VIT Vidyalankar Institute of Technology
(Accredited A+ by NAAC)
(Autonomous Institute Affiliated to University of Mumbai)

Mid Semester Examination

Branch	Date	Sem.	Roll No. / Exam Seat No.	Subject	Student's Signature	Junior Supervisor's Name and Sign
AI	2/4/24	8		Text web SMA		

Question No.	A	B	C	D	E	F	G	H	Total	Total out of (20 /30 / 40)
1										
2										
3										
4										

Examiners Signature	Student's Sign (After receiving the assessed answer sheet)

Q-1

a) Text Mining - It is also known as text data Mining or textual data Mining. It is a proces of extracting valuable information from text. It involve the discovery by computer of new previously unknown information by automatically extracting information from diffent written resources
→ Knowledge Discovey
→ efficient information retrival
→ Sentimant Analysis
⇒ Enhancing customer support

b) Tokenization is technique in text Mining & NLP that involve dividing text into smaller units called token. Then token can be words, phrases, symbols or other meaningful elemt that constitute a piece of text. The proces of tokenization help in preparing text data for further analysis or

or procesig text data for further analysis or procesig. such as parsig, part of speech taggig etc.

c) Social Media Analysis involve collecting data from SM platform & analyzing it to gain insights into various aspects, such as public opinion, market trends, brand presence, user behaviour. This analysis can be applied accross various domain inluding marketig, customer servic, public relation.

d) feature selection is crucial step in text clustering as it significantly impacts the performance & outcome of clusteig process. Text clusteig involve grouping a set of texts into cluster grouping a set of texts into cluster, that contain similar text.
1) Term frequency inverse docunt frequency
2) feature pruning
3) latent Semantic Analysis
4) word embedding
5) chi Square

e) Rule based classifier - It is a type of ML algo that makes classification decision based on set of predifined rules. These rules are often created by human experts, who understand the domain well or can be automatically generated through algo.
characteristics:-
- Transparency
- Simplicity
- Domain expertise

f) Text Mining - It is a process of extracting meaningful information & insight from textual data. It involve techniques from linguistics, statics, & ML to analysis understand & interpret large volume of text. Common tasks in text Mining include sentimant analysis, topic Modeling, NER, & text classification.

g) NER stands for Named entity Recognition. It is a subtask of NLP that aims to identify & classify named entities within text into predifine categories such as names of person, organization, location, date, Numerical expression & more

4) Hidden Markov Model - They are Statistical Model used to describe sequence of observation events, where each event in the sequence corresponds to an underlying state that is not directly observed. hence "hidden". HMM are widly used in various fields, including speech recognition, NLP, bioinformatics & finance.

Q=2

(a) N-gram are contigous sequence of N item (word, character or token) extracted from given text or sequence. They are widly used in NLP & other field for various task such as language modeling, text generation & sentiment analysis.

Type of N gram-

1) Unigram - (N=1)

- unigram are single word or token in sequence
- example - "Apple", "orange", "banana"

2) Bigram (N=2)
- Bigram are sequence of 2 adjacent word or token
- ex - "Apple orange"

3) Trigram (N=3)
- Trigram consist of 3 adjacent word or token
ex - "Apple orange, banana"

4) N-gram (N>3
- N gram represent sequence of N adjacent word or token
- ex - "Apple organee, banana, pear"

Q-2

(b) Bayesian N/w - It is probabilistic graphical model that represent a set of random variable & their conditional dependency in a directed acyclic graph (DAG)

Component of Bayesian Network

1) Edge - Directed edge bet" Node indicate the dependencies bet" variable.
An edge from node A to node B means B is conditionally dependent on A

2) Node - Node represent random variable, event, or attribute in the domain of interest.

Advantage of BN

1. Probabilistic Reasoning
2. graphical Representation
3. Modularity & Scalability
4. Casual Interfere
5. efficient inference
6. Handling Missig data
7. Decision Support

Q=3 Recommendation Algorithm are techique used
(a) to suggest item or action to users based on their prefernce, behaviour or characteristic.

Their are Mainly 2 Major type

(i) Collaborative filtering: Recommand to the user to a based on the prefernce of user with Similar tastes It identifies user with Similar item prefernce & suggest item liked by those user. This is mainly the type of collaborative named as User based collaborative filtering

(ii) Item based collaborative filtering. This method focuses on the Similarity between item rather than user. The Similarity bet" item is calculated based on the rating or interaction they recive from user.

2) Content based filtering → The Approaches recommend item by analyzing the content or feature of items & profile of user's prefrence.

Step -

1) feature extraction - Analyze items to extract feature

2) Profile Building - Construct a user profile based on feature of item

3) Item scoring - for each item, compute a score based on Similarity profile

4) Recommend item - Recommend item with the Highest score that user hasn't interacted with

Q.3

(b) Distance based clustering algorithm partition a set of object into cluster. based on Similarity measure often using distance bet^n data points.

1) K-mean clustering :- partitioning approach. divide the dataset into K cluster by iteratively updating cluster centroid & assigning data point to the nearest centroid.

- require specifying the no. of cluster.

- Assume cluster are spherical & equally sized

- More Scalable with large no of variable & data point.

(2) Hierarchical - It is approch where build a tree of cluster. either by starting with indivual data points & merging them.
- Does not require specify no of cluster.