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 I**I) 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 II**I) 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

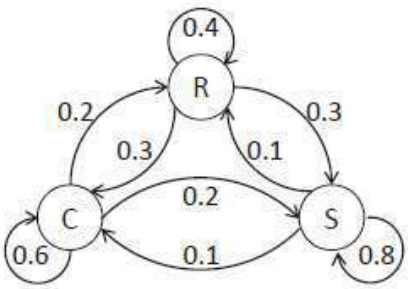
Maximum of all these sequences correspond to T2 T1 T3.

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

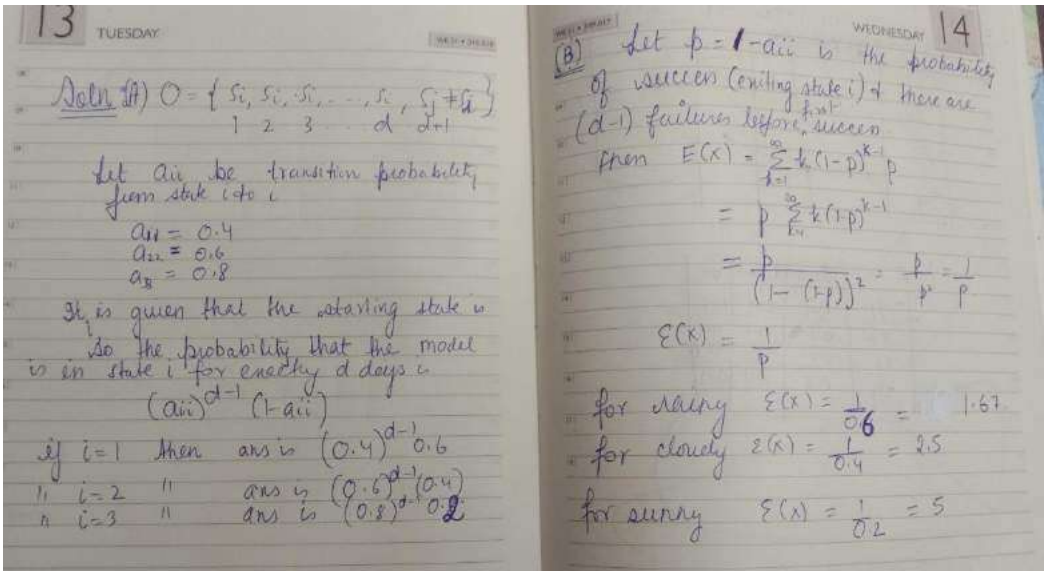
The joint probability for the most probable sequence is 0.0086

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).



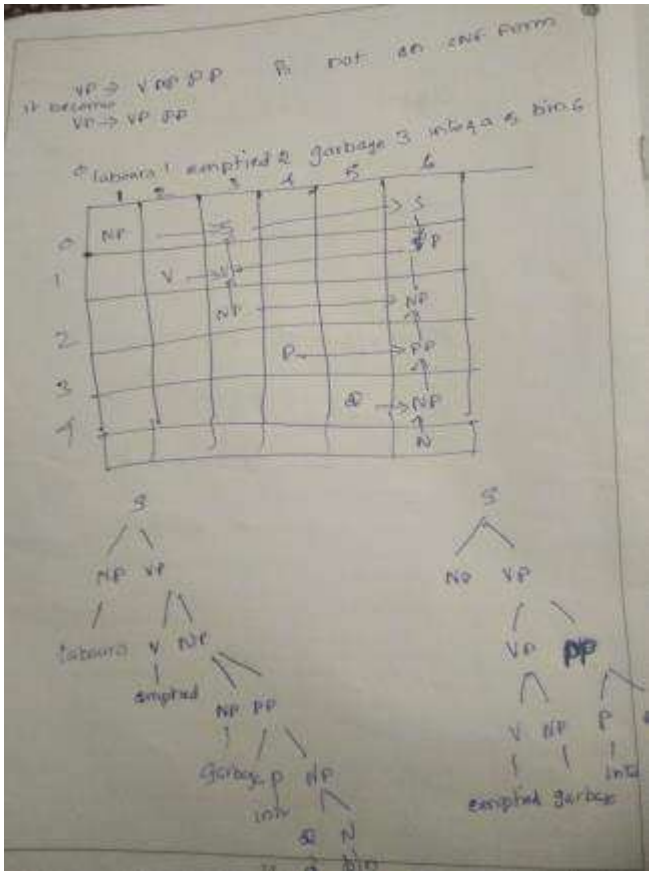
Question 3. [Marks 5+2+5=12 marks]

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"
```

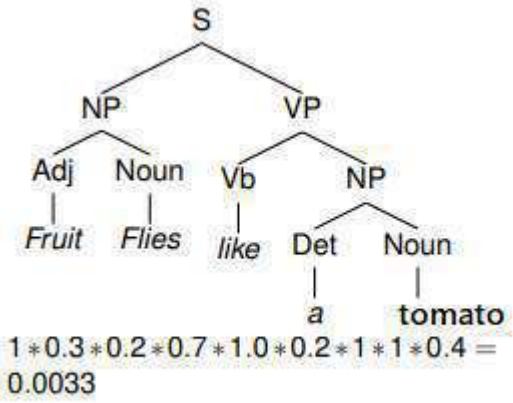


c) Find the probability of the sentence “Fruits flies like a tomato” using PCFG parsing method

[2 Marks]

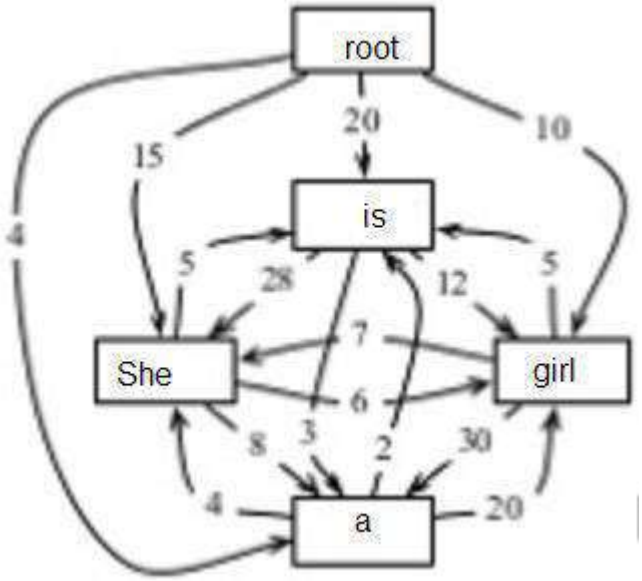
1.0 S → NP VP
0.3 NP → Adj Noun
0.7 NP → Det Noun
1.0 VP → Vb NP
-
0.2 Adj → fruit
0.2 Noun → flies
1.0 Vb → like
1.0 Det → a
0.4 Noun → banana
0.4 Noun → tomato
0.8 Adj → angry

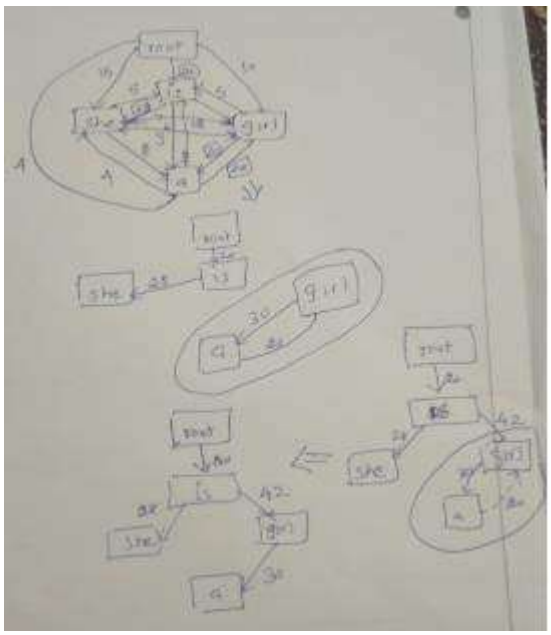
Solution:



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

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)

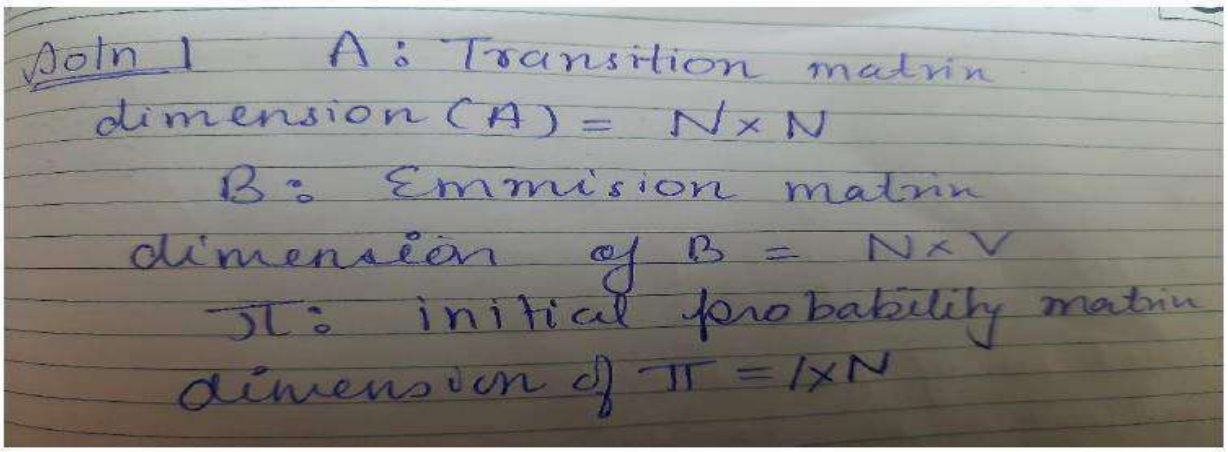
Trigram

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.

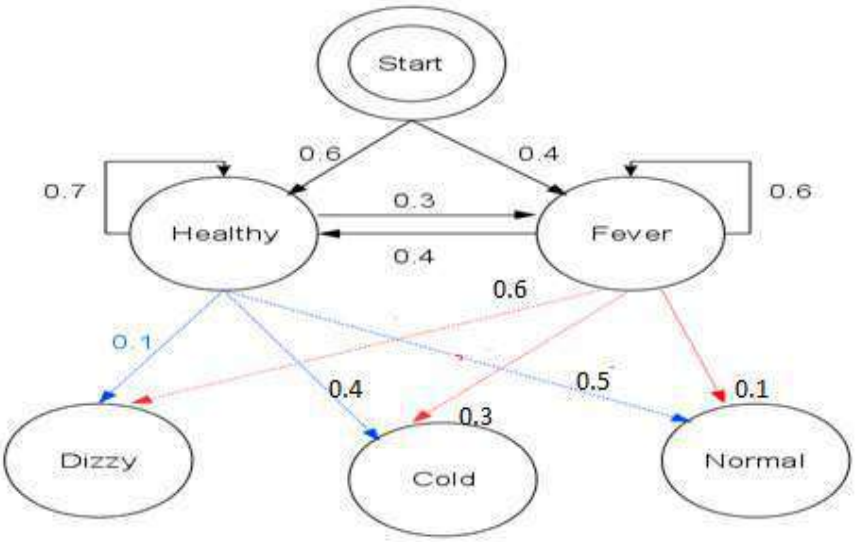


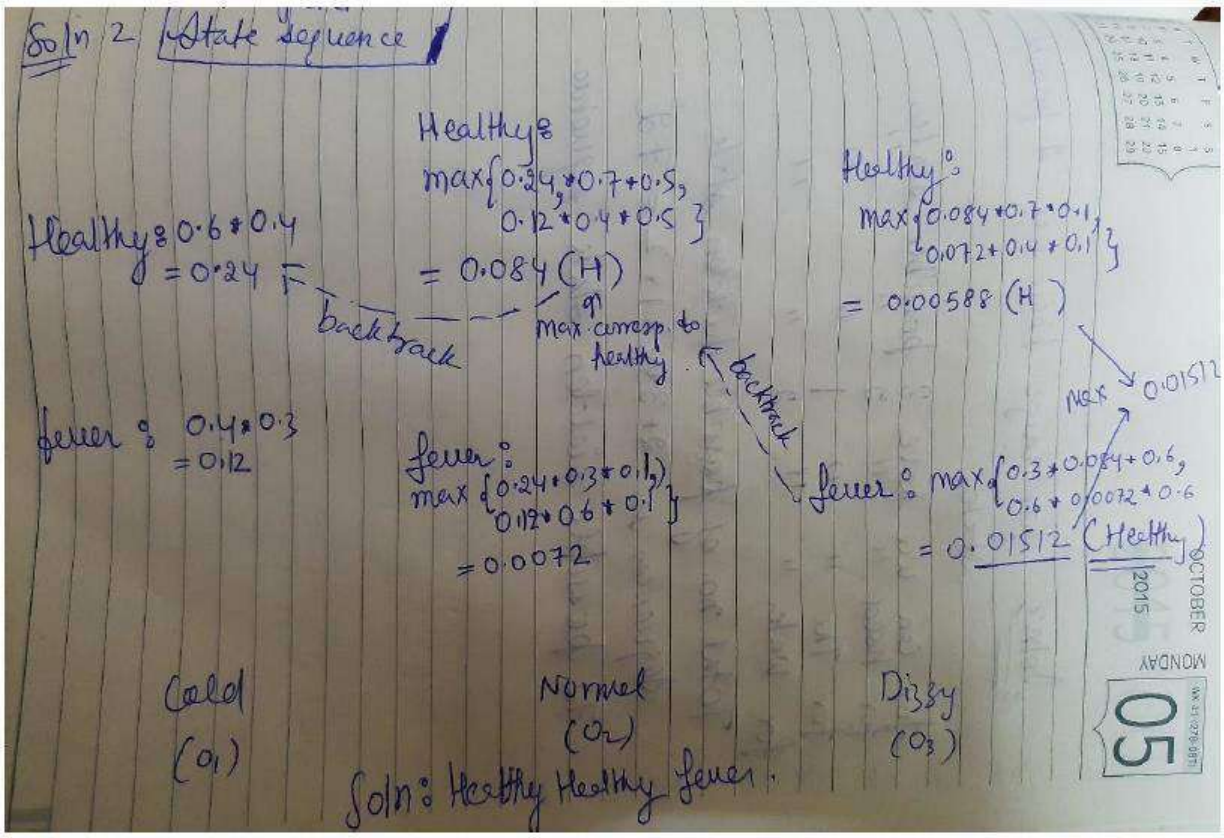
b) Consider an apartment where all residents are either healthy or have a fever and only the doctor can determine whether each has a fever. The doctor diagnoses fever by asking patients how they feel. The residents may only answer that they feel normal, dizzy, or cold.

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.

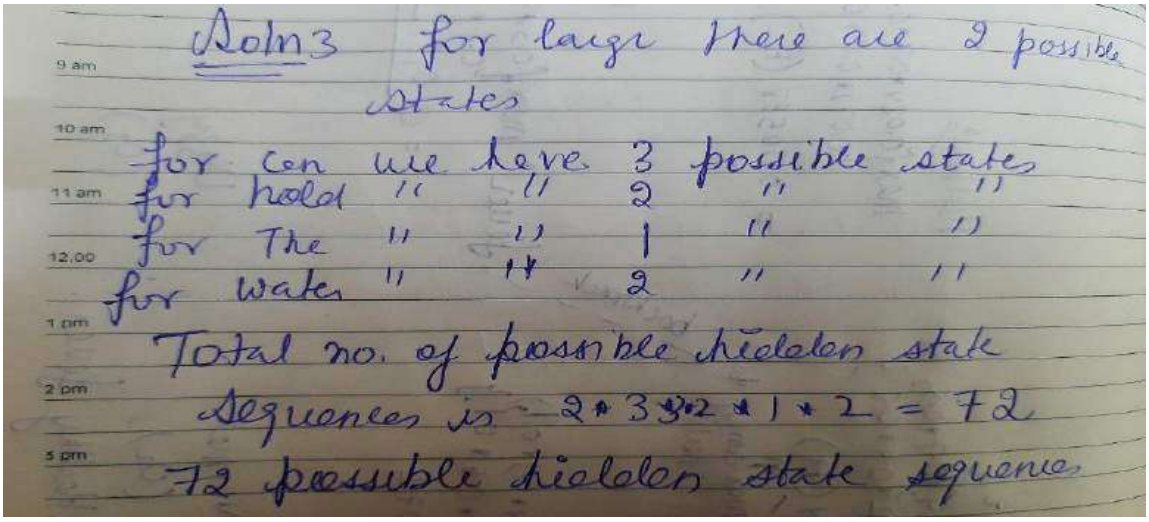
Set of Observations: {Dizzy, cold, Normal}, Set of states={Healthy, fever}

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

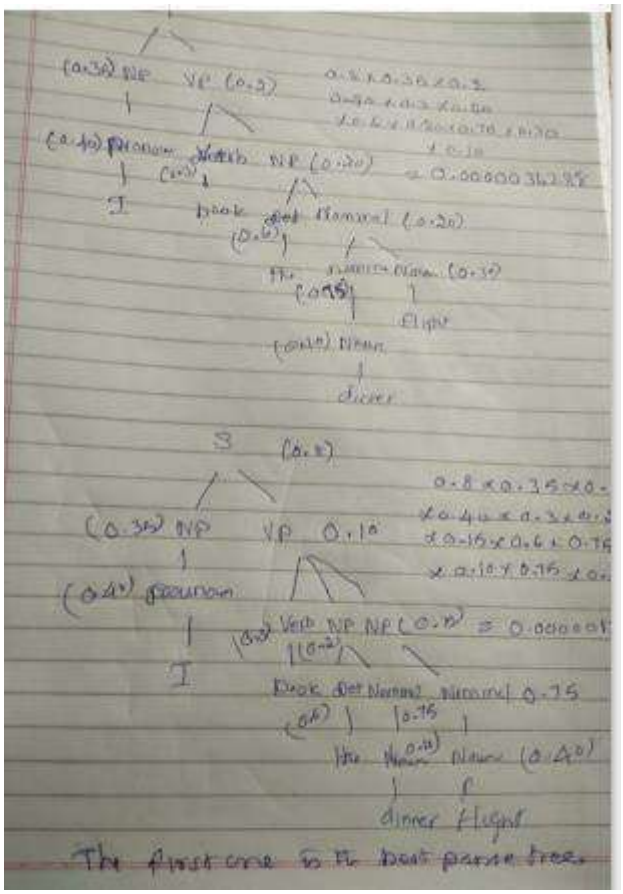
State	Backup State	Action
1.((S) 1)		
2.((NP VP) 1)		
3.(DT N VP) 1)		matches the
4.((N VP) 2)		matches guy
5.((VP)3)		
6.((V NP) 3)		matches ate
7.((Det N) 4)		matches the
8.((N))5		matches pizza

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

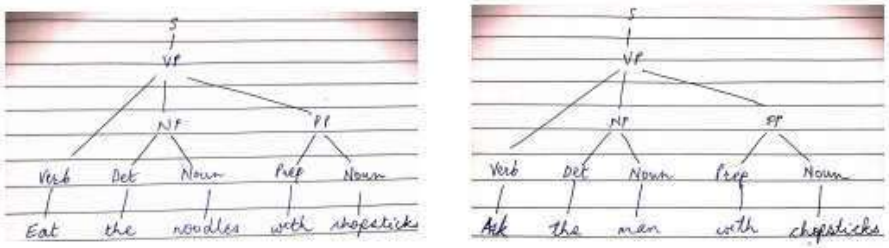
(a).Draw the top-ranked parse tree for the sentence below by applying the given PCFG. Does the result seem reasonable to you? Why or why not? Eat the noodles with chopsticks

(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

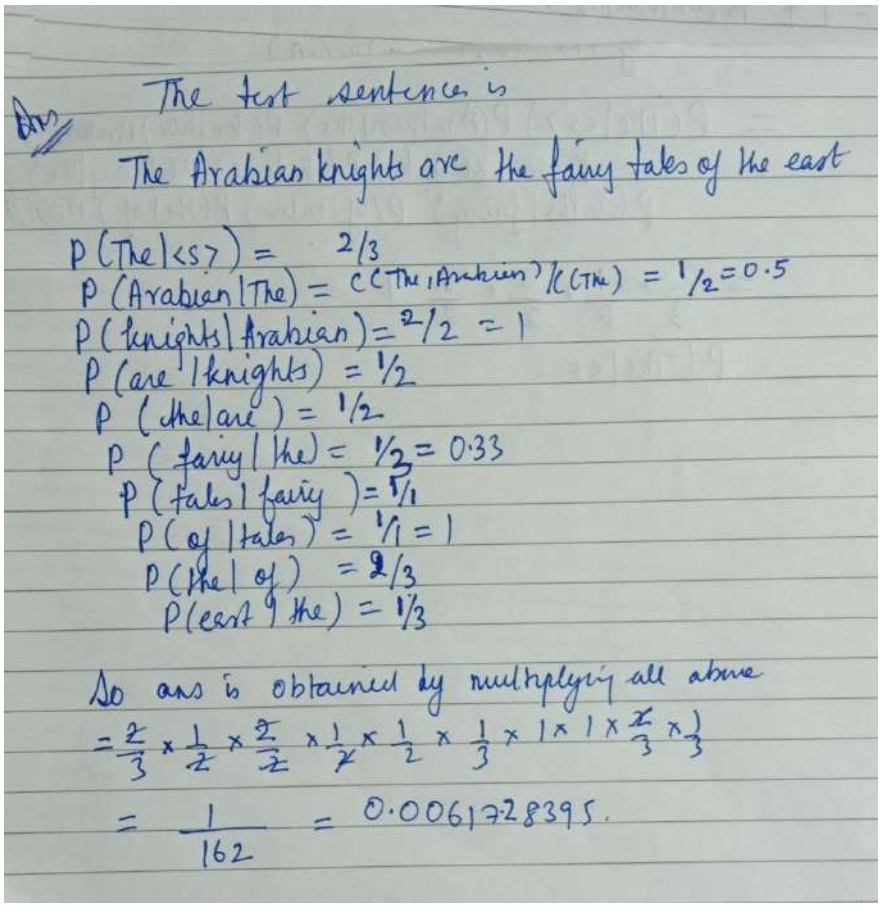
Note: Assumptions made if any, should be stated clearly at the beginning of your answer.

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*			

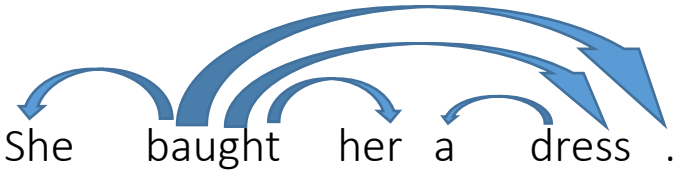
Question 2.

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"/>
```

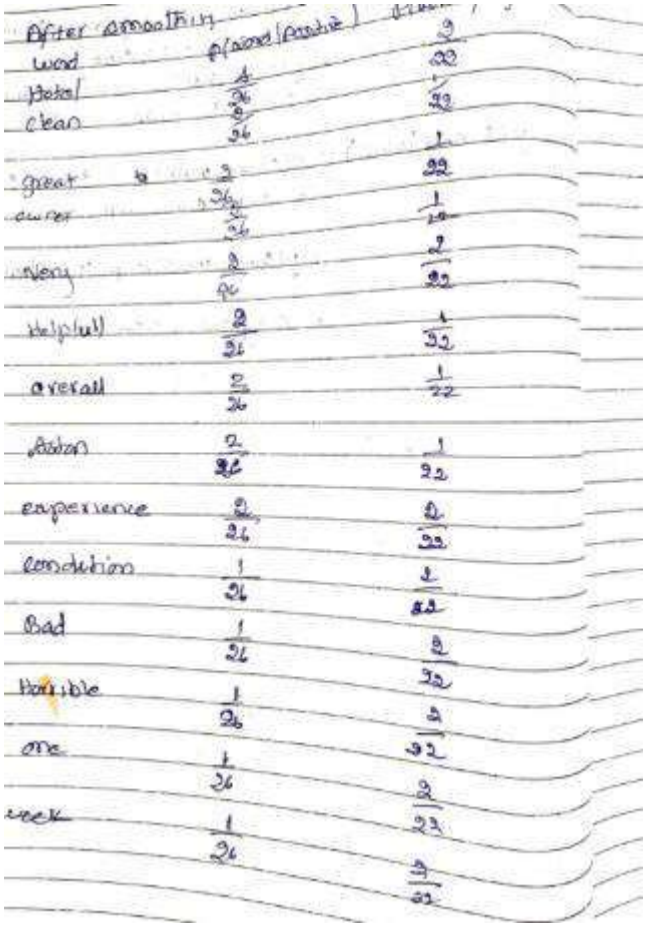
```
  <rdfs:range  rdf:resource="#Animal"/>
</owl:ObjectProperty>
```

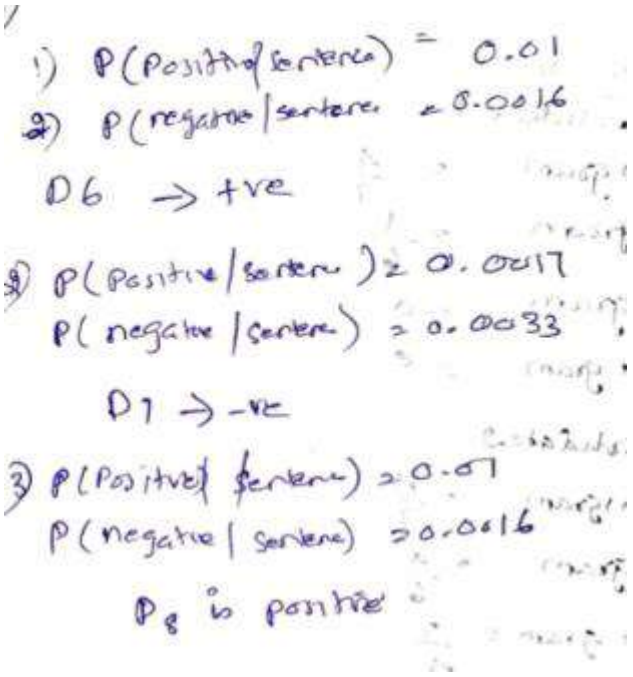
Question 4.

