SRN | | | | | | | | | | | |

PES UNIVERSITY

PES University, Bengaluru-85
(Established under Karnataka Act No. 16 of 2013)

UE16CS333

May 2019: B.TECH, VI-SEMESTER
ESA
UE16CS333 - NATURAL LANGUAGE PROCESSING

Time: 3 Hrs | Answer All Questions | Max Marks: 100

		There will not be any part marking unless explicitly stated in the question paper. Each question's part (a) and part (b) has 5 marks each while part (c) has 10 marks.	
1	a)	Using Dynamic Programming approach (show the memoization table), calculate the minimum edit distance when the source string is "stall" and target string is "table" . Given: cost (insert) = cost(delete)= **1** ; cost (substitute) =**2**	05
	b)	You are probabilistically generating a word containing characters from vocabulary **{p,q,r,s,t}**. You always start by generating "**p**" . Then you keep generating any number of characters till you generate "**t**". Once you generate "**t**", you stop. The probability of generating "**t**" is **x**. 1) Draw an FSA with start and stop states that accepts only the above words 2) Suppose the length of the generated string is n. Write an equation for **P(n)** signifying the probability of the generated sequence. For **x=0.3** and **n =4**, what will be the value of P(n) ?	02 03
	c)	**Indicate the correct answer for each of the question(each question carries 2 marks). In your answer script, write the chosen answer option clearly stating the question number:** 1) Considering meronymy and holonymy for sentence : **"The CPU is part of computer"** I. CPU is meronym of computer and computer is holonym of CPU II. CPU is holonym of computer and computer is meronym of CPU 2) "Just because it is there on the table does not mean that I would read The Da Vinci Code before my ESA" I. The above sentence is an example of anaphora II. The above sentence is an example of cataphora 3) Because of the way they work I. Lemmatization is faster than stemming II. Stemming is faster than lemmatization 4) Recursion in natural language is responsible for I. The hierarchical structure of the language II. The lack of upper bound on sentence length 5) Zipf's Law states I. The frequency of a word type in a large corpus is proportional to its rank by frequency II. The frequency of a word type in a large corpus is inversely proportional to its rank by frequency	10
2	a)	Consider an HMM that is spitting out visible symbol sequence : "**Bat hits the ball**". There are only two tags in this system i.e. q and r. The initial state vector is provided: **{ q=1; r=0}** Calculate (i) probability of getting the above sequence (ii) the most probable tag sequence for the same	03 02
	b)	<s> The boy walks </s> <s> The girl runs </s> <s> The boy runs </s> P(boy/<s>, the)= P(girl/<s>,the) = 0.5 P(runs/the, girl) = P(walks/the, boy) = P(runs/the, boy) = 1 P(</s>/boy,walks)= P(</s>/boy,runs)= P(</s>/girl,runs)=0.5	05

