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

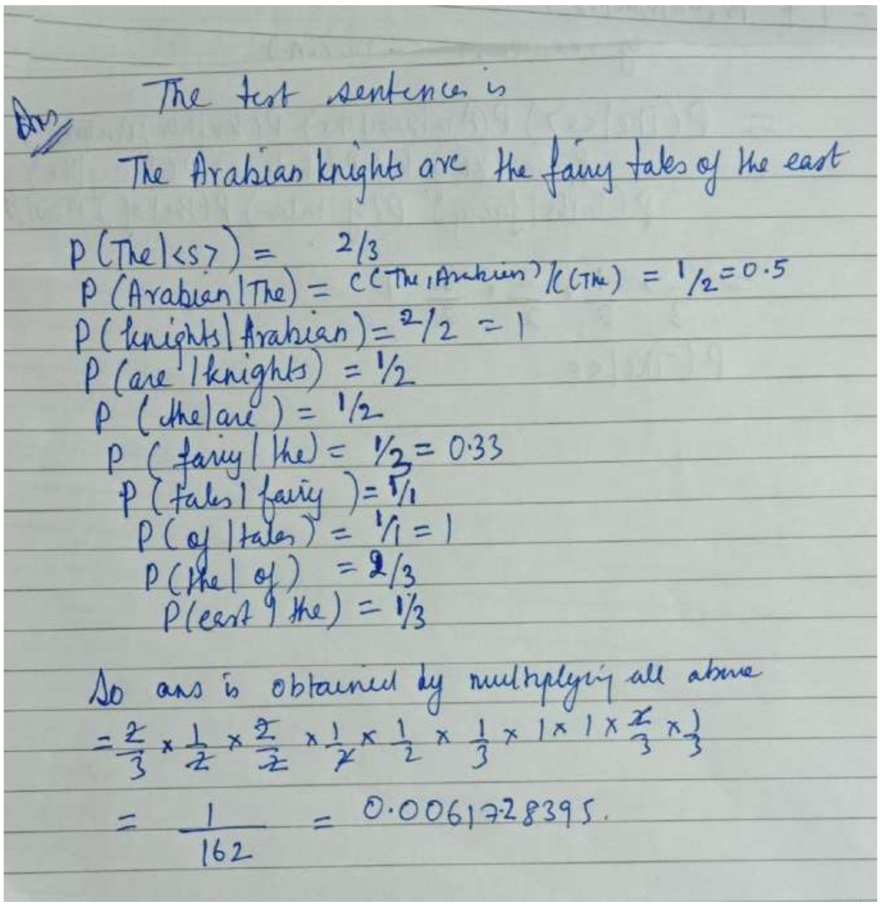
Note: Assumptions made if any, should be stated clearly at the beginning of your answer.

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*			

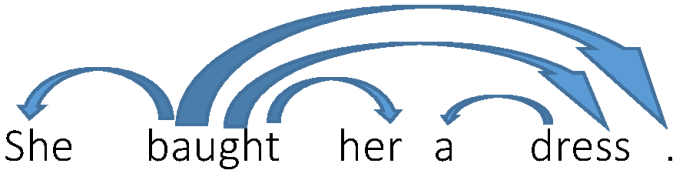
Question 2.

