VIT Vidyalankar Institute of Technology

End Semester Examination (MU)

Evaluation of University Question Paper Solution Prepared by Faculty

Subject (Write in full) : Text, web & SMA (Regular ✓ / KT)

Exam : May 20 24 **/ Nov. 20** ____ **Exam Date :** 29 | May | 2024 **Q. Paper Code :** ____

Department : All **Year : FE/SE/TE/BE/ME/MMS Semester :** 8 **Scheme : CBSGS/CBCS**

Handwritten Solution Prepared by : Amit Aylani

Name of the Subject Cluster : AI ML

Name of the Cluster Mentor / Assessor : Prof. Avinash Shrivas

1st Assessment :

Question No.	Marks Obtained						Total
1	(a)	(b)	(c)	(d)	(e)	(f)	
2	(a)	(b)	(c)	(d)	(e)	(f)	
3	(a)	(b)	(c)	(d)	(e)	(f)	
4	(a)	(b)	(c)	(d)	(e)	(f)	
5	(a)	(b)	(c)	(d)	(e)	(f)	
6	(a)	(b)	(c)	(d)	(e)	(f)	
Total (Out of marks)							

Signature of Assessor / Cluster Mentor : _____

2nd Assessment :

Question No.	Marks Obtained						Total
1	(a)	(b)	(c)	(d)	(e)	(f)	
2	(a)	(b)	(c)	(d)	(e)	(f)	
3	(a)	(b)	(c)	(d)	(e)	(f)	
4	(a)	(b)	(c)	(d)	(e)	(f)	
5	(a)	(b)	(c)	(d)	(e)	(f)	
6	(a)	(b)	(c)	(d)	(e)	(f)	
Total (Out of marks)							

Signature of Assessor / Cluster Mentor : _____

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	Q. 1 (a)	Named entity recognition (NER) is a process in NLP that identifies & classifies key elements in text into predefined categories such as name of people, organ. location, date & many more. for example - In the sentence "Apple Inc was founded by Steve Jobs, Steve Wozniak, a NER System would identify "Apple Inc" as an organization. "Steve Jobs" "Steve Wozniak" & Ronald Wayne" as person, "Cupertino, California as a location & April 1, 1976 as a date. One effective approch to NER is using Conditional Random fields (CRFs) which are Statistical Model that predict the sequence of label. (entity catories) for a sequence of word. CRFs consider the content of each word & use feature like word itself part of speech tags, orthographic feature & neighbouring word to accurately predict entities. > Feature Extraction > Training the CRF Model > Interface

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
		Advantage :-
		> CRFs Consider the Context of each word by taking into account neighboring words & their feature, making them well suited for Sequential labeling task.
		> Unlike Hidden Markov Model, CRFs don not make the independence assumption bet" the observed feature allowing for more flexibility & accuray in Modeling
	Q=2 (5)	Probabilistic Docunt Clustering → It is a technique used to group documents into clusters. based on probabilistic Model that captre the underlyping data distribution This Method assumes that docunt are generated by a Mixture of probabistic distribution, it aims to find the parameters of these distribution that best explain the data. Common Model used in probabilistic docun document clustering include Latent Dirichelt Allocation (LDA) & (GMM)
		> Assignment of Docunt :

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

1) Distance based clustering - In method like K Mean or hierarchical clustering. Each document is assigned to a single document is assigned to a single cluster based on a distance metrics. (eg. Euclidean distance or cosine similarity). Docunt are grouped by Minimizing the within cluster variance.

