BACHELOR OF TECHNOLOGY (CBCS) (2021-COURSE)
B. Tech. Sem - VI Computer Science & Engineering : WINTER : 2024
SUBJECT: NATURAL LANGUAGE PROCESSING

Day : Saturday
Date : 07/12/2024

W-25597-2024

Time : 10:00 AM-01:00 PM
Max. Marks : 60

N.B :
1) All questions are **COMPULSORY.**
2) Figures to the right indicate **FULL** marks.
3) Draw neat and labeled diagrams **WHEREVER** necessary.
4) Assume suitable data if necessary.

Q.1 What will be the next probable word predicted by the model for the following word sequence; (10)

Training corpus:
1. <s> I am Henry</s>
2. <s>I like College</s>
3.<s>Do Henry like college</s>
4.<s>Henry I am</s>
5.<s>Do I like Henry</s>
6.<s>Do I like College</s>
7.<s> I do like Henery</s>

Test Data:
1. Calculate perplexity of given sentence using Bigram model
<s>I like College</s>
2. Calculate perplexity of given sentence using Trigram model
<s>I like College</s>

OR

Q.1 How zero probability problem can be resolve on N-gram model explain with suitable example (10)

Q.2 Define Concept of vector space model. Also solve below example to fine cosine similarity of documents. (10)
D1= (0.5, 0.8, 0.3, 0.6)
D2= (0.9, 0.4, 0.2, 0.7)
Q=(1.5,1.0,0,0)

OR

Q.2 Explain the difference between tokens and terms in natural language processing. Discuss the process of extracting meaningful terms from tokenized text data. Describe techniques such as term frequency-inverse document frequency (TF-IDF) and named entity recognition (NER) for identifying important terms in text corpora. (10)

Q.3 Describe the challenges and techniques involved in automatic part-of-speech tagging. Evaluate the performance of part-of-speech tagging systems using evaluation metrics such as accuracy and precision. Describe the concept of maximum Entropy model for POS with suitable example.

P.T.O.

OR

Q.3 How correct POS tag is identified for the word using HMM Viterbi Algorithm. (10)
Consider the below example.
Emission Probability Matrix:

	DT	NN	VB
That	0.40	0.0	0.00
girl	0.0	0.15	0.0031
smiles	0.0	0.0004	0.20

State Transition Matrix:			
	DT	NN	VB
<S>	0.50	0.40	0.1
DT	0.1	0.99	0.00
NN	0.30	0.30	0.40
VB	0.40	0.40	0.20

Test Data: That girl smiles			
	DT	NN	VB
That	0.40	0.0	0.00
girl	0.0	0.15	0.0031
smiles	0.0	0.0004	0.20

Q.4 Define the concept of Parsing. Also, explain the concept of Shift Reduce Parser. (10)
Solve the below examples using shift-reduce parser.
1. Grammar:
S→S+S
S→S-S
S→(S)
S→a

Input String
$a_1-(a_2+a_3)$

2. Grammar:
E→2E2
E→3E3
E→4

Input String
32423

OR

Q.4 Define context-free grammars (CFGs) and explain their significance in natural language processing. Discuss how context-free rules are utilized to generate syntactically valid sentences and parse trees in CFGs. Provide examples to illustrate the construction of context-free rules and derivation trees for simple linguistic structures. (10)

Q.5 Explain the concept of matrix factorization and its application in natural language processing. Describe Singular Value Decomposition (SVD) as a technique for decomposing a matrix into its constituent parts. Discuss how SVD is used for dimensionality reduction and latent semantic analysis in text data. (10)

OR

Q.5 Apply the concept of singular value decomposition on below matrix (10)

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

Q.6 How linguistic data is manage with the help of GATE, Explain it in detail. (10)

OR

Q.6 Discuss the use of XML (extensible Mark-up Language) in managing linguistic resources and data. Explain how XML facilitates the representation, storage, and interchange of linguistic annotations, corpora, and lexicons. Analyse the advantages of using XML for organizing linguistic data in a structured and hierarchical format. (10)

* * * * *