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"/>
```

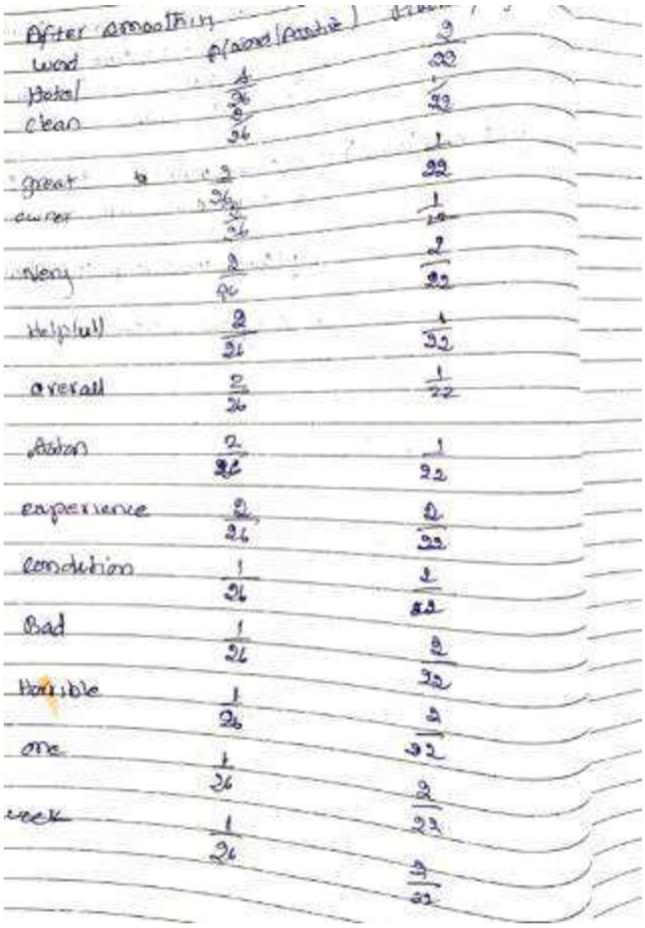
```
    <rdfs:range  rdf:resource="#Animal"/>
  </owl:ObjectProperty>
```

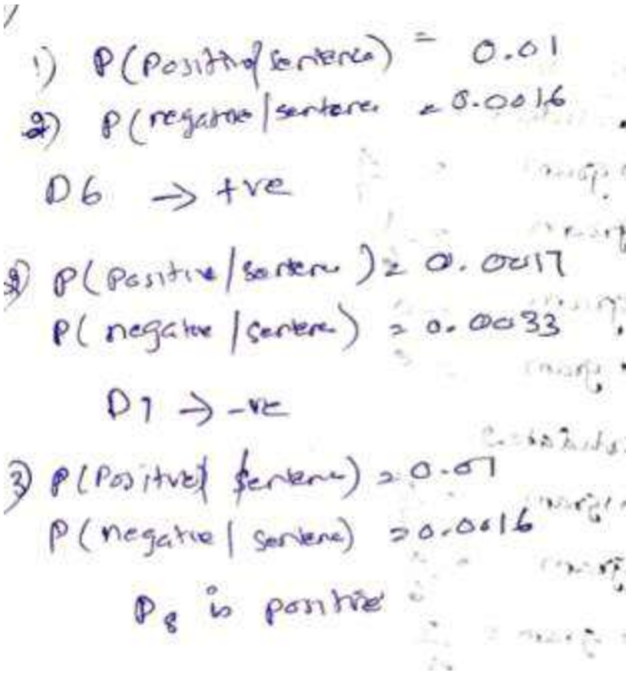
Question 4.

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 :

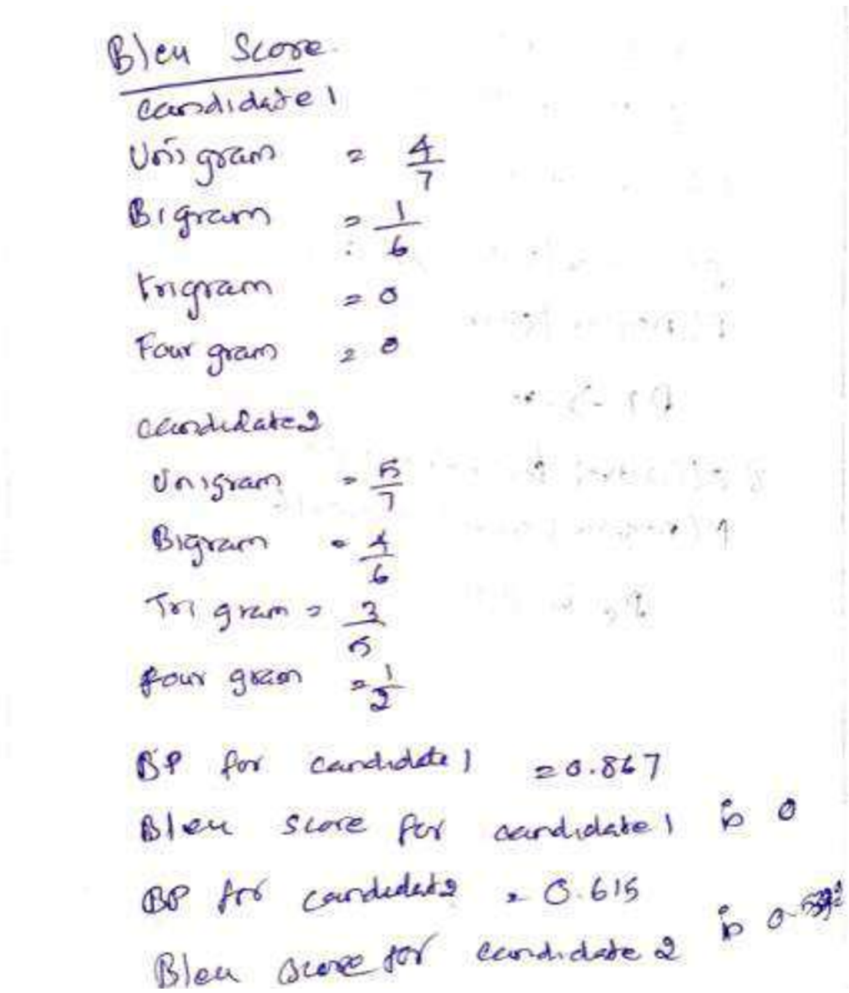




b. Compute the BLEU score for the below translations (candidate1, candidate2).Consider 1gram,2 gram,3 gram,4 gram and Brevity-Penalty for calculating BLUE score .

