Total No. of Questions : 8]

SEAT No. : ____

PE-2184

[Total No. of Pages : 3

[6584]-83

B.E. (Computer Engineering)

NATURAL LANGUAGE PROCESSING

(2019 Pattern) (Semester - VIII) (410252 A) (Elective-V)

Time : 2½ Hours] ***[Max. Marks : 70***

Instructions to the candidates :

1) ***Answer Q1 or Q2, Q3 or Q4, Q5 or Q6, Q7 or Q8.***
2) ***Figures to the right indicate full marks.***
3) ***Neat diagrams must be drawn whenever necessary.***
4) ***Make suitable assumptions whenever necessary.***

Q1) a) Given the following sentence corpus: **[9]**

S1: “language models learn patterns”

S2: “models learn from data”

S3: “data helps improve language models”

1) Construct a bigram model from the above sentences.

2) Compute the smoothed bigram probability of the sentence “language models learn from data” using Add-1 smoothing.

3) Show all intermediate steps including vocabulary size, unigram and bigram counts.

b) Explain the concept of probabilistic language modeling using Markov assumptions. Construct a trigram language model using a given text corpus and compute the probability of a given sentence using Maximum Likelihood Estimation (MLE). Illustrate your answer with appropriate equations and examples. **[9]**

OR

P.T.O.

Q2) a) Explain how Non-negative Matrix Factorization (NMF) differs from Latent Dirichlet Allocation (LDA) in topic modeling. Use a document-term matrix to show how NMF factorizes the matrix and interprets the result. **[9]**

b) Consider the following three documents: **[9]**

- Dl: “neural networks are powerful”
- D2: deep learning powers neural models”
- D3: “networks and models are important”

1) Compute the TF-IDF score of the term “neural” in all three documents.

2) Show how document frequency (DF) and inverse document frequency (IDF) are computed.

3) Interpret the results to determine in which document “neural” is more significant?

Q3) a) Explain the Vector Space Model in the context of Information Retrieval. How are documents and queries represented and how is relevance computed between them? **[9]**

b) Compare and contrast Entity Extraction, Relation Extraction and Coreference Resolution. How do these tasks contribute to building knowledge from unstructured text? Illustrate with examples. **[8]**

OR

Q4) a) Describe the Named Entity Recognition (NER) system building process. Given a sample sentence, outline the steps you would take to build an NER system using a supervised learning approach. **[9]**

b) What is Cross-Lingual Information Retrieval (CLIR)? Discuss the challenges involved and evaluate the different approaches to implementing CLIR systems. **[8]**

Q5) a) Describe Walker's algorithm for word sense disambiguation. How does it differ from other disambiguation techniques like Lesk's Algorithm and what are the scenarios where it can be most effective? **[9]**

b) List the tools available for the development of NLP applications. Write about the significant features of NLTK and Gensim libraries. **[9]**

OR

Q6) a) Compare the IndoWordNet with the traditional WordNet. What are the key advantages of IndoWordNet, especially in the context of Indian languages? **[9]**

b) Which types of tasks are performed by the Gensim library? Give an example. **[9]**

Q7) a) Compare the architectures of Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT). How does each approach handle syntactic and semantic ambiguities? **[9]**

b) Explain Natural Language Generation with reference architecture. **[8]**

OR

Q8) a) Explain the architecture and challenges of building a Question Answering (QA) system. **[9]**

b) Discuss the key components of a conversational agent, such as chatbots or virtual assistants. How do natural language generation and understanding play a role in creating effective conversational agents? **[8]**