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 :

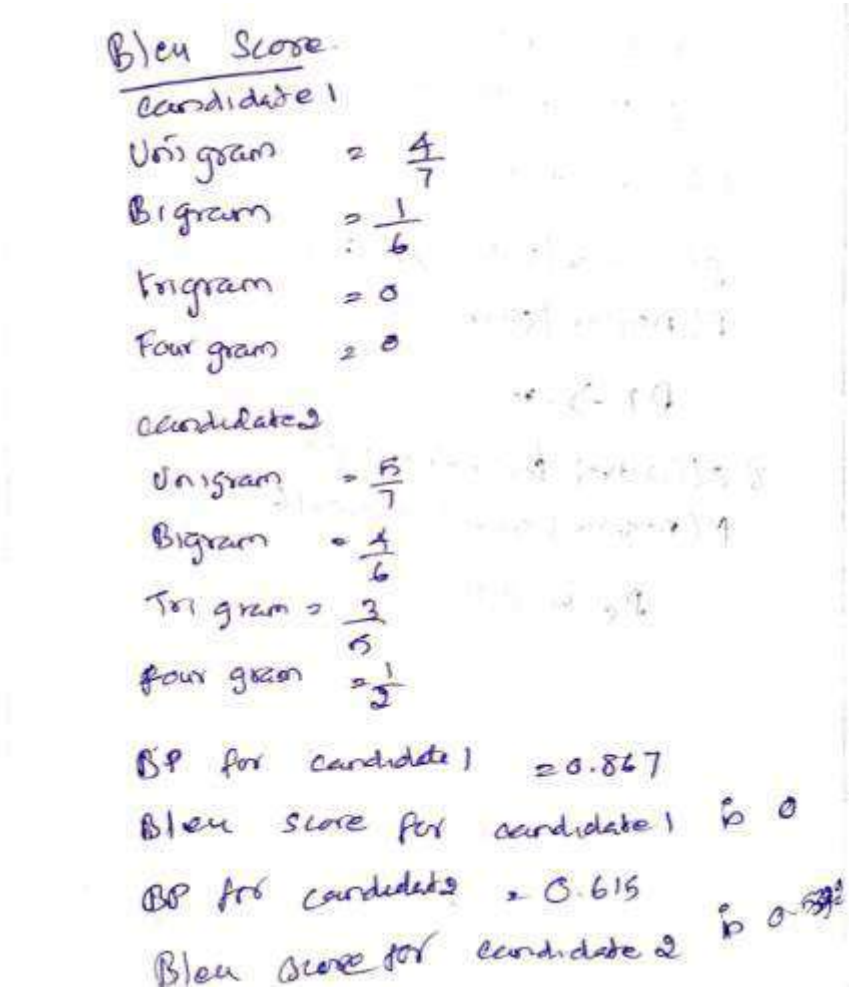




b. Compute the BLEU score for the below translations (candidate1, candidate2).Consider 1gram,2 gram,3 gram,4 gram and Brevity-Penalty for calculating BLUE score .

Reference: The teacher arrived late because of the traffic

Candidate 1: The teacher was late due to the traffic
Candidate 2: A teacher arrived late because of transportation

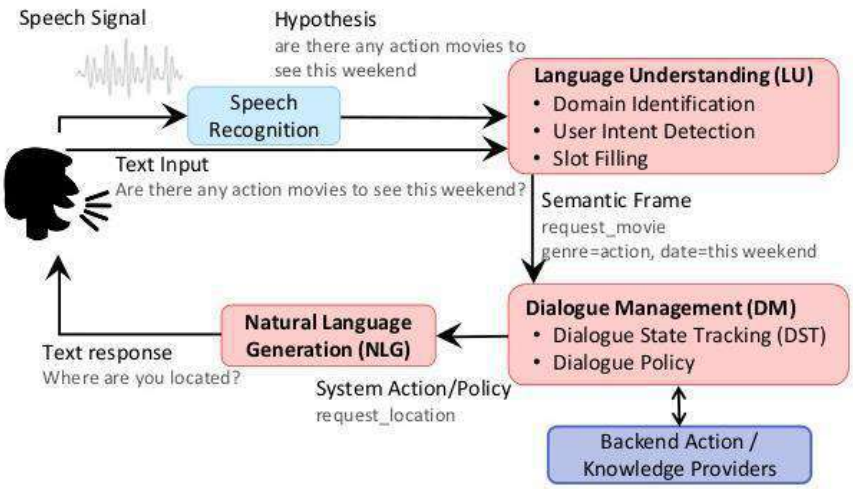


1. Identify the type of the dialogue agent from the following conversation. Also explain the working of that type of dialogue agent. (5 marks)

   System: How may I help you?
   User: I want to book dinner at a restaurant that serves good Chinese food
   System: Berkos is a restaurant that serves good Chinese food.
   User: ok
   System: Should I book a dinner?
   User: yes, please
   System: ok.

   Soln: Above is the task oriented dialogue system

The diagram shows the working of the task oriented dialogue system. The student should explain briefly each step as mentioned in the slides



2. Find the intent, domain and slots for the following: (2 marks)
   Book an appointment on 12th Feb 2021 at 10 am for a ECG Test.
   SOLN:
   DOMAIN: Medical
   INTENT: Book an Appointment

   Slots

   - Services: ECG TEST
   - Date: 12th Feb 2021
   - Time: 10 AM

3. In a collection of 10000 document, the following words occur in the following number of documents: (3 marks)
   Oasis occurs in 400 documents, Place occurs in 3500 documents, Desert occurs in 800 documents, Water occurs in 800 documents, Comes occur in 800 documents
   Beneath occurs in 200 documents, Ground occurs in 900 documents
   Calculate TF-IDF term vector for the following document:
   Oasis Place Desert Water Comes Beneath Ground Place

Doc	(TF) Term freq.	IDF	TF*IDF
Oasis	1/8	log(10000/400)	0.1747
Place	2/8	log(10000/3500)	0.11398
Desert	1/8	log(10000/800)	0.137114
Water	1/8	log(10000/800)	0.137114
Cenes	1/8	log(10000/800)	0.137114
Beneath	1/8	log(10000/200)	0.212371
Ground	1/8	log(10000/900)	0.13072

TF-IDF vector (0.1747, 0.11398, 0.137114, 0.137114, 0.137114, 0.212371, 0.13072).

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 Regular)

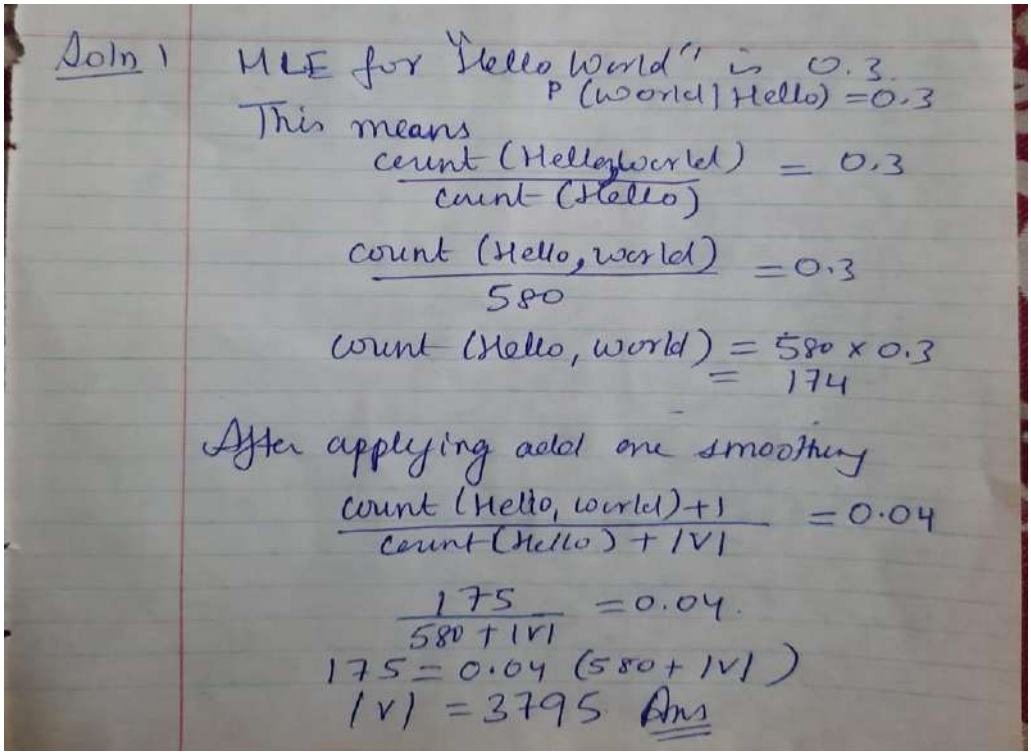
Course No.	: DSECLZG525
Course Title	: Natural Language Processing
Nature of Exam	: Open Book
Weightage	: 50%
Duration	: 2 hours

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

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

Question 1.

a) Given a corpus C, the maximum likelihood estimation (MLE) for the bigram "Hello World" is 0.3 and the count of occurrence of the word "Hello" is 580 for the same corpus, the likelihood of ""Hello World" after applying the add-one smoothing is 0.04. What is the vocabulary size of Corpus C.
(3 marks)



b) What are the challenges in the Natural Language Processing? (3 marks)

Natural Language Processing has following challenges:

- Contextual words and phrases and homonyms

The same words and phrases can have different meanings according the context of a sentence and many words – especially in English – have the exact same pronunciation but totally different meanings.

- Synonyms

Synonyms can lead to issues similar to contextual understanding because we use many different words to express the same idea.

- Irony and sarcasm

Irony and sarcasm present problems for machine learning models because they generally use words and phrases that, strictly by definition, may be positive or negative, but actually connote the opposite

- Ambiguity

Lexical ambiguity: a word that could be used as a verb, noun, or adjective.
Semantic ambiguity: the interpretation of a sentence in context. For example: I saw the boy on the beach with my binoculars. This could mean that I saw a boy through my binoculars or the boy had my binoculars with him
Syntactic ambiguity: In the sentence above, this is what creates the confusion of meaning. The phrase with my binoculars could modify the verb, "saw," or the noun, "boy."

- Errors in text or speech

Misspelled or misused words can create problems for text analysis. Autocorrect and grammar correction applications can handle common mistakes, but don't always understand the writer's intention.

- Colloquialisms and slang

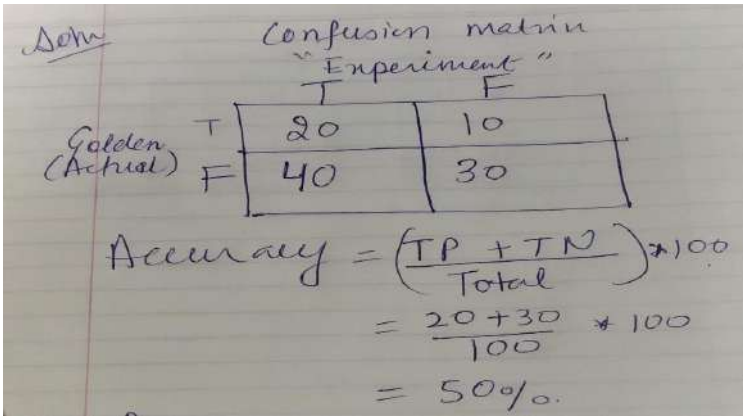
Informal phrases, expressions, idioms, and culture-specific lingo present a number of problems for NLP – especially for models intended for broad use.

- Domain-specific language

Different businesses and industries often use very different language. An NLP processing model needed for healthcare, for example, would be very different than one used to process legal documents.

- Lack of research and development

c) There were 100 documents and each document contained one word. 30 of these documents contained the word "hello". I asked Bob to separate all the documents containing the word "hello". He showed me 60 but "hello" was not in 40 of them. Construct the confusion matrix and calculate the accuracy. (4 marks)



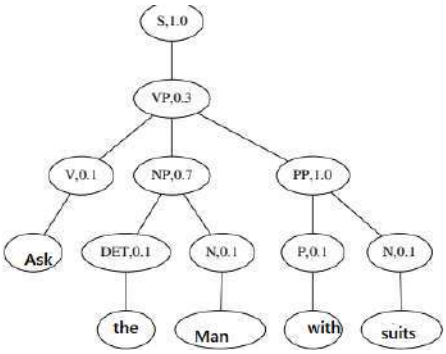
Question 2.

Given the following PCFG, find the parse trees for the given sentence and their probabilities .And find out that the word 'suits' is attached with' ask 'or 'man' and why? [10 marks]

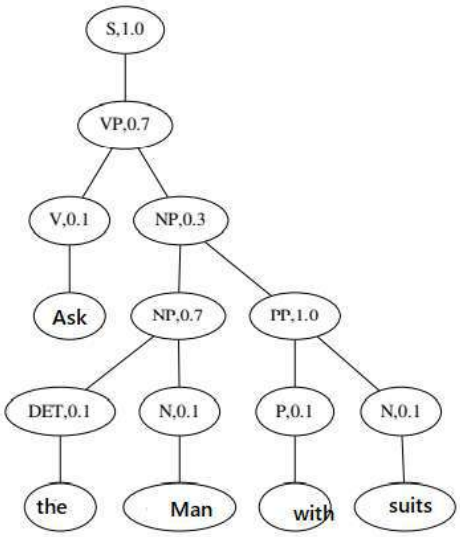
Ask the man with suits

Rule	probability
S → V P	1.0
V P → V NP	0.7
V P → V NP P P	0.3
NP → NP P P	0.3
NP → DET N	0.7
P P → P N	1.0
DET → the	0.1
V → ask	0.1
P → with	0.1
N → man | suits	0.1

Soln:



Probability= $0.3 \times 0.7 \times 0.1^{5} = 21 \times 10{-}7$



Probability= $0.3 \times 0.7 \times 0.7 \times 0.1^{5} = 14.7 \times 10{-}7$

The first tree has higher probability and it is the correct parse since 'with suits' should attach to 'ask' rather than 'man'.

Question 3. Word sense disambiguation and ontology-

a) How can the Simple Lesk algorithm be applied to disambiguate the exact meaning of "**bass**" in following sentence **[5 marks]**
The **bass** guitar, is the lowest pitched member of the guitar family of instruments.

S:(n) bass (the lowest part of the musical range)
S: (n) sea bass, bass (the lean flesh of a saltwater fish of the family Serranidae)
S: (n) bass (the member with the lowest range of a family of musical instruments)
S: (adj) bass, deep (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

b) Build a small part of ontology for MTech DSE program in OWL syntax with following concepts **[3 marks]**

- Professor
- Student
- Courses
  Also include following relations/constraints:
  - Domain
  - Range
  - subClassOf
  - disjointWith

  How are the ontology languages OWL and RDF different from each other. Can you express the same constraints using RDF? If not which one cannot be expressed using RDF? [2 **marks]**

  ```
  <rdfs:Class rdf:ID=" Professor">
    <rdfs:subClassOf rdf:resource="# AcademicStaff "/>
  </rdfs:Class>
  <rdfs:Class rdf:ID="Professor">
    <owl:disjointWith rdf:resource="#AssistantProfessor"/>
  </rdfs:Class>
  ```

  OWL is more advanced and has inferencing capability since owl is based on description logic. Some contraints like disjoint with cannot be expressed using RDF

Question 4.

1. Given the two machine translation systems output and reference given below, find the best machine translation system using BLEU score with Brevity penality. [5marks]

[Hint: Assume 1-gram, 2-gram, 3 -gram and 4- gram for calculating BLEU score)

System A: Israeli official's responsibility of airport safety

System B: Airport security Israeli officials are responsible

Reference: Israeli officials are responsible for airport security

2. Given the following documents and their sentiment polarities [5 marks]

Document	Sentiment words	Polarity
D1	Great, Enjoy, Great	Positive
D2	Poor, Unpleasant	Negative
D3	Enjoy ,amazing	Positive
D4	Great, Lovely	Positive
D5	Great, Poor, Rude	Negative
D6	Great ,amazing	?

Determine the sentiment polarity of document D6 using the multinomial naïve Bayes classification (with add1 smoothing) approach. Show your step in detail.

Solution:

P (Positive) =3/5

P (Negative) =2/5

P (Great /Positive) =3+1/7+7=4/14 P (Great/Negative) =1+1/5+7=2/12

P (Amazing/Positive) =1+1/7+7 =2/14 P (Amazing/Negative) =0+1/5+7=1/12

For the document 6

P (Positive/Great, Amazing) = 4/14*2/14*3/5

=0.29*0.14*0.6

=0.024

P (Negative/ Great, Amazing) =2/12*1/12*2/5

=0.16*0.083*0.4

=0.005

Sentiment polarity of document D6 is Positive

Question 5.

a) Let there be two questions and let there be 4 candidate answers for each question. Also Question Answering System chooses the best answer for question1 and second best answer for question 2. **Calculate the Mean Reciprocal Rank to evaluate the Question Answering System (1 marks**)

Soln: MMR = (1+1/2)/2=3/4

b) Let there be four documents given by

D1: the best American restaurant enjoys the best burger

D2: Indian restaurant enjoys the best dosa

D3: Chinese restaurant enjoys the best Manchurian

D4: the best the best Indian restaurant

Compute the BOW for D1, D2, D3 and D4 in the table. (2 Marks)

	the	best	American	Restaurant	enjoys	burger	dosa	manchurian	Chinese	Indian
D1										
D2										
D3										
D4										

Soln b)

	the	best	American	Restaurant	enjoys	burger	dosa	manchurian	Chinese	Indian
D1	2	2	1	1	1	0	0	0	0	0
D2	1	1	0	1	1	0	1	0	0	1
D3	1	1	0	1	1	0	0	1	1	0
D4	2	2	0	1	0	0	0	0	0	1

a) Also find out TF-IDF vector for D1, D2, D3, D4 for the above documents in b. (3 marks)

Soln c)

WORDS	TF (NORMALISED FREQUENCY)					Tf*idf			
	D1	D2	D3	D4	Idf	D1	D2	D3	D4
the	2/8	1/6	1/6	2/6	log(4/4)=0	0	0	0	0
best	2/8	1/6	1/6	2/6	Log(4/4)=0	0	0	0	0
American	1/8	0	0	0	Log(4/1)=0.6	0.6/8=0.075	0	0	0
Restaurant	1/8	1/6	1/6	1/6	Log(4/4)=0	0	0	0	0
enjoys	1/8	1/6	1/6	0	Log(4/3)=0.12	0.12/8=0.015	0.02	0.02	0
burger	1/8	0	0	0	Log(4/1)=0.6	0.6/8=0.075	0	0	0
dosa	0	1/6	0	0	Log(4/1)=0.6	0	0.1	0	0
manchurian	0	0	1/6	0	Log(4/1)=0.6	0	0	0.1	0
Chinese	0	0	1/6	0	Log(4/1)=0.6	0	0	0.1	0
Indian	0	1/6	0	1/6	Log(4/2)=0.3	0	0.3/6=0.05	0	0.3/6=0.05

b) Find Domain, Intent and Define Slots for each of the following Sentences: (4 marks)

1) Book a taxi at 6:00 PM from India Gate to Ambience Mall
2) I want to deposit 100 Dollars in my savings account.

solution

1) Book a taxi at 6:00 PM from India Gate to Ambience Mall
   - DOMAIN: Cab or Taxi
   - INTENT: Taxi-BOOKING
   - Slots
     - SOURCE-LOCATION: India Gate

- DESTINATION-LOCATION: Ambience Mall
- PICKUP TIME: 6:00 PM

2) I want to deposit 100 Dollars in my savings account.
   - DOMAIN: Banking
   - INTENT: Deposit-Account
   - Slots
   - Account Type: Savings Account
   - Transaction: Deposit
   - Amount: 100 dollars

Online Assessment anytime.anyw

wilp.wheebox.com/WAT-3/Subjective_detailResultOverview.obj?detailResult=view

Apps | Coinbase - Your Ho...

Applicant Dashboard

Send Re-Eval Request

Back To List

Home

GAURAV ROY .

Next User

Date of Exam

2021-07-18

Obtain Marks

6

7.5

8

8

ted soon

Qtext :

a) Draw the top-ranked parse tree for the sentence below by applying the PCFG given in below table. Does the results are good? Provide your comments. **(5+1=6 marks)**

Sentence: Write the notebooks with pencil.

Consider the following PCFG.

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 → Write | Ask | Find| ...	0.1
Prep → with | in | ...	0.1
Noun → notebooks | teacher | pencil | college | bike | summer |	0.1

b) Which of the three Noun Phrases (1. Pronoun, 2. Proper Noun, 3. Common Noun) to be the most difficult to handle computationally while performing top-down parsing. Explain why?

User Answer :

Q1 a. Answer sheet uploaded.

The results look good. This is because the deeper the tree, grows, more likely the probability of the tree goes down due to fractional multiplication of probabilities.

Therefore, a simpler tree with less depth is better to get the top ranked parse tree.

ENG

Online Assessment anytime.anyw

wilp.wheebox.com/WAT-3/Subjective_detailResultOverview.obj?detailResult=view

Apps | Coinbase - Your Ho...

Applicant Dashboard

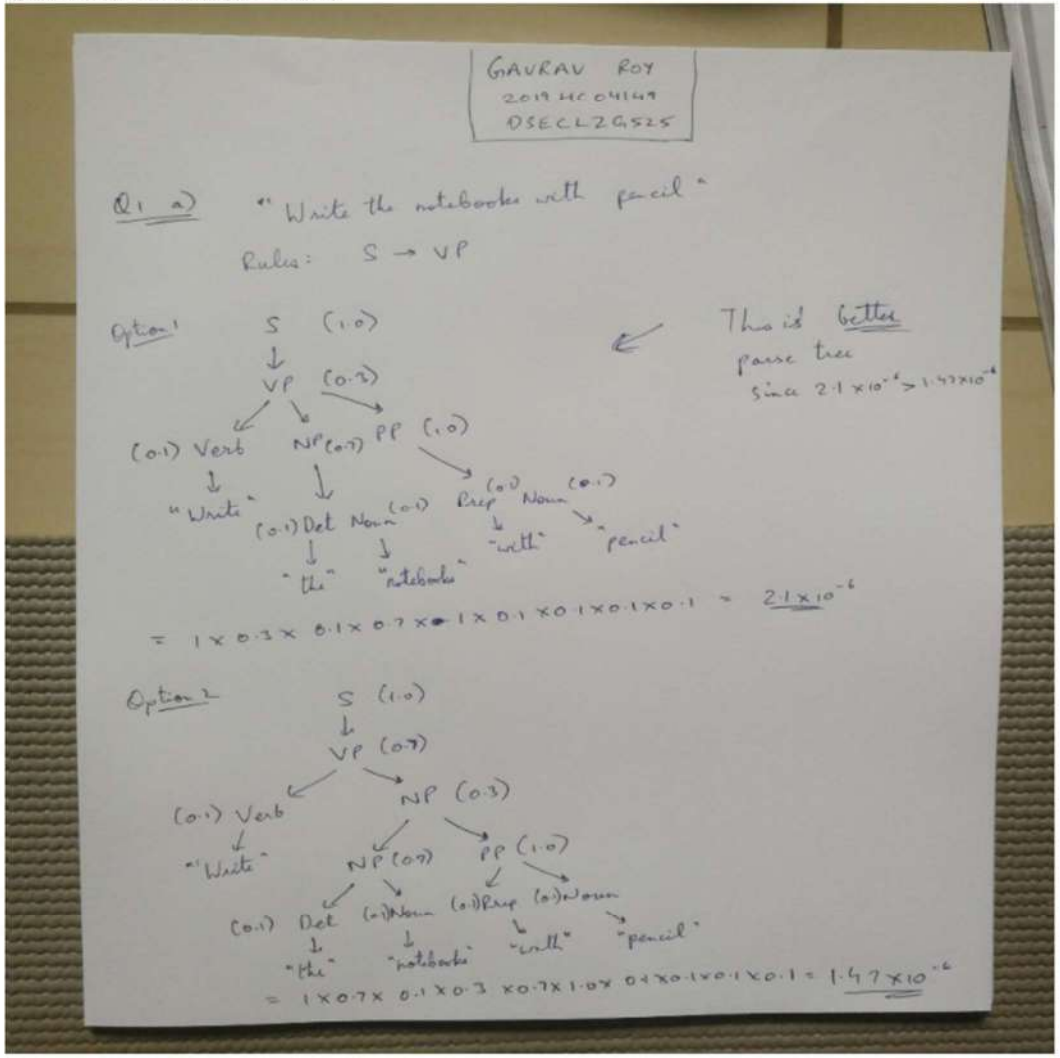
Send Re-Eval Request

Back To List

Home

probabilities.
Therefore, a simpler tree with less depth is better to get the top ranked parse tree.