	Bat	hits	the	ball
q	0.4	0.3	0.2	0.1
r	0.2	0.4	0.1	0.3

	q	r
q	0.7	0.3
	0.6	0.4

SRN | | | | | | | | | | | | |

		A trigram language model is built on a corpus of the above three sentences and the trigram probabilities as provided above can be assumed. Evaluate this model by calculating the perplexity up to 2 decimal points.	
	c)	**Indicate the correct answer for each of the question(each question carries 2 marks). In your answer script, write the chosen answer option clearly stating the question number:** 1) Some WordNet based semantic relatedness algorithms use the idea of Information Content of a concept based on concept probability. Hence lower the node in the WordNet tree I. Higher is this probability II. Lower is this probability 2) For a unigram language model with uniform prior probability and vocabulary M I. Model is $\frac{1}{M}$ way uncertain and perplexity is M II. Model is M way uncertain and perplexity $2^{\log M}$ 3) In a noisy channel, what user types at transmission end is comprehended differently at the receiving end. If P(Y) is the model for true language model at transmission end and P(X) is what we experience at the receiving end. I. P(Y/X) models the noise of the noisy channel and P(X/Y) is used to estimate Y II. P(X/Y) models the noise of the noisy channel and P(Y/X) is used to estimate Y 4) In a sentiment classification exercise based on text reviews (corpus with vocabulary size of 10^4), n-grams are the features. If you use 4-gram instead of trigram, how many more features you will have ? I. 10^4 II. 10^{12} 5) Wu-Palmer similarity and Lin similarity are both measures of lexicon based semantic similarity. I. Both use **information content(IC) of a concept** but do not use **lowest superordinate (lso)** II. Both use **lowest superordinate (lso)** but do not use **information content (IC) of a concept**	10
3	a)	**Document 1** : I don't like probability or mathematics **Document 2** : I hate doing my mathematics **Document 3** : I love food and I hate dog Three single sentence documents are provided. First do the stop words removal followed by stemming. The stop word list has terms such as **'I, or, do, don't, my, and'**. Then, find the TF IDF representation of each document	05
	b)	You are trying to use MaxEnt to tag the sentence **"The green contour"**. The context for MaxEnt feature function is defined as **{ w_i, w_{i-1}, w_{i+1}, t_{i-1} }**. Each feature function has a uniform weight of 1.0. $f_1 \rightarrow t_{i-1}$=det and t_i=adj $f_2 \rightarrow t_{i-1}$=noun and t_i=verb $f_3 \rightarrow t_{i-1}$=adj and t_i=noun $f_4 \rightarrow w_{i-1}$=the and t_i=adj $f_5 \rightarrow w_{i-1}$=the, w_{i+1}=contour and t_i=adj $f_6 \rightarrow w_{i-1}$=green and t_i=noun $f_7 \rightarrow w_{i-1}$=green and t_i=det $f_8 \rightarrow w_{i-1}$=null and t_i=noun The possible POS tags for the words are : the → det, noun green → verb, adj contour → verb, noun (i) How many tag sequences are possible ? (ii) Which are the two most likely tag sequences (use beam size 2) for the **first two words** in the sentence? Assume e = 2.72	01 04
	c)	**Indicate the correct answer for each of the question (each question carries 2 marks). In your answer script, write the chosen answer option clearly stating the question number :** 1) On average, if any sentence in a test language model can be coded in 100 bits, the model perplexity is : I. 2^{100} per sentence II. 2^{-100} per sentence	10

SRN | | | | | | | | | | | |

		2) If we compare the probability function representing tag sequence given sequence of observation in MEMM and Conditional Random Field (CRF), the summations in the numerator and the normalization in the denominator are different. I. Uncommon between CRF and MEMM is sum over all feature functions in denominator II. Uncommon between CRF and MEMM is sum over all possible label sequences in denominator 3) Latent Semantic Induction(LSI), PMI (Pointwise Mutual Information), TF IDF and Neural Embedding are four methods of vectorization of text. Out of these I. LSI and TF IDF are sparse vectors whereas PMI and Neural Embedding are dense vectors II. PMI and TF IDF are sparse vectors whereas LSI and Neural Embedding are dense vectors 4) (a) "no consideration of the context" and (b) "no consideration of global context" are two common weaknesses we encounter with respect to TF IDF, PPMI and Word2Vec. I. (a) is a weakness of TF IDF and (b) is a weakness of PPMI and Word2Vec II. (a) is a weakness of TF IDF and PPMI and (b) is a weakness of Word2Vec 5) Shallow auto-encoder and matrix factorization of global co-occurrence matrix are two main ways to extract low dimensional embedding for words in text corpus. I. Glove and Fasttext use matrix factorization approach whereas Word2Vec uses shallow autoencoder II. Word2Vec and Fasttext use Shallow autoencoder whereas Glove uses matrix factorization	
4	a)	You are using CNN for text feature extraction in sentiment analysis task to predict positive or negative sentiment. You have created a text embedding layer. Your embedding vector length is 6 and let us assume that your text data is 15 words. You are using convolution filters or kernels of 5,4,3 and 2 words respectively and you have three filters of each type. You are then doing max pooling from each feature map generated by convolution and finally concatenating these max-pooling outputs into a single vector. On this vector, softmax with binary cross entropy error function is applied. I. What are the sizes of the four sets of kernel matrices ? II. What are the sizes of the four sets of feature maps after convolution ? III. What is the size of feature vector after concatenating the output of max pooling ? IV. What is the output vector size after softmax ?	1 2 1 1
	b)	You are doing multi class classification with text data and comparing the number of weights to train for feedforward network, simple RNN and LSTM based RNN. The following dimensions(#nodes) of the input, hidden and output vectors are provided to you : Input = 100 ; Hidden = 20 ; output = 2 . You can assume only one hidden layer for this assessment. Find the number of weights to train for I. Simple feedforward network II. Simple RNN III. LSTM based RNN	01 02 02
	c)	**Indicate the correct answer for each of the question(each question carries 2 marks). In your answer script, write the chosen answer option clearly stating the question number:** 1. **Q :** "Do you like Scrambled Egg ?" **A :** "Do you like Poached Egg ? " In the Information Retrieval Based Chatbot architectures, this Question Answer belongs to the following category : I. Response to most similar turn II. Similar turn 2. Changing and obtaining details about your flight reservation in an airline company is mostly handled by a chatbot which can provide you details about your reservation, update the reservation and can handle almost all passenger queries. This is an example of a chatbot that is I. Open domain, corpus based II. Closed domain, Information based 3. You are using cross entropy loss function in your multiclass text classification problem. Here total loss is I. Sum of logarithmic loss for target label of each class II. Logarithm of sum of loss for target label of each class	10