2) Modeling data -

> They Method rely on predefined notion of distance or similarity. They do not explicitly model the underlying distribution of the data

> They method use probabilistic model to represent the data, They captur the Statistical properties & relationship within the data, providing a generation process of document

3) Handling overlap -
Typically assume that cluster are distinct & Non- overlapping. Each document belongs belongs to ~~exactly~~ Exactly one cluster.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	5)>	Complexity & Computation -
		→ often involve simpler computation, like distance calculation. & centroid update. Then ~~mean~~ method. Can be faster & easier to implement
	Q=2	Rule Based classifier :-
	(a)	Rule Based classifier use a set of predefined rules to assign labels to text document. These rule are often constructed based on the presence or absence of specific keyword, phrases. or pattern within text. Each rule typically consist of condition & an action
		ex - Rule-1 : IF the email contain the phase "Congratulation you have won then label it as Spam",
		Rule-2 : IF the email contain the word "free" more than three times then label it a "spam"
		Rule-3 :- IF the email contain the phrase "unsubscribe" Then label it as "spam"
		Rule-4 :- IF email contain the word "Urgent" AND it is form a known contact

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
		then label it as "not spam".
		> Strength of Rule Based classifier:
		> Simplicity & Interpretability:- The rules are straightforward & easy to understand. Each decision can be traced back to a specific rule. making the classification process transparent.
		> Customization - Rule can be tailored to specific needs & updated easily as new patterns emerge.
		> Efficiency - for small to medium sized dataset, rule based system can be very efficient since, they do not require extensive training.
	Q-2 (b)	Distance based clustering Algo.
		> K-Mean:- K-mean is a popular distance based clustering algo that partition data into K cluster. It aims to Min. the variance within each cluster by assigning each doaml to the nearest cluster centroid.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

Advantage -
- → Simiplicity
- → effectively
- → Scalability

Disadvantage -
1) fixed no of clusters.
2) Sensitivity to initialization
3) Assumption of spherical clusters

2) Hierarchical clustering :-

Hierarchical clustering build a tree of clusters by either agglomerative bottom up or top down approches

Advantage -
- No need to specify no of clusters
- flexibility
- Dendrogram Visualization

Disadvantage -
- Computationally intensive
- Not suitable for very large dataset
- Sensitivity to Noise & Outliers.

> Comparsion :-
(1) Kmean is more efficient & Scalable for large dataset but require the no of clusters to be predefined &

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

START WRITING HERE

Space for Marks | Question No.

Assumes Spherical clusters.

> Hierachical clustering - It provide more flexibility & does not require the no. of cluster to be specific in advance but it is computationally intensive & less suitable for very large dataset.

Q-3

(a) Web Spam employs various techique

1. Cloaking :- Serving diffent Content to Search Engine & users. A web Page might show high quality, keyword rich content to Search Engine while displaying irrelevant or Spammy content to users.

2. Keyword Stuffing :- Overloading a web Page with keyword to manipulate Search Engine ranking.

3. Hidden Text & Link :- Using techique to hide text & links from users while making them visible to Search engine

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

→ Challenges Associated with detecting & Combating Spam

1) Evolving technique :- Spammer continually develop new teque to evade detection

2) false positive & Negativ : Legitimate Content may be mistakenly classified as spam or spam may go undected.

3) Resource Intensiveness :-
effective spam detection often requir Significant Computational resources & Sophisticated algorithm.

(b) Challenges in Multilingual Sentiment Analysis

1) Language Variability :-
— Diffent language have unique Syntax, grammar & idiomatic expression that affect how sentiment is Conveyed.

2) Resource availibity - There is often a lack of annotated dataset & linguistic resoure (like Sentiment lexicon) for many langua

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

3) Translating Issue :
- Translating Issue for Sentiment analysis can introduce error & may not preserve the orignal Sentiment context.

> Sentiment can be expressed diffently across cultur might have a neutral or Negatin connotation in another.

4) Addrssing challenges in Multilingual Sentiment analysis

> Multilengual Pre trained Model

> Cross Lingual Embedding.

> Transfer learning.

> Data Augmented & Crowd sourcing

> cultural Adaptation

> Translation aware Model

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	Q.4 (a)	Session & visitor analysis in web Mining involve examing the interaction & behaviors of users on a website over a specfic period. A session typically represent a single visit by a user, capturing all interaction from time they enter the website until they leave. visitor analysis considers both the session level & broder behaviour of user, including return visit & overall engagement patterns

Important :-

> Understanding user behaviour :-
Analyzing session help in understanding how user interact with website, what page they visit & HOW long they stay on each page.