Q1 b.
Pronoun is the most difficult to handle computationally. PCFG expands based on position and not due to structural context. Therefore, contextual tags like pronouns are not categorized properly.



GAURAV ROY

Next User

Date of Exam

2021-07-18

Obtain Marks

6

7.5

8

8

sted soon.

Online Assessment anytime.anyw

wilp.wheebox.com/WAT-3/Subjective_detailResultOverview.obj?detailResult=view

Apps | Coinbase - Your Ho...

Applicant Dashboard

Send Re-Eval Request

Back To List

Home

GAURAV ROY

Next User

Date of Exam

2021-07-18

Obtain Marks

6

7.5

8

8

ted soon.

Qtext :

a) **Given, the following training corpus, Using a bigram language model with and without add-one smoothing, what is P(Delhi is beautiful)? (6+2=8 marks)**

<s> Delhi is the capital of India </s>

<s> Delhi is cold </s>

<s> Delhi has beautiful gardens </s>

b) **Suppose the sentence consists of random alphabets (A, a, B, b, ...,Z, z) and each of the 26 letters in upper and lower case occurs with equal probability. What is the perplexity of this sentence?**

User Answer :

Q2 a. Sheet uploaded.

Q2 b.

There are 26 letters in the alphabet.

If both upper case and lower case are considered, the total is 26*2 = 52 letters

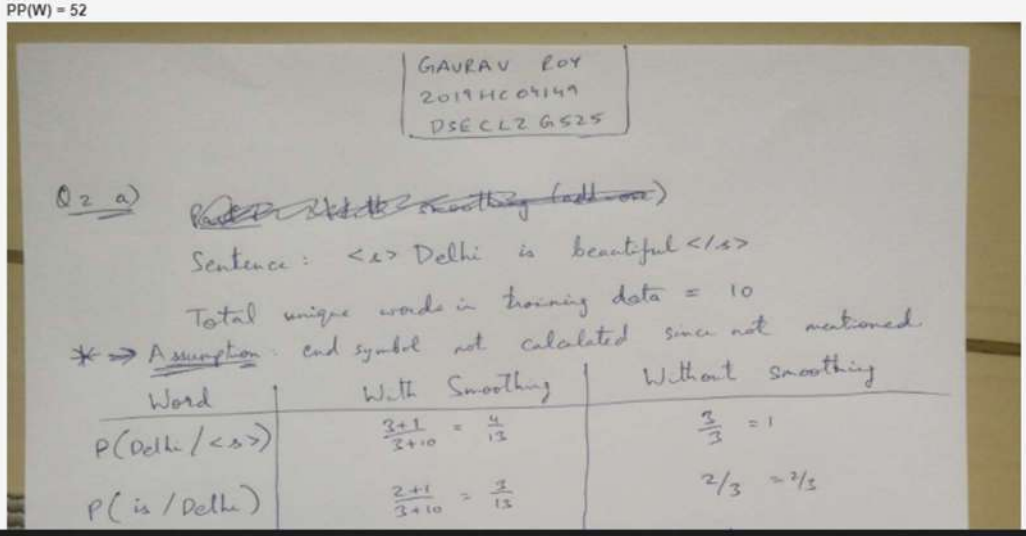
Since all of them can occur with equal probability, each of the given letters has a probability of (1/52)

Therefore, for a sentence of length N,

Perplexity = P(w1,w2,...,wn)^(-1/N)

PP(W) = ((1/52)^N)^(-1/N)

PP(W) = 52



Online Assessment anytime.anyw

wilp.wheebox.com/WAT-3/Subjective_detailResultOverview.obj?detailResult=view

Apps | Coinbase - Your Ho...

Applicant Dashboard

Send Re-Eval Request

Back To List

Home

PP(W) = 52

GAURAV ROY
2019HC09149
DSECL2 GS25

Q2 a) ~~Part 2 Without smoothing (add one)~~

Sentence : <s> Delhi is beautiful </s>

Total unique words in training data = 10

* => Assumption : end symbol not calculated since not mentioned.

Word	With Smoothing	Without smoothing
P(Delhi / <s>)	$\frac{3+1}{3+10} = \frac{4}{13}$	$\frac{3}{3} = 1$
P(is / Delhi)	$\frac{2+1}{3+10} = \frac{3}{13}$	$2/3 = 2/3$
P (beautiful / is)	$\frac{0+1}{2+10} = \frac{1}{12}$	$0/2 = 0$

∴ With smoothing,

$P(\text{Delhi is beautiful}) = \frac{4}{13} \times \frac{3}{13} \times \frac{1}{12} = \frac{1}{169} = 5.917 \times 10^{-3}$

Without smoothing,

$P(\text{Delhi is beautiful}) = 1 \times 2/3 \times 0 = 0$

Close

GAURAV ROY

Next User

Date of Exam

2021-07-18

Obtain Marks

6

7.5

8

8

ted soon.

Online Assessment anytime.anyv

wilp.wheebox.com/WAT-3/Subjective_detailResultOverview.obj?detailResult=view

Apps | Coinbase - Your Ho...

Applicant Dashboard

Send Re-Eval Request

Back To List

Home

GAURAV ROY

Next User

Date of Exam

2021-07-18

Obtain Marks

6

7.5

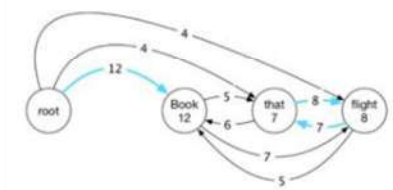
8

8

ted soon.

Qtext :

a). Does the following stage of a Edmond algorithm parsing has an MST ? If not, continue the algorithm for one more step with an Explanation. Obtain MST. (6+2=8 marks)



b). What are the basic differences between syntactic parsing and dependency parsing.

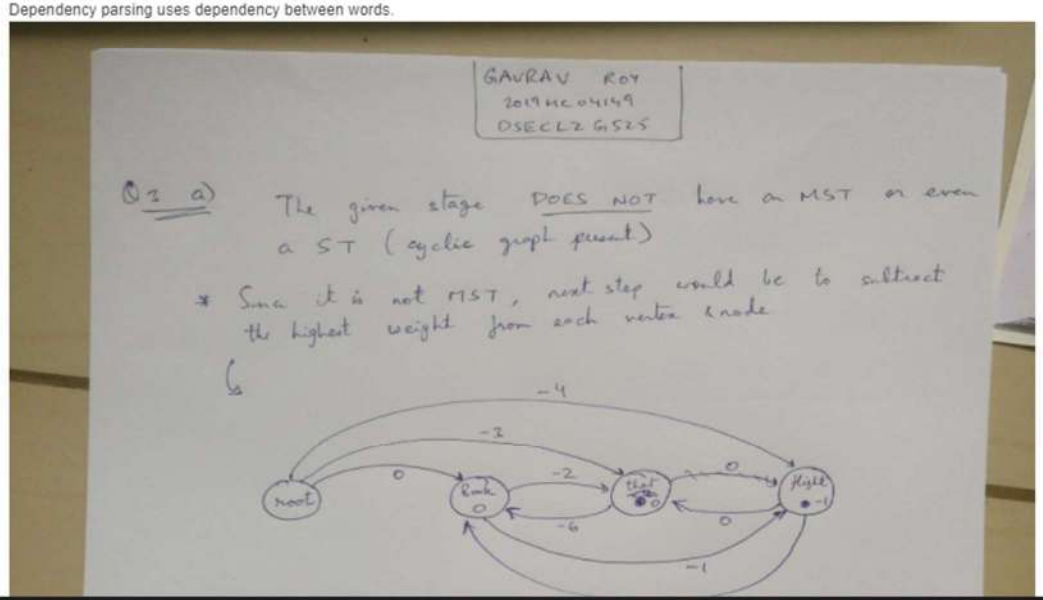
User Answer :

Q3 a. Sheet uploaded.

Q3 b.

Syntactic parsing is based on constituents of words and how the grammar is formed for it

Dependency parsing uses dependency between words.



Online Assessment anytime.anyw

wilp.wheebox.com/WAT-3/Subjective_detailResultOverview.obj?detailResult=view

Apps | Coinbase - Your Ho...

Applicant Dashboard

Send Re-Eval Request

Back To List

Home

GAURAV ROY

Next User

Date of Exam

2021-07-18

Obtain Marks

6

7.5

8

8

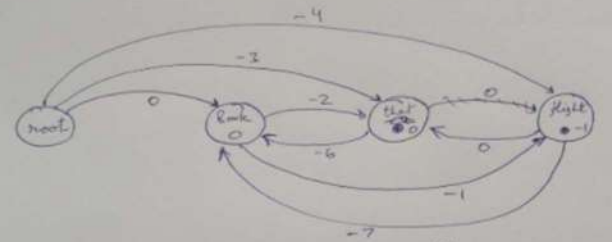
ted soon.

GAURAV ROY
2019HC04149
DSECL2 G525

Q3 a) The given stage DOES NOT have an MST or even a ST (cyclic graph present)

* Since it is not MST, next step would be to subtract the highest weight from each vertex / node

b



→ Here, the new highest weight @ "Book" is 0

∴ ~~Final Cost~~ as a consequence the new highest weight @ "that" is -2

→ because, "that" has already been covered by an edge between "Book" & "that", we can remove the "that" → "flight" edge

→ Therefore, the highest weight @ "flight" is -1 "that" → "flight"

→ Updated weight @ "that" is 0 since we have

∴ Final MST

Close

Online Assessment anytime.anyv

wilp.wheebox.com/WAT-3/Subjective_detailResultOverview.obj?detailResult=view

Apps Coinbase - Your Ho...

Applicant Dashboard

Send Re-Eval Request

Back To List

Home

GAURAV ROY

Next User

Date of Exam

2021-07-18

Obtain Marks

6

7.5

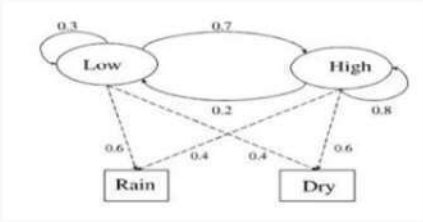
8

8

ted soon.

Qtext :

The following diagram describes HMM model with two hidden states: Low and High and the observations are rainy and dry. Both the states are equally probable to be initial states (3+5=8 marks)



a. **Construct transition state matrix and emission matrix.**

b. **Let the observation sequence be given as Dry, Rain. Give the corresponding Hidden state sequence.**

User Answer :

Q 4 a.

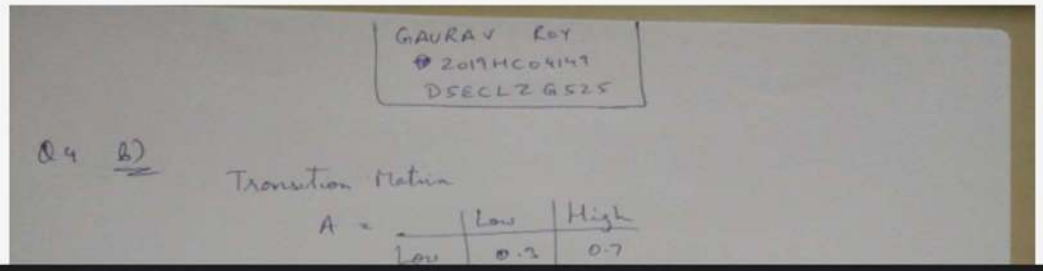
Transition State Matrix

A	Low	High
Low	0.3	0.7
High	0.2	0.8

Emission Matrix

B	Rain	Dry
Low	0.6	0.4
High	0.4	0.6

Q 4 b. Sheet uploaded.



Online Assessment anytime.anyw

wilp.wheebox.com/WAT-3/Subjective_detailResultOverview.obj?detailResult=view

Apps | Coinbase - Your Ho...

Applicant Dashboard

Send Re-Eval Request

Back To List

Home

Q 4 b. Sheet uploaded.

GAURAV ROY

2019HC04149

DSECLZG525

Q4 b)

Transition Matrix

A =

	Low	High
Low	0.3	0.7
High	0.2	0.8

Emission Matrix

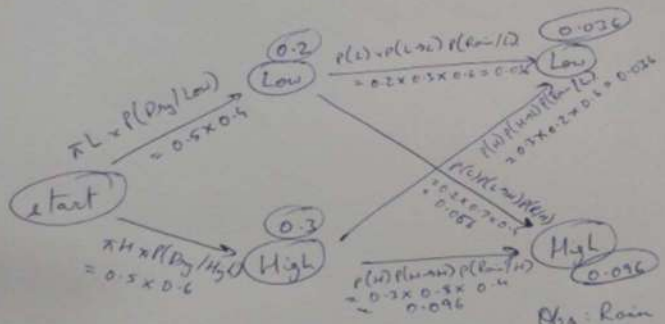
B =

	Rain	Dry
Low	0.6	0.4
High	0.4	0.6

Initial State (π)

Low = 0.5

High = 0.5



Obs: Dry

Obs: Rain

∴ Hidden state sequence = {High, High}

GAURAV ROY

Next User

Date of Exam

2021-07-18

Obtain Marks

6

7.5

8

8

ted soon.

Test Submitted Successfully.

For security reasons, please exit your browser.

View Uploaded Answer Sheets

Qtext:-

a) B is a corpus which only contains one single bitstring:
1 1 0 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 0 0 0
Calculate the following bigram probabilities from the corpus B using MLE (Maximum Likelihood Estimation). **[2 marks]**
i) P (0 | 1)
ii) P (0 | 0)

b) Write the formulae to calculate the unigram, bigram and trigram probabilities of the sentence: "It is health that is real wealth and not pieces of gold and silver." **[3 marks]**

c) What will be the perplexity value if you calculate the perplexity of an unsmoothed language model on a test corpus with unseen words? Explain. [1 marks]

d) We use the following (part of) lexicon: - [4 marks]

adult	JJ	has	VBZ
adult	NN	just	RB
daughter	NN	my	PRP$
developed	VBD	programs	NNS
developed	VBN	programs	VBZ
first	JJ	tooth	NN
first	RB	whose	WP$

Consider the following sentence: "My daughter whose first adult tooth has just developed programs" With this lexicon, how many different PoS tagging does this sentence have? Justify your answer.

It seems like you have not uploaded any images/files for this question.

Online Assessment anytime.anywhere. - Google Chrome

wilp.wheebox.com/WET-2/EndTestV2.obj?login_id=2019hc04149&submitBy=autoSubmit

Test Submitted Successfully.

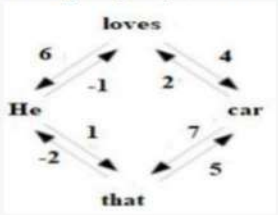
For security reasons, please exit your browser.

- Question 1
- Question 2
- Question 3
- Question 4
- Question 5
- Question 6
- Question 7

View Uploaded Answer Sheets

Qtext:-

In the below weighted graph, find the edge weights between car-loves and that ? car using maximum spanning tree [Use Chu Liu Edmond algorithm] [5Marks]



ENG 12:42 PM

Online Assessment anytime.anywhere. - Google Chrome

wilp.wheebox.com/WET-2/EndTestV2.obj?login_id=2019hc04149&submitBy=autoSubmit

Test Submitted Successfully.

For security reasons, please exit your browser.

- Question 1
- Question 2
- Question 3
- Question 4
- Question 5
- Question 6
- Question 7

View Uploaded Answer Sheets

Qtext:-

i) You are designing a frame-based dialog system for movie booking. [5 marks]

a). What are the different slots in your design? Mention along with their corresponding entity types and questions that the system would ask a user.

b). Show a finite-state dialog manager for the system

c). What changes would you make to the design to change it from a single initiative system to multi initiative system?

ii) Find Domain, Intent and Define Slots for each of the following Sentences: [4 marks]

a) Book me a table at Mariott hotel.

b) Search the list of movies directed by Satyajit Roy

iii) 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 [4 marks]

a) inverseOf
b) domain
c) range
d) Cardinality
e) disjointWith
f) subClassOf


ENG 12:42 PM

Online Assessment anytime.anywhere. - Google Chrome

wilp.wheebox.com/WET-2/EndTestV2.obj?login_id=2019hc04149&submitBy=autoSubmit

Test Submitted Successfully.

For security reasons, please exit your browser.

- Question 1
- Question 2
- Question 3
- Question 4
- Question 5
- Question 6
- Question 7

View Uploaded Answer Sheets

Qtext:-

Find TF-IDF score for the following 2 documents and then find the cosine similarity score [5 marks]

study parsing algorithm article NLP blog

study POS tagging article NLP blog

ENG 12:42 PM

Online Assessment anytime.anywhere. - Google Chrome

wilp.wheebox.com/WET-2/EndTestV2.obj?login_id=2019hc04149&submitBy=autoSubmit

Test Submitted Successfully.

For security reasons, please exit your browser.

- Question 1
- Question 2
- Question 3
- Question 4
- Question 5
- Question 6
- Question 7

View Uploaded Answer Sheets

Qtext:-

Machine translation

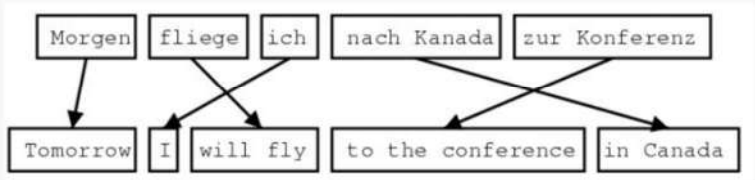
a) Calculate Final BLEU Score for below Examples (Use Unigram and Bigram Precision and Brevity penalty) [5 marks]

Reference: The NASA Opportunity rover is battling a massive dust storm on Mars

Candidate 1: The Opportunity rover is combating a big sandstorm on Mars

Candidate 2: A NASA rover is fighting a massive storm on Mars

b) Find the alignment vector for the following, where the first sentence is the source and the second sentence is the target. [2 marks]



ENG 12:42 PM

Online Assessment anytime.anywhere. - Google Chrome

wilp.wheebox.com/WET-2/EndTestV2.obj?login_id=2019hc04149&submitBy=autoSubmit

Test Submitted Successfully.

For security reasons, please exit your browser.

- Question 1
- Question 2
- Question 3
- Question 4
- Question 5
- Question 6
- Question 7

View Uploaded Answer Sheets

Qtext:-

a) How can the Simple Lesk algorithm be applied to disambiguate the exact meaning of "bank" in following sentence. (4 marks)

the coin bank was empty at home

- S: (n) bank (a long ridge or pile) "
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers)
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- ***S: (n)*** *savings bank, coin bank, money box, bank (a container (usually with a slot in the top) for keeping money at home)*

b) For the following sentence pairs identify the various lexical relations existing between the common word ***(2 marks)***

i) *The* ***school*** *is organizing painting competition*
She is studying in a ***school.***
ii) *Ministry has issued a* ***tender*** *for the furniture*
The meat should be well cooked and ***tender***

It seems like you have not uploaded any images/files for this question.

ENG 12:42 PM

Online Assessment anytime.anywhere. - Google Chrome

wilp.wheebox.com/WET-2/EndTestV2.obj?login_id=2019hc04149&submitBy=autoSubmit

Test Submitted Successfully.

For security reasons, please exit your browser.

- Question 1
- Question 2
- Question 3
- Question 4
- Question 5
- Question 6
- Question 7

View Uploaded Answer Sheets

Qtext:-

Give a practical example of aspect based sentiment analysis and state problems associated with it.

[4 Marks]

It seems like you have not uploaded any images/files for this question.

ENG 12:42 PM

NLP End Term exam

The English language has 26 characters. The unigram probabilities of the characters A, N and T are 0.1, 0.05 and 0.01. What will be the perplexity of the sequences "NAN" and "ANT". How will the perplexity change for the two sequences if the probabilities were equal for all the 26 characters? (4 marks)

Solution:
Perplexity ("NAN") = P("NAN")$^{-1/3}$= (0.05 * 0.1* 0.05)$^{-1/3}$
Perplexity ("ANT") = (0.1 * 0.05* 0.01)$^{-1/3}$

If the probabilities were the same, then
Perplexity ("NAN") = ((1/26)*(1/26)*(1/26)$^{-1/3}$
Perplexity ("ANT") = ((1/26)*(1/26)*(1/26)$^{-1/3}$

A sentiment classifier predicts the labels as given below. Find the precision, recall and accuracy of the model. (3 marks)

Expected	**Predicted**
Positive	Negative
Positive	Positive
Negative	Negative
Positive	Positive
Negative	Positive
Negative	Negative
Negative	Negative
Positive	Positive
Positive	Negative
Negative	Negative

Solution:

	Pos Actual	Neg Actual
Pos Pred	3	1
Neg Pred	2	4

Accuracy = 7/10
Precision = ¾
Recall = 3/5

Find the tf-idf vectors and cosine similarity between the following documents.
d1= The best team plays the finals
d2= India won a medal in the finals (7 marks)

Solution:

TF (2 marks) IDF (2 marks) TF-IDF (1 mark), Cosine Similarity (2 marks)

	d1	d2	IDF	TF-IDF d1	TF-IDF d2
the	0.33	0.14	0.00	0.00	0.00
best	0.17	0.00	1.00	0.17	0.00
team	0.17	0.00	1.00	0.17	0.00
plays	0.17	0.00	1.00	0.17	0.00
India	0.00	0.14	1.00	0.00	0.14
finals	0.17	0.14	0.00	0.00	0.00
won	0.00	0.14	1.00	0.00	0.14
a	0.00	0.14	1.00	0.00	0.14
medal	0.00	0.14	1.00	0.00	0.14
in	0	0.14	1	0.00	0.14

Cosine Similarity = 0

For the sentence, “Play this year’s French radio-hit pop songs”, a chatbot determines the following slots, Genre: Pop, Language: French, Year: 2020. Determine the intent accuracy and slot error rate. (1 mark)

Solution:
Slot Error Rate = ⅓
Intent Accuracy = ⅔

Q4. 10 Marks

Solve word sense disambiguation problem for below example. Recommend the best approach to solve and justify how?

Example: Anaconda is one of the popular frameworks to do python programming.

1. **Word: Anaconda. Actual Semantic Meaning: Type of snake.**
2. **Word: Anaconda, Expected Semantic Meaning: Programming framework.**
3. **Word: python. Actual Semantic Meaning: Type of snake.**
4. **Word: Python, Expected Semantic Meaning: Programming language.**

Solution: Anaconda has 2 meanings. Lesk algorithm will disambiguate. The given sentence and the second meaning of the word Anaconda have common word Programming framework.
Similarly for the word python.
NEED TO BRIEFLY EXPLAIN FOR BOTH THE WORDS ANACONDA AND PYTHON

Q6 (5+5=10 marks)

a)"These earphones are a good choice at this price. Connected with laptop for office calls and these are working well although there is no noise cancellation. Quality of wires are a bit thin and look delicate, though neckband is ok. Bass will seem ok if you have not used good quality earphones earlier."
You have been given product review data like the one shown above. You are asked to design a sentiment analysis model for this data. What would be your approach? Describe the different components of your solution. State any assumptions that you are making and pros/cons (if any) of your approach.

