * 0.2 * 1 * 1 * 0.4 = 0.0033$

d) Find the dependency parse tree using Chu Lieu Edmonds algorithm **[5 marks]**





Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2020-2021
M.Tech (Data Science and Engineering)
Mid-Semester Test (EC-2 Regular)

Course No.	: DSECL ZG565
Course Title	: Natural Language Processing
Nature of Exam	: Open Book
Weightage	: 30%
Date of Exam	: 1st November, 2020

No. of Pages = 2
No. of Questions = *6*

Question 1. [3+2+3=8 Marks]

a) For each of the following sentences, please identify whether they are lexically, syntactically semantically and pragmatically correct

Solution:

1. Eats Ice-cream I in summer. - lexically correct
2. The fruits are flying in the blue sky. - lexically and syntactically correct
3. The baby is eating the chocolate wrapper. Lexically, syntactically and semantically correct

b) How many trigrams phrases can be generated from the following sentence, after replacing punctuations by a single space?

"Natural Language processing is very interesting, though not easy."

Solution: (Any one from 2 options correct)

Number of trigrams=8

<s> Natural Language, Natural Language processing, Language processing is, processing is very, is very interesting, very interesting though, interesting though not, though not easy

OR

Number of trigrams=9

<s> Natural Language, Natural Language processing, Language processing is, processing is very, is very interesting, very interesting though, interesting though not, though not easy, not easy </s>

c) Write the formulae to calculate the unigram, bigram and trigram probabilities of the below sentence

"Life should be great rather than long".

Solution:

Unigram

P (**"Life should be great rather than long"**)

=P(Life)P(should)P(be)P(great)P(rather)P(than)P(long)

Bigram

P (**"Life should be great rather than long"**)

=P(Life|<s>)P(should|Life))P(be|should)P(great|be)P(rather|great)P(than|rather)P(long|than)

Trigram

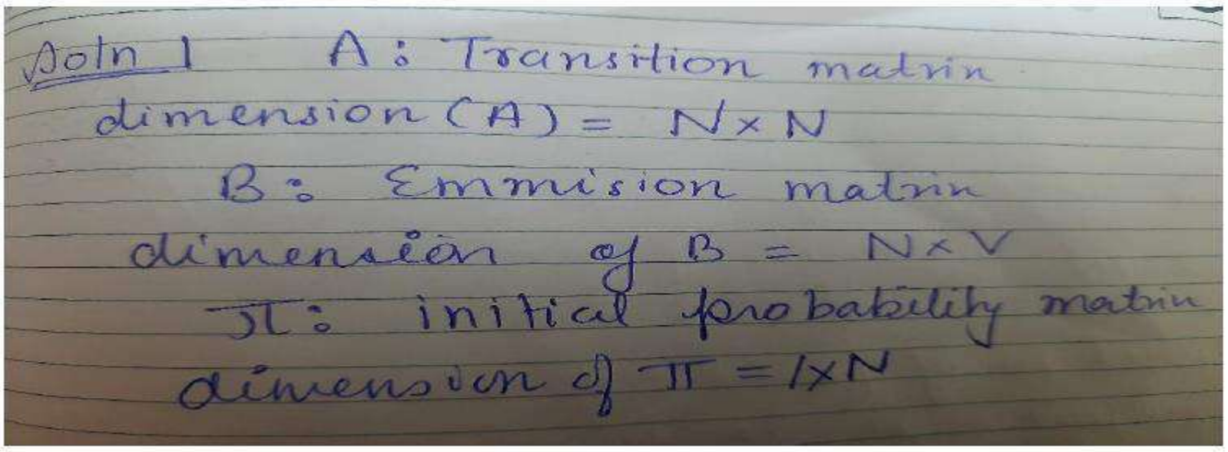
P (**"Life should be great rather than long"**)

=P(Life|<s>,<S>)P(should|Life,<s>))P(be|should,life)P(great|be, should)P(rather|great,be)

P(than|rather,great)P(long|than,rather)

Question 2. [2+5+3 =10 Marks]

a) For an HMM MODEL with N hidden states, V observations, what are the dimensions of parameter matrices A, B, and π? A: Transition matrix, B: Emmision matrix, and π: Initial Probability matrix.

