SEMESTER - II

DATA MINING AND PATTERN DISCOVERY

1. Course Description

Programme:	**M. Sc. Data Science**	**Max. Hours:**	**60**
Course Code:	**P26/CDS/DSC/201**	**Hours per week:**	**4**
Type of Course:	**DSC**	**Max. Marks:**	**100**
No. of Credits:	**4**		

2. Course Objectives

- Understand data mining concepts, data characteristics, and preprocessing techniques.
- Apply association rule mining for pattern discovery
- Examine classification methods for predictive data analysis.
- Analyze data using clustering and anomaly detection methods.

3. Course Outcomes

After the successful completion of the course, the student will be able to:

CO1: Describe the concepts of data mining, data types, and preprocessing techniques. **(L II)**

CO2: Apply association rule mining to discover patterns in data. **(L III)**

CO3: Apply classification techniques for predictive data analysis. **(L III)**

CO4: Analyze datasets using clustering and anomaly detection methods. **(L IV)**

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

P. V. Sudha

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

4. Course content

MODULE I: **15 Hours**

Data Mining Foundations

Data Mining Fundamentals – fundamentals of data mining; patterns and techniques used in data mining; data mining tasks; applications of data mining; various data sets used for data mining.

Data and Data Preprocessing – data types; statistics of data; similarity and distance measures; data quality; data cleaning; data integration; data transformation; dimensionality reduction.

MODULE II: **15 Hours**

Pattern Mining

Pattern Mining Methods – market basket analysis; frequent itemsets and association rules; frequent itemset mining using the Apriori method; pattern evaluation methods, mining various kinds of patterns

MODULE III: **15 Hours**

Classification

Classification Methods – basic concepts of classification; decision tree induction; Bayesian classification methods; learning from your neighbours (k-nearest neighbour), Bayesian belief networks; support vector machines; rule-based classification

MODULE IV: **15 Hours**

Cluster Analysis and Anomaly Detection

Cluster Analysis – basic concepts of clustering; partitioning methods; hierarchical methods; density-based and grid-based methods; evaluation of clustering.

Outlier Detection - types of outliers, statistical approaches, distance and density-based outlier detection, clustering and classification-based approaches

5. Reference Books

1. Jiawei Hän, Micheline Kamber, Jin Pei, Data Mining: Concepts & Techniques, 3rd Edition., Morgon Koffman ,2011
2. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Education, 2008.
3. Vikram Pudi, P. Radha Krishna, Data Mining, Oxford University Press,1st Edition, 2009.
4. Robert Layton, Learning Data Mining with Python, Packt, 2nd Edition, 2017

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

6. Syllabus Focus

a) Relevance to Local, Regional, National and Global Development Needs

S. No	Student Centric Methods Adopted	Type/Description of Activity
1.	National	Analyze Indian indigenous knowledge datasets using clustering and association rules.
2.	Global	Mine global IKS datasets to find patterns and correlations.

b) Components on Skill Development/Entrepreneurship Development/Employability

SD/ED/EMP	Syllabus Content	Description of Activity
Skill Development	Module I & II	Hands-on pattern mining and preprocessing exercises
Employability	Module III & IV	Classification and Clustering Mini Project
Empowerment	Module I–IV	Data Mining on Indian Knowledge Systems and Global Datasets

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

7. Pedagogy

S. No	Student Centric Methods Adopted	Type / Description of Activity
1.	Participative Learning	Presentations and group discussions on data mining concepts and patterns
2.	Experiential Learning	Hands-on practical exercises on data preprocessing, pattern mining, and evaluation
3.	Problem solving	Mini projects on classification, clustering, outlier detection, and pattern analysis

8. Course Assessment Plan

a) Weightage of Marks in Continuous Internal Assessments and End Semester Examination

CO	Continuous Internal Assessments CIA - 40%	End Semester Examination-60%
CO1	CIA 2 – Test 1: MCQ's, Quiz test or subjective	Written Exam
CO2	CIA 1 - Subjective	
CO3		
CO4	CIA 2 – Test 2: MCQ's or Presentation	

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

CBCS 2026

b) Model Question Paper – End Semester Exam Theory

DATA MINING AND PATTERN DISCOVERY

Course Code: P26/ CDS/ DSC/201 **MAX MARKS: 60**
Credits: 4 **TIME: $2^{1/2}$ hours**

Note: This question paper consists of Section A and B. The answer to Section A & B must be written in the answer book given.