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

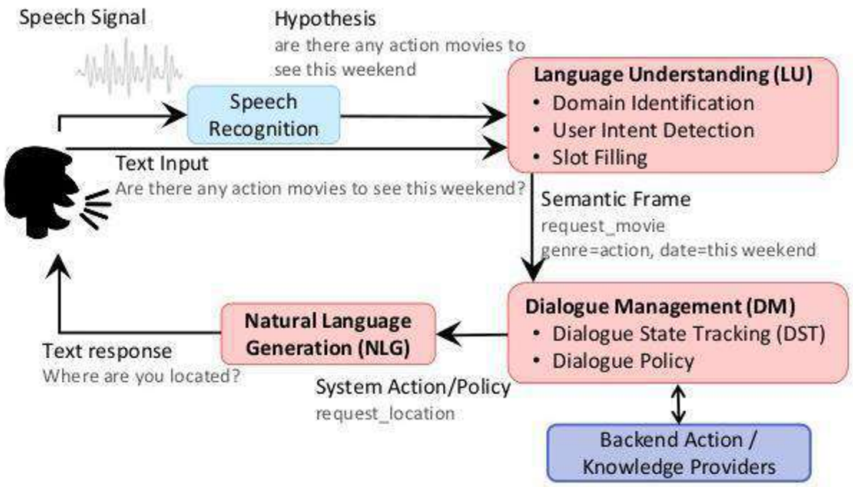


1. Identify the type of the dialogue agent from the following conversation. Also explain the working of that type of dialogue agent. (5 marks)

   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

Term	(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).

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

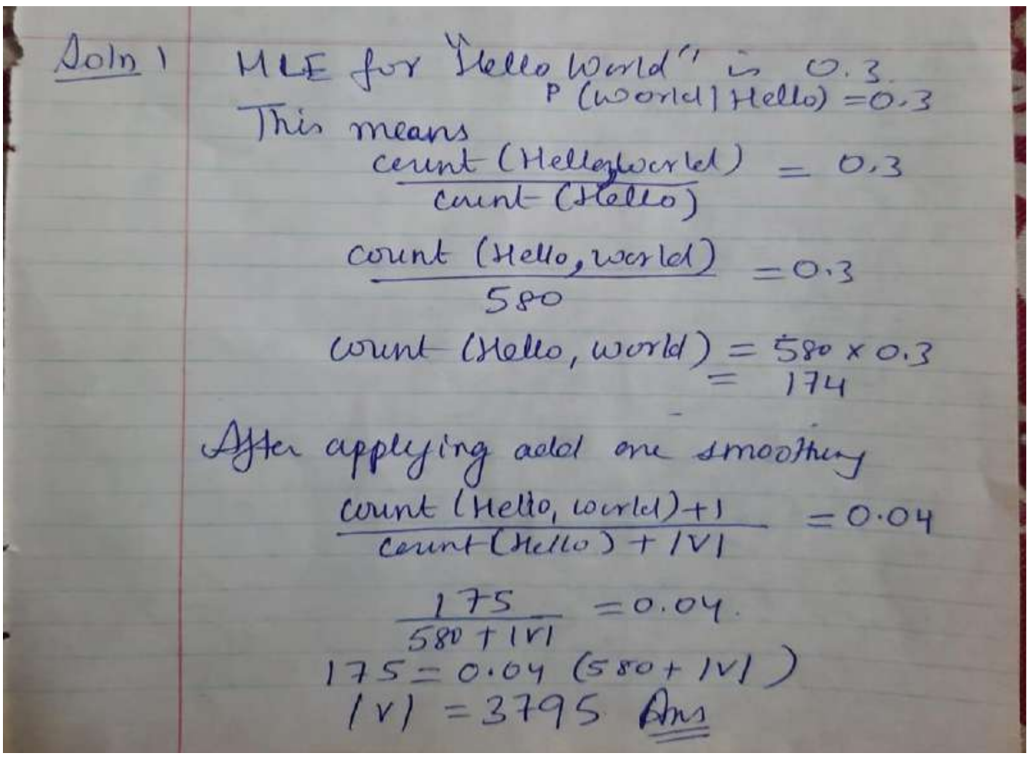
No. of Pages = 3
No. of Questions = 5