SRN | | | | | | | | | | | | |

4. For all neural networks (feedforward, RNN, bi-directional RNN, LSTM), the computational complexity is :
   I. O(no of edges)
   II. O(no of input and output)
5. In simple sequence2sequence model for Neural Machine Translation, the context vector is built out of encoder's last hidden state. In attention based variant of the model , the context vector is:
   I. Sum of hidden states of the encoder
   II. Weighted sum of hidden states

5 a)

A → a	S→ AB
B → b	S → BC
C → a	A → BA
	B → CC
	C → AB

The grammar and lexicon are provided.

(i) You need to parse the sentence **"b a a b a"** using CYK algorithm clearly showing all entries in the chart. 4

(ii) Draw the parse tree found. 1

b)

Terminals : with, saw, Peter, ears, stars, telescope

Non terminals : S, PP, P, NP. VP, V

Starting symbol for parse tree : S

S → NP VP 1.0	PP → P NP 1.0
VP → V NP 0.7	VP → VP PP 0.3
NP → NP PP 0.4	P→ with 1.0
V → saw 1.0	NP → ears 0.18
NP →saw 0.04	NP →stars 0.18
NP → Peter	0.1
NP → telescope	0.1

The grammar and lexicon are provided with corresponding probabilities. For the sentence **"Peter saw stars with ears",**

Intuitively (no need to use CYK etc.) find two possible parse trees and then answer the followings:

(a) What is the probability of the more likely parse. 3

(b) What is the probability of the sentence. 2

c)

Indicate the correct answer for each of the question. In your answer script, write the chosen answer option clearly stating the question number:

1. Machine learning models used in NLP can be both discriminative and generative.
   I. N-gram, HMM, PCFG, Naïve Bayes are generative whereas MaxEnt, MEMM , CRF are discriminative
   II. N-gram, HMM, PCFG are generative whereas MaxEnt, MEMM, CRF, Naïve Bayes are discriminative
2. In Lappin and Leass algorithm for coreference resolution, only ways to handle recency :
   I. One salience factor is for recency and overall salience value of an entity is reduced by half if a sentence boundary is crossed
   II. Overall salience value of an entity is reduced by half if a sentence boundary is crossed
3. HMM and PCFG are both generative models and are close parallels. A PCFG figuring out the parse tree for a sentence is closely similar to
   I. How likely is a certain observation given the model in HMM
   II. Choosing a state sequence that best supports the observation in HMM
4. A well-formed dependency tree is supposed to be projective. In that case, if word A depends on word B, then all words between A and B
   I. Are subordinate to A
   II. Are subordinate to B
5. The generative Hidden Markov Model (HMM) and discriminative Maximum Entropy Markov Model (MEMM) both address the same problem of predicting a hidden state given the observation. For N states and M unique observation, unlike HMM, MEMM has only one probability matrix of the dimension
   I. (N.M) . M
   II. (N.M) . N

10