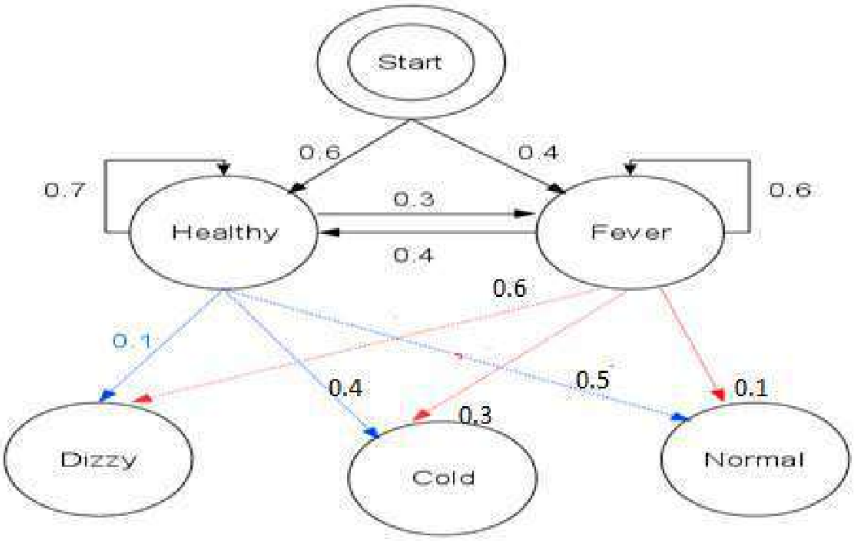


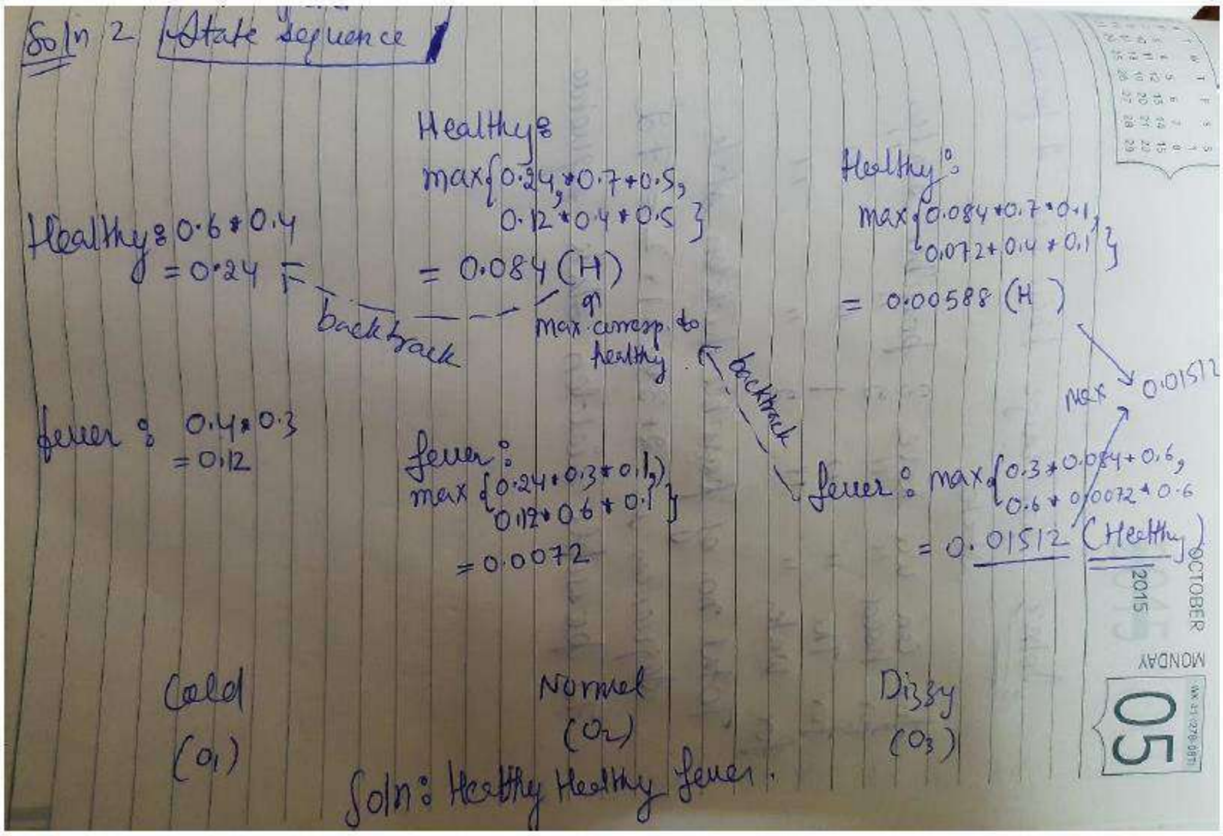
b) Consider an apartment where all residents are either healthy or have a fever and only the doctor can determine whether each has a fever. The doctor diagnoses fever by asking patients how they feel. The residents may only answer that they feel normal, dizzy, or cold.