SECTION – A (Long Essay Type)

Answer ALL questions: **Marks: 4 x 10 = 40**

1. a. Discuss in detail the functionalities of Data Mining.
   b. Explain various data preprocessing techniques. Describe how data reduction contributes to effective preprocessing.

OR

2. a. Explain the major issues in Data Mining.
   b. Describe the problem of data quality with examples. Explain feature subset selection in preprocessing. Discuss the importance of data pre-processing, including data cleaning, integration, transformation, and dimensionality reduction.

3. a. State Apriori principle. Write Apriori algorithm for frequent itemsets. Explain with an example.
   b. For the given transaction dataset, illustrate the process of Market Basket Analysis and demonstrate how association rules are discovered. Assume minimum support = 60% and minimum confidence = 80%

TID	Items
1	{Milk, Bread}
2	{Bread, Butter}
3	{Milk, Butter, Eggs}
4	{Bread, Eggs}
5	{Milk, Bread, Butter}
6	{Butter, Eggs}
7	{Milk, Bread, Eggs}

OR

4. a. Show how frequent itemsets are extracted from an FP-tree by constructing conditional FP-trees and traversing them step-by-step

   b. Apply the following transaction data set that shows few transactions and list of items using FP Growth Approach to find frequent itemset with min-support =3

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST. FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

TID	Items
1	{a, b}
2	{b, c, e}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

5. Apply decision tree induction to construct a classification model and explain the steps involved in building the tree.

OR

6. Demonstrate how classification can be performed using Bayesian classification and support vector machines.

7. Compare partitioning, and grid-based clustering methods. Distinguish their characteristics and prioritize situations where each method is most suitable.

OR

8. Categorize outlier detection methods into statistical, distance-based, density-based, and clustering/classification-based approaches. Compare their effectiveness in detecting anomalies.

SECTION –B (Short Essay Type)

II. Write short notes on any **FIVE** of the following: **Marks: 4 x 5 = 20**

9. Explain different types of data used in data mining.
10. Describe similarity and distance measures and their role in data mining..
11. Illustrate how sequential pattern mining is used to discover patterns in sequential data.
12. Apply the concepts of support and confidence to evaluate an association rule.
13. Demonstrate how a classification model can be used to predict class labels.
14. Use rule-based classification to determine the class labels of data instances.
15. Analyze how the DBSCAN algorithm identifies clusters and noise.
16. Classify the different types of outliers in data mining.

P.V.Sridhar

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

Question Paper format – Blooms Taxonomy Level

<table>
<tr><th colspan="5">SECTION A - INTERNAL CHOICE 4Q X 10 M = 40 M</th></tr>
<tr><th>Question Number</th><th>Module Covered</th><th>Question</th><th>CO</th><th>BTL (Blooms Taxonomy Level)</th></tr>
<tr><td>1</td><td>Module 1</td><td>a. Discuss in detail the functionalities of Data Mining.
b. Explain various data preprocessing techniques. Describe how data reduction contributes to effective preprocessing.</td><td>CO 1</td><td>Level II</td></tr>
<tr><td>2</td><td>Module 1</td><td>a. Explain the major issues in Data Mining.
b. Describe the problem of data quality with examples. Explain feature subset selection in preprocessing.</td><td>CO 1</td><td>Level II</td></tr>
<tr><td>3</td><td>Module 2</td><td>a. State Apriori principle. Write Apriori algorithm for frequent itemsets. Explain with an example.
b. For the given transaction dataset, illustrate the process of Market Basket Analysis and demonstrate how association rules are discovered. Assume minimum support = 60% and minimum confidence = 80%
<table>
<tr><th>TID</th><th>Items</th></tr>
<tr><td>1</td><td>{Milk, Bread}</td></tr>
<tr><td>2</td><td>{Bread, Butter}</td></tr>
<tr><td>3</td><td>{Milk, Butter, Eggs}</td></tr>
<tr><td>4</td><td>{Bread, Eggs}</td></tr>
<tr><td>5</td><td>{Milk, Bread, Butter}</td></tr>
<tr><td>6</td><td>{Butter, Eggs}</td></tr>
<tr><td>7</td><td>{Milk, Bread, Eggs}</td></tr>
</table></td><td>CO 2</td><td>Level III</td></tr>
<tr><td>4</td><td>Module 2</td><td>a.Show how frequent itemsets are extracted from an FP-tree by constructing conditional FP-trees and traversing them step-by-step
b. Apply the following transaction data set that shows few transactions and list of</td><td>CO 2</td><td>Level III</td></tr>
</table>