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

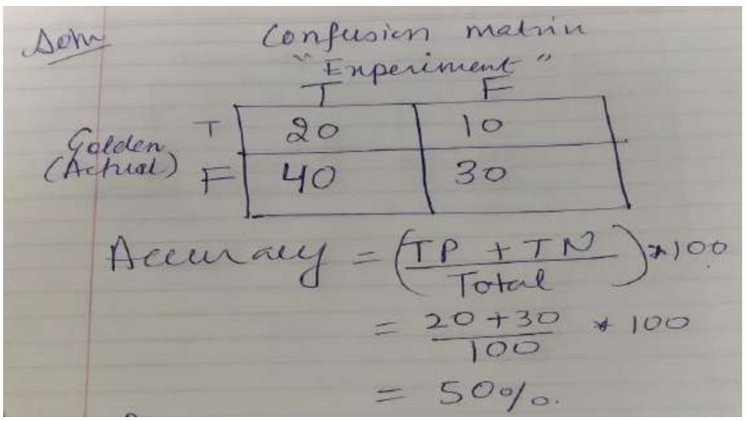
Informal phrases, expressions, idioms, and culture-specific lingo present a number of problems for NLP – especially for models intended for broad use.

- 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)



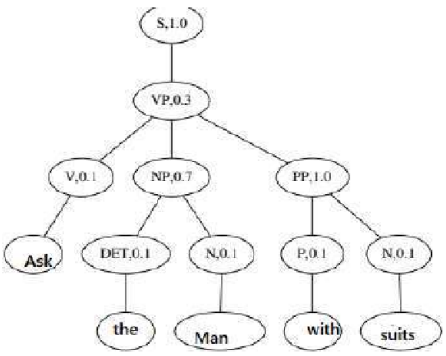
Question 2.

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]

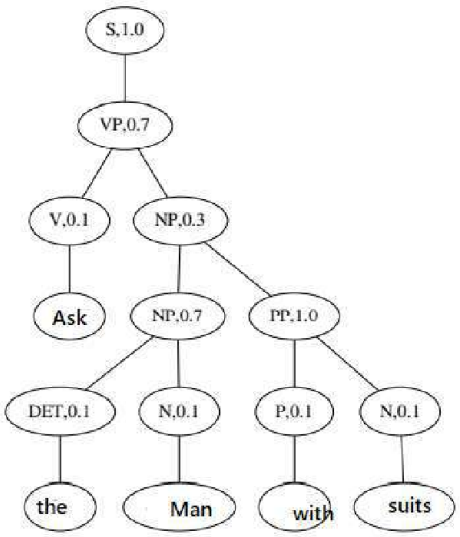
Ask the man with suits

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.