The doctor believes that the health condition of his patients operates as a discrete Markov chain. There are two states, "Healthy" and "Fever", but the doctor cannot observe them directly; they are hidden from him. On each day, there is a certain chance that the patient will tell the doctor he is "normal", "cold", or "dizzy", depending on their health condition.

Set of Observations: {Dizzy, cold, Normal}, Set of states={Healthy, fever}

If the observation sequence is [cold normal dizzy]. Use Viterbi Algorithm to compute the corresponding state sequence.





c) Suppose you have a sentence “Large can can hold the water”. And you know the possible tags for each word in the sentence.

Large: N, V

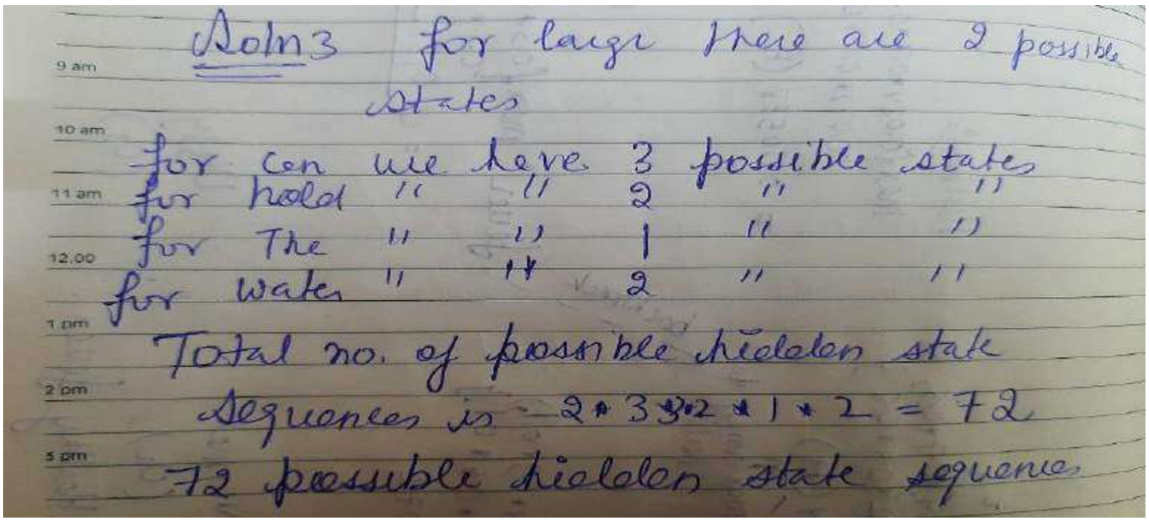
Can: V, Aux, N

Hold: N, V

The: article

Water: V, N

How many possible hidden state sequences are possible for the above sentence?



Question 3. [Marks 3+5+4=12 marks]

a) Given the grammar and lexicon below dervie the parse tree using top down parsing method for the sentence [3 marks]

S :The guy ate pizza

S->NP VP

VP->VNP

NP->Det N

N->pizza

N->guy ,Det ->the

V->ate

Solution:

1The 2 guy 3 ate 4 the 5pizza 6

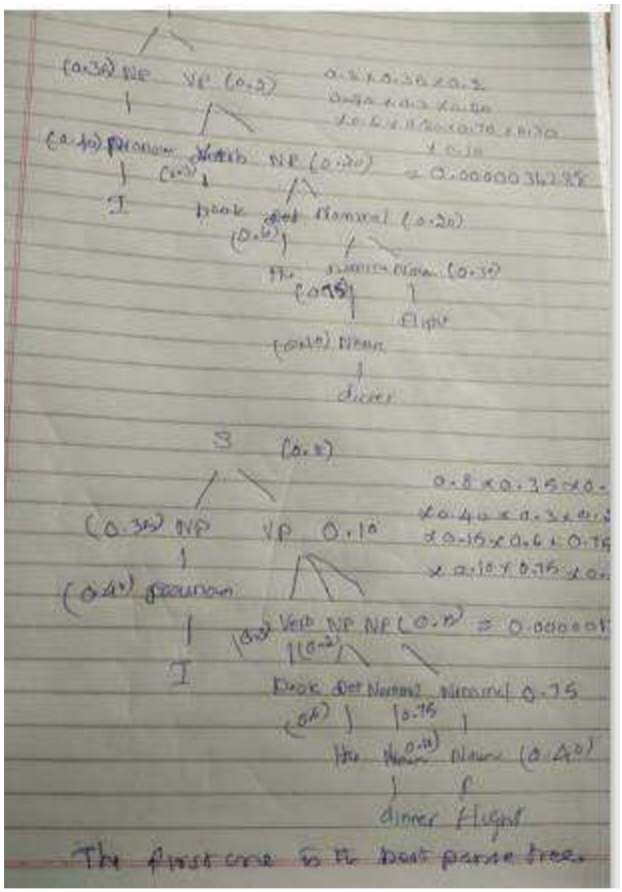
State	Backup State	Action
1.((S) 1)		
2.((NP VP) 1)		
3.(DT N VP) 1)		matches the
4.((N VP) 2)		matches guy
5.((VP)3)		
6.((V NP) 3)		matches ate
7.((Det N) 4)		matches the
8.((N))5		matches pizza

b) Given the grammar and lexicon below find the probabilty of the best parse tree using PCFG for the below sentence [5 marks]