Solution: Its aspect based sentiment analysis.
The different components are
quintuple (*e*, *a*, *s*, *h*, *t*)
e is the earphones
(earphones, price ,good, opinion holder=Me, t=1)
(earphones, quality of wires, thin and delicate, opinion holder=me, t=1)
(earphones, NOISE CANCELLATION, NO, opinion holder=me, t=1)

(earphones, neckband, ok, opinion holder=me, t=1)
(earphones, bass, ok, opinion holder=me, t=1)

Assumption: If you have not used good quality earphones earlier."

b) Describe how ontology plays role in semantic web. Explain one example of Resource description framework (RDF) triple and RDFS.
Refer slides

Q7. [5 marks]
Given the grammar below, show how it would be used to derive a parse tree for the sentence below. Show the order in which rules would be applied if using a bottom up chart parsing. (5 MARKS)

The book covers key ideas

$S \rightarrow NP\ VP$
$NP \rightarrow DET\ NOM$
$NOM \rightarrow ADJ\ NOM$
$NOM \rightarrow NOUN$
$VP \rightarrow VERB\ NP$
$VP \rightarrow VERB$

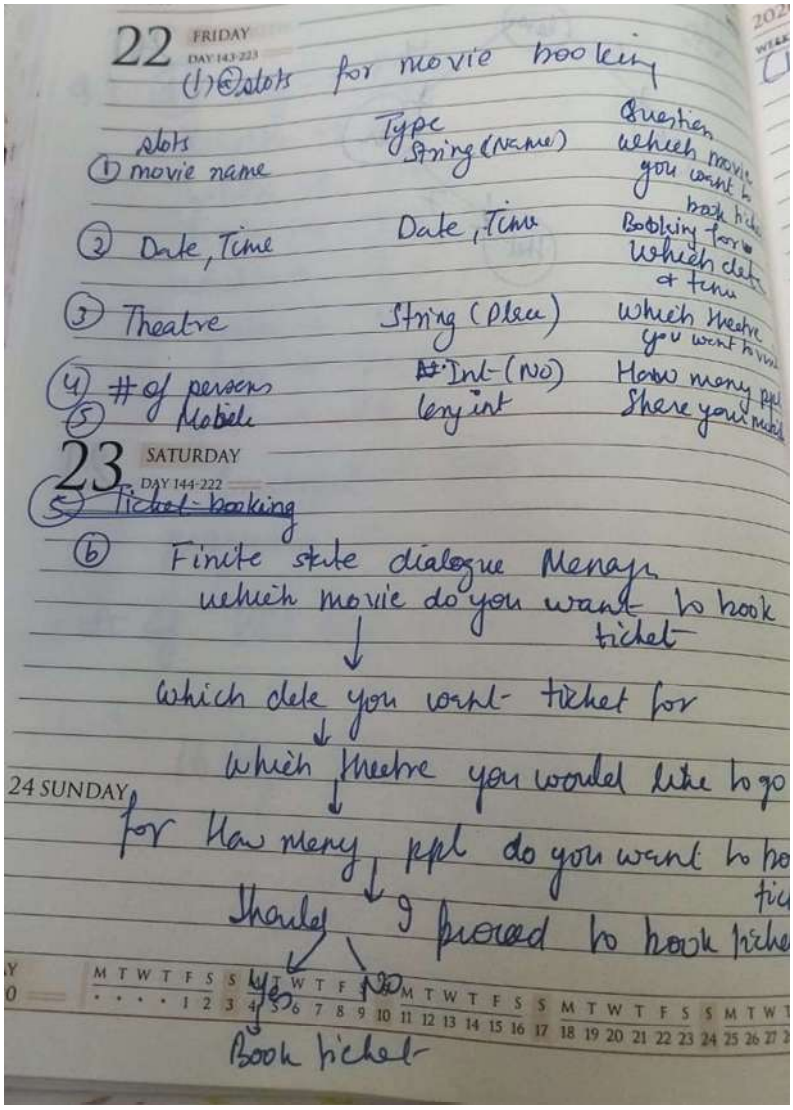
$DET \rightarrow the$
$ADJ \rightarrow book$
$ADJ \rightarrow key$
$VERB \rightarrow covers$
$NOUN \rightarrow book$
$NOUN \rightarrow covers$
$NOUN \rightarrow ideas$
$NOUN \rightarrow key$

Ques:

i) You are designing a frame-based dialog system for movie booking. [5 marks]
a). What are the different slots in your design? Mention along with their corresponding entity types and questions that the system would ask a user.
b). Show a finite-state dialog manager for the system
c). What changes would you make to the design to change it from a single initiative system to multi initiative system?

ii) Find Domain, Intent and Define Slots for each of the following Sentences: [4 marks]
a) Book me a table at Mariott hotel.
b) Search the list of movies directed by Satyajit Roy

iii) 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 [4 marks]
a) inverseOf
b) domain
c) range
d) Cardinality
e) disjointWith
f) subClassOf



MAY

27 WEDNESDAY

(i) (c) To make it a mixed initiative system, we can follow (Genial Understander System)

It is a kind of mixed initiative system. The conversation shifts b/w user & system. The structure of frame guides, dialogue system asks question from user, filling any slots that user specifies. When form is filled, do database query.

28 THURSDAY

If the user answers 3 questions at once, the system can fill the slots & not asks those questions again.

MAY

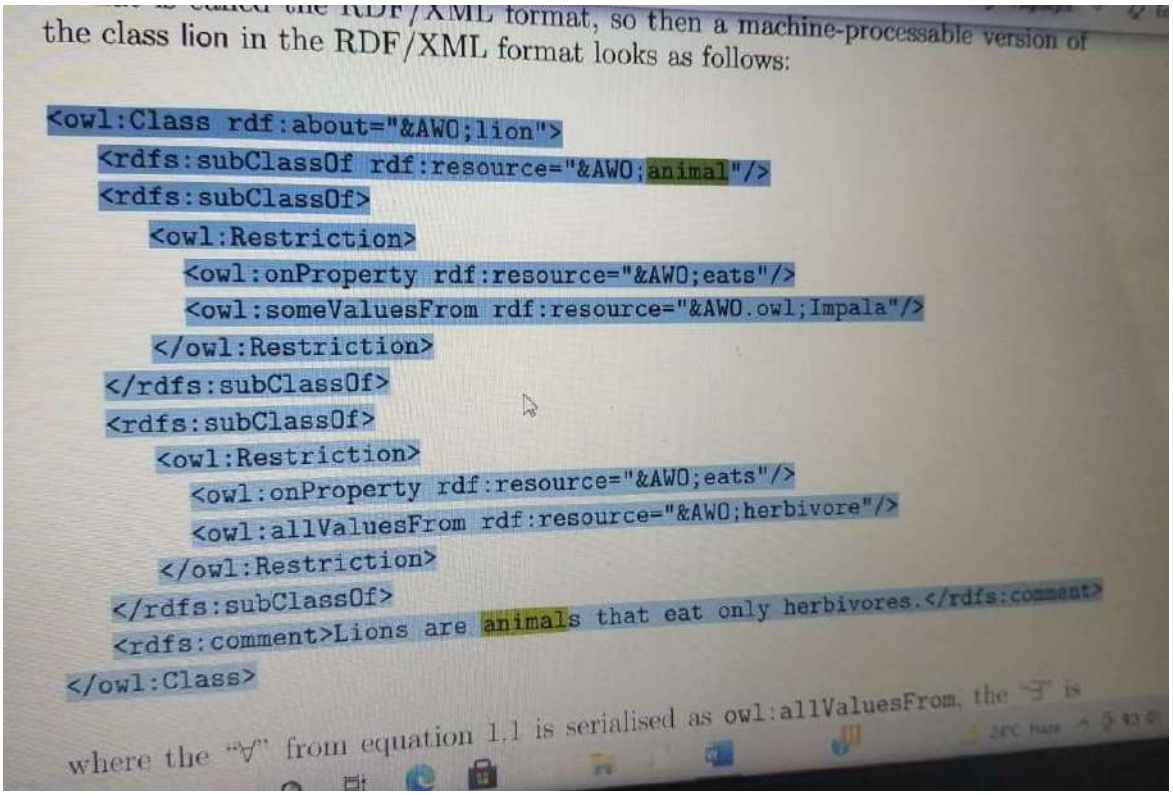
MONDAY 25

(ii) (a) Book me a table at Marriott

Domain: Hotel Restaurant or Hotel
Intent: Book a table
Hotel name: Marriott

(b) Search list of movies directed by Satyajit Ray.

Domain: Movie
Intent: Searching a movie
Director: Satyajit Ray.



the class lion in the RDF/XML format looks as follows:

```
<owl:Class rdf:about="&AWO;lion">
    <rdfs:subClassOf rdf:resource="&AWO;animal"/>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="&AWO;eats"/>
            <owl:someValuesFrom rdf:resource="&AWO.owl;Impala"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="&AWO;eats"/>
            <owl:allValuesFrom rdf:resource="&AWO;herbivore"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:comment>Lions are animals that eat only herbivores.</rdfs:comment>
</owl:Class>
```

where the "∀" from equation 1.1 is serialised as owl:allValuesFrom, the "∃" is

MAY 2020 WEEK 20

13 WEDNESDAY DAY 134-232

Makeup Exam

Q4 TF-IDF score

Document1 : study parsing algorithm article NLP blog

Document 2 : study POS tagging article NLP blog

words	TF (doc1)	TF (doc2)	IDF	TF-IDF	TF-IDF
study	1/6	1/6	log(2/2)=0	0	0
parsing	1/6	0	log(2)=0.3	0.3×1/6=0.05	0
algorithm	1/6	0	log(2)=0.3	0.3×1/6=0.05	0
article	1/6	1/6	log(2/2)=0	0	0

14 THURSDAY DAY 135-231

NLP	1/6	1/6	log(2/2)=0	0	0
blog	1/6	1/6	log(2/2)=0	0	0
pos	0	1/6	log(2)=0.3	0	0.05
tagging	0	1/6	log(2)=0.3	0	0.05

TF-IDF (doc1) = [0 0.05 0.05 0 0 0 0.0]

TF-IDF (doc2) = [0 0 0 0 0 0 0.05 0.05]

Cosine similarity b/w them

= 0

M T W T F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S S
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Q1(a) Top down Parsing (Depth first strategy)

$_1$The $_2$young $_3$women $_4$cried $_5$

Step	Current state	Backup state	Comment
1	((S) 1)		
2	((NP VP) 1)		
3	(CART N VP) 1)	((ART ADJ N VP) 1)	
4	((N VP) 2)	((ART ADJ N VP) 1)	'The': Art
5	((VP) 3)	~~((V NP) 3)~~ ((ART ADJ N VP) 1)	
6	((V) 3)	((V NP) 3) ((ART ADJ N VP) 1)	
7.	(() 4)	((V NP) 3) ((ART ADJ N VP) 1)	
(8)	((V NP) 3)	((ART ADJ N VP) 1)	
(9)	((NP) 4)	((ART ADJ N VP) 1)	
(10)	((ART N) 4)	(CART ADJ N) 4) ((ART ADJ N VP) 1)	
(11)	((ART ADJ N) 4)	(CART ADJ N VP) 1)	
(12)	((ART ADJ N VP) 1)		
(13)	((ADJ N VP) 2)		
(14)	((N VP) 3)		
(15)	((VP) 4)		
(16)	((V) 4)	((V NP) 4)	
(17)	(() 5)	success.	

Grammar

S → NP VP
NP → ART N
NP → ART ADJ N
VP → V
VP → V NP
Cried: V
dogs: N, V
The: Art
young: ADJ, N
women: N, V

Ans 1(b) CKY Parsing

the man hit the dog

S → np vp

np → det n

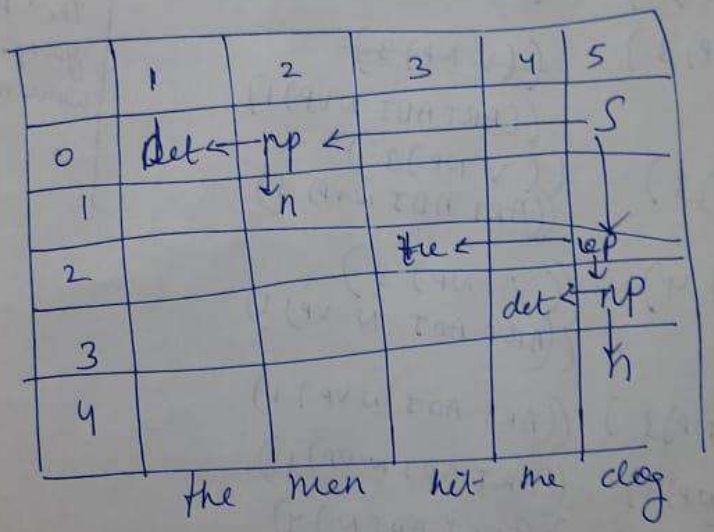
det → the

n → man

n → dog

vp → tv, np

the → hit



Ans 2 (a) He gave her a pen

Stack	Buffer	Arcs.	Operation
$[\,]_S$	$[He, gave, her, a, pen, .]_B$		SH.
$[He]_S$	$[gave, her, a, pen, .]_B$	$He \xleftarrow{SBJ} gave$	LA
$[\,]_S$	$[gave, her, a, pen, .]_B$		SH
$[gave]_S$	$[her, a, pen, .]_B$	$gave \xrightarrow{IOBJ} her$	RA
$[gave, her]_S$	$[a, pen, .]_B$		SH
$[gave, her, a]_S$	$[pen, .]_B$	$a \xleftarrow{Det} pen$	LA
$[gave, her]$	$[pen, .]_B$		RE
$[gave]_S$	$[pen, .]_B$	$gave \xrightarrow{DOBJ} pen$	RA
$[gave, pen]_S$	$[.]_B$		RE
$[gave]_S$	$[.]_B$	$gave \xrightarrow{punc} .$	RA
$[gave, .]_S$	$[\,]_B \rightarrow$ empty		

2b) See the slides

Ans 3

3 gram	count	2 gram	count	1 gram	count
A beautiful day	5	beautiful day	7	day	20
A beautiful night	0	beautiful night	0	night	5

- P_{bd} (day | a beautiful) $= \frac{5}{5} - \frac{1}{8} = 1 - \frac{1}{8} = \frac{7}{8}$

P_{nb} (night | a beautiful) $= \lambda_1$ P (night | beautiful)

λ P (night | beautiful) $= \lambda_2$ P (night)

P (night) $= \frac{5}{25} - \frac{1}{8} = \frac{3}{40} = 0.075$

P(night | beautiful) + P (day | beautiful) $= 1$

$\frac{3\lambda_2}{40} + \left(1 - \frac{1}{8}\right) = 1$

$\frac{3\lambda_2}{40} = \frac{1}{8}$

$\lambda_2 = 5/3 =$ ~~1.33~~

~~λ_2 P (night~~ P (night | beautiful) $= \lambda_2$ P(night)

$= \frac{5}{3}\left[\frac{5}{25} - \frac{1}{8}\right]$

$= \frac{5}{3} \times \left[\frac{3}{40}\right] = \frac{1}{8}$

PTO

Ans 3

3gram
A beautiful day
^

Ans 3 contd

$P(\text{day} \mid \text{a beautiful}) + P(\text{night} \mid \text{a beautiful}) = 1$

$\frac{7}{8} + \frac{d_1}{8} = 1$

$7 + d_1 = 8$

$d_1 = 1$

$P_{nd}(\text{night} \mid \text{a beautiful}) = d_1 \, P(\text{night} \mid \text{beautiful}) = \frac{1}{8}$

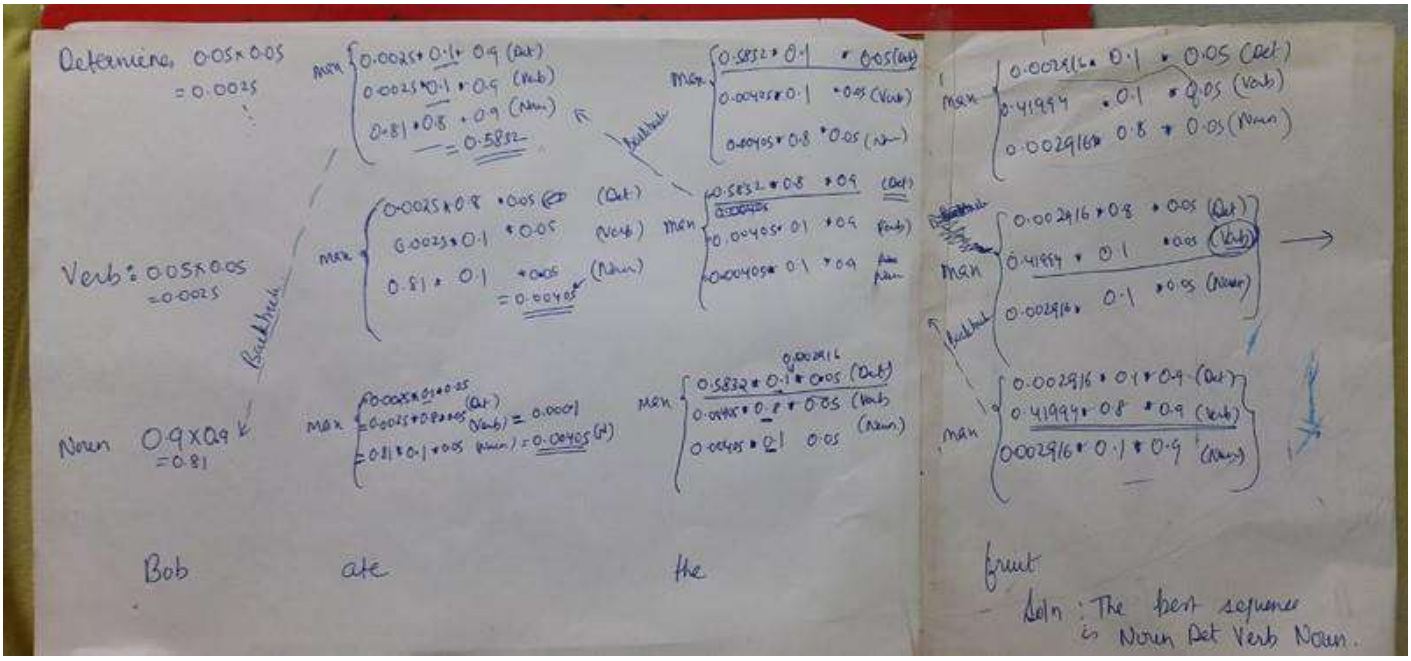
$P(\text{day} \mid \text{a beautiful}) = \frac{7}{8}$

next word should be "day"

Ans 3

(i) A stone smelled the color blue :- Lexically & syntactically correct but semantically incorrect.

(ii) Its kind of a → lexically incorrect



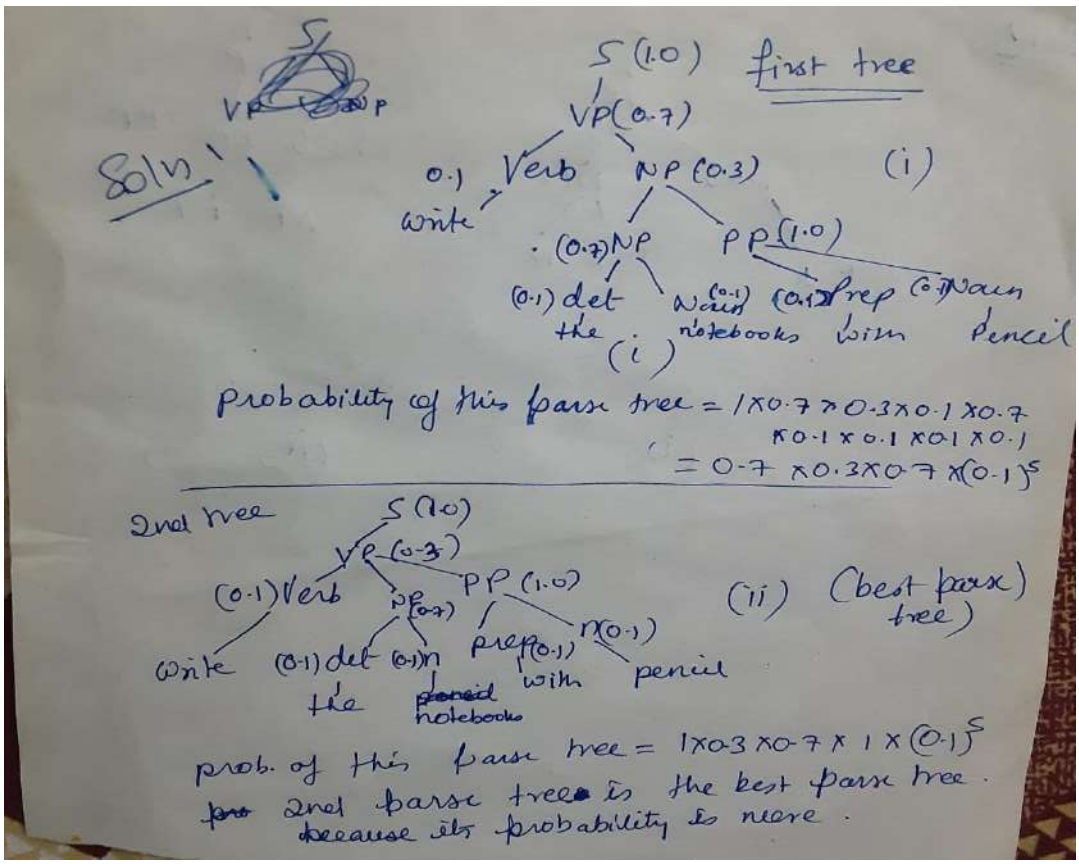
Q.1 (5+1=6 marks)

a) Draw the top-ranked parse tree for the sentence below by applying the PCFG given in below table. Does the results are good? Provide your comments.

Sentence: Write the notebooks with pencil.

Consider the following PCFG

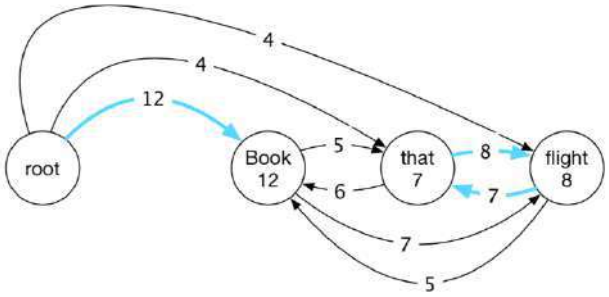
S → VP	1.0
VP → Verb NP	0.7
VP → Verb NP PP	0.3
NP → NP PP	0.3
NP → Det Noun	0.7
PP → Prep Noun	1.0
Det → the	0.1
Verb → Write | Ask | Find| ...	0.1
Prep → with | in | ...	0.1
Noun → notebooks | teacher | pencil | college | bike | summer |.....	0.1

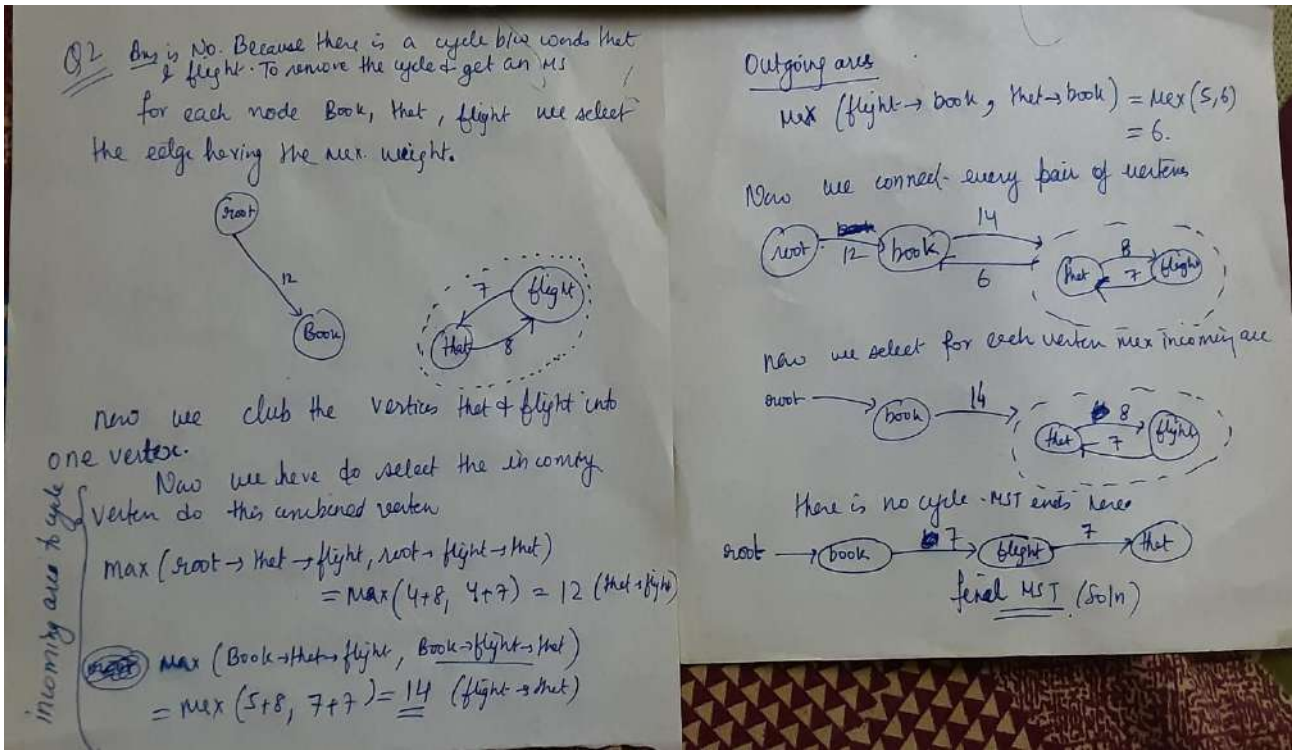


b) Which of the three Noun Phrases (1. Pronoun, 2. Proper Noun, 3. Common Noun) to be the most difficult to handle computationally while performing top-down parsing. Explain why?