> Improving User Experience :-
by studing visitor behaviour journey can identify pain points in the user journey such as drop off rates on specific pap or during particular process

> Personalization :-
Insight from session & visitor analysis can be used to tailor content & offer to individual users, enhancing personalizati-

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	*	Marketing & Conversion optimization > understanding the path to each conversion point; making purchase Sign up for new letter
		Insigned Cainced
		> Traffic source & Referely
		> User engagent & Content effectivny
		> Customer Journey mapping
		> Usr idemographic & Segmentation

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	Q-4 (b)	Analysis of Sequence & Navigational pattern in web Mining

> Sequential & Navigation pattern analysis in web Mining focus on understanding the order & structure of user action as they navigate through a website. This analysis aims to uncover common path, Sequence & pattern in user navigation.

> Sequence Mining Algorithm.

1) Apriori Algorithm:- It used for Market analysis, it can be adapted to identify frequent navigation Sequence by finding frequent itemsets. & then generating association rules.

2) prefixSpan Algo - It focus on finding frequent Sequential pattern without generating candidate Sequence making it more efficient for large datasets

3) GSP - Extends Apriori by considering time constraint & allowing for gaps bet" Sequences.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

Application:-

They algorithm can applied to web logs to identify frequent sequence of pages that user visit helping to understand common navigation path & user behaviours.

> Navigational pattern path.

1) Path analysis

2) clickstream analysis

⇒ Insights gained

> User profile & Interest

> optimization & Navigation Structure

> Improving user flow

> predictive Analytics.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	Q.5 (a)	**Behaviour Analysis in SM Mining :-** Behavious analytics in SM Mining involves the study & interpretation of user action & interaction on SM platform. This include tracking metric like, share comment, retweet, follow click, & time spend on content. By analyzing their behaviour organization can gain insights into user enagent prefere & emerging trends. 1) User enagent :- Analyzing metric such as like, share comment & retweet helps measure how user interact with content 2) User prefere :- By tracking which type of content reain the most engagent duiney can infer user prefere 3) Trend identification - Monitoring the volum & velocity of specific keywords hashtags & topic over time 4) Audience & Segmentation :- Analyzing demographic data & behaviour pattern to segment user to distinct group.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
	Q.5 (a)	**Behaviour Analysis in SM Mining :-** Behaviour analytics in SM Mining involves the study & interpretation of user action & interaction on SM platform. This include tracking metric like, share comment, retweet, follow click, & time spend on content. By analyzing their behaviour organization can gain insights into user engagent prefer & emerging trends. 1) User engagent : Analyzing metric such as like, share comment & retweet helps measure how user interact with content 2) User prefere :- By tracking which type of content reain the most engagent drivers can infer user prefer 3) Trend identification - Monitoring the volum & velocity of specific keywords hashtags & topic over time 4) Audience & Segmentation : Analyzing demographic data & behaviour pattern to segment user to distinct group.

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE
		5) Sentiment Analysis - Analzing the tone & sentiment of user comments review & post
		6) Influence identification - Tracking the engagent & reach of individual user to identify influential figure within a week
	(b)	Collaboratin filtering - It is a recommendation approach can be based on the assumption that user with similar behaviour & prefrence will like similar item
		> User based Collaboratin filtering : > recommed item by finding similar usu based on their rating & behaviour
		> Item based Collaboratin filtering :- - It recommend item by finding similar item based on usu interation with them

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE

> Content Based filtering
It recommend item based on characteristics of item & user past interaction

> Analyzes the content of item & matches them with user preference derived from their past behaviour

> Difference bet" Collabrative & Content based filtering

1) Data depency	Dependent on user item interaction ex rating	Depend on item feature & user profile.
2) personalization	provide recommendation based on Similarity. user item	provide recommendation based on individual user preference
3) Cold start problem	Suffer from Cold start problem for new user & new item	> Less affected by cold start problem for new item recomdation

Total Marks of Question no.		Examiner	
		Moderator	
		Re-Assessor	

Space for Marks	Question No.	START WRITING HERE