S: I book the dinner flight

Grammar		Lexicon
$S \rightarrow NP\ VP$	[.80]	$Det \rightarrow that$ [.10] | a [.30] | the [.60]
$S \rightarrow Aux\ NP\ VP$	[.15]	$Noun \rightarrow book$ [.10] | $flight$ [.30]
$S \rightarrow VP$	[.05]	| $meal$ [.15] | $money$ [.05]
$NP \rightarrow Pronoun$	[.35]	| $flights$ [.40] | $dinner$ [.10]
$NP \rightarrow Proper\text{-}Noun$	[.30]	$Verb \rightarrow book$ [.30] | $include$ [.30]
$NP \rightarrow Det\ Nominal$	[.20]	| $prefer$; [.40]
$NP \rightarrow Nominal$	[.15]	$Pronoun \rightarrow I$ [.40] | she [.05]
$Nominal \rightarrow Noun$	[.75]	| me [.15] | you [.40]
$Nominal \rightarrow Nominal\ Noun$	[.20]	$Proper\text{-}Noun \rightarrow Houston$ [.60]
$Nominal \rightarrow Nominal\ PP$	[.05]	| NWA [.40]
$VP \rightarrow Verb$	[.35]	$Aux \rightarrow does$ [.60] | can [40]
$VP \rightarrow Verb\ NP$	[.20]	$Preposition \rightarrow from$ [.30] | to [.30]
$VP \rightarrow Verb\ NP\ PP$	[.10]	| on [.20] | $near$ [.15]
$VP \rightarrow Verb\ PP$	[.15]	| $through$ [.05]
$VP \rightarrow Verb\ NP\ NP$	[.05]	
$VP \rightarrow VP\ PP$	[.15]	
$PP \rightarrow Preposition\ NP$	[1.0]	

Solution:



c) Give the correct sequence of arc eager parsing operations for the given sentence [2marks]

a) Provide a modified transition sequence where the parser mistakenly predicts the arc cat → slept, but gets the other dependencies right. [2marks]

Solution:

c) **SH,SH,LA,LA,SH,LA,RA**

[ ]	[The lazy cat slept]	[]
[The]	[lazy cat slept]	[Shift]
[The ,lazy]	[cat slept]	[Shift]
[The ,lazy]	[cat slept]	[LA]
[The]	[cat slept]	[LA]
[cat]	[slept]	[SH]
[ ]	[slept]	[LA]
[slept]	[]	[RA]

OR

[ ]	[The lazy cat slept]	[]
[Root,The]	[lazy cat slept]	[Shift]
[Root ,the ,Lazy]	[cat slept]	[Shift]
[Root ,the ,Lazy]	[cat slept]	[LA]
[Root the]	[cat slept]	[LA]
[Root,Cat]	[slept]	[SH]
[ ]	[slept]	[LA]
[Root,Slept]	[]	[RA]
[Root]	[]	[RE]

d)

[ ]	[The lazy cat slept]	[]
[Root,The]	[lazy cat slept]	[Shift]
[Root ,the ,Lazy]	[cat slept]	[Shift]
[Root ,the ,Lazy]	[cat slept]	[LA]
[Root the]	[cat slept]	[LA]
[Root,Cat]	[slept]	[SH]
[Cat]	[]	[RA]
[Root,Cat ,Slept]	[]	[RE]
[Root,cat]	[]	[RE]
[Root]	[]	[RE]

Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
M. Tech. (Data Science Engineering)-Cluster program.
Second Semester 2020- 2021
Mid-Sem Examination
(Batch3 -3rd semester-Cluster program)

Course No	**: DSECLZG525**
Course Title	**: Natural Language processing**
Nature of Exam	**: Closed Book**
Weightage	**: 30%**
Duration	**: 2hours**
Date of Exam	**: 30-June-2021**

No. of Pages = 3
No. of Questions = 5

Session AN: 10 am 12 pm

1.

i) Find out the type of ambiguity (lexical or syntactic) and justify your answer [2marks]

a. Nicole saw the people with binoculars. (Syntactic)

b. They went to the bank. (Lexical)

Note: identification 1marks, Justification 1 mark

ii) a. Compute the bigram probability for P(read|books) and P(loves|she). You are given mini-corpus of seven sentences: [3 marks]