P. V. Sridevi
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

		items using FP Growth Approach to find frequent itemset with min-support =3 **TID** | **Items** 1 | {a, b} 2 | {b, c, e} 3 | {a, c, d, e} 4 | {a, d, e} 5 | {a, b, c} 6 | {a, b, c, d} 7 | {a} 8 | {a, b, c} 9 | {a, b, d} 10 | {b, c, e}		
5	Module 3	Apply decision tree induction to construct a classification model and explain the steps involved in building the tree.	CO 3	Level III
6	Module 3	Demonstrate how classification can be performed using Bayesian classification and support vector machines.	CO 3	Level III
7	Module 4	Compare partitioning, and grid-based clustering methods. Distinguish their characteristics and prioritize situations where each method is most suitable.	CO 4	Level IV
8	Module 4	Categorize outlier detection methods into statistical, distance-based, density-based, and clustering/classification-based approaches. Compare their effectiveness in detecting anomalies.	CO 4	Level IV
SECTION B - ANSWER ANY 5 OUT OF 8 (To compulsorily have ONE question from each module)				**4Q X 5 M = 20 M**
9	Module 1	Explain different types of data used in data mining.	CO 1	Level II
10	Module 1	Describe similarity and distance measures and their role in data mining.	CO 1	Level II
11	Module 2	Illustrate how sequential pattern mining is used to discover patterns in sequential data.	CO 2	Level III

P.V Sudha
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

12	Module 2	Apply the concepts of support and confidence to evaluate an association rule.	CO 2	Level III
13	Module 3	Demonstrate how a classification model can be used to predict class labels.	CO 3	Level III
14	Module 3	Use rule-based classification to determine the class labels of data instances.	CO 3	Level III
15	Module 4	Analyze how the DBSCAN algorithm identifies clusters and noise.	CO 4	Level IV
16	Module 4	Classify the different types of outliers in data mining.	CO 4	Level IV

c) Question Paper Blueprint

Modules	Hours Allotted in the Syllabus	CO Addressed	Section A (No. of Questions)	Total Marks	Section B (No. of Questions)	Total Marks
1	15	CO-1	2	4x10 = 40	8 (By taking two questions from each Module)	5x4 = 20
2	15	CO-2	2			
3	15	CO-3	2			
4	15	CO-4	2			

9. CO-PO Mapping

CO	PO	Cognitive Level	Classroom sessions (hrs)
1	1	Understand	15
2	1,2	Apply	15
3	1, 2	Apply	15
4	1, 4	Analyze	15

P. V. Sridevi
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

DATA MINING AND PATTERN DISCOVERY
Practical Syllabus

1. Course Description

Programme:	**M. Sc. Data Science**	**Max. Hours:**	**40**
Course Code:	**P26/CDS/DSC/201/P**	**Hours per week:**	**2**
Type of Course:	**DSC**	**Max. Marks:**	**50**
No. of Credits:	**2**		

2. Course Objectives

1. Develop practical NLP skills using Python.
2. Apply NLP to Indian and global text datasets.

3. Course Outcomes

After the successful completion of the course, the student will be able to:

CO1: Pre-process and analyze text data effectively.