Q2. (6+2=8 marks)

a). Does the following stage of a Edmond algorithm parsing has an MST ? If not, continue the algorithm for one more step with an Explanation. Obtain MST.





b) What are the basic differences between syntactic parsing and dependency parsing.

Ans: See from the slides

Q3 (6+2=8 marks)

a) **Given, the following training corpus, Using a bigram language model with and without add-one smoothing, what is P(Delhi is beautiful)?**

<s> Delhi is the capital of India </s>

<s> Delhi is cold </s>

<s> Delhi has beautiful gardens </s>

Solution:

P(Delhi is beautiful) = P(Delhi | <s>) * P(is|Delhi) * P(beautiful | is) * P(</s> | beautiful)

P(wn|wn−1) = C(wn−1 wn)/ C(wn−1)

Without Smoothing
P(Delhi | <s>) = 3/3 = 1
P(is|Delhi) = ⅔ = 0.676
P(beautiful | is) = 0/2 = 0
P(</s> | beautiful) = 0/1 = 0
Unique words = 10
With Smoothing
P(Delhi | <s>) = (3+1)/(3+10) = 0.31
P(is|Delhi) = (2+1)/(3+10) = 0.23
P(beautiful | is) = (0+1)/(2+10) = 0.08
P(</s> | beautiful) = (0+1)/(1+10) = 0.09
P(Delhi is beautiful) = 0.31*0.23*0.08*0.09 = 5.13 *10^-4

b) Suppose the sentence consists of random alphabets (A, a, B, b, ….,Z, z) and each of the 26 letters in upper and lower case occurs with equal probability. What is the perplexity of this sentence?

PP(W) = P(w1w2 ...wN) ^*(− 1 /N)

Since both upper and lower cases are considered 52 letters have equal probability.

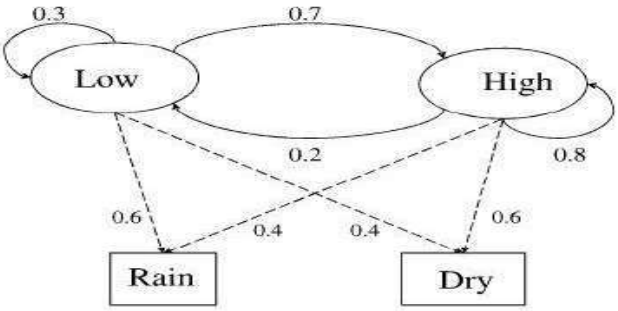
Perplexity is $\left(\left(\frac{1}{52}\right)^{52}\right)^{-1/52}$

= 52

Q4. (3+5=8 marks)

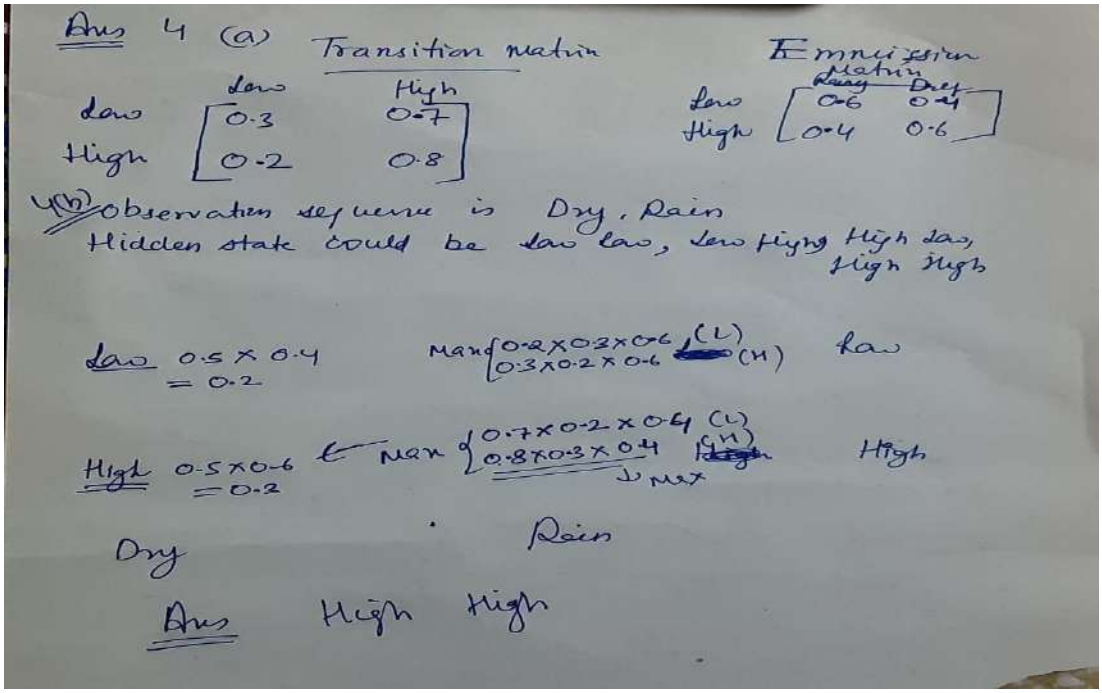
The following diagram describes HMM model with two hidden states: Low and High

and the observations are rainy and dry. Both the states are equally probable to be initial states



a) **Construct transition state matrix and emission matrix.**

b) **Let the observation sequence be given as Dry, Rain. Give the corresponding Hidden state sequence.**



Ques:

7 Marks

Compute the BLEU score for the following example.[Hint :Use unigram,bigram,trigram and brevity penalty also]
Reference 1: the dog is on the grass
Reference 2: there is a dog on the grass
Candidate: the dog dog on the grass grass

Ques:

[2+3=5 marks]

a) There are different types of questions in Modern systems. Which type is the following question:
"How many variations of the COVID Vaccine are available?"

b) Find Domain, Intent and Define Slots for each of the following Sentences:

i) Find me a cheap South Indian restaurant in Delhi
ii). Book an Appointment on Sunday 10:00 AM for Hair spa
iii). What will be the weather tomorrow morning in New Delhi?

2020
APRIL
WEEK 16
MONDAY 13
DAY 104-262

Soln 3

Ref 1: the dog is on the grass
Ref 2: there is a dog on the grass
Candidate: the dog dog on the grass grass

Unigram

word	count	Ref 1	Ref 2	max(Ref1, Ref2)	min(count, R)
the	2	2	1	2	2
dog	2	1	1	1	1
on	1	1	1	1	1
grass	2	1	1	1	1
	Total = 7				5 (total)

TUESDAY 14
DAY 105-261

Unigram = 5/7

Bigram

Words	count	Ref 1	Ref 2	max(Ref1, Ref2)	min(count, R)
the dog	1	1	0	1	1
dog dog	1	0	0	0	0
dog on	1	0	1	1	1
on the	1	1	1	1	1
the grass	1	1	1	1	1
grass grass	1	0	0	0	0
	6				4

Bigram = 4/6 = 2/3

MAY 2020

APRIL

15 WEDNESDAY

word	Trigram count	Ref1	Ref2	min(R1,R2)	min(R1,R2)
the dog dog	1	0	0	0	0
dog dog on	1	0	0	0	0
dog on the	1	0	1	1	1
on the grass	1	1	1	1	1
the grass grass	1	0	0	0	0
Total = 5					2

Trigram = 2/5.

|C| = 7

16 THURSDAY

|Ref1| = 6

|Ref2| = 7

if students assume (Ref1 ∩ Ref2) like best answer then

~~BP = exp(1-~~

BP = 1

Bleu score = BP × P

$= 1 \times \left(\frac{5}{7} \times \frac{2}{3} \times \frac{2}{5}\right)^{1/3}$

2020 APRIL

WEEK 16

Ans (a) factoid question.

17 FRIDAY

(b) I find a cheap south Indian restaurant in Delhi

Domain : Restaurant
Intent : to look for restaurant
food type : South Indian
place : Delhi
cost : cheap

Book an appointment on Sunday 10a.m. for Hair spa.

18 SATURDAY

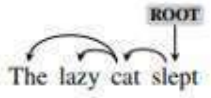
Domain : Salon
Intent : Book an appointment
Service : Hair Spa.
Day : Sunday
Time : 10am.

What will be weather tomorrow morning in New Delhi

Domain : Weather forecasting
Intent : Show weather
Day : tomorrow
Place : Delhi
time : Morning.

MAY 2020

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



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

Solution:

a)

[Root,]	[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]
[Root,]	[slept]	[LA]
[Root,Slept]	[]	[RA]
[Root]	[]	[RE]

SH,SH,LA,LA,SH,LA,RA

b)

[Root,]	[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]
[Root, Cat]	[]	[RA]
[Root,Cat ,Slept]	[]	[RE]
[Root,cat]	[]	[RE]
[Root]	[]	[RE]

b) 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
9.()		

2. Given the grammar and lexicon below show the final chart for the following sentence after applying the bottom up chart parser.[5 marks]

S: Book the flight on airasia

S->VP

VP->V NP

NP->NP PP

NP->Det Noun

PP ->Prep Noun

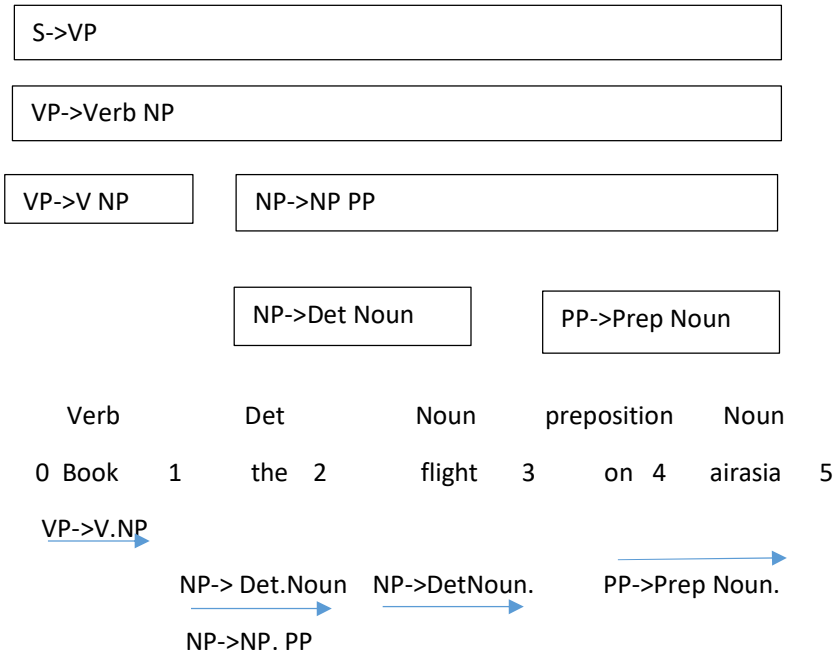
Det ->the

Verb->Book

Prep->on

Noun->flight|airasia

Solution:



3. 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:

Attached the screenshot

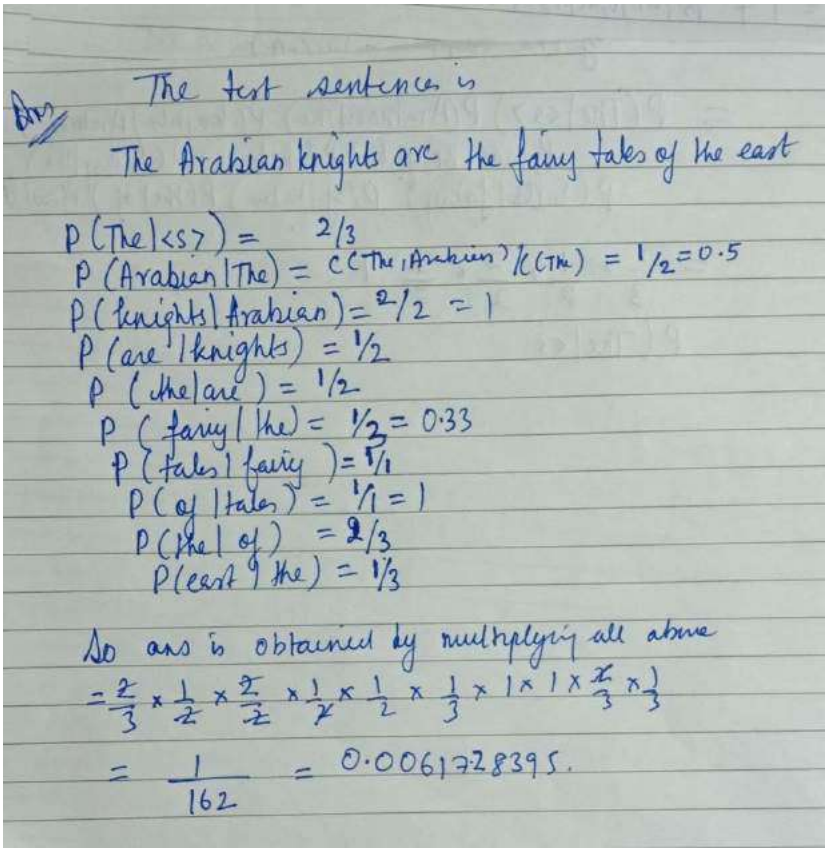
Language model

1. Consider the training set:

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

Soln:



---

2. You are an English class teacher and to make your course interesting you hold a write like Shakespeare competition where you ask each student to write a play in the style of Shakespeare. You try to score the plays automatically by using a trigram model where the probability distribution for the trigrams is

calculated using all of Shakespeare's plays as the corpus. While scoring the student plays however, you find that the data was not enough and most sentences in the student's plays have a score of 0. What options do you have to come out of your predicament if you still want to score the plays automatically?

Ans:

The options are the same as when we have inadequate language models. We back-off to simpler models e.g. bigram and unigram models and use smoothing to compute probabilities where the corpus does not have instances of the relevant n-gram. Another option is to use a weighted linear combination of multiple n-gram models for different values of n (e.g. n=1, 2, 3).

Part of speech tagging and HMM

1. 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]

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*			

Soln:

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]

2.

Part of speech tagging using Hidden Markov Model (HMM) – M2

The following three sentences (S1, S2 and S3) and their corresponding tag sequences (T1, T2 and T3) are given as training data for implementing HMM. **Answer questions A-D.**

S1: John cut the paper .	S2: Mary asked for a hair cut .	S3: Sharon asked for a pay cut .
T1: N V D N STOP	T2: N V I P N N STOP	T3: N V I P N N STOP

What will be size of Emission and Tag translation matrices?
Note: Include the start and the stop symbols and assume we are working with a bigram model.

What will be the Emission Probability P(asked | V) ?

What will be the Tag Translation Probability P(I|V) ?

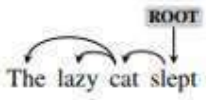
If a brute force approach is employed to find the tags for the test sentence "**I had a deep cut.**", how many possible tag sequences need to be evaluated?

Parsing

1. 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 → 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

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



Solution:

[Root,]	[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]
[Root,]	[slept]	[LA]
[Root,Slept]	[]	[RA]

SH,SH,LA,LA,SH,LA,RA

3. 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

Soln:

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) the		matches
4.((N VP) 2) guy		matches
5.((VP)3)		

6.((V NP) 3) matches
ate

7.((Det N) 4)
matches the

8.((N))5
matches pizza

9.()

4. Given the grammar and lexicon below show the final chart for the following sentence after applying the bottom up chart parser.[5 marks]

S: Book the flight on airasia

S->VP

VP->V NP

NP->NP PP

NP->Det Noun

PP ->Prep Noun

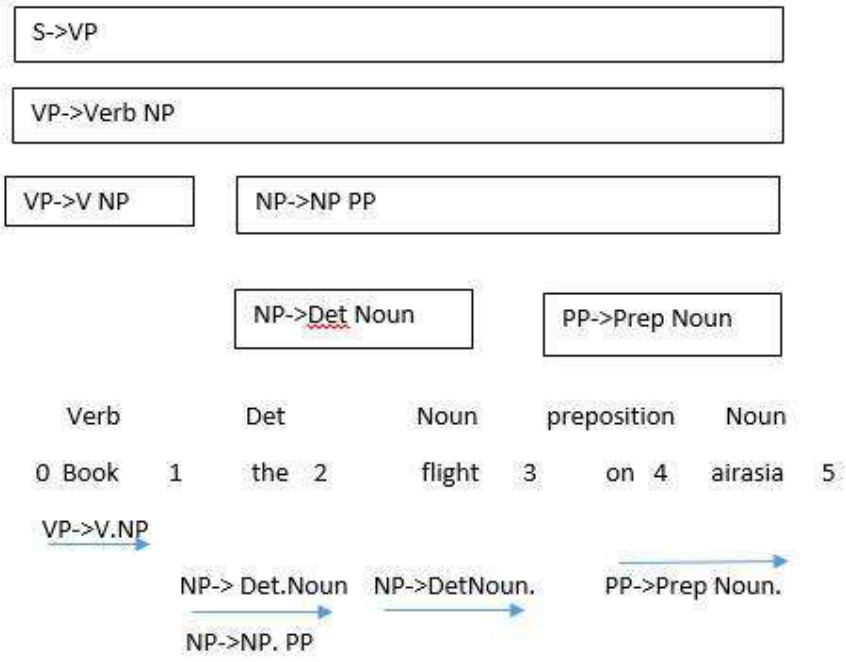
Det ->the

Verb->Book

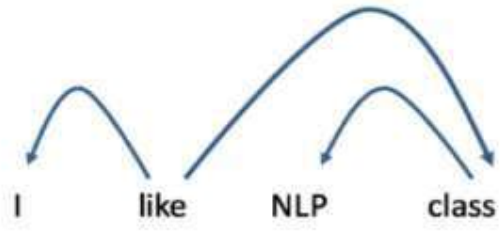
Prep->on

Noun->flight|airasia

Soln:



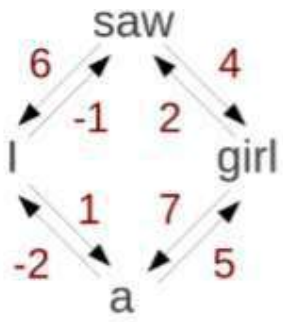
5. Correct sequence of actions that generates the following parse tree of the sentence “I like NLP class" using Arc-Eager Parsing is [5marks]



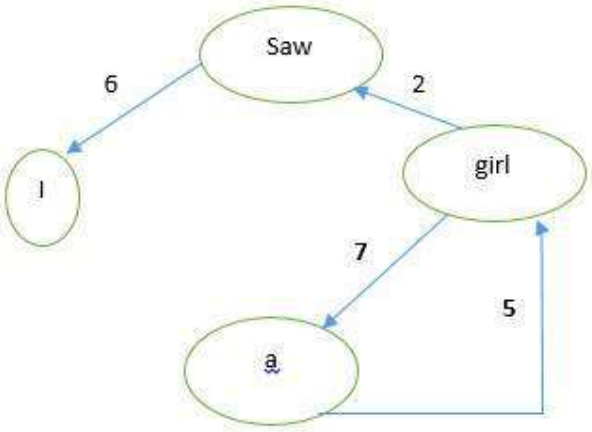
Soln:

SH->LA->RA->SH->LA->RA->RE->RE->RE

6. In the below weighted graph, the edge weights between girl-saw and a-girl in the maximum spanning tree are: [5]



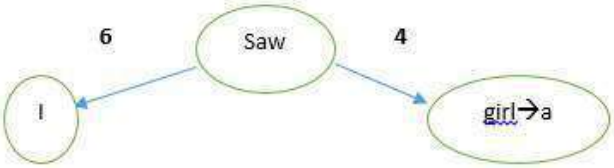
Soln:



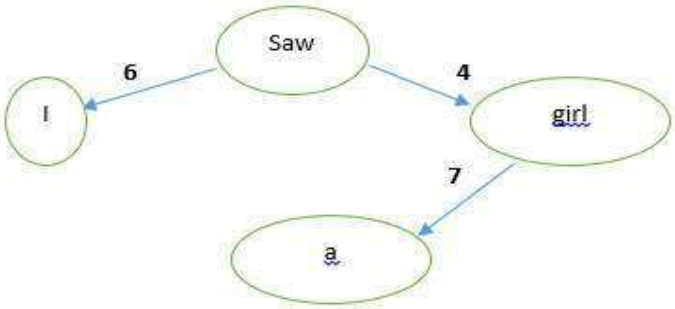
A → girl → forms a cycle

As per Chu Liu Edmond algorithm we can remove the cycle by contracting and expansion

Contracting step



Expanding



Because saw ->girl -> a gives more weightage than a->girl->saw

So the final weight for saw ->girl is 4 and a-> girl is 7

.

7. Given a treebank how would you determine probabilities for the grammar rules given in Question 1 (for use with a basic PCFG parser)?

Let's take the VP rule. There are three VP rules. I would count the total number of VP rules in the

Treebank. Then, for each rule, I would count the number of times that rule occurs and divide by the total number of VP rules. That would yield the probability for each rule. I would follow a similar procedure for each rule where the same non-terminal appeared on the left-hand side.

8.

1. Given the grammar and lexicon below (which is the same as that of question 5), show one possible **top-down derivation** for the sentence:

   Run the Detroit marathon

S → NP VP	Det → the
S → VP	Noun → run — marathon
NP → Det NP	Verb → run
NP → Proper-Noun Noun	Proper-Noun → Detroit
VP → Verb NP	

Answer

S → VP
S → Verb NP
S → Verb Det NP
S → run Det NP
S → run the Proper-Noun Noun
S → run the Detroit Noun
S → run the Detroit marathon

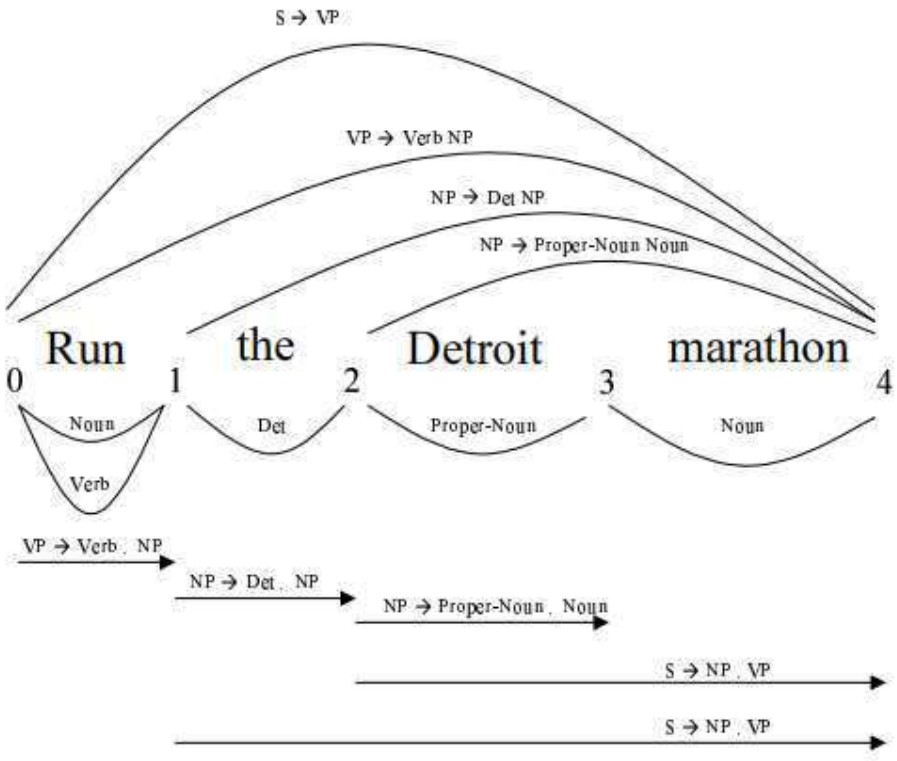
9.

Given the grammar and lexicon below, show the **final chart** for the following sentence after applying the bottom-up chart parser from class:

Run the Detroit marathon

Remember that the final chart contains all edges added during the parsing process. You may use either the notation from class (i.e. nodes/links) or the notation from the book to depict the chart.

S → NP VP	Det → the
S → VP	Noun → run | marathon
NP → Det NP	Verb → run
NP → Proper-Noun Noun	Proper-Noun → Detroit
VP → Verb NP	


S → VP
VP → Verb NP
NP → Det NP
NP → Proper-Noun Noun
Run
the
Detroit
marathon
0
1
2
3
4
Noun
Verb
Det
Proper-Noun
Noun
VP → Verb . NP
NP → Det . NP
NP → Proper-Noun . Noun
S → NP . VP
S → NP . VP

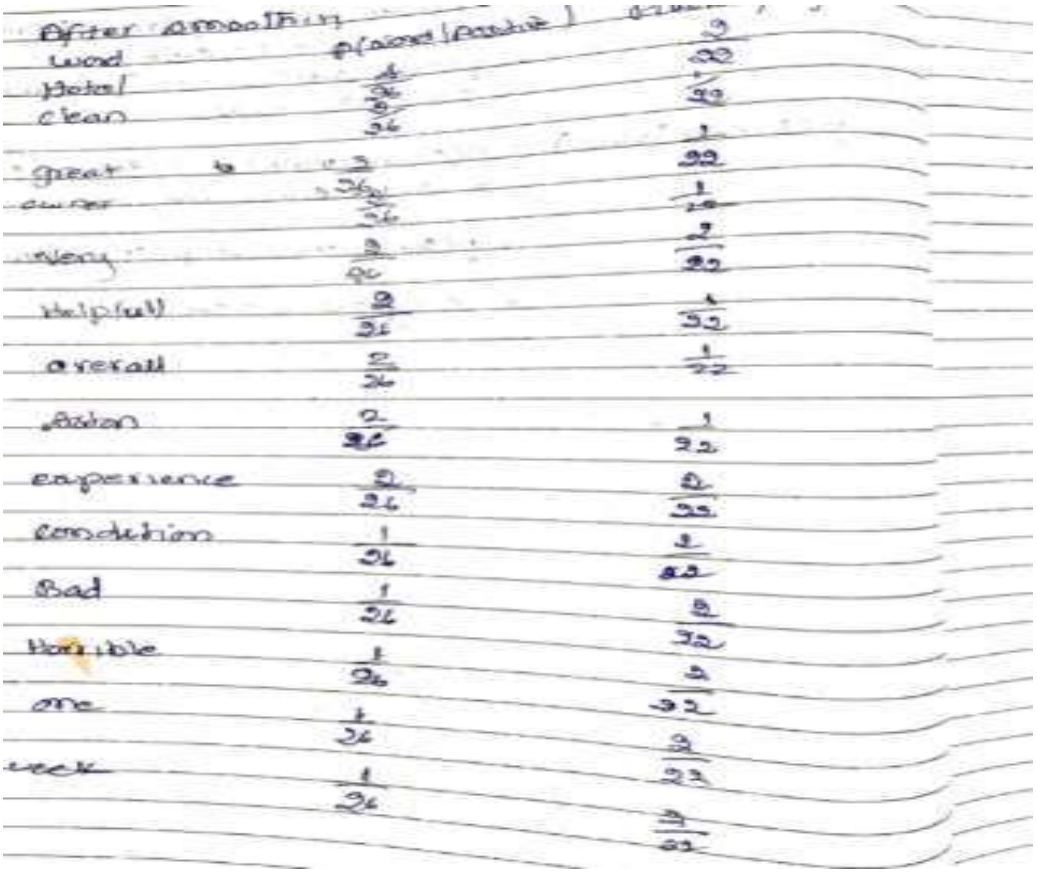
_____*********_____

Sentiment analysis

1.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 Positive	?

Soln:



1) P(Positive/sentence) = 0.01

2) P(negative/sentence) = 0.0016

D6 → +ve

3) P(Positive/sentence) = 0.0017

P(negative/sentence) = 0.0033

D7 → -ve

3) P(Positive/sentence) = 0.07

P(negative/sentence) = 0.0016

P_8 is positive

2. Given the following documents and their sentiment polarities

Document	Sentiment words	Polarity
D1	Good, Enjoy, Good	Positive
D2	Poor, Unpleasant	Negative
D3	Enjoy ,Wonderful	Positive
D4	Good, Lovely	Positive
D5	Good, Poor, Rude	Negative
D6	Good ,Wonderful	?

Determine the sentiment polarity of document D6 using the multinomial naïve Bayes classification (with add1 smoothing) approach. Show your step in detail.

Solution:

P (Positive) =3/5

P (Negative) =2/5

P (Good /Positive) =3+1/7+7=4/14 P (Good/Negative) =1+1/5+7=2/12

P (Wonderful/Positive) =1+1/7+7 =2/14 P (Wonderful/negative) =0+1/5+7=1/12

For the document 6

P (Positive/Good, Wonderful) = 4/14*2/14*3/5

=0.29*0.14*0.6

=0.024

P (Negative/ Good, Wonderful) =2/12*1/12*2/5

=0.16*0.083*0.4

=0.005

Sentiment polarity of document D6 is Positive

3. For sentiment analysis of twitter data, Which classifier did you choose among SVM and Naïve bayes and why? Justify your answer.[3m]

Ans:

Option -1 mark

Justification (2marks) since, you are given only the data of tweets and no other information, which means there is no target variable present. One cannot train a supervised learning model, both svm and naive bayes are supervised learning techniques.

b) ***I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought it. She also thought the phone was too expensive, and wanted me to return it to the shop.*** **What are the problems associated with this type of sentiment analysis?[3Marks]**

Ans:

Identification of Implicit aspects.

Multiple sentiments for same opinion phrase for different aspects

Association of corresponding sentiments to aspects in a multi-aspect review

4. How is sentiment calculated or scored? [2M]

5. Compare Rule based approach and machine learning approach in sentiment analysis [5].

6.

Given a list of positive and negative seed sentiment words and Table-1 providing the details of word occurrence and co-occurrences:
positive seeds (Pwords): good, nice, excellent, positive, fortunate, correct, superior
negative seeds (Nwords): bad, nasty, poor, negative, unfortunate, wrong, inferior
Compute the polarity of the phrase "excellent and outstanding".

Word 1	Word 2	Count of word 1	Count of word 2	count of co-occurrences
excellent	outstanding	5578	2749	1384
poor	outstanding	283891	3293296	3347

Table-1

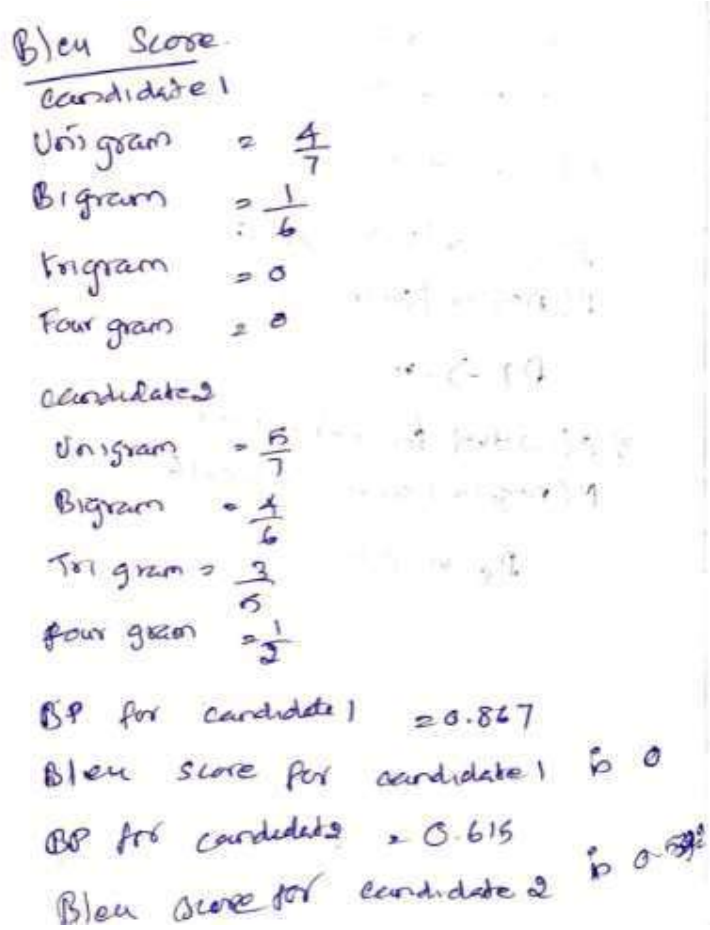
Machine translation

1. Compute the BLEU score for the below translations (candidate1, candidate2).Consider 1gram,2 gram,3 gram,4 gram and Brevity-Penalty for calculating BLUE score .

Reference: The teacher arrived late because of the traffic

Candidate 1: The teacher was late due to the traffic
Candidate 2: A teacher arrived late because of transportation

Soln:



2. Compute the BLEU score for the below translations (candidate1, candidate2).Consider 1gram,2 gram,3 gram,4 gram and **Brevity-Penalty for calculating BLUE score .**

Reference: The NASA Opportunity rover is battling a massive dust storm on Mars.
Candidate 1: The Opportunity rover is combating a big sandstorm on Mars.
Candidate 2: A NASA rover is fighting a massive storm on Mars.

Soln:

Metric	Candidate 1	Candidate 2
precision1 (1gram)	8/11	9/11
precision2 (2gram)	4/10	5/10
precision3 (3gram)	2/9	2/9
precision4 (4gram)	0/8	1/8
Brevity-Penalty	0.83	0.83
BLEU-Score	0.0	0.27

3. Compute the BLEU score for the machine transaltion output .

Assume 1gram ,2 gram and *brevity penalty*

German: Ich bin zur Zeit nicht im Büro
Reference English: I am currently out of the office
MT English: I am currently not in the office

Unigram

$$p_1 = \frac{1+1+1+0+0+1+1}{7} = \frac{5}{7} = 0.71$$

Bigram

English: (I am), (am currently), (currently out), (out, of), (of the), (the office)
MT English: (I am), (am currently), (currently not), (not in), (in the), (the office)

P1=1+1+1+1+1+1/6 =4/6=0.67

brevity penalty =1

*Final BLUE Score =1*0.71*0.67=0.48*

4. Compute the BLEU score for the MT output.

(**Assume 1gram,2 gram.3 gram.4 gram and brevity penality)**

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: The cat the cat on the mat.

Question and answer system and Chat bot

1. Identify the type of the dialogue agent from the following conversation. Also explain the working of that type of dialogue agent.
(5 marks)

System: How may I help you?

User: I want to book dinner at a restaurant that serves good Chinese food

System: Berkos is a restaurant that serves good Chinese food.

User: ok

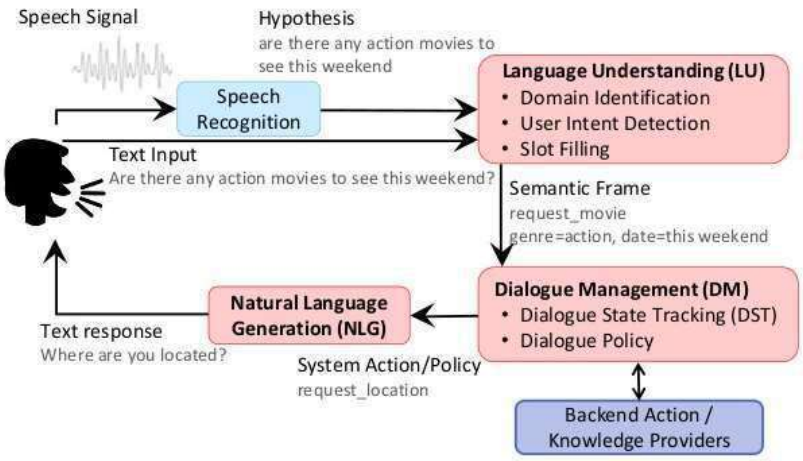
System: Should I book a dinner?

User: yes, please

System: ok.

Soln: Above is the task oriented dialogue system

The diagram shows the working of the task oriented dialogue system.
The student should explain briefly each step as mentioned in the slides



2. Find the intent, domain and slots for the following:
(2 marks)

Book an appointment on 12th Feb 2021 at 10 am for a ECG Test.

SOLN:

DOMAIN: Medical
INTENT: Book an Appointment

Slots

- Services: ECG TEST
- Date: 12th Feb 2021
- Time: 10 AM

3. In a collection of 10000 document, the following words occur in the following number of documents:

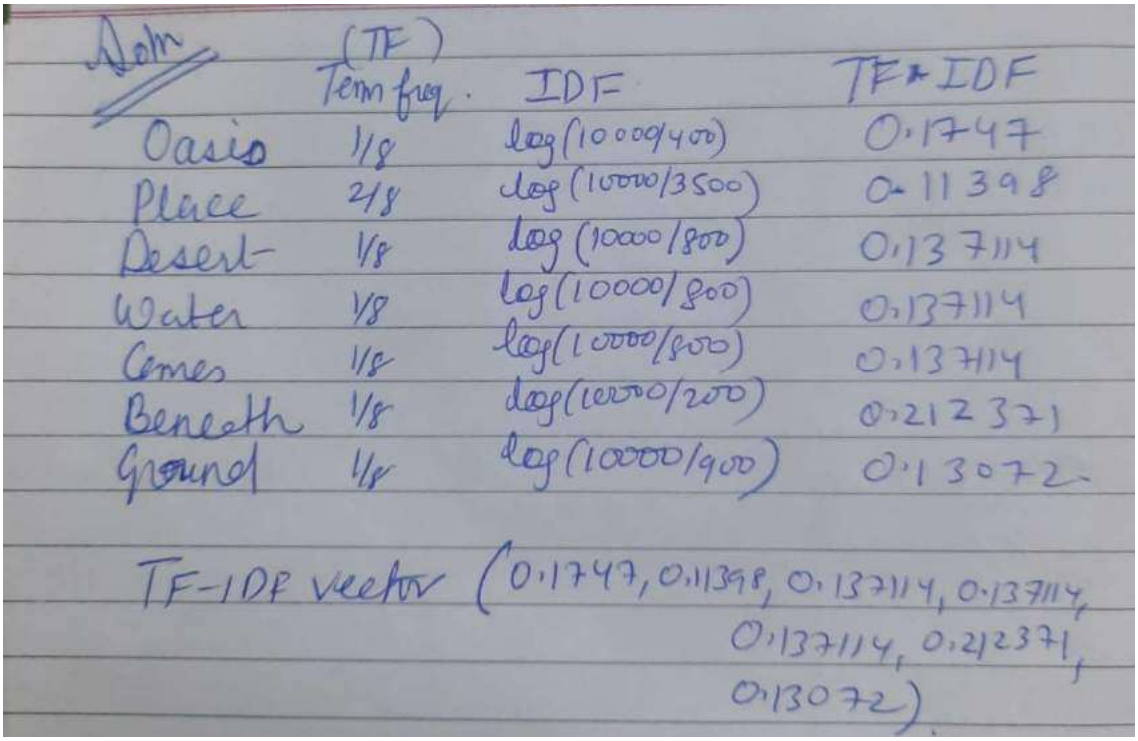
(3 marks)

Oasis occurs in 400 documents, Place occurs in 3500 documents, Desert occurs in 800 documents, Water occurs in 800 documents, Comes occur in 800 documents

Beneath occurs in 200 documents, Ground occurs in 900 documents

Calculate TF-IDF term vector for the following document:

Oasis Place Desert Water Comes Beneath Ground Place



Dom	(TF) Term freq.	IDF	TF*IDF
Oasis	1/8	log(10000/400)	0.1747
Place	2/8	log(10000/3500)	0.11398
Desert	1/8	log(10000/800)	0.137114
Water	1/8	log(10000/800)	0.137114
Comes	1/8	log(10000/800)	0.137114
Beneath	1/8	log(10000/200)	0.212371
Ground	1/8	log(10000/900)	0.13072

TF-IDF vector (0.1747, 0.11398, 0.137114, 0.137114, 0.137114, 0.212371, 0.13072)

Word sense disambiguation and ontology

1) 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).

2. 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

Ans:

- **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.**

3. 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 word net can be extracted for these words and senses with close relations can be extracted 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.

4. Consider below three types of movie reviews and convert it into Bag of words model:

- Review 1: This movie is very scary and long
- Review 2: This movie is not scary and is slow
- Review 3: This movie is spooky and good

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good
Review 1	1	1	1	1	1	1	1	0	0	0	0
Review 2	1	1	2	0	0	1	1	0	1	0	0
Review 3	1	1	1	0	0	0	1	0	0	1	1

2.

- *the cat sat*
- *the cat sat in the hat*
- *the cat with the hat*

.

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

Semantic web

1. How is Syntactic web different from the Semantic web? What is URI in semantic web ontology?

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 likecarnivorous, 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>
```

Midsem Question structure

A Vijayalakshmi

All Sections

Course Title Natural Language Processing

Weightage : 30%

Q1) 8 marks

Module 1: Introduction (Theory) - 3 Marks

- The Different Levels of Language Analysis
- Representations and Understanding.
- The Organization of Natural Language Understanding Systems

Module 2: Language Models (Problem) - 5 Marks

- N-Grams
- Evaluating Language Models
- Generalization and Zeros
- Smoothing
- The Web and Stupid Backoff

Q2) 10 Marks

Module 3: Hidden Markov Models (Theory/Problem) - 5 Marks

- Markov Chains
- The Hidden Markov Model

- Likelihood Computation: The Forward Algorithm
- Decoding: The Viterbi Algorithm

Module 4: POS Tagging (Problem) - 5 Marks

- (Mostly) English Word Classes
- The Penn Treebank Part-of-Speech Tag set
- Part-of-Speech Tagging
- HMM Part-of-Speech Tagging

Q3) 12 Marks

Module 5: Parsing (Problems)

- Grammars and Sentence Structure.
- What Makes a Good Grammar
- A Top-Down Parser.
- A Bottom-Up Chart Parser.
- Chart Parsing.

Module 6: Statistical Constituency Parsing (Theory or Problems)

- Probabilistic Context-Free Grammars
- Probabilistic CKY Parsing of PCFGs
- Ways to Learn PCFG Rule Probabilities
- Problems with PCFG
- Probabilistic Lexicalized CFGs
- Probabilistic CCG Parsing

Module7:Dependency parsing(Theory or problems)

- Arc eager parsing


Search entries or author
Unread

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

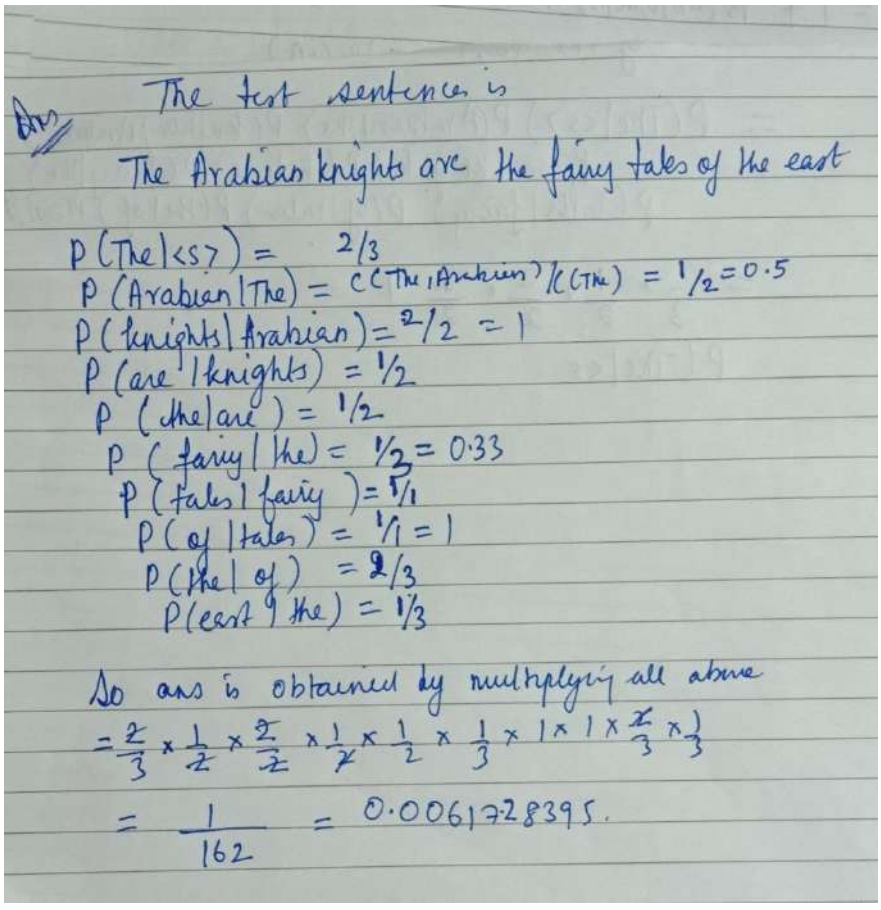
Note: Assumptions made if any, should be stated clearly at the beginning of your answer.

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*			

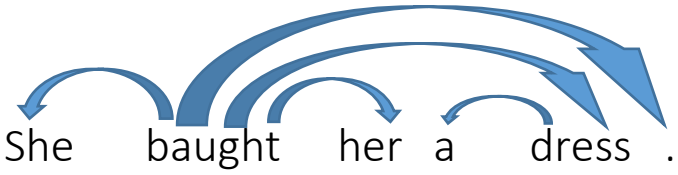
Question 2.

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"/>
```

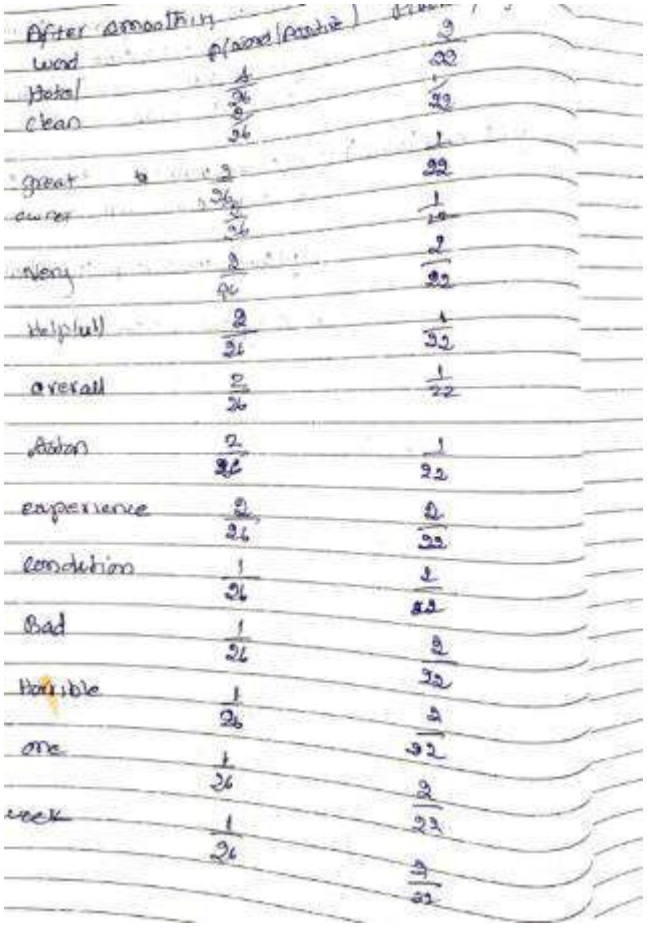
```
  <rdfs:range  rdf:resource="#Animal"/>
</owl:ObjectProperty>
```

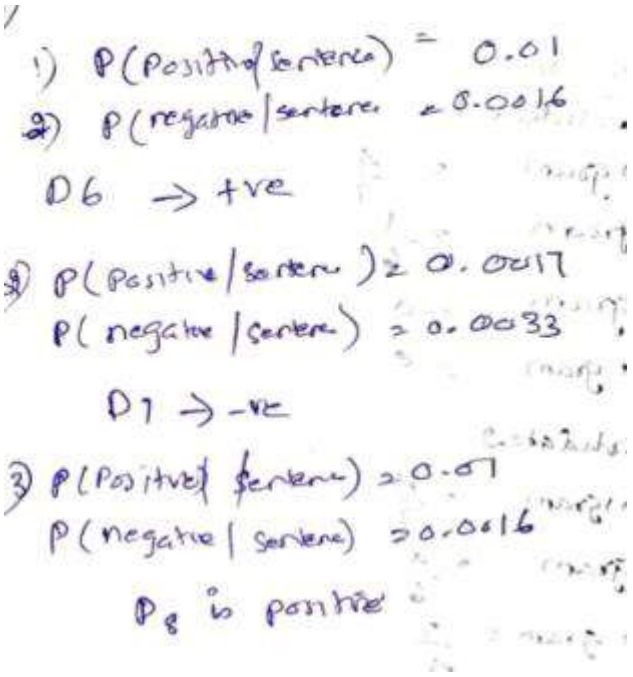
Question 4.

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 :

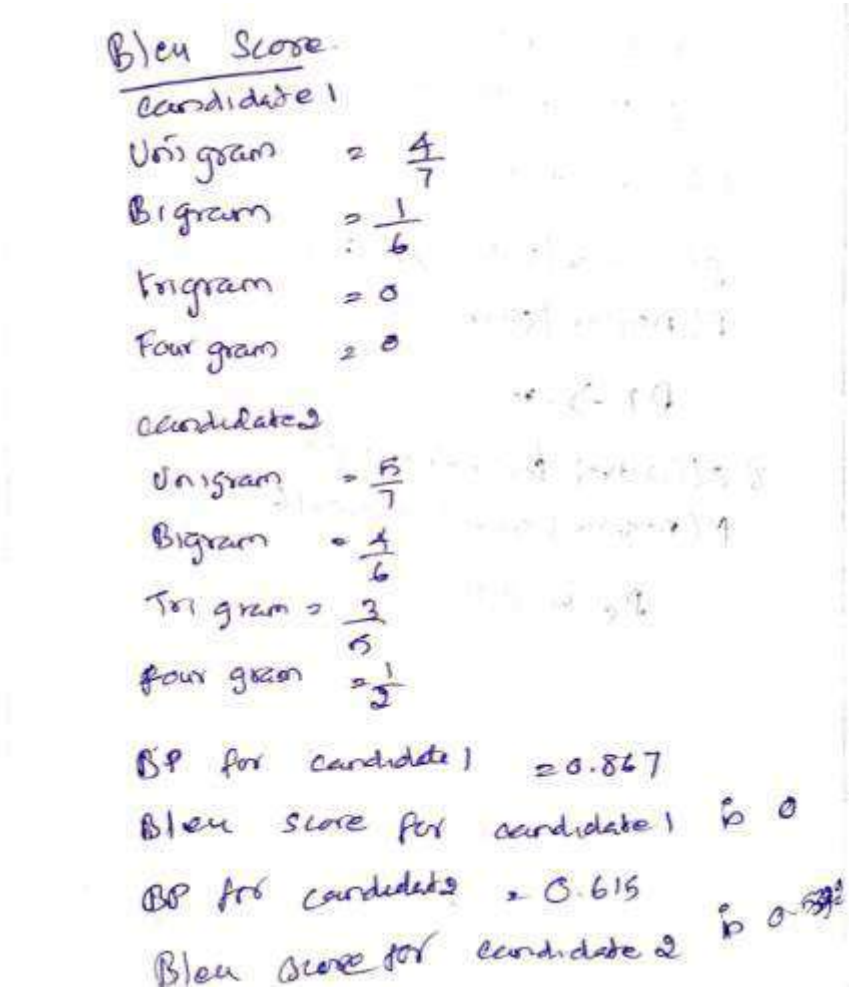




b. Compute the BLEU score for the below translations (candidate1, candidate2).Consider 1gram,2 gram,3 gram,4 gram and Brevity-Penalty for calculating BLUE score .

Reference: The teacher arrived late because of the traffic

Candidate 1: The teacher was late due to the traffic
Candidate 2: A teacher arrived late because of transportation

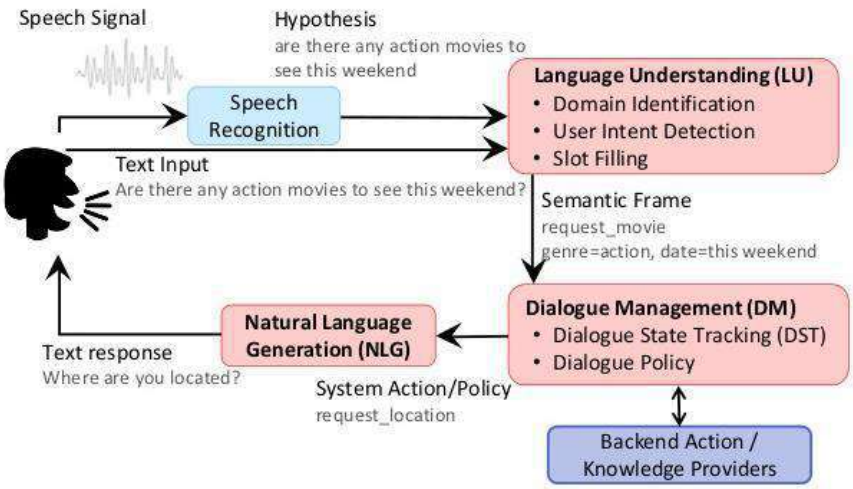


1. Identify the type of the dialogue agent from the following conversation. Also explain the working of that type of dialogue agent. (5 marks)

   System: How may I help you?
   User: I want to book dinner at a restaurant that serves good Chinese food
   System: Berkos is a restaurant that serves good Chinese food.
   User: ok
   System: Should I book a dinner?
   User: yes, please
   System: ok.

   Soln: Above is the task oriented dialogue system

The diagram shows the working of the task oriented dialogue system. The student should explain briefly each step as mentioned in the slides



2. Find the intent, domain and slots for the following: (2 marks)
   Book an appointment on 12th Feb 2021 at 10 am for a ECG Test.
   SOLN:
   DOMAIN: Medical
   INTENT: Book an Appointment

   Slots

   - Services: ECG TEST
   - Date: 12th Feb 2021
   - Time: 10 AM

3. In a collection of 10000 document, the following words occur in the following number of documents: (3 marks)
   Oasis occurs in 400 documents, Place occurs in 3500 documents, Desert occurs in 800 documents, Water occurs in 800 documents, Comes occur in 800 documents
   Beneath occurs in 200 documents, Ground occurs in 900 documents
   Calculate TF-IDF term vector for the following document:
   Oasis Place Desert Water Comes Beneath Ground Place

Doc	(TF) Term freq.	IDF	TF*IDF
Oasis	1/8	log(10000/400)	0.1747
Place	2/8	log(10000/3500)	0.11398
Desert	1/8	log(10000/800)	0.137114
Water	1/8	log(10000/800)	0.137114
Cenes	1/8	log(10000/800)	0.137114
Beneath	1/8	log(10000/200)	0.212371
Ground	1/8	log(10000/900)	0.13072

TF-IDF vector (0.1747, 0.11398, 0.137114, 0.137114, 0.137114, 0.212371, 0.13072).

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 Regular)

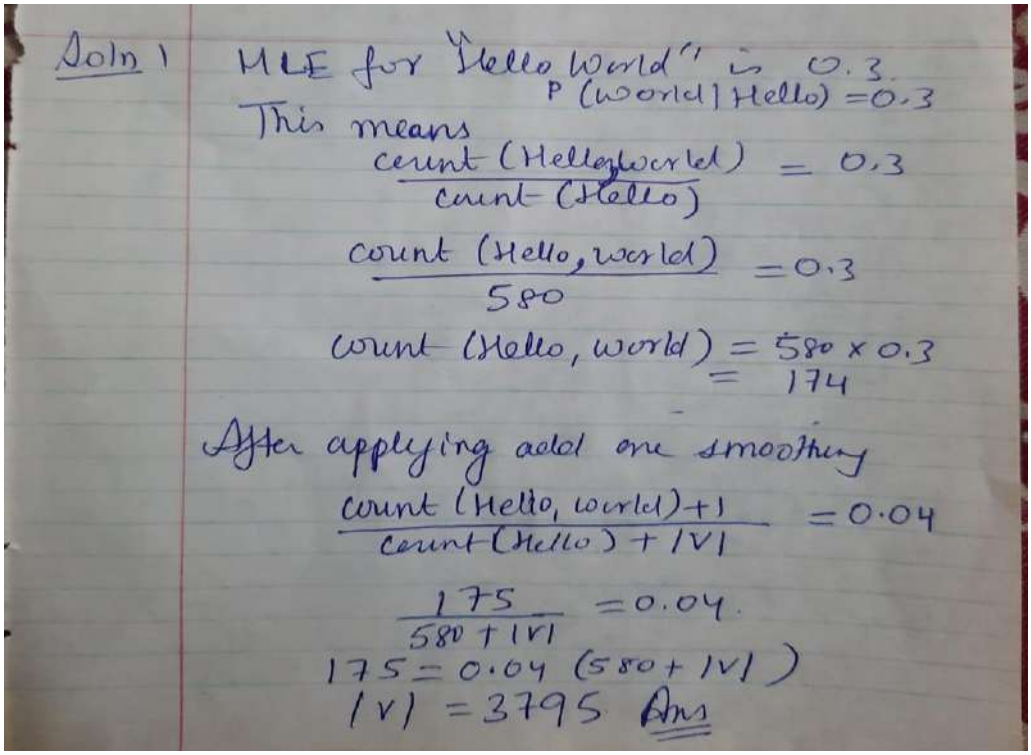
Course No.	: DSECLZG525
Course Title	: Natural Language Processing
Nature of Exam	: Open Book
Weightage	: 50%
Duration	: 2 hours

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

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

Question 1.

a) Given a corpus C, the maximum likelihood estimation (MLE) for the bigram "Hello World" is 0.3 and the count of occurrence of the word "Hello" is 580 for the same corpus, the likelihood of ""Hello World" after applying the add-one smoothing is 0.04. What is the vocabulary size of Corpus C.
(3 marks)



b) What are the challenges in the Natural Language Processing? (3 marks)

Natural Language Processing has following challenges:

- Contextual words and phrases and homonyms

The same words and phrases can have different meanings according the context of a sentence and many words – especially in English – have the exact same pronunciation but totally different meanings.

- Synonyms

Synonyms can lead to issues similar to contextual understanding because we use many different words to express the same idea.

- Irony and sarcasm

Irony and sarcasm present problems for machine learning models because they generally use words and phrases that, strictly by definition, may be positive or negative, but actually connote the opposite

- Ambiguity

Lexical ambiguity: a word that could be used as a verb, noun, or adjective.
Semantic ambiguity: the interpretation of a sentence in context. For example: I saw the boy on the beach with my binoculars. This could mean that I saw a boy through my binoculars or the boy had my binoculars with him
Syntactic ambiguity: In the sentence above, this is what creates the confusion of meaning. The phrase with my binoculars could modify the verb, "saw," or the noun, "boy."

- Errors in text or speech

Misspelled or misused words can create problems for text analysis. Autocorrect and grammar correction applications can handle common mistakes, but don't always understand the writer's intention.

- Colloquialisms and slang

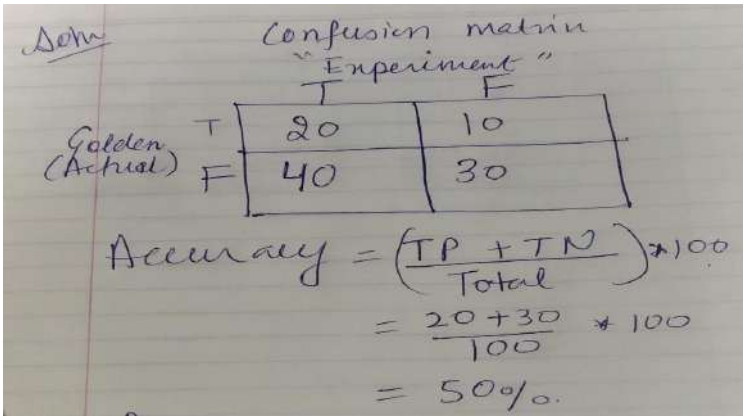
Informal phrases, expressions, idioms, and culture-specific lingo present a number of problems for NLP – especially for models intended for broad use.

- Domain-specific language

Different businesses and industries often use very different language. An NLP processing model needed for healthcare, for example, would be very different than one used to process legal documents.

- Lack of research and development

c) There were 100 documents and each document contained one word. 30 of these documents contained the word "hello". I asked Bob to separate all the documents containing the word "hello". He showed me 60 but "hello" was not in 40 of them. Construct the confusion matrix and calculate the accuracy. (4 marks)



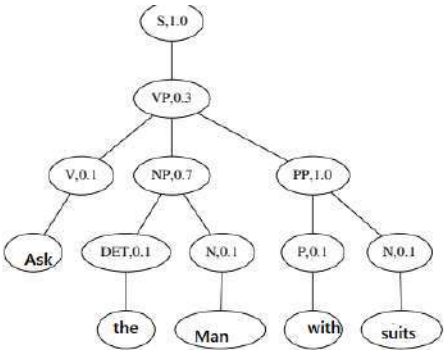
Question 2.

Given the following PCFG, find the parse trees for the given sentence and their probabilities .And find out that the word 'suits' is attached with' ask 'or 'man' and why? [10 marks]

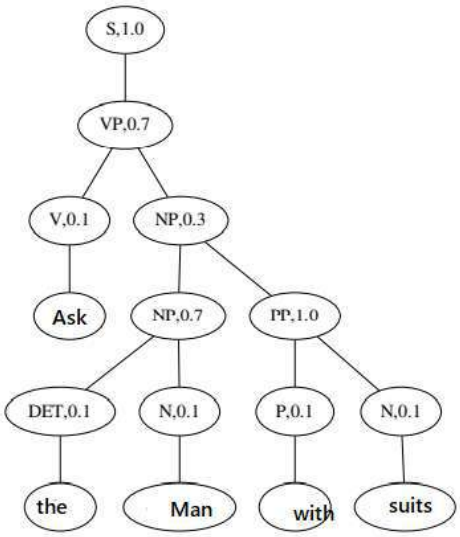
Ask the man with suits

Rule	probability
S → V P	1.0
V P → V NP	0.7
V P → V NP P P	0.3
NP → NP P P	0.3
NP → DET N	0.7
P P → P N	1.0
DET → the	0.1
V → ask	0.1
P → with	0.1
N → man | suits	0.1

Soln:



Probability= $0.3 \times 0.7 \times 0.1^{5} = 21 \times 10{-}7$



Probability= $0.3 \times 0.7 \times 0.7 \times 0.1^{5} = 14.7 \times 10{-}7$

The first tree has higher probability and it is the correct parse since 'with suits' should attach to 'ask' rather than 'man'.

Question 3. Word sense disambiguation and ontology-

a) How can the Simple Lesk algorithm be applied to disambiguate the exact meaning of "**bass**" in following sentence **[5 marks]**
The **bass** guitar, is the lowest pitched member of the guitar family of instruments.

S:(n) bass (the lowest part of the musical range)
S: (n) sea bass, bass (the lean flesh of a saltwater fish of the family Serranidae)
S: (n) bass (the member with the lowest range of a family of musical instruments)
S: (adj) bass, deep (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

b) Build a small part of ontology for MTech DSE program in OWL syntax with following concepts **[3 marks]**

- Professor
- Student
- Courses
  Also include following relations/constraints:
  - Domain
  - Range
  - subClassOf
  - disjointWith

  How are the ontology languages OWL and RDF different from each other. Can you express the same constraints using RDF? If not which one cannot be expressed using RDF? [2 **marks]**

  ```
  <rdfs:Class rdf:ID=" Professor">
    <rdfs:subClassOf rdf:resource="# AcademicStaff "/>
  </rdfs:Class>
  <rdfs:Class rdf:ID="Professor">
    <owl:disjointWith rdf:resource="#AssistantProfessor"/>
  </rdfs:Class>
  ```

  OWL is more advanced and has inferencing capability since owl is based on description logic. Some contraints like disjoint with cannot be expressed using RDF

Question 4.

1. Given the two machine translation systems output and reference given below, find the best machine translation system using BLEU score with Brevity penality. [5marks]

[Hint: Assume 1-gram, 2-gram, 3 -gram and 4- gram for calculating BLEU score)

System A: Israeli official's responsibility of airport safety

System B: Airport security Israeli officials are responsible

Reference: Israeli officials are responsible for airport security

2. Given the following documents and their sentiment polarities [5 marks]

Document	Sentiment words	Polarity
D1	Great, Enjoy, Great	Positive
D2	Poor, Unpleasant	Negative
D3	Enjoy ,amazing	Positive
D4	Great, Lovely	Positive
D5	Great, Poor, Rude	Negative
D6	Great ,amazing	?

Determine the sentiment polarity of document D6 using the multinomial naïve Bayes classification (with add1 smoothing) approach. Show your step in detail.

Solution:

P (Positive) =3/5

P (Negative) =2/5

P (Great /Positive) =3+1/7+7=4/14 P (Great/Negative) =1+1/5+7=2/12

P (Amazing/Positive) =1+1/7+7 =2/14 P (Amazing/Negative) =0+1/5+7=1/12

For the document 6

P (Positive/Great, Amazing) = 4/14*2/14*3/5

=0.29*0.14*0.6

=0.024

P (Negative/ Great, Amazing) =2/12*1/12*2/5

=0.16*0.083*0.4

=0.005

Sentiment polarity of document D6 is Positive

Question 5.

a) Let there be two questions and let there be 4 candidate answers for each question. Also Question Answering System chooses the best answer for question1 and second best answer for question 2. **Calculate the Mean Reciprocal Rank to evaluate the Question Answering System (1 marks**)

Soln: MMR = (1+1/2)/2=3/4

b) Let there be four documents given by

D1: the best American restaurant enjoys the best burger

D2: Indian restaurant enjoys the best dosa

D3: Chinese restaurant enjoys the best Manchurian

D4: the best the best Indian restaurant

Compute the BOW for D1, D2, D3 and D4 in the table. (2 Marks)

	the	best	American	Restaurant	enjoys	burger	dosa	manchurian	Chinese	Indian
D1										
D2										
D3										
D4										

Soln b)

	the	best	American	Restaurant	enjoys	burger	dosa	manchurian	Chinese	Indian
D1	2	2	1	1	1	0	0	0	0	0
D2	1	1	0	1	1	0	1	0	0	1
D3	1	1	0	1	1	0	0	1	1	0
D4	2	2	0	1	0	0	0	0	0	1

a) Also find out TF-IDF vector for D1, D2, D3, D4 for the above documents in b. (3 marks)

Soln c)

WORDS	TF (NORMALISED FREQUENCY)					Tf*idf			
	D1	D2	D3	D4	Idf	D1	D2	D3	D4
the	2/8	1/6	1/6	2/6	log(4/4)=0	0	0	0	0
best	2/8	1/6	1/6	2/6	Log(4/4)=0	0	0	0	0
American	1/8	0	0	0	Log(4/1)=0.6	0.6/8=0.075	0	0	0
Restaurant	1/8	1/6	1/6	1/6	Log(4/4)=0	0	0	0	0
enjoys	1/8	1/6	1/6	0	Log(4/3)=0.12	0.12/8=0.015	0.02	0.02	0
burger	1/8	0	0	0	Log(4/1)=0.6	0.6/8=0.075	0	0	0
dosa	0	1/6	0	0	Log(4/1)=0.6	0	0.1	0	0
manchurian	0	0	1/6	0	Log(4/1)=0.6	0	0	0.1	0
Chinese	0	0	1/6	0	Log(4/1)=0.6	0	0	0.1	0
Indian	0	1/6	0	1/6	Log(4/2)=0.3	0	0.3/6=0.05	0	0.3/6=0.05

b) Find Domain, Intent and Define Slots for each of the following Sentences: (4 marks)

1) Book a taxi at 6:00 PM from India Gate to Ambience Mall
2) I want to deposit 100 Dollars in my savings account.

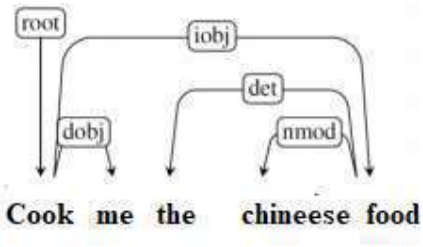
solution

1) Book a taxi at 6:00 PM from India Gate to Ambience Mall
   - DOMAIN: Cab or Taxi
   - INTENT: Taxi-BOOKING
   - Slots
     - SOURCE-LOCATION: India Gate

- DESTINATION-LOCATION: Ambience Mall
- PICKUP TIME: 6:00 PM

2) I want to deposit 100 Dollars in my savings account.
   - DOMAIN: Banking
   - INTENT: Deposit-Account
   - Slots
   - Account Type: Savings Account
   - Transaction: Deposit
   - Amount: 100 dollars

1. Give the correct sequence of arc eager parsing operations for the given sentence



2. Consider the grammar G given below:

1 S → NP VP
2 VP → VT NP
3 NP → D N
4 N → ADJ N
5 VT → saw
6 D → the
7 D → a
8 N → dragon
9 N → boy
10 ADJ → young

(a) You are given the sentence below with the positions marked:

0 the 1 young 2 boy 3 saw 4 the 5 dragon 6

Using the CYK parsing algorithm fill in the table/chart that indicates whether the above sentence has been parsed or not.

(b) Using the table above extract the parse in the form of a derivation of the sentence starting from the start symbol

3. Design a sample ontology for the 'real estate' domain. Clearly mention the

a. Classes
b. Properties
c. Relations
d. Axioms / constraints

e.g. House, Price with 'hasPrice' relation.

The ontology should contain about 10 classes with associated properties, relations and axioms and presented in RDF triple format.

4. You are required to design a word sense disambiguation (WSD) model using WordNet as the background knowledgebase.

a.What are the different features that you would leverage in your model?
b.How would you model the solution and why? Are there any pros/cons of your modeling choice?

5 ."These earphones are a good pick at this price. Connected with laptop for office calls and these are working well although there is no noise cancellation. Quality of wires are a bit thin and look delicate, though neckband is ok. Bass will seem ok if you have not used good quality earphones earlier."
You have been given product review data like the one shown above. You are asked to design a sentiment analysis model for this data. What would be your approach? Describe the different components of your solution. State any assumptions that you are making and pros/cons (if any) of your approach.

6. Compute the BLEU score for the following candidates. Based on this, what can you say about the effectiveness of the BLEU score? Can you suggest ways to make the scoring more effective?

Source: Le professeur est arrive en retard a cause de la circulation
Reference 1: The teacher arrived late because of the traffic
Reference 2: The teacher was delayed due to traffic
Candidate 1: The professor was delayed due to the congestion
Candidate 2: The teacher was held up by the traffic

7. Consider a document *d* containing 200 words wherein the word 'covid' appears 5 times. The document is part of a collection of 100 thousand documents, of which, 10,000 documents contain the word 'covid'. Compute the TF-IDF weight for the word in the document *d.*

8. You are designing a frame-based dialog system for 'cab booking'.

a. What are the different slots in your design? Mention along with their corresponding entity types and questions that the system would ask a user.
b. Show a finite-state dialog manager for the system
c. What changes would you make to the design to change it from a single initiative system to multi-initiative system?

Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division

Second Semester 2020-2021

Mid-Semester Test

(EC-2 Regular)

Course No.	: DSECLZG525
Course Title	: Natural language processing
Nature of Exam	: Open Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: 16/01/2022 FN

No. of Pages = 3

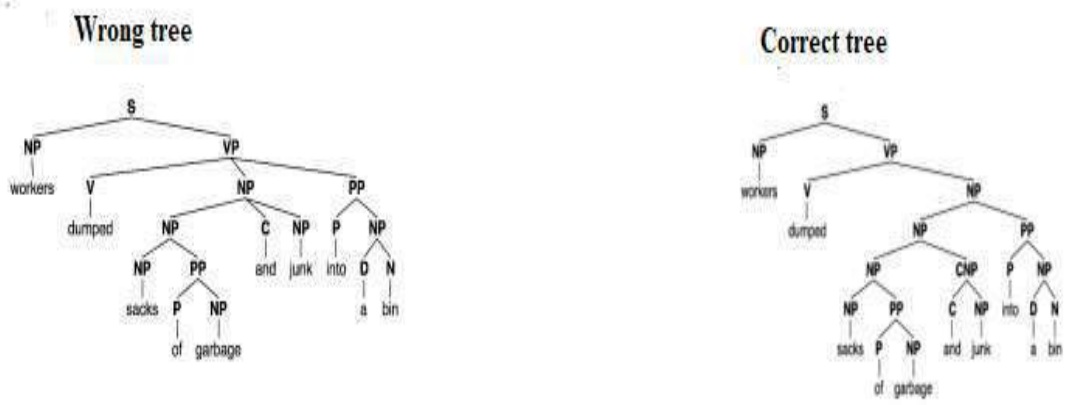
No. of Questions = 4

Note to Students:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1.

a)What is the precision and recall for the following tree model [2 marks]



Solution:

precision = 17/17 (there are 17 nodes in (d), and all of them are correct) -**1mark**

recall = 17/19 (there are 19 nodes in (b), and 17 of them were found) -**1mark**

b) Consider the following training data [6marks]

<s>I am Geeta</s>

<s> Geeta I am</s>

<s>Geeta I like</s>

<s>Geeta I do like </s>

<s> do I like Geeta</s>

What is the most probable next word predicted by the bigram model for the following data

1. <s>Geeta..