<s> Savita likes to read books</s>
<s> She read fictional books</s>
<s> Savita enjoy reading comic books also </s>
<s> She also likes to sing songs </s>
<s> She does not like EDM songs but she loves opera</s>
<s> The series of books she loves are Harry Potter and Game of Thrones</s>
<s> She is very possessive of her books and her song albums</s>

Ans - P(read|books) = 1/5
P(loves|she) = 2/6 = 1/3

b. Compute P (I want Thai) . P (I have to eat Chinese tomorrow) – from these bigram fragments [2+3=5M]

<start> I	.25	Want Thai	.01
Chinese tomorrow	.01	To eat	.26
Eat Chinese	.02	To have	.14
I want	.32	To spend	.09
I don't	.29	To be	.02
I have	.08	British food	.60

I have	.04	British restaurant	.15
Want to	.65	British cuisine	.01

Ans - P(I|<start>) * P(want|I) * P(Thai|want) = .25*.32*.01 = .0008

2. Assuming the grammar below, show how it would be used to derive the parse tree for the sentence using the top-down search strategy. (5 Mark)

The small fluffy cat went under the table

S –> NP VP

VP –> VP PP

VP –> VERB NP

VP –> VERB

NP –> DET NOM

NOM –> ADJ NOM

NOM –> NOUN

PP –> PREP NP

DET –> the

ADJ –> small

ADJ –> fluffy

NOUN –> cat

VERB –> went

PREP –> under

NOUN –> table

Ans: Refer to the Table 1 The final parse tree:
(S ((NP ((DET the) (NOM ((ADJ small) (NOM ((ADJ fluffy) (NOM (NOUN cat)))))))) (VP ((VP (VERB went)) (PP ((PREP under) (NP ((DET the) (NOM (NOUN table))))))))))

Step	Current State	Backup state	Comment
1	((S)1)		
2	((NP VP)1)		S to NP VP
3	((DET NOM VP)1)		
4	((NOM VP)2)		DET → the
5	((ADJ NOM VP)2)	((NOUN VP)2)	
6	((NOM VP)3)	((NOUN VP)2)	
7	((ADJ NOM VP)3)	((NOUN VP)2) ((NOUN VP)3)	
8	((NOM VP)4)	((NOUN VP)2) ((NOUN VP)3)	
9	((ADJ NOM VP)4)	((NOUN VP)2) ((NOUN VP)3) ((NOUN VP)4)	
10	((NOUN VP)4)	((NOUN VP)2) ((NOUN VP)3)	backtrack
11	((VP)5)	((NOUN VP)2) ((NOUN VP)3)	
12	((VP PP)5)	((NOUN VP)2) ((NOUN VP)3) ((VERB NP)5) ((VERB)5)	
13	((VERB PP)5)	((NOUN VP)2) ((NOUN VP)3) ((VERB NP)5)	
14	((PP)6)	((NOUN VP)2) ((NOUN VP)3) ((VERB NP)5)	
15	((PREP NP)6)	((NOUN VP)2) ((NOUN VP)3) ((VERB NP)5)	
16	((NP)7)	((NOUN VP)2) ((NOUN VP)3) ((VERB NP)5)	
17	((DET NOM)7)	((NOUN VP)2) ((NOUN VP)3) ((VERB NP)5)	
18	((NOM)8)	((NOUN VP)2) ((NOUN VP)3) ((VERB NP)5)	
19	((ADJ NOM)8)	((NOUN VP)2) ((NOUN VP)3) ((VERB NP)5) ((NOUN)8)	
20	((NOUN)8)	((NOUN VP)2) ((NOUN VP)3) ((VERB NP)5)	backtrack
21	(()9)		success!

Table 1: Top down search for *The small fluffy cat went under the table*

3. Given the sentence “I love to ride” and the HMM model shown in the table 1 below, compute the most probable POS tag sequence for the sentence using the Viterbi algorithm.[5 marks]

	I	*love*	*to*	*ride*
VB	0	.0093	0	.00008
TO	0	0	.99	0
NN	0	.0085	0	.00068
PPSS	.37	0	0	0

(a) Observation likelihoods

	VB	TO	NN	PPSS
<s>	.19	.0043	.041	.067
VB	.0038	.035	.047	.0070
TO	.83	0	.00047	0
NN	.0040	.016	.087	.0045
PPSS	.23	.00079	.0012	.00014

(b) Tag transition probabilities

Table 1: Viterbi model

Ans:

(b) $v_1(PPSS) = P(PPSS| < s >)P(I|PPSS) = 0.067 * 0.37 = 0.02479$
$v_2(VB) = v_1(PPSS) * P(VB|PPSS) * P(love|VB) = 0.02479 * .23 * .0093 = 0.00005302581$
$v_2(NN) = v_1(PPSS) * P(NN|PPSS) * P(love|NN) = 0.02479 * .0012 * .0085 = 2.52858e-7$
$v_3(TO) = max(v_2(VB) * P(TO|VB) * P(to|TO), v_2(NN) * P(TO|NN) * P(to|TO)) = max(0.00005302581 * .035 * .99, 2.52858e-7 * .016 * .99) = max(0.00000183734, 4.00527072e-9) = 0.00000183734$
$v_4(VB) = v_3(TO) * P(VB|TO) * P(ride|VB) = 0.00000183734 * .83 * .00008 = 1.2199938e-10$
$v_4(NN) = v_3(TO) * P(NN|TO) * P(ride|NN) = 0.00000183734 * .00047 * .00068 = 5.8721386e-13$
So the HMM sequence is I/PPSS love/VB to/TO ride/VB

4. Describe the following for PCFG (2 Marks + 2 Marks = 4 Marks)

(a) Given a corpus, how would you compute the probability for the rule: VP –> VERB NP PP?

Ans:

(a) If we have access to a corpus of parsed sentences, we can compute the probability of the given rule by counting the number of times that rule appears in any parse and then normalizing it.

$$P(\text{VP} \rightarrow \text{VERB NP PP}|\text{VP}) = \frac{count(\text{VP} \rightarrow \text{VERB NP PP})}{count(\text{VP})}$$

(b) How is the probability of a parse tree computed in a PCFG?

(b) The probability of a parse tree T is computed as the product of the probabilities of all n non-terminal nodes in the parse tree, where each rule i can be expressed as $LHS_i \rightarrow RHS_i$:

$$P(T) = \prod_{i=1}^{n} P(RHS_i|LHS_i)$$

Note: If students write it I in theory without formula also , give marks.

5. Consider the following PCFG (Refer to table 2): (3 Marks + 3 Marks = 6 Marks)

production rule	probability
S → VP	1.0
VP → Verb NP	0.7
VP → Verb NP PP	0.3
NP → NP PP	0.3
NP → Det Noun	0.7
PP → Prep Noun	1.0
Det → the	0.1
Verb → cut | eat | ask	0.1
Prep → with | in	0.1
Noun → noodles | grandma | chopsticks | man | suits | summer | ...	0.1

Table 2: PCFG

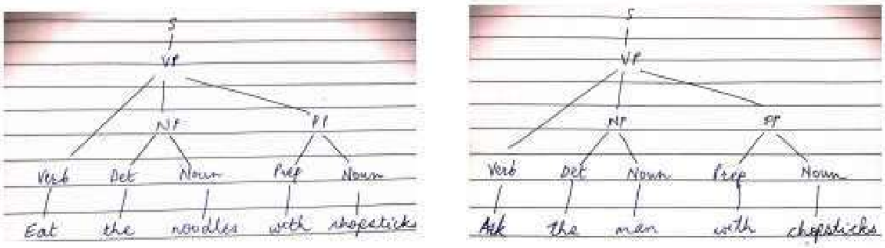
(a).Draw the top-ranked parse tree for the sentence below by applying the given PCFG. Does the result seem reasonable to you? Why or why not? Eat the noodles with chopsticks

(b) Draw the top-ranked parse tree for the sentence below by applying the given PCFG. Does the result seem reasonable to you? Why or why not? Ask the man with chopsticks

Ans:

(a) Refer to the parse tree in Figure 1a. This shows the top-ranked parse with probability $1.0 \times 0.3 \times 0.7 \times 1.0 \times (0.1)^5$ which is greater than the probability of the other possible parse with VP $\rightarrow$ Verb NP. Semantically, "with chopsticks" should attach to verb and hence the resulting parse tree is a reasonable one.

(b) Refer to the parse tree in Figure 1b. This shows the top-ranked parse with probability $1.0 \times 0.3 \times 0.7 \times 1.0 \times (0.1)^5$ which is greater than the probability of the other possible parse with VP $\rightarrow$ Verb NP. Here, "with chopsticks" should attach to the noun phrase and hence, this is not a reasonable parse.