CO 2: Build end-to-end NLP pipelines for real-world tasks.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

P. V. Subbareddy

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007

4. Course Content

1. Perform exploratory data analysis on multiple datasets using Python (Pandas, Matplotlib, Seaborn), R, or Tableau
   Sample Case Study: Retail sales trends and customer segments
2. Perform data preprocessing including handling missing values, outliers, and data transformation using Python (Scikit-learn, NumPy), KNIME, or RapidMiner
   Sample Case Study: Clean and prepare healthcare patient records
3. Apply feature engineering and dimensionality reduction (normalization, standardization, PCA) using Python
   Sample Case Study: Reduce dimensionality of facial recognition dataset
4. Implement market basket analysis using Python (mlxtend), R (arules), or Weka
   Sample Case Study: Identify product associations in supermarket transactions
5. Perform frequent pattern mining and compare Apriori, FP-Growth, and Eclat algorithms using Python or Apache Spark
   Sample Case Study: Build product recommendations for e-commerce data
6. Mine sequential patterns and evaluate rules using Python or SPMF library
   Sample Case Study: Analyze web clickstream navigation patterns
7. Build and visualize decision trees and random forests using Python (Scikit-learn) or R
   Sample Case Study: Predict loan defaults in banking data
8. Implement Naive Bayes and KNN classifiers and compare performance using Python or Weka
   Sample Case Study: Email spam detection
9. Experiment with SVM kernels and hyperparameters using Python or MATLAB
   Sample Case Study: Classify breast cancer as malignant or benign
10. Apply ensemble methods, cross-validation, and model evaluation using Python (XGBoost, AdaBoost)
    Sample Case Study: Predict telecom customer churn
11. Implement K-means and K-medoids clustering and determine optimal K using Python or R
    Sample Case Study: Segment retail customers for targeted marketing
12. Apply hierarchical and density-based clustering (DBSCAN, Agglomerative) and create dendrograms using Python
    Sample Case Study: Identify crime hotspots in city data
13. Perform anomaly detection using statistical, distance-based methods in Python
    Sample Case Study: Detect credit card fraud
14. Perform advanced clustering and evaluate clustering metrics using Python or TensorFlow
    Sample Case Study: Detect communities in social networks

Mini Project: End-to-end data mining project incorporating multiple techniques on a chosen dataset

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

5. Model Question Paper – End Semester Exam Practical

DATA MINING AND PATTERN DISCOVERY

Programme : M.Sc. Data Science
Course Code : P26/CDS/DSC/201/P
Type of Course: DSC
No. of credits : 2

Duration : 2 Hours
Max. Marks: 50

Answer the following

- Implement an end-to-end data mining workflow using Python and relevant libraries (Pandas, Scikit-learn, mlxtend, PyOD, etc.).
- Demonstrate a mini-project case study using a real-world dataset (e.g., retail sales, healthcare records, financial transactions, or social media data).

Prepared by	Checked &Verified by	Approved by
Padma **Ms. Padmashree** **Teaching faculty**	Dr. Sujatha Yeruva **Dr. Sr. Sujatha Yeruva** **HoD**	Uma Joseph. **Prof. Uma Joseph** **Principal**

Dr. Sujatha Yeruva
HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

P.V. Sudhakar
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

SEMESTER - II

NATURAL LANGUAGE PROCESSING

1. Course Description

Programme: M.Sc. Data Science
Course Code: P26/CDS/DSC/202
Type of Course: DSC
No. of Credits: 4

Max. Hours: 60
Hours per week: 4
Max. Marks: 100

2. Course Objectives

- Understand fundamentals of NLP and text analytics.
- Apply Python to pre-process and transform text data.
- Apply text analysis techniques to extract patterns and insights.
- Compare and develop Python applications for sentiment analysis, topic modeling, and visualization.

