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)

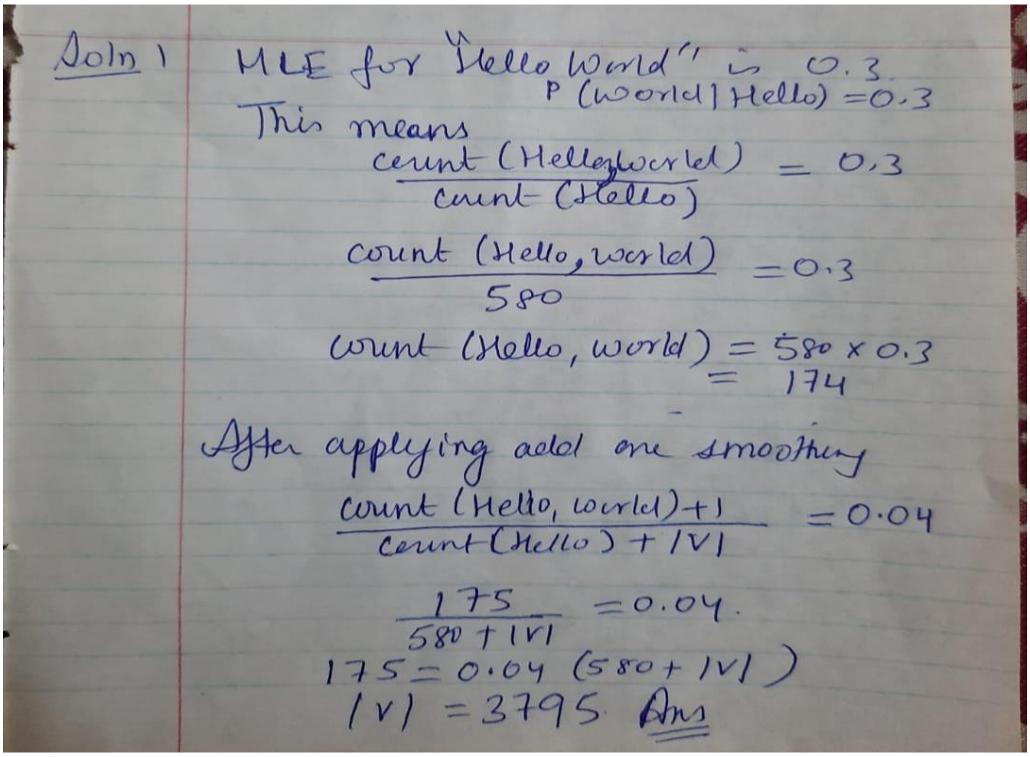
Course No.	: DSECLZG525
Course Title	: Natural Language Processing
Nature of Exam	: Open Book
Weightage	: 50%
Duration	: 2 hours

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)



Soln 1 MLE for "Hello World" is 0.3

$P(\text{world} \mid \text{Hello}) = 0.3$

This means

$$\frac{\text{count}(\text{Helloworld})}{\text{count}(\text{Hello})} = 0.3$$

$$\frac{\text{count}(\text{Hello, world})}{580} = 0.3$$

$$\text{count}(\text{Hello, world}) = 580 \times 0.3 = 174$$

After applying add one smoothing

$$\frac{\text{count}(\text{Hello, world}) + 1}{\text{count}(\text{Hello}) + |V|} = 0.04$$

$$\frac{175}{580 + |V|} = 0.04.$$

$$175 = 0.04\,(580 + |V|)$$

$$|V| = 3795$$ Ans

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

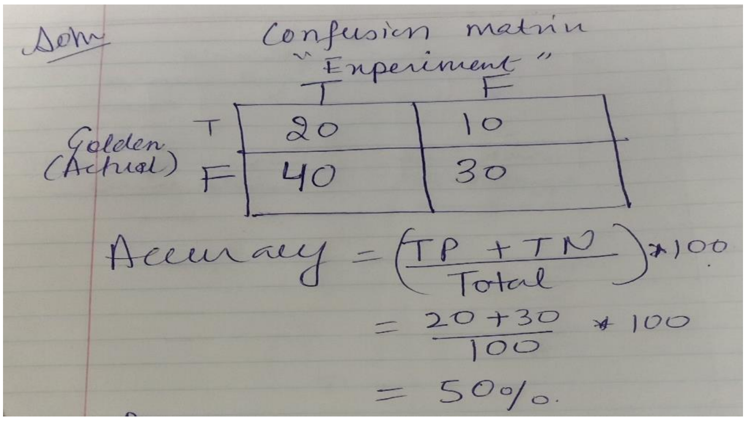
Informal phrases, expressions, idioms, and culture-specific lingo present a number of problems for NLP – especially for models intended for broad use.

- 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)



Soln

Confusion matrix

“Experiment”

Golden (Actual)	T	F
T	20	10
F	40	30

$$Accuracy = \left(\frac{TP + TN}{Total}\right) * 100$$

$$= \frac{20+30}{100} * 100$$

$$= 50\%.$$

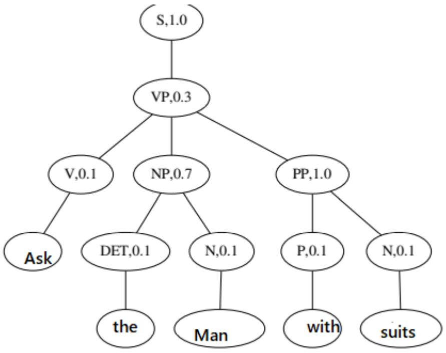
Question 2.

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]

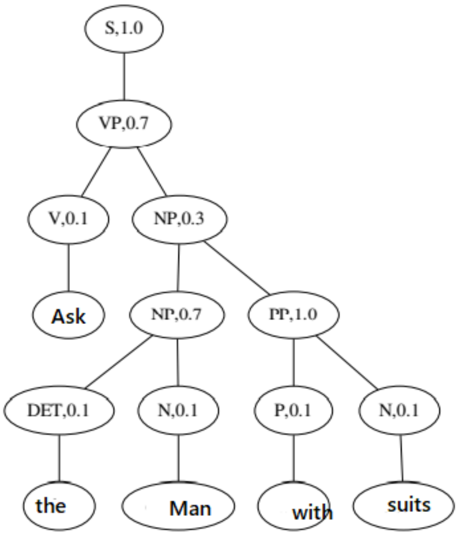
Ask the man with suits

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 × 0.7 × 0.1 ^{5} = 21 × 10−7



Probability= 0.3 × 0.7 × 0.7 × 0.1 ^{5} = 14.7 × 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

o SOURCE-LOCATION: India Gate

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

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