(a) *Eat the noodles with chopsticks*

(b) *Ask the man with chopsticks*

Figure 1: Parse trees

Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2020-2021
M.Tech (Data Science and Engineering)
End-Semester Test (EC-3 Makup)

Course No. : DSECLZG525
Course Title : Natural Language Processing
Nature of Exam : Open Book
Weightage : 50%

No. of Pages = 4
No. of Questions = 5

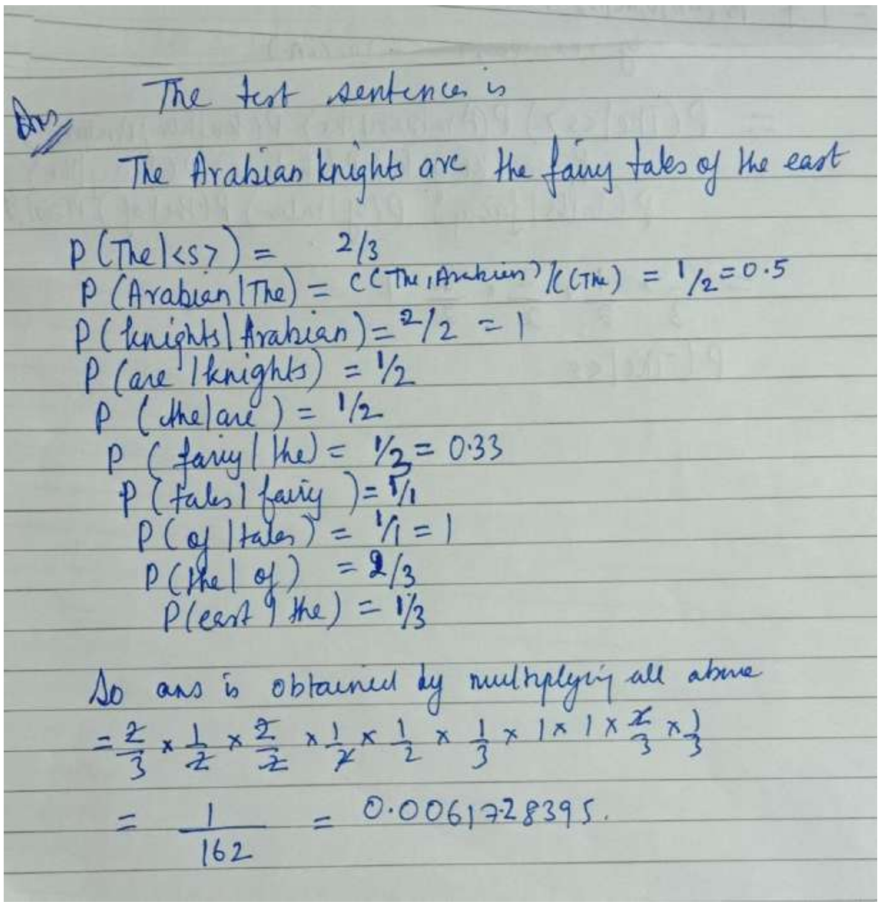
Note: Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1.

a) Consider the training set: (4 marks)

The Arabian knights
These are the fairy tales of the east
The stories of the Arabian knights are translated in many languages

Compute using the bigram model the probability of the sentence. Include start and end symbol in your calculations.
The Arabian knights are the fairy tales of the east



b) Using Penn Tree bank, find the POS tag sequence for the following sentences: [6 Marks]

1. The actor was happy he got a part in a movie even though the part was small. [2 marks]
2. I am full of ambition and hope and charm of life. But I can renounce everything at the time of need [3 marks]
3. When the going gets tough, the tough get going. [1 mark]

Solution

The/DT actor/NN was/VB happy/JJ he/PRP got/VB a/DT part/NN in/IN a/DT movie/NN "even though"/CC the/DT part/NN was/VB small/ADV. [2 marks]

I//PRP am/VB full/JJ of/IN ambition/NN and/CC hope/NN and/CC charm/JJ of/IN life/NN. But/CC I/PRP can/VB renounce/VB everything/JJ at/IN the/DT time/NN of/IN need/NN [3 marks]

When/WDT the/DT going/NN gets/VB tough/RB, the/DT tough/NN get/VB going/RB.[1 mark]

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

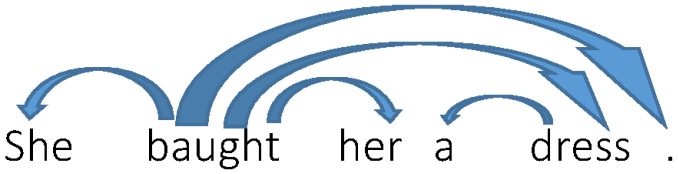
Question 2.

a) Build a parse tree for the sentence "She loves to visit Goa" using Probabilistic Parsing [5marks]

S → NP VP 1.0
VP → V PP 0.4
VP → V NP 0.6
PP → P NP 1.0
NP -> V NP 0.1
NP → NP PP 0.3
NP → N 0.3
N → visit 0.3
V → visit 0.6

N → Goa 0.3
N → She 0.5
V → loves 1
P → to 1
DT→ a 1

a) State the correct sequence of actions that generates the following parse tree of the sentence “She baught her a dress" using Arc-Eager Parsing [5marks]



Solution:
Transitions: SH-LA-SH-RA-SH-LA-RE-RA-RE-RA
Arcs:
She <- baught
baught _> her
a <- dress
baught -> dress
baught -> .