3. Course Outcomes

After the successful completion of the course, the student will be able to:

CO1: Interpret NLP and text processing techniques. **(Level II)**

CO2: Implement text pre-processing pipelines in Python. **(Level III)**

CO3: Perform sentiment analysis and topic modeling on datasets. **(Level III)**

CO4: Compare and develop end-to-end text analytics workflows with visualization. **(Level IV)**

P. V. Subbaiah
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

4. Course content

MODULE I: **15 Hours**

Regular Expressions and Text Processing

Natural Language basics : Linguistics, language syntax and structure – words, phrases, clauses, grammar – dependency grammars, constituency grammars, word order typology; language semantics – lexical semantic relations, representation of semantics, Text Corpora – Corpora annotations and utilities, popular corpora, accessing text corpora, applications of natural language processing; Working with Text – String literals, string operations and methods, regular expressions, text analytics frameworks

Processing and Understanding Text: Text tokenization – sentence tokenization, work tokenization; Text normalization – cleaning text, tokenizing text, removing special characters, expanding contractions, case conversions, removing stopwords, correcting words, stemming, lemmatization.

MODULE II: **15 Hours**

POS Tagging and Text Classification

Understanding text syntax and structure – installing necessary dependencies, importing machine language concepts, Parts of Speech (POS) tagging – recommended POS taggers, building POS taggers, shallow parsing, dependency-based parsing, constituency-based parsing

Text classification – automated text classification, text classification blueprint, text normalization, feature extraction – Bag of Words model, TF-IDF model, advanced word vectorization models – averaged word vectors, TF-IDF weighted average vector words, classification algorithms – multinomial Naïve Bayes, support vector machines, evaluating classification models, building a multi class classification system, applications and uses

MODULE III: **15 Hours**

Feature Extraction

Text Summarization: feature matrix, singular value decomposition, feature extraction, key phrase extraction – collocations, weighted tag-based phrase extraction, topic modelling – latent semantic indexing, latent Dirichlet allocation, non-negative matrix factorization, extracting topics from product reviews, automated document summarization – latent semantic analysis, TextRank, summarizing a product description

Text Similarity and Clustering: similarity measures, analyzing text similarity – Hamming distance, Manhattan distance, Euclidean distance, Levenshtein Edit distance, Cosine distance and similarity, analyzing document similarity – Cosine similarity, Hellinger – Bhattacharya

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

distance, Okapi BM25 ranking, document clustering, clustering greatest movies of all time – K-means clustering, affinity propagation, Ward's Agglomerative Hierarchical clustering.

MODULE IV: **15 Hours**

Semantic and Sentiment Analysis

Semantic Analysis : Exploring WordNet – understanding Synsets, analyzing semantic relations – entailments, homonyms and homographs, synonyms and antonyms, hyponyms and hypernyms, holonyms and meronyms, semantic relationships and similarity, word sense disambiguation, named entity recognition, analyzing semantic representation – propositional logic, First Order logic, sentiment analysis, sentiment analysis of IMDb Movie reviews- setting up dependencies, preparing datasets, supervised machine learning technique, unsupervised Lexicon-based techniques - AFINN lexicon, Sent WordNet, VADER lexicon, Pattern lexicon

5. Reference Books

1. Dipanjan Sarkar, Text Analytics with Python - A Practical Real-World Approach to Gaining Actionable Insights from your Data, Apress, 2016
2. Hobson Lane and Maria Dyshel, Natural Language Processing in Action, Second Edition, Manning, 2025
3. Steven Bird, Ewan Klein, and Edward Loper, Natural Language Processing with Python, O'Reilly, 2009
4. Daniel Jurafsky and James H. Martin, Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, Second Edition, Pearson, 2013

6. Syllabus Focus

a) Relevance to Local, Regional, National and Global Development Needs

S. No	Student Centric Methods Adopted	Type/Description of Activity
1.	National	Students will apply NLP techniques to Indian Knowledge System texts, including Sanskrit and regional literature, for pre-processing, POS tagging, and semantic analysis
2.	Global	Students will apply NLP techniques to global English corpora, such as IMDB reviews and Wikipedia, for classification, summarization, and sentiment analysis.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

b) Components on Skill Development/Entrepreneurship Development/Employability

SD/ED/EMP	Syllabus Content	Description of Activity
Skill Development	Module I & II	Hands- on Practicals
Employability	Module III & IV	Text Summarization & Clustering Mini Project
Empowerment	Module I–IV	NLP on Indian Knowledge System & Global Corpora

