Reg. No															

B.Tech DEGREE EXAMINATION, NOVEMBER 2023

Fifth Semester

18AIE329T - INFORMATION RETRIEVAL

(For the candidates admitted during the academic year 2020 - 2021 & 2021 - 2022)

Note:

i. **Part - A** should be answered in OMR sheet within first 40 minutes and OMR sheet should be handed over to hall invigilator at the end of 40th minute.

ii. **Part - B** and **Part - C** should be answered in answer booklet.

Time: 3 Hours **Max. Marks: 100**

PART - A (20 × 1 = 20 Marks)

Answer **all** Questions

		Marks	BL	CO
1.	The process of removing most common words (and, or, the, etc.) by an information retrieval system before indexing is known as (A) Lemmatization (B) Stop word removal (C) Inverted indexing (D) Normalization	1	1	1
2.	A metric used measure the importance of a term in a text document collection is called (A) Inverse Document Frequency (B) Term Frequency (C) Inverse Term Frequency (D) Document Frequency	1	2	1
3.	Which of the following is the local method for improving recall of an information retrieval system? (A) Query expansion (B) Relevance feedback (C) Ontology based model (D) None of the above	1	2	1
4.	______ is affected by the number of false positive errors. (A) Precision (B) Recall (C) Both precision and recall (D) Neither precision nor recall	1	4	1
5.	Steps in Indexing are performed in following order : (A) Stop word Elimination, Tokenization, Stemming (B) Tokenization, Stemming, Stop word Elimination (C) Tokenization, Stop word Elimination, Stemming (D) Stemming, Tokenization Stop word Elimination	1	1	2
6.	The vocabulary size (unique words) of a text can be estimated using (A) Zipf's law (B) Scientific law (C) Heaps' law (D) Inverted index rule	1	4	2
7.	What is the disadvantage of Boolean retrieval model? (A) Easy to implement (B) Difficult to rank output (C) Difficult to process a query (D) It is one of the complex retrieval models	1	2	2
8.	For a query, a retrieval system retrieves 42 relevant documents and 34 irrelevant documents from a document collection that consists of 95 relevant documents. What is the precision of the retrieval system (A) 0.40 (B) 0.50 (C) 0.62 (D) 0.55	1	5	2
9.	Spam Classification is an example for ____________. (A) Naive Bayes (B) Probabilistic condition (C) Random Forest (D) All the Above	1	4	3

10. What is the main assumption of the K-Nearest Neighbors (KNN) algorithm? 1 2 3

(A) The data instances are independent and identically distributed.
(B) The data instances are linearly separable.
(C) The data instances have equal variance.
(D) The data instances have a similar probability distribution.

11. Which of the following is the main objective of Support Vector Machines (SVM) algorithm? 1 1 3

(A) Minimize the misclassification rate
(B) Maximize the margin between classes
(C) Reduce the dimensionality of the input features
(D) Optimize the bias-variance tradeoff

12. Which of the following is finally produced by Hierarchical Clustering? 1 3 3

(A) Final estimate of cluster centroids
(B) Tree showing how close things are to each other
(C) Assignment of each point to clusters
(D) All of the mentioned

13. The generalized form of Bayesian network that represents and solve decision problems under uncertain knowledge is known as ? 1 1 4

(A) Directed Acyclic Graph
(B) Table of conditional probabilities
(C) Influence diagram
(D) Sequential diagram

14. The proportion of non-relevant items that has been retrieved in a given search is 1 2 4

(A) Precision
(B) Recall
(C) Generality
(D) Fallout

15. Suppose the frequency of the most frequent word in a corpus of Tamil documents is 10000. What would be the estimated frequency of second most frequent in the given corpus as per Zipf's law? 1 5 4

(A) 10000
(B) 2500
(C) 5000
(D) Can not be determined

16. In WordNet, the lexemes that share same form but have unrelated meanings is 1 3 4

(A) Homonym
(B) Hypernym
(C) Hyponym
(D) Meronym

17. Which of the following is a disadvantage of click relevance feedback method? 1 4 5

(A) Easy availability
(B) Less noisy
(C) Very noisy
(D) Very expensive to obtain

18. Which search technique allows you to specify how close two or more words must be to each other in order to register as match ? 1 1 5

(A) Proximity search
(B) Truncation
(C) Phrase searching
(D) Parenthesis

19. Which of the following is not an element to measure the relevance of search results in ad hoc information retrieval? 1 4 5

(A) A benchmark document collection
(B) A benchmark suit of queries
(C) An assessment of either relevant or non-relevant for each query and each document
(D) A set of trained users to check the result

20. A _____ is the term used when a search engine returns a web page that matches the search. 1 2 5

(A) Blog
(B) Hit
(C) Link
(D) View

PART - B (5 × 4 = 20 Marks) **Marks BL CO**

Answer **any 5** Questions

21.	Develop the mechanism in which the speed of the Information Retrieval process can be increased in real world applications.	4	3	1
22.	Discuss and design the system architecture for web-based search engine.	4	2	1
23.	How do you process a query using an inverted index and the basic Boolean Retrieval model?	4	2	2
24.	How do the Naive Bayes algorithm used for text classification?	4	4	3
25.	Why product recommendation engines are not good product search engines?	4	4	3
26.	How the documents are ranked using Probability Ranking Principle (PRP)?	4	1	4
27.	Demonstrate the process involved in Content-based image retrieval system.	4	3	5

PART - C (5 × 12 = 60 Marks)

Answer **all** Questions

		Marks	BL	CO
28.	(a) Describe the various information retrieval model in IR system. Analyze the performance of each model. **(OR)** (b) How Porter's Stemmer algorithm and Lovins Stemmer algorithm used in NLP? Implement Porter Stemmer in NLTK.	12	2	1
29.	(a) Demonstrate the various wild card characters used in pattern matching. How Permuterm index method is used to handle wild card queries? **(OR)** (b) Identify and describe the various methods for Spell Check and Hyphenation.	12	3	2
30.	(a) You are asked to design a text classification engine to process all queries raised by the employees of your organization. Elaborate in detail about the steps involved and the various factors to be considered while designing. **(OR)** (b) Discuss in detail about the various feature selection techniques in machine learning.	12	2	3
31.	(a) How do the relevance feedback used in IR system? Describe probabilistic relevance feedback model with a suitable example. **(OR)** (b) Demonstrate the various process involved in query expansion. How do the query drift issue is resolved during query expansion?	12	1	4
32.	(a) How do the eigen values and singular decomposition are related? Discuss in detail about the significance of low rank approximation in minimization problem. **(OR)** (b) Develop a framework for content-based image retrieval. How to compare any two images in a content-based image retrieval? What are the challenges faced in Image retrieval?	12	3	5

* * * * *