Question 3. Word sense disambiguation and ontology-

b) What are lexical sample task and all word task in word sense disambiguation? How can sources like Wikipedia be used for word sense disambiguation [2 marks]

Solution

What are lexical sample task and all word task in word sense disambiguation?

Lexical sample task and all word task are 2 variants of word sense disambiguation

- Lexical sample task -Small pre-selected set of target words
- All-words task - System is given an all-words entire texts and lexicon with an inventory of senses for each entry. We have to disambiguate every word in the text (or sometimes just every content word).

How can sources like Wikipedia be used for word sense disambiguation

Wikipedia can be used as training data for word sense disambiguation using supervised learning techniques

- Concept is mentioned in a Wikipedia: article text may contain an explicit link to the concept’s Wikipedia page, which is named by a unique identifier (can be used as a sense annotation)
- These sentences can then be added to the training data for a supervised system.

How can WordNet relations be used for word sense disambiguation in following sentences: [3 marks]

1. A bat is not a bird, but a mammal.
2. Jaguar reveals its quickest car ever
3. Raghuram Rajan was the 23rd Governor of the Reserve Bank of India

Solution

Nouns and verbs can be extracted from the sentences. The senses in wordnet can be extracted for these words and senses with close relations can be extacted as correct sense.

1. Bat can be sports bat or mammal. But looking at nouns bat, bird and mammal, correct sense of bat as MAMMAL can be found using WordNet relations.
2. Jaguar can be a car or animal. Looking at nouns Jaguar, correct sense of Jaguar as CAR can be found using WordNet relations.
3. Bank can be river bank or financial bank.: Search senses of nouns Bank,"Raghuram Rajan", Governer. The correct sense of BANK as FINANCIAL sense can be found using WordNet relations.

c) How is Syntactic web different from the Semantic web? What is URI in semantic web ontology? [2 marks]

Syntactic web consist of huge data on net connected by hyperlinks which is rendered by machines but machines cannot process it due to inability to understand the meaning of the content.

The semantic Web identifies a set of technologies, tools, and standards which form the basic building blocks of an infrastructure to support the vision of the Web associated with meaning.

A Universal Resource Identifier (URI) is a formatted string that serves as a means of identifying abstract or physical resource. A URI can be further classified as a locator, a name, or both. Every resource is identified with unique URI in ontology.

Develop an OWL ontology using the following for animal kingdom for classes like carnivorous, herbivorous and omnivorous. Use following Property characteristics, restrictions and Class expressions [3 marks]

- inverseOf
- domain
- range
- Cardinality
- disjointWith
- subClassOf

```
<rdfs:Class rdf:ID="Carnivorous">
  <rdfs:subClassOf rdf:resource="#Animal"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Herbivorous">
  <rdfs:subClassOf rdf:resource="#Animal"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Omnivorous">
  <rdfs:subClassOf rdf:resource="#Animal"/>
</rdfs:Class>

<rdfs:Class rdf:ID="Carnivorous">
  <owl:disjointWith rdf:resource="#Herbivorous"/>
</rdfs:Class>
<owl:Restriction>
   <owl:onProperty rdf:resource="#hasLegs" />
   <owl:cardinality
 rdf:datatype="&xsd;nonNegativeInteger">4</owl:cardinality>
 </owl:Restriction>
 <owl:ObjectProperty rdf:ID="Eats">
   <rdfs:domain rdf:resource="#Carnivorous"/>
```

```
    <rdfs:range  rdf:resource="#Animal"/>
  </owl:ObjectProperty>
```

Question 4.

a) In this modern age where the internet is growing rapidly, the existence of the internet can make it easier for tourist to find information about hotels. Tourists usually tell the experience during the hotel by writing reviews on the internet. Hence many hotel's reviews are found on the internet. With the availability of reviews on the internet with large numbers, tourists can't understand all the reviews they read whether they contain positive or negative opinions. It takes a sentiment analysis to quickly detect if the reviews is a positive or negative reviews. Using the Multinomial Naïve Bayes Classifier method find out that the given hotel reviews are positive or negative.

D1	The hotel is clean and great	Positive
D2	The hotel owner is very helpful	Positive
D3	Overall Aston Hotel's experience was great	Positive
D4	The condition of the hotel was very bad	Negative
D5	A HORRIBLE EXPERIENCE FOR ONE WEEK	Negative
D6	The hotel view was great	?
D7	My holiday experience stay in usa so horrible	?
D8	Overall the hotel in aston very clean and great	?

Soln :