7. Pedagogy

S. No	Student Centric Methods Adopted	Type / Description of Activity
1.	Participative Learning	Presentations and Group Discussions on NLP concepts and Indian Knowledge System texts
2.	Experiential Learning	Hands-on Practicals using Python for tokenization, POS tagging, text classification, and feature extraction
3.	Problem solving	Mini Projects on text summarization, clustering, semantic analysis, and sentiment analysis

8. Course Assessment Plan

a) Weightage of Marks in Continuous Internal Assessments and End Semester Examination

CO	Continuous Internal Assessments CIA -40%	End Semester Examination-60%
CO1	CIA 2 – Test 1: MCQ's, Quiz test or subjective	Written Exam
CO2	CIA 1 - Subjective	
CO3		
CO4	CIA 2 – Test 2: MCQ's or Presentation	

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

P. V. Sudha
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

CBCS 2026

b) Model Question Paper – End Semester Exam Theory

NATURAL LANGUAGE PROCESSING

Course Code: P26/CDS/DSC/202 **MAX MARKS: 60**
Credits: 4 **TIME: 2 ½ Hours**

Note: This question paper consists of Section A and B. The answer to Section A & B must be written in the answer book given.

SECTION – A (Long Essay Type)

Answer ALL questions: **Marks: 4 x 10 = 40**

1. a. Describe dependency grammars and constituency grammars in NLP, and explain how they represent sentence structure.

b. For the sentence: "She enjoys reading books", explain its structure using both dependency grammar and constituency grammar representations.

OR

2. a. Explain the concept of regular expressions and their role in text processing.

b. Given the text: "Contact us at info123@gmail.com or support_45@yahoo.com", explain how regular expressions can be used to identify and extract email addresses.

3. a. Explain the concept of automated text classification and describe the text classification blueprint, including steps such as text normalization and feature extraction.
b. Given a small set of sentences related to sports and politics, demonstrate how you would preprocess the text and represent it using the Bag of Words model.

OR

4. a. Explain feature extraction techniques in text classification including Bag of Words, TF-IDF, and word vectorization methods.

b. Given the sentences: "NLP is interesting" and "NLP is powerful," explain how TF-IDF values are computed and how they help distinguish documents.

5. a. Explain topic modeling techniques, including latent semantic indexing, latent Dirichlet allocation, and non-negative matrix factorization.
b. Given product review sentences, demonstrate how topics can be extracted using one topic modeling approach.

OR

6 a. Describe key phrase extraction techniques in text summarization, including collocations and weighted tag-based phrase extraction.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

b. Given the sentence: "Natural Language Processing enables machines to understand human language efficiently," identify key phrases using collocations and weighted tags.

7. a. Explain WordNet and its role in NLP, including the concepts of synsets and semantic relations.
b. Given the words "bank" (financial institution) and "bank" (river edge), demonstrate how WordNet can be used to identify synsets and perform word sense disambiguation.

OR

8. a. Explain named entity recognition and its role in extracting structured information from unstructured text.
b. Given the sentence: "Apple released the new iPhone in California," demonstrate how named entities can be identified and categorized.

SECTION –B (Short Essay Type)

II. Write short notes on any **FIVE** of the following: **Marks: 5 x 4 = 20**

9. Define tokenization and explain its types in NLP.
10. Describe the use of regular expressions in text cleaning and preprocessing.
11. Represent the sentences "NLP is fun" and "Machine learning is fun" using the Bag of Words model.
12. Compute TF-IDF values for the words in the sentences "AI is the future" and "AI is powerful."
13. Describe how Cosine similarity and Euclidean distance can be used to compare documents.
14. Explain Ward's Agglomerative Hierarchical clustering and its use in clustering textual data.
15. Analyze the importance of named entity recognition in extracting meaningful information from text.
16. Explain the process of word sense disambiguation and its role in resolving ambiguity in sentences.

