

IV. B. Tech - I sem - End Examination - Dec - 2025

Advanced Text and multimedia Analytics  
(CSD)

Date: 5/12/2025

Scheme of valuation &amp; answers

PART - A

- 1 a) unigram - 1 word [1 m]  
Bigram - 2 words  
Trigram - 3 words [1 m]
- b) It is the process of splitting a text into smaller units like words or tokens.
- c) PageRanking identifies the most relevant and authoritative pages in web mining [1 m]
- d) Accuracy and F1 score [1 m]
- e) URL selection, fetching, parsing, and storing [1 m]
- f) Techniques for automated web page retrieval such as BFS, DFS and focus crawling [1 m]
- g) It measures how important or influential a node is within the n/w [1 m]
- h) Graph based model [1 m]  
probabilistic model

- 1) Fraud detection [1m]  
personalized recommendations
- 2) Data sparsity [1m]  
dynamic user preferences

### PART-B

2. 1) Converts unstructured text into numerical [10m]  
format for algorithm processing
- 2) Enables similarity measurement  
between documents
- 3) supports machine learning tasks  
like classification and clustering
- 4) Common methods include bag-of-words,  
TF-IDF, and word embeddings
- 5) Reduce ambiguity by representing  
semantic meaning
- 3 NER and its application [10m]
- 1) Identifies and classifies entities like  
names, places, dates in text
- 2) Improves information extraction and  
text understanding

- 3) support question answering and summarization systems
- 4) used in biomedical text mining and customer analytics
- 5) Help organize large text datasets efficiently

#### 4 a) Feature Extraction & Dimensionality Reduction: [5m]

- 1) converts text into informative features  
(eg: TF-IDF, pos tags)
- 2) Removes irrelevant or noisy attributes
- 3) Techniques include PCA, LSA/SVD and chi-square feature selection
- 4) Reduces computational cost and improves classifier performance
- 5) Helps avoid overfitting

#### 4 b) Working & Advantages of HDP [5m]

- 1) Non-parametric Bayesian model that discovers an unknown number of topics
- 2) Allows topic to be shared across multiple groups of documents.

3) Automatically adjusts complexity based on data size

4) Handles hierarchical structure better than LDA

5) produces more flexible and accurate topic distributions.

5. Compare and contrast LDA, PLSI, and HDP [10m]

1. LDA uses Dirichlet priors; PLSI lacks priors and overfits easily

2. HDP extends LDA by allowing infinite topics (non-parametric)

3. LDA works well with fixed topic counts; HDP discovers topic count automatically

4. PLSI is less scalable and less generalizable

5. HDP best for large, evolving datasets

6. Visitor segmentation in web analytics & its role in personalization [10m]

→ Groups website visitors based on behavior, demographics, or interests

→ use metrics like session duration, click pattern, and purchase history

→ Helps targets users with customized content or ads

→ Improves user experience and engagement

→ supports better marketing and decision making

7. session identification in web usage mining [10 m]

→ separates user activity logs into meaningful sessions.

→ use time based heuristics (eg 30-min timeout)

→ important for analyzing user behavior patterns

→ important for analyzing user behavior patterns

→ Employs navigational based and referral-based rules

→ Helps build user profiles for recommendation systems

8. Types of social n/w models & Their significance [10 m]

\* Graph model :

Represents users as nodes and connections as edges

\* Random n/w model :

studies probability-based connectivity

\* small-world model :

Explains short path lengths in real networks

\* scale-free model :

models power-law degree distribution

Helps understand flow, influence, and connectivity

9. Information diffusion in social media (Example) [10 m]

\* Describes how information spreads across users over time.

• Influencers act as sources for faster propagation.

\* Influencers act as sources for faster propagation

\* follows models like SIR, SIS, or cascade models

\* Example: viral spread of hashtag on Twitter

\* Helps in marketing, rumor detection and trend prediction.

## 10. Behavioral Analytics in social media [10m]

\* Analyze patterns in user action (likes, shares, comments)

\* Detects anomalies, preferences, and engagement trends

\* Helps personalize content and improve targeting

\* supports fraud detection and sentiment analysis.

\* uses statistical and ML models.

11 a) Handling comparative sentences in Sentiment analysis [5m]

- \* Detects comparison keywords (eg. better, worse)

- \* Identifies entities being compared

- \* determines sentiment orientation relative to comparison

- \* uses dependency parsing for accurate interpretation.

- \* Helps in product reviews and competitive analysis

11 b) Brief notes on Behavioral Analytics [5m]

- \* studies user behavior patterns to derive insights

- \* uses logs, clicks, and interactions as data sources

- \* Helps predict future behavior and personalize services

- \* supports fraud detection and risk management

- \* widely used in marketing and cybersecurity