U-5 → Big Data Analytics & Model Evaluation.

Page No.

* Clustering –

cluster → a no. of similar things that occur together.

technique in which the data points are arranged in similar groups dynamically without any pre-assignment of groups

partioning a set of data in a set of meaningful subclasses.

- properties of a cluster –
  - all data points in a cluster should be similar to each other
  - data points from different clusters should be as different as possible.

Types → Hard clustering → each data point → only one cluster
soft clustering → each data point → separate cluster
↳ probability or likelihood of data.

Distance = $p = (p_1, p_2, \ldots)$ $q = (q_1, q_2, \ldots)$.

$$d = \sum_{i=1}^{k} (p_i - q_i)^2.$$

Centroid is a point whose coordinates are averages of corresponding coordinates of a given set of points

Appl^n → • Customer Segmentation • Image Processing • Recommendation Engine
• Marketing • Insurance • Seismology
• Land Use • Urban planning • Healthcare.

Page No.
Date

K-Means-

- heuristic method.
- unsupervised learning algo.
- solve clustering problems.
- groups are unlabelled dataset into different clusters.
- K → defines no. of predefined clusters. need to be created.
  - K=2 → two clusters
  - K=3 → three clusters.

- ~~divides~~ each dataset belongs only one group has similar projects.
- minimizes sum of distances between data point & their corresponding clusters.
- distance calculation → Euclidean Distance.

Adv →
- efficient in computation
- easy to implement.

Disadv →
- Applicable only when mean is defined.
- need to specify K, no. of clusters, in advance
- Trouble with noisy data & outliers
- not suitable to discover clusters with non-convex shapes.

Page No.
Date

- **Hierarchical clustering -**
  - Hierarchical cluster analysis or HCA.
  - method of cluster analysis
  - data points are arranged in hierarchy of clusters.
- **Dendrogram -**
  - diagram representing a tree or hierarchy.

Hierarchical Clustering Strategies (Algorithms)

Agglomerative

- bottom-up approach
- each item in its own cluster
- Iteratively clustered are merged together.

steps → Initialization, similarity measurement, merge clusters, creating dendrograms, cut the dendrogram.

- no need of specifying no. of clusters.
- expensive → large dataset
- handling noise or outliners is difficult

Divisive

- top-down approach
- all items in one cluster.
- large clusters are successively divided.

steps → Initialization, split cluster, recursive division, creating dendrogram, cut the dendrogram.

- expensive → large dataset.
- recursively splitting clusters
- overcome limitation of agglo
- determining appropriate no. of clusters & interpreting result

Page No.
Date

Time Series Analysis –

- attemps to model underlying structure of observations taken over time.

- Appln → Retail Sales
  - → Spare parts planning
  - → Stock Trading

- Characteristics → Trend
  - → Seasonality
  - → Cyclic
  - → Random

ARIMA → (Auto Regressive Integrated Moving Average).

- forecasting technique that projects future values of series
- short term forecasting → 40 historical data points.
- Parameters of ARIMA model –
  - p (lag order) – no. of lag observations.
  - d (degree of differencing) → no. of times raw observations are differenced.
  - q (order of moving average) → size of moving avg window.

Page No.
Date

- Text Analysis -
  - also called text Analytics
  - refers to representation, processing & modelling of textual data to derive useful insights.

- CORPUS → large collection of texts used for NLP.

→ Challenges → • High Dimensionality
• Unstructured Data.

- Steps in Text Analysis - (Preprocessing)
  1) Tokenization
  2) Lowercasing
  3) Stop word Removal (noise in text)
  4) Punctuation Removal
  5) Lemmatization or Stemming.
  6) Removing HTML tags or special characters.
  7) Removing Numbers
  8) Removal of frequent words.
  9) Removal of rare words.

Term Frequency → measures frequency of a word in given document.

$$TF = \frac{\text{count of terms in document (t)}}{\text{total no. of terms in document (d)}}$$

Page No.
Date

* Social Network Analysis (SNA)

process of investigating social structures through use of networks & graph theory.

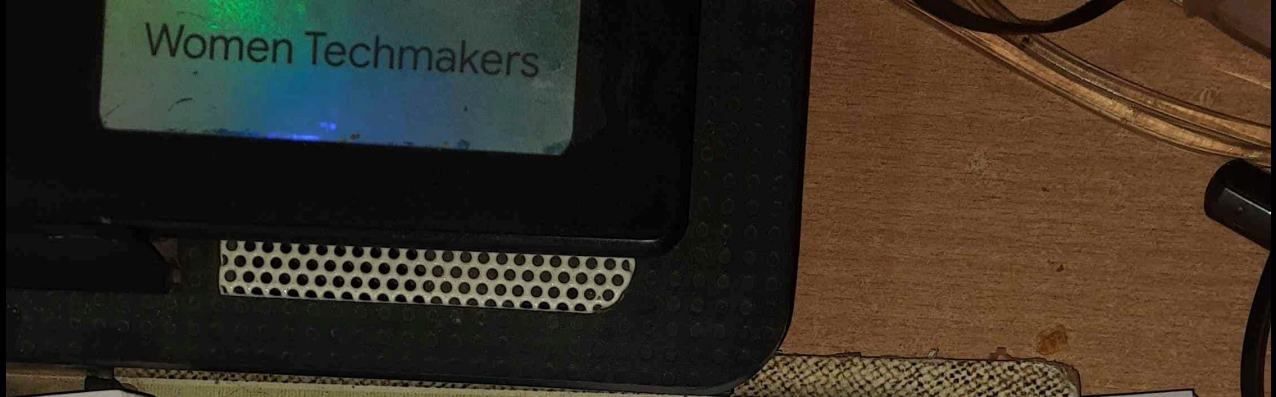
- graph theory & mathematical models.
- examines patterns of interactions, connections & dependencies for insights.
- properties of graph
  - Degree of Node → Density
  - Path Length → closeness
  - Centrality → Betweeness