P.V. Sridhar
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

Question Paper format – Blooms Taxonomy Level

SECTION A - INTERNAL CHOICE M = 40 M			4Q X 10	
Question Number	**Module Covered**	**Question**	**CO**	**BTL (Blooms Taxonomy Level)**
1	Module 1	Explain the process of text normalization in NLP. Discuss its importance and list the typical steps involved.	CO 1	Level II
2	Module 1	Define regular expressions and describe their role in text processing. Give two examples of patterns and explain what they match.	CO 1	Level II
3	Module 2	Compare dependency-based parsingandconstituency-based parsingin NLP. Include examples of each.	CO 2	Level III
4	Module 2	Explain the Bag-of-Words and TF-IDF models for text feature extraction. How are they used in text classification?	CO 2	Level III
5	Module 3	Explain Latent Dirichlet Allocation (LDA) and its application in topic modeling. Illustrate with a simple example.	CO 3	Level III
6	Module 3	Discuss text similarity measures. Compare Cosine similarity and Euclidean distance in document analysis.	CO 3	Level III
7	Module 4	Explain semantic relationships in NLP such as synonyms, antonyms, hypernyms, and hyponyms. How does word sensedisambiguation help in text understanding?	CO 4	Level IV
8	Module 4	Compare supervisedand lexicon-based sentiment analysis methods. Give examples of when each approach is preferable.	CO 4	Level IV

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

P.V Sudha

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

SECTION B - ANSWER ANY 5 OUT OF 8 = 20 M (To compulsorily have ONE question from each module)				**5Q X 4 M**
9	Module 1	Define tokenization and explain its types in NLP.	CO 1	Level II
10	Module 1	Give two examples of string operations or methods in Python used for text pre-processing.	CO 1	Level II
11	Module 2	What is POS tagging? Name any two POS taggers.	CO 2	Level III
12	Module 2	Differentiate between Bag-of-Words and TF-IDF in one sentence each.	CO 2	Level III
13	Module 3	What is Latent Semantic Analysis (LSA)? Mention one application.	CO 3	Level III
14	Module 3	Name any two text similarity measures and briefly state when each is used.	CO 3	Level III
15	Module 4	What is a synonym and antonym? Give one example each.	CO 4	Level IV
16	Module 4	Name any two lexicon-based sentiment analysis techniques used in NLP.	CO 4	Level IV

P. V. Subbaiah
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

c) Question Paper Blueprint

Modules	Hours Allotted in the Syllabus	CO Addressed	Section A (No. of Questions)	Total Marks	Section B (No. of Questions)	Total Marks
1	15	CO-1	2	4x10=40	8 (By taking two questions from each Module)	5x4=20
2	15	CO-2	2			
3	15	CO-3	2			
4	15	CO-4	2			

9. CO-PO Mapping

CO	PO	Cognitive Level	Classroom sessions (hrs)
1	1, 2	Understand	15
2	1,2	Apply	15
3	1, 2	Apply	15
4	1, 4	Analyze	15

P.V. Sudha
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

NATURAL LANGUAGE PROCESSING
Practical Syllabus

1. Course Description

Programme:	**M. Sc**	**Max. Hours:**	**40**
Course Code:	**P26/CDS/DSC/202/P**	**Hours per week:**	**2**
Type of Course:	**DSC**	**Max. Marks:**	**50**
No. of Credits:	**2**		

2. Course Objectives

1. Develop practical NLP skills using Python.
2. Apply NLP to Indian and global text datasets.

3. Course Outcomes

After the successful completion of the course, the student will be able to:

CO1: Pre-process and analyze text data effectively.

CO 2: Build end-to-end NLP pipelines for real-world tasks.

4. Course Content

Regular Expressions and Text Processing

1. Social Media Text Pre-processing: Clean