3 7 0

B.Tech. DEGREE EXAMINATION, NOVEMBER 2018
3rd to 7th Semester

15IT330E – TEXT MINING
(For the candidates admitted during the academic year 2015-2016 to 2017-2018)

Note:
(i) **Part - A** should be answered in OMR sheet within first 45 minutes and OMR sheet should be handed over to hall invigilator at the end of 45th minute
(ii) **Part - B** and **Part - C** should be answered in answer booklet.

Time: Three Hours

Max. Marks: 100

PART – A (20 × 1 = 20 Marks)
Answer ALL Questions

1. The two major components of NLP are
   (A) Natural language understanding Natural language generation
   (B) Natural language unit Natural language generation
   (C) Natural language user Natural language generation
   (D) Natural language understanding Natural language analysis

2. _____ is an example of morpheme formed by cliticization.
   (A) will
   (B) 'll
   (C) caught
   (D) ate

3. Which character is used for specify the end of sentence in RE?
   (A) ^
   (B) ?
   (C) *
   (D) $

4. In which phase N-gram tiling is used in question answering system?
   (A) Question processing
   (B) Answer processing
   (C) Query formulation
   (D) Passage retrieval

5. _____ is a function word.
   (A) Work
   (B) Noun
   (C) Of
   (D) Delhi

6. CRF is base on _____.
   (A) Vector model
   (B) Statistical modeling
   (C) Scalar model
   (D) Deterministic model

7. The small tagset is used in text mining is _____.
   (A) Brown corpus
   (B) C5 tagest
   (C) Pen Treebank tagset
   (D) Switchboard corpus

8. SV, SVO and SVOU specify the _____ pattern for the verb.
   (A) Generalized
   (B) Subcategorization
   (C) Prioir probability
   (D) Likelihood

9. _____ is the fraction of retrieved documents that are relevant
(A) Sensitivity
(B) Precision
(C) Recall
(D) Fscore

10. Term document matrix is used for _____.
(A) Finding similarity between documents
(B) Finding pos in document
(C) Finding TF-IDF
(D) Finding summarization of documents

11. Principal component analysis works on _____ data.
(A) Quantitative multivariate
(B) Linear data
(C) Qualitative data
(D) Digital data

12. LSA is limited to _____ problem.
(A) Sematic
(B) Synonymy
(C) Syntax
(D) Allocation

13. _____ technique is used for finding the topic of the document.
(A) Principal component analysis
(B) Linear discriminant analysis
(C) Latent dirichlet allocation
(D) Generalized discriminant analysis

14. _____ algorithm merges and splits nodes to help modify nonoptimal partitions.
(A) Agglomerative clustering
(B) Expectation maximization
(C) Conceptual clustering
(D) k-mean clustering

15. EM algorithm finds _____.
(A) Probability
(B) Maximum likelihood
(C) TF
(D) IDF

16. Document classification is based on _____.
(A) Machine learning
(B) Deterministic approach
(C) Probabilistic approach
(D) Statistical approach

17. Which of the following will be Euclidean distance between two points A(1,3) and B(2,3)
(A) 1
(B) 2
(C) 4
(D) 8

18. _____ clustering technique stats with all records in one cluster and then try to split that into small pieves.
(A) Agglomerative
(B) Divisive
(C) Partition
(D) Numeric

19. The threshold function is replaced by continuous functions called _____ functions.
(A) Activation
(B) Deactivation
(C) Dynamic
(D) Standard

20. With Bayes theory the probability of hypothesis H-specified by P(H) is referred to
(A) A priori probability
(B) A conditional probability
(C) A posterior probability
(D) Bidirectional probability

PART – B (5 × 4 = 20 Marks)
Answer ANY FIVE Questions

21. Define ambiguity in natural language processing. Example lexical ambiguity with example.

22. What are the two types of kleen operator in regular expression? Explain with example.

23. In the context of mathematically modeling a system, how would define the term hypothesis? Give an example.

24. Write short note on feature identification and feature selection.

25. Consider the following data; that consist in the following three document
    Reference document 1: "Some tigers liver in the zoo"
    Reference document 2: "Green is a color"
    Reference document 3: "Go to New York city"
    The new document contains the word "orange in color".
    Apply document classification algorithm for classify the new document.

26. What are the different types of clustering?

27. Write short note on neural network.

PART – C (5 × 12 = 60 Marks)
Answer ALL Questions

28. a. Using dynamic programming find the minimum edit distance for two strings, also *backtrack* the algorithm to explain individual operation
    String 1: piece (target)
    String 2: peace (source)

(OR)

b. Explain the process of question answering system in detail with the challenges.

29. a.i. Draw the graphical model for CRF based text modeling, explain with example. (8 Marks)

ii. What are the advantages of CRF over HMM? (4 Marks)

(OR)

b. Consider the following sentences
    (i) Rama killer Ravana
    (ii) Ravana was the king of Srilanka
    (iii) Srilanka is a beautiful place
    (iv) Rama crosses the sea to reach Island
    (1) Pos tag the sentences with states {noun, verb, adverb, adjectives, propositions, adverb, conjunctions, det, pronoun, unknown}
    (2) Build A,B matrices of HMM with token as observations and pos tags as states.

30. a.i. How to measure the effectiveness of information retrieval process? (4 Marks)

ii. Explain precision and recall measure with example.

(OR)

b. Explain in detail, how to perform plagiarism detection using latent semantic analysis.

31. a. Explain in detail about expectation-maximization algorithm with example.

(OR)

b. Consider the following set of sentences

"I eat fish and vegetables"
"fish are pets"
"My kitten eats fish"

Apply LDA techniques to discover the topics in the given document. Topics per document and words per topic are found.

32. a. Explain in detail about support vector machine classification algorithm with example.

(OR)

b. Consider the given data points

A_1 (2,10)	A_2 (2,5)	A_3 (8,4)
B_1 (5,8)	B_2 (7,5)	B_3 (6,4)
C_1 (1,2)	C_2 (4,9)	

The distance function is Euclidean distance initially assign A_1, B_1 and C_1 as center as each cluster respectively. (ie k =3). Use the k-means algorithm to find the final three clusters.

* * * * *