- vertex or Node → represent an object in graph.
- connection between → nodes → edge or a link.

* Need of SNA.

- Identifying Influencers
- Human Resource Management (HRM)
- Contact Tracing
- Identify themes & connections.
- Fraud Detection.

Fundamental concepts
- Actor
- Relational tie
- Dyad
- Triad
- Subgroup
- Group
- Relation.



Page No.
Date

Business Analysis –

involves identifying business needs, analyzing requirements, & finding solutions to organisation problems.

- Business analyst → intermediaries b/w stakeholders & technology terms to ensure that business goals.

- understand business process
- areas for improvement & propose solutions
- documenting business requirements.
- analyzing & modeling business processes.
- facilating communication & collaboration b/w stakeholders.
- managing change & ensuring smooth implementation of solutions
- defining & validating soln requirements.

Page No.
Date

Q18-

1) Leave P- Out Cross Validation
2) Sub- Sampling

→ • variant of k-fold cross validation → evaluate performance of ML models.
• doesn't divide dataset into k folds
• divides dataset into all possible combinations of leaving p samples. ↗ valid^n set.
& remaining for training

steps → • value of p (choose)
• create possible combinations
• each combin^n → train model on ~~training~~ remaing sample & evaluate
• calculate avg performance
• more comprehensive evaluation.
• small datasets or order of sample matters.

2) Sub- Sampling → down sampling or undersampling
↳ ml to address class imbalance problems.
↳ no. of class → significantly uneven.
↳ Biased mode performance.
↳ Balance class distribution → reduce samples.

Process steps → Identify class → large no. of samples.
↳ Randomly select subset of sample.
↳ combine selected subset from majority class to minority class.
↳ train ml model using balanced dataset.

Page No.
Date

Q13 - Stemming & Lemmatization.

↳ techniques used to in NLP to reduce words to their base or root forms.
↳ normalize words & reduce inflectional or derivational variations.

- Stemming -
- process of removing prefixes & suffixes from words to obtain word's base forth.
- resulting stem may not always be a valid word.

original words - fishing, fishes, fished
stemmed words - fish, fish, fish.

In this case, the stem 'fish' is derived by removing the suffixes "-ing", "-es" & "-ed" from original words.

- Lemmatization -
- aims to transform words to their base form or lemma, considering word's meaning & part of speech.
- uses linguistic rules & morphological analysis to acheive this.
- resulting lemma is a valid word represent base form

original words - walking, walks, walked
stemmed words - walk, walk, walk.

stemming → faster
lemmatization → more accurate results

Q14 -

Q14. Preprocessing Techniques -

essential steps in preparing raw text data for NLP tasks.

- Transforming & cleaning text to improve its quality & facilitate effective analysis.

1) Tokenization -
   ↳ process of breaking a text into smaller units called token.
   - Tokens → words, sentences, or even subword units.
   - segmenting text & forms basis for subsequent analysis.

Input Text → "I love to play soccer."
Tokenized Output → ["I", "love", "to", "play", "soccer", "."]

2) Lowercasing -
   ↳ involves converting all text to lowercase.
   It helps in standardizing text & treating words with diff cases (eg. "hello" & "Hello") as same token.

Input Text → "I love NLP".
Lowercased output → "i love nlp."

3) Stop Word Removal -
   common words → no significant meaning.

Input Text - "I love to read books & watch movies."
Stop word removed Output - "love read books watch movies.".

4) Punctuation Removal -
5) Lemmatization or Stemming -
6) Removing HTML tags or special characters.

8q Apriori Algorithm

↳ association rule mining algorithm → discover frequent items in a dataset

Components → support, confidence & Lift

Principle → If an itemset is a frequent itemset then its subsets must also be frequent.

1) Generating itemsets of size 1 (singletons)

2) Generating Frequent itemsets of size k (k>1):

3) Iteratively Repeating Step 2-

min support → 3 (occurs atleast 3 transactions).

frequent itemsets of size 1 would be {Milk}, {Bread}, {Butter}, & {Eggs}

→ Handles large dataset
→ describe rel^n b/w items
→ meaniful associations & patterns in transcational data.