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

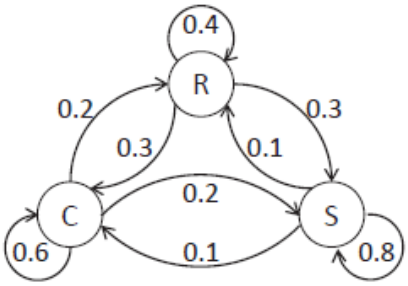
Maximum of all these sequences correspond to T2 T1 T3.

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

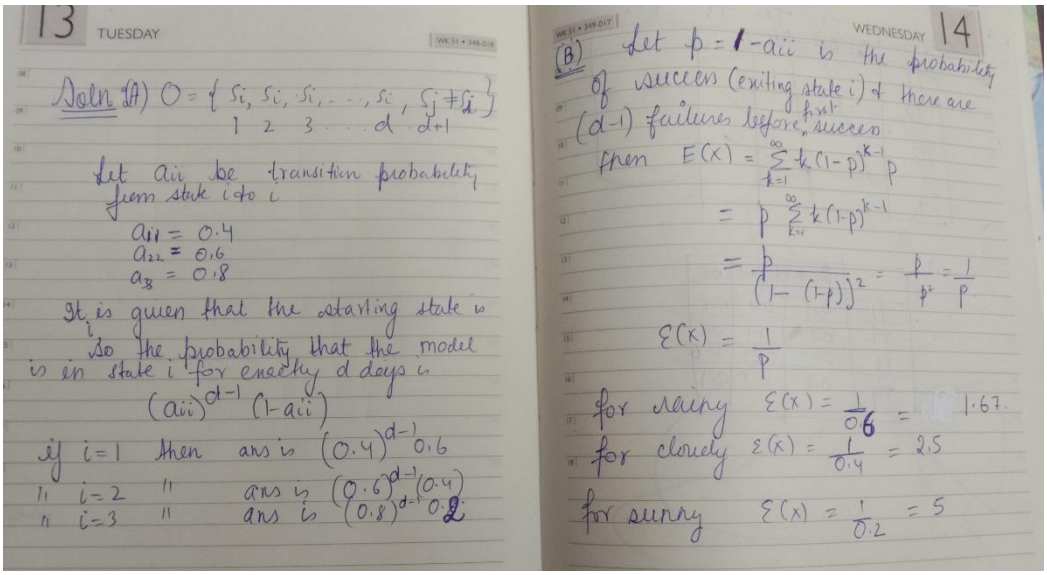
The joint probability for the most probable sequence is 0.0086

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).



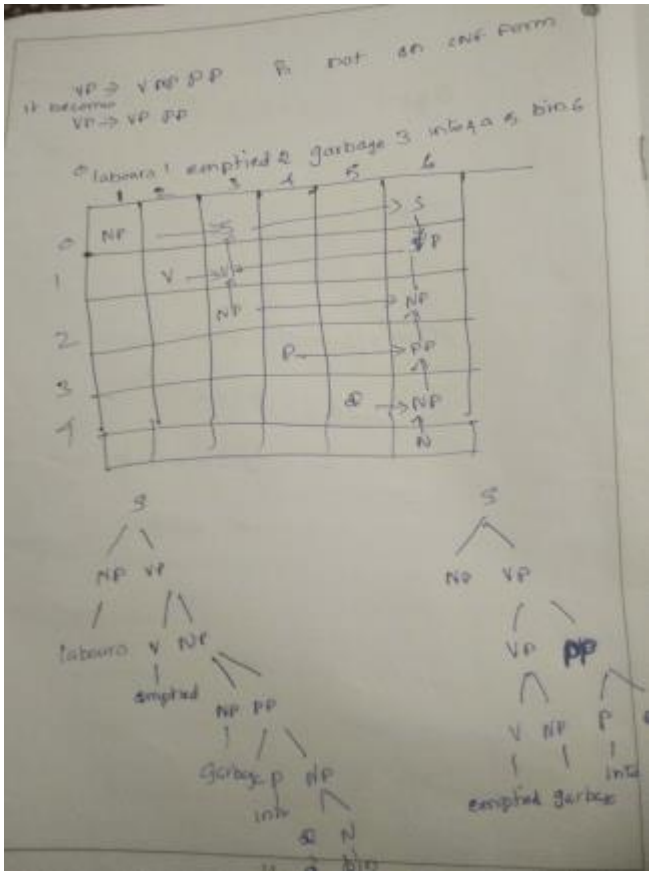
Question 3. [Marks 5+2+5=12 marks]

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"
```

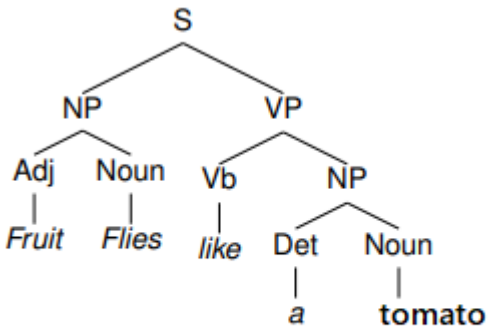


c) Find the probability of the sentence “Fruits flies like a tomato” using PCFG parsing method

[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]**