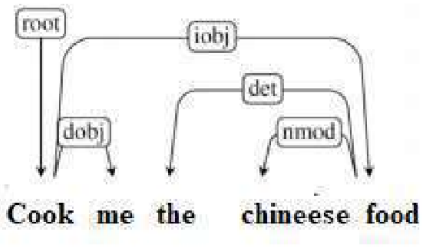
solution

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

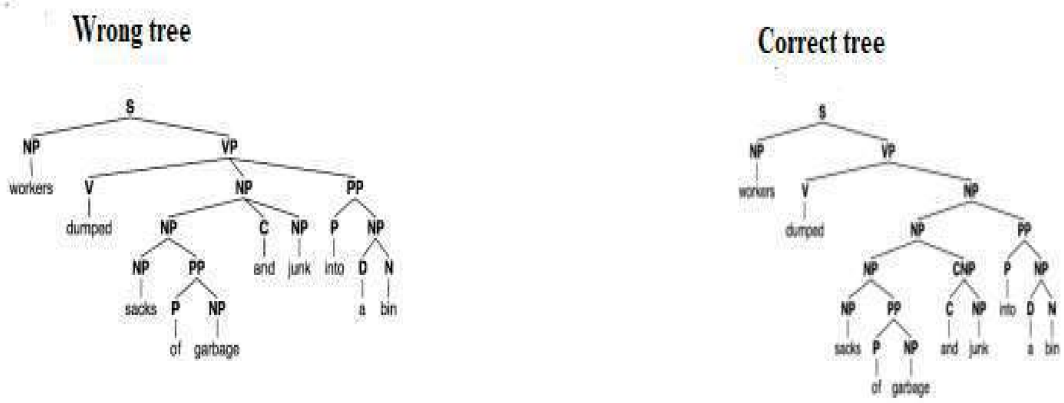
No. of Questions = 4

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,

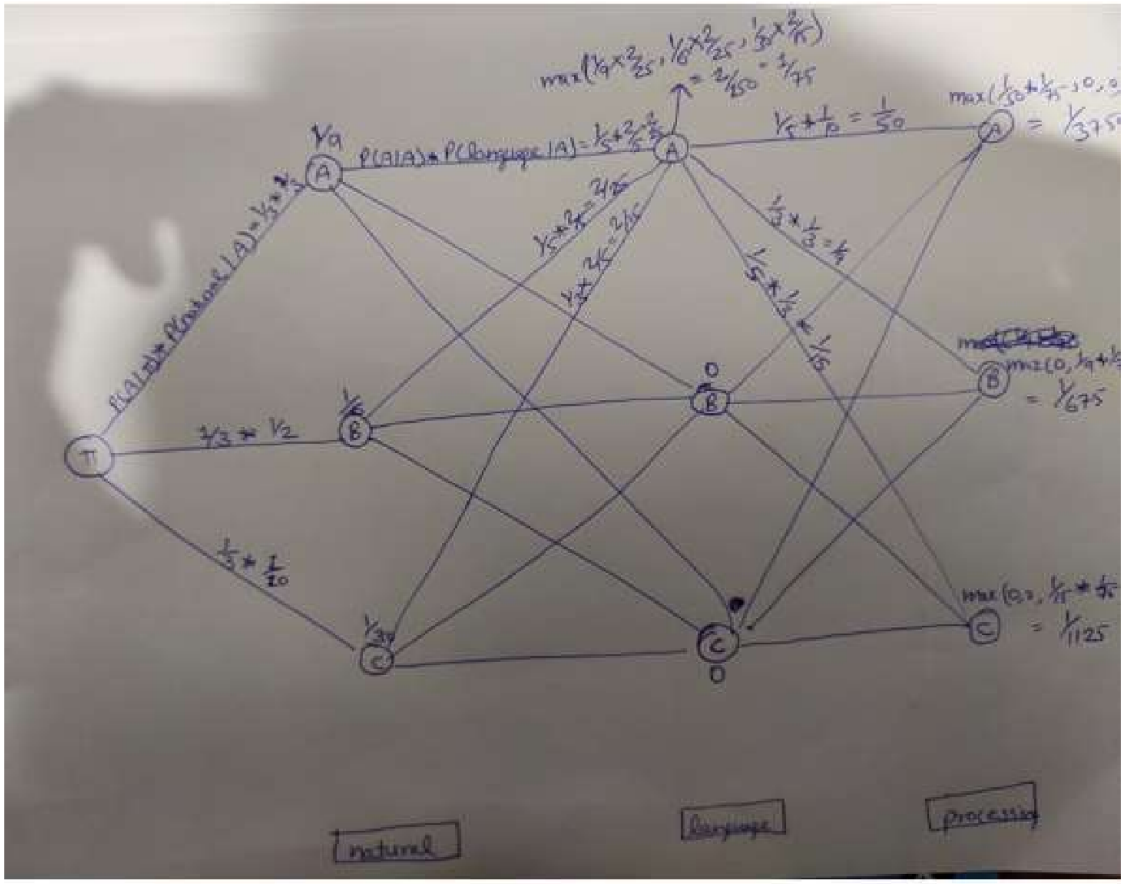
P(processingl|A)=1/10, P(processing|B)=1/3, P(processing|C)=1/3

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]

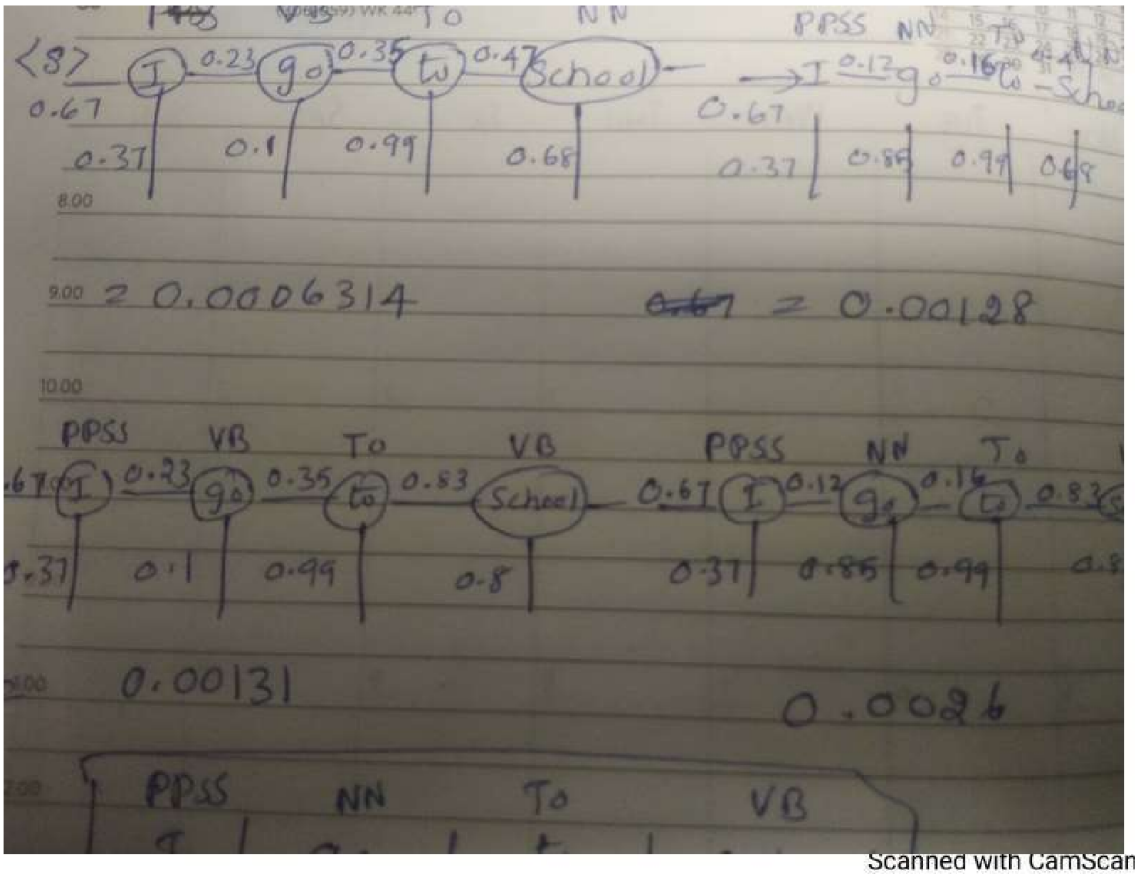
"I go to school"

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:

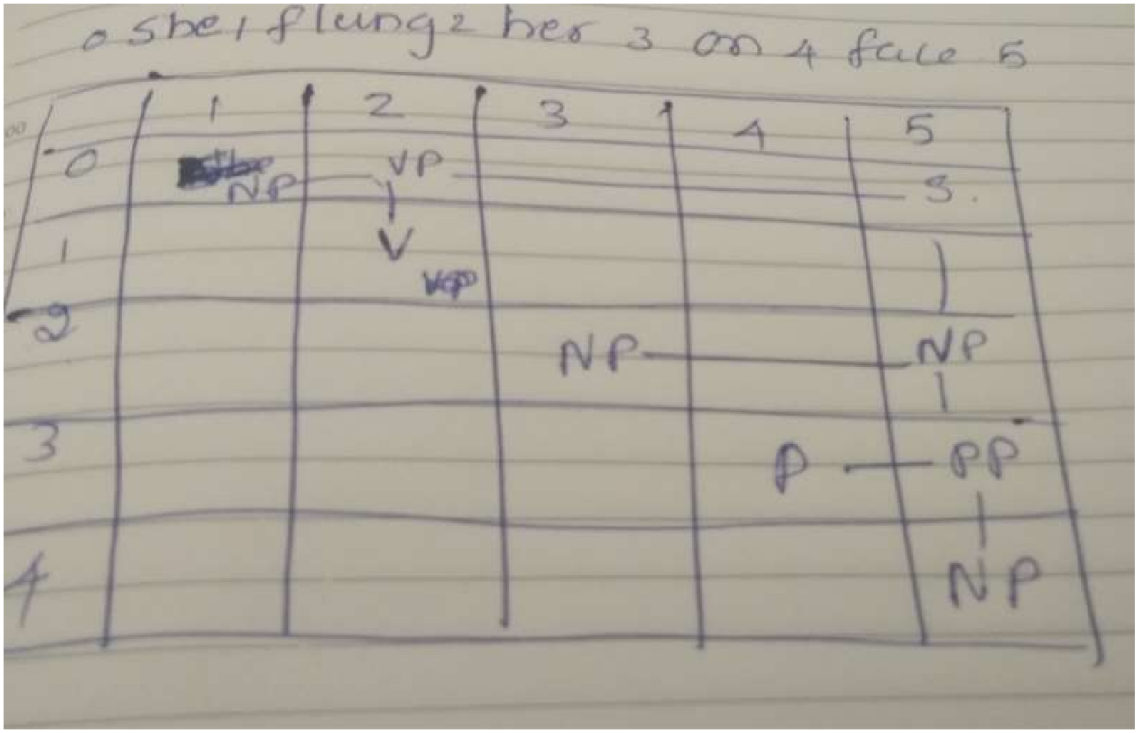
If the chart Is not correct don't give marks

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

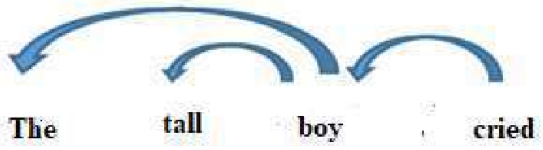
"She flung her on face" given the following grammar and lexicon

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]	[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