2. <s>Geeta I am Geeta..

3.<s>Geeta I do ..

4.<s> do I …

Solution: attached pdf

Marking scheme:

Calculating all the probabilities 2 marks

1. <s>Geeta.. **1mark**

2. <s>Geeta I am Geeta.. **1mark**

3.<s>Geeta I do .. **1mark**

4.<s> do I … **1mark**

c) Obtain all the n-gram probabilities P((I| <s>), P(NLP|<s>), P(am|I) P(do|I) P(NLP|am) from the following set of sentences[2 marks]

<s> I am NLP </s>

<s> NLP I am </s>

<s> I do not like Exams and Marks </s>

Solution:

P((I| <s>)=2/3=0.67; **0.5 marks**

P(NLP|<s>)=1/2=0.5; **0.5 marks**

P(am|I)=2/3=0.67; **0.5 marks**

P(do|I)=1/3=0.33; **both together 0.5 marks**

P(NLP|am)=1/2=0.5

Q2) Suppose that an NLP Engine want to tag the sequence,"natural language processing" using 3 possible tag A, B and C. The engine has the following propbabilities information from training data: P(natural|A)=1/3, P(natural|B)=1/2, P(natural|C)=1/10

P(languagel|A)=2/5, P(language|B)=0, P(language|C)=0,

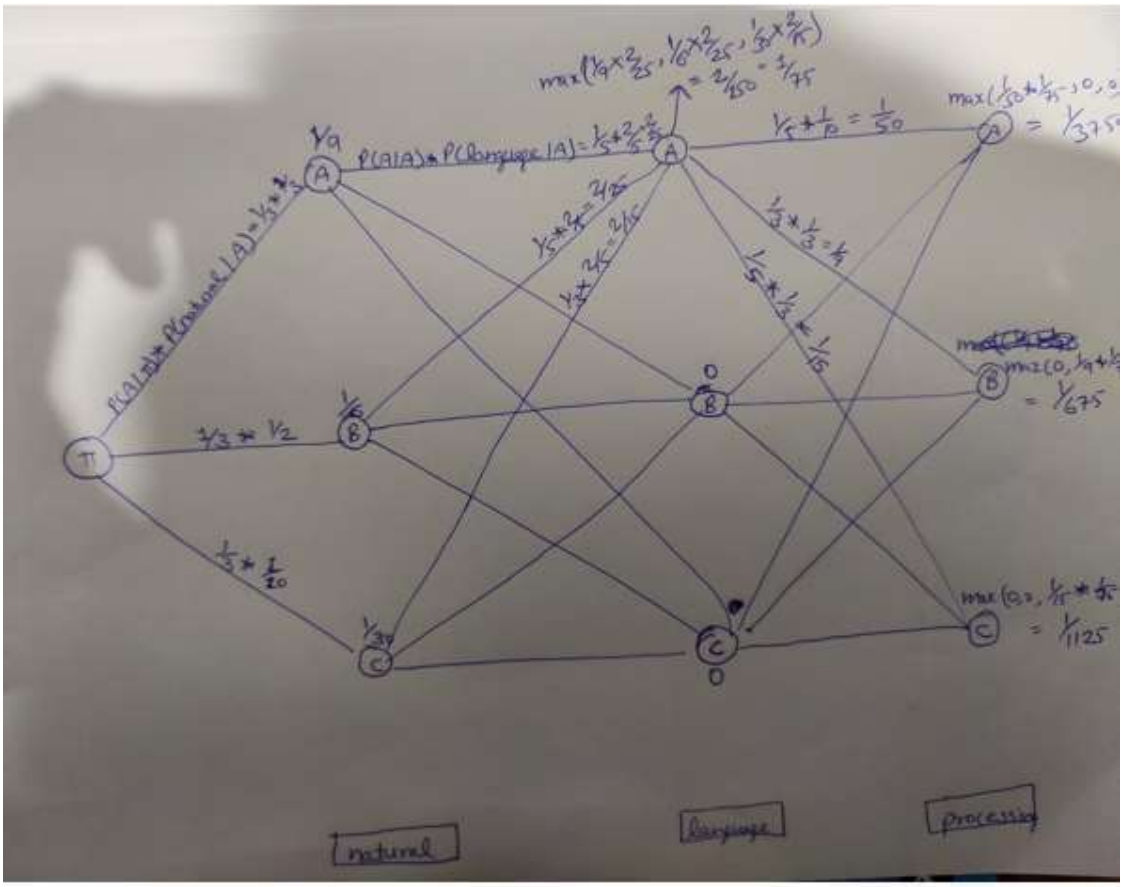
P(processingl|A)=1/10, P(processing|B)=1/3, P(processing|C)=1/3

P(A|A)=1/5,P(B|A)=1/3,P(C|A)=1/5

P(A|B)=1/5,P(B|B)=1/10,P(C|B)=0

P(A|C)=1/3,P(B|C)=1/5,P(C|C)=0

Assume that all the tags have the same probabilities at the beginning of the sentence (and that is 1/3 each). Find out the best tag sequence using Viterbi algorithm along with value at each vertex. [5marks]



nitial state can be represented by 'pi' or by any other symbol
n each layer, we will take the vertex which has the highest value, so best tag sequence is B,
l, B

Marking scheme:

All the node vale should be correct

Best tag sequence should be BAB Then you can give full marks

If only tag sequence is correct and node value are wrong then acc to the no of correct values of node give 1 or 2 marks .dont give full marks.

b) Suppose in our training corpus [2marks]

- **girl** appears 8 times as a noun and 4 times as a verb

- **sleep** appears twice as a noun and 6 times as a verb what is the Emission probabilities of the below sentence

  girl sleep

  Solution:

Noun

P(girl|noun) 0.8 -**0.5marks**

P(sleep| noun) 0.2 -**0.5 marks**

Verb

P(girl| verb) 0.4 -**0.5 marks**

P(sleep| verb) 0.6 -**0.5 marks**

c. Given the emission probabilities and transition probabilities find the correct pos tag for the sentence using HMM [3marks]

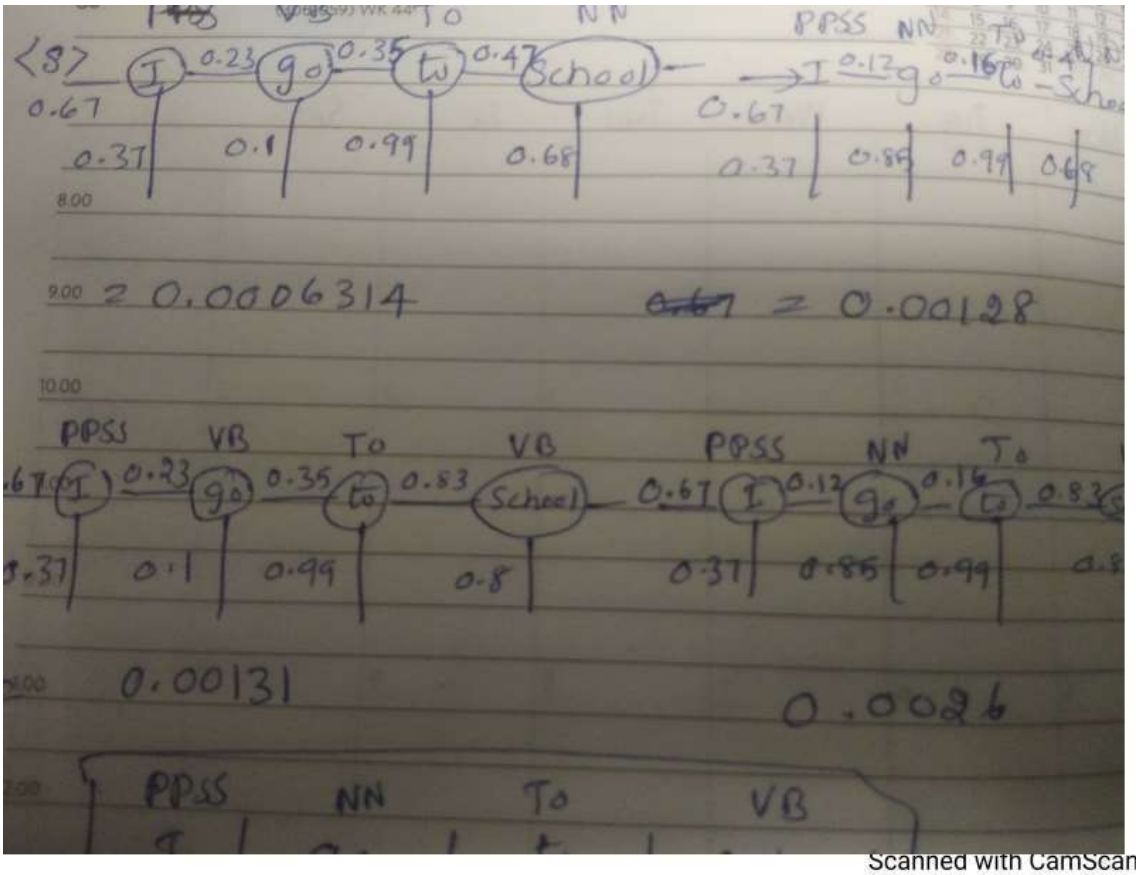
"I go to school"

Emission probabilities

	I	GO	TO	SCHOOL
VB	0	0.1	0	0.8
TO	0	0	0.99	0
NN	0	0.85	0	0.68
PPSS	0.37	0	0	0

Transition probabilities

	VB	TO	NN	PPSS
<s>	0.19	0.43	0.41	0.67
VB	0.38	0.35	0.47	0.70
TO	0.83	0	0.47	0
NN	0.40	0.16	0.87	0.45
PPSS	0.23	0.79	0.12	0.14



Marking scheme:

For calculating probabilities 0.5 each (if it is correct)

Correct answer 1 mark

Q3.

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

S : The cat caught the rat

S->NP VP VP->VNP NP->Det N

N->rat, N->cat ,Det ->the V->caught

Solution:

1The 2 cat 3 caught 4 the 5 rat 6

State	Backup	Action
1. ((S) 1		
2.((NP VP) 1)		
3.(DT N VP) 1)		matches the
4.((N VP) 2)		Matches cat
5.((VP)3)		
6.((V NP) 3)		Matches caught
7.((Det N) 4)		Matches the
8.((N))5		Matches rat

Marking scheme:

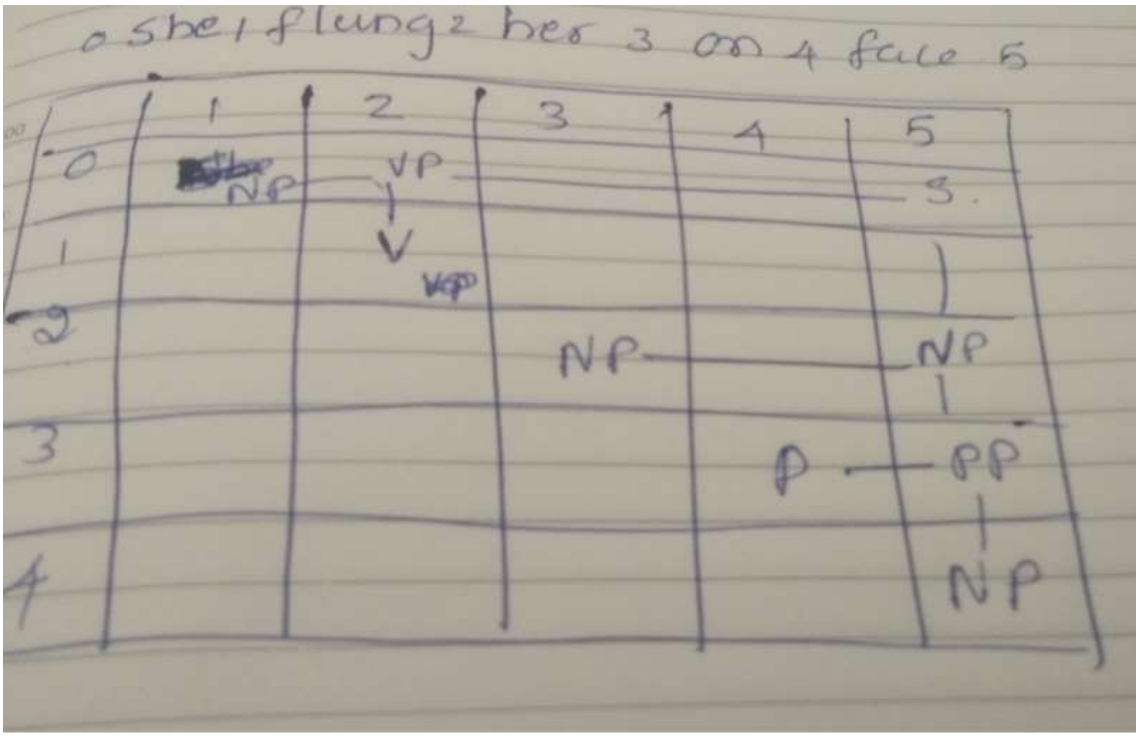
If the chart Is not correct don't give marks

b) Use the CKY parser to parse the sentence[3marks]

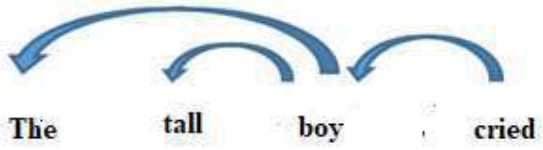
"She flung her on face" given the following grammar and lexicon

S -> NP VP
VP -> V NP
VP -> VP PP
V - > flung
VP -> flung
NP -> NP PP
NP -> She
NP -> her
Face -> NP
P -> On

PP -> P NP



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



[ ]	[Tha tall boy cried]	[]
[The]	[tall boy cried]	[Shift]
[The , tall]	[boy cried]	[Shift]
[The , tall]	[boy cried]	[LA]
[The]	[boy cried]	[LA]
[boy]	[. cried]	[SH]
[ ]	[cried]	[LA]
[cried]	[]	[RA]

OR

[ ]	[The tall boy cried]	[]
[Root,The]	[tall boy cried]	[Shift]
[Root ,the , tall]	[boy cried	[Shift]
[Root ,the , tall]	[boy cried]	[LA]
[Root the]	[boy cried]	[LA]
[Root, boy]	[cried]	[SH]
[ ]	[cried]	[LA]
[Root, cried]	[]	[RA]
[Root]	[]	[RE]

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

d)

[ ] [The tall boy cried]	[]	
[Root,The]	[tall boy cried]	[Shift]
[Root ,the , tall ·]	[boy cried]	[Shift]
[Root ,the , tall]	[boy cried]	[LA]
[Root the l	[boy cried]	[LA]
[Root, boy]	[cried]	[SH]
[· boy] []	[RA]	
[Root, boy cried]	[]	[RE]
[Root, boy]	[]	[RE]
[Root]	[]	[RE]

Note:Without root is also correct.

Marking scheme:

Please scheck the operation sequence .If it is correct give full marks otherwise give 0 marks

****

$$P(I/S) = \frac{C(\langle S\rangle\ I)}{C(\langle S\rangle)} = \frac{1}{5}$$

8.00

9.00 $P(Jack/\langle S\rangle) = C(\langle S\rangle/Jack)/C(\langle S\rangle) = \frac{3}{5}$

10.00 $P(do/\langle S\rangle) = C(do, \langle S\rangle)/C(\langle S\rangle) = \frac{1}{5}$

11.00 $P(am/I) = C(I/am)/C(I) = \frac{2}{5}$

12.00 $P(like/I) = C[I/like]/C(I)] = \frac{2}{5}$

1.00 $P(do/I) = C(I/do)/C(I) = \frac{1}{5}$

2.00 $P(\langle S\rangle/Jack) = C(Jack, \langle S\rangle)/C(Jack) = \frac{2}{5}$

3.00 $P(\langle S\rangle/like) = C(like/\langle S\rangle)/C(like) = \frac{2}{3}$

4.00 $P(\langle S\rangle/am) = C(am/\langle S\rangle)/C(am) = \frac{1}{2}$

5.00 $P(I/Jack) = C(Jack\ I)/C(Jack) = \frac{3}{5}$

6.00 $P(like/do) = C(do, like)/C(do) = \frac{1}{2}$

$P(I/do) = C(do, I)/C(do) = \frac{1}{2}$

$P(Jack/like) = C(like, Jack)/C(like) = \frac{1}{3}$

$P(Jack, am) = C(am, Jack)/C(am) = \frac{1}{2}$

① <S> Geeta
P(<S> | Geeta) = 0.4
P(I, Geeta) = 0.6 ✓
am : <S> Geeta I

<S> Geeta I do
P(like | do) = 1/2
P(I | do) = 1/2
am <S> Geeta I do I or like

<S> Geeta I am Geeta
P(<S> | Geeta) = 0.4
P(I | Geeta) = 0.6
: Geeta I am Geeta I

DECEMBER - 2018

M	T	W	T	F	S	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

THURSD
NOVEM
WK 46 (319

do I ...

$P(\text{do} \mid I) \quad \frac{1}{5} = 0.2$

$P(\text{like} \mid I) \quad \frac{2}{5} = 0.4$

$P(\text{am} \mid I) \quad \frac{2}{5} = 0.4$

Ans: do I like or am

a) Suppose that an NLP Engine want to tag the sequence,?natural language processing? using 3 possible tag A, B and C. The engine has the following propbabilities information from training data: P(natural|A)=1/3, P(natural|B)=1/2, P(natural|C)=1/10

P(languagel|A)=2/5, P(language|B)=0, P(language|C)=0,

P(processingl|A)=1/10, P(processing|B)=1/3, P(processing|C)=1/3

P(A|A)=1/5,P(B|A)=1/3,P(C|A)=1/5

P(A|B)=1/5,P(B|B)=1/10,P(C|B)=0

P(A|C)=1/3,P(B|C)=1/5,P(C|C)=0

Assume that all the tags have the same probabilities at the beginning of the sentence (and that is 1/3 each). Find out the best tag sequence using Viterbi algorithm along with value at each vertex. **[5marks]**

b) Suppose in our training corpus **[2marks]**

- **girl** appears 8 times as a noun and 4 times as a verb
- **sleep** appears twice as a noun and 6 times as a verb what is the Emission probabilities of the below sentence

girl sleep

c) Given the emission probabilities and transition probabilities find the correct pos tag for the sentence using HMM **[3marks]**

?I go to school?

Emission probabilities

	I	GO	TO	SCHOOL
VB	0	0.1	0	0.8
TO	0	0	0.99	0
NN	0	0.85	0	0.68
PPSS	0.37	0	0	0

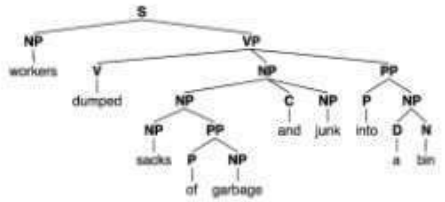
Transition probabilities

	VB	TO	NN	PPSS
<s>	0.19	0.43	0.41	0.67
VB	0.38	0.35	0.47	0.70
TO	0.83	0	0.47	0
NN	0.40	0.16	0.87	0.45
PPSS	0.23	0.79	0.12	0.14

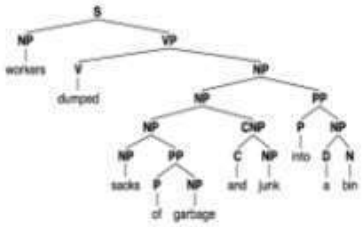
Qtext:-

a) What is the precision and recall for the following tree model **[2 marks]**

Wrong tree



Correct tree



b) Consider the following training data **[6marks]**

<s>I am Geeta</s>

<s> Geeta I am</s>

<s>Geeta I like</s>

<s>Geeta I do like </s>

<s> do I like Geeta</s>

What is the most probable next word predicted by the bigram model for the following data

1. <s>Geeta..
2. <s>Geeta I am Geeta..
3. <s>Geeta I do ..
4. <s> do I ?

c) Obtain all the n-gram probabilities P((I| <s>), P(NLP|<s>), P(am|I) P(do|I) P(NLP|am) from the following set of sentences **[2 marks]**

<s> I am NLP </s>

<s> NLP I am </s>

<s> I do not like Exams and Marks </s>

Qtext:-

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

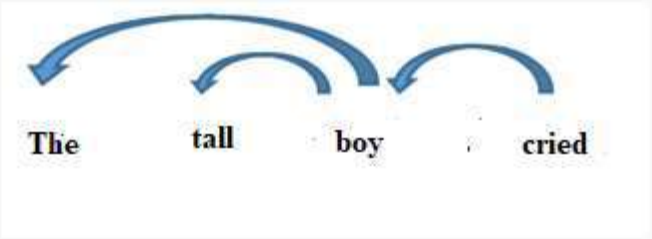
S : The cat caught the rat
S->NP VP VP->VNP NP->Det N
N->rat, N->cat ,Det ->the V->caught

b) Use the CKY parser to parse the sentence [3marks]

?She flung her on face? given the following grammar and lexicon

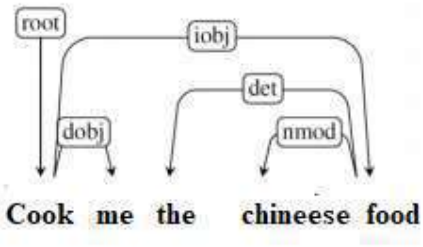
S -> NP VP
VP -> V NP
VP -> VP PP
V - > flung
VP -> flung
NP -> NP PP
NP -> She
NP -> her
Face -> NP
P -> On
PP -> P NP

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



d) Provide a modified transition sequence where the parser mistakenly predicts the arc boy? cried, but gets the other dependencies right. **[2marks]**

1. Give the correct sequence of arc eager parsing operations for the given sentence



2. Consider the grammar G given below:

1 S → NP VP
2 VP → VT NP
3 NP → D N
4 N → ADJ N
5 VT → saw
6 D → the
7 D → a
8 N → dragon
9 N → boy
10 ADJ → young

(a) You are given the sentence below with the positions marked:

0 the 1 young 2 boy 3 saw 4 the 5 dragon 6

Using the CYK parsing algorithm fill in the table/chart that indicates whether the above sentence has been parsed or not.

(b) Using the table above extract the parse in the form of a derivation of the sentence starting from the start symbol

3. Design a sample ontology for the 'real estate' domain. Clearly mention the

a. Classes
b. Properties
c. Relations
d. Axioms / constraints

e.g. House, Price with 'hasPrice' relation.

The ontology should contain about 10 classes with associated properties, relations and axioms and presented in RDF triple format.

4. You are required to design a word sense disambiguation (WSD) model using WordNet as the background knowledgebase.

a.What are the different features that you would leverage in your model?
b.How would you model the solution and why? Are there any pros/cons of your modeling choice?

5 ."These earphones are a good pick at this price. Connected with laptop for office calls and these are working well although there is no noise cancellation. Quality of wires are a bit thin and look delicate, though neckband is ok. Bass will seem ok if you have not used good quality earphones earlier."
You have been given product review data like the one shown above. You are asked to design a sentiment analysis model for this data. What would be your approach? Describe the different components of your solution. State any assumptions that you are making and pros/cons (if any) of your approach.

6. Compute the BLEU score for the following candidates. Based on this, what can you say about the effectiveness of the BLEU score? Can you suggest ways to make the scoring more effective?

Source: Le professeur est arrive en retard a cause de la circulation
Reference 1: The teacher arrived late because of the traffic
Reference 2: The teacher was delayed due to traffic
Candidate 1: The professor was delayed due to the congestion
Candidate 2: The teacher was held up by the traffic

7. Consider a document *d* containing 200 words wherein the word 'covid' appears 5 times. The document is part of a collection of 100 thousand documents, of which, 10,000 documents contain the word 'covid'. Compute the TF-IDF weight for the word in the document *d.*

8. You are designing a frame-based dialog system for 'cab booking'.

a. What are the different slots in your design? Mention along with their corresponding entity types and questions that the system would ask a user.
b. Show a finite-state dialog manager for the system
c. What changes would you make to the design to change it from a single initiative system to multi-initiative system?