****

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

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

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

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

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

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

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

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

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

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

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

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

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

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

① <S> Geeta
P(<S> | Geeta) = 0.4
P(<S> I) 1
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

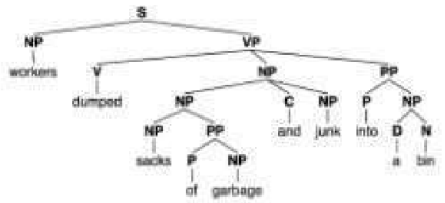
Transition probabilities

	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

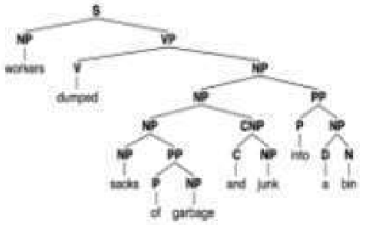
Qtext:-

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]**

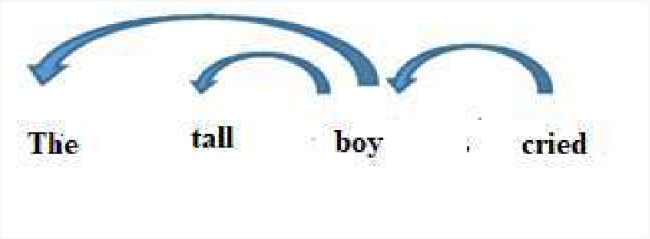
S : The cat caught the rat
S->NP VP VP->VNP NP->Det N
N->rat, N->cat ,Det ->the V->caught

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

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

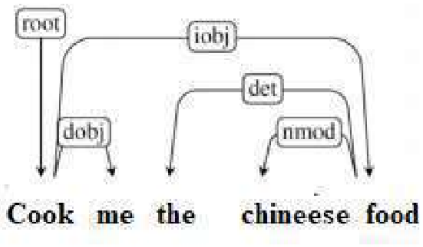
S -> NP VP
VP -> V NP
VP -> VP PP
V - > flung
VP -> flung
NP -> NP PP
NP -> She
NP -> her
Face -> NP
P -> On
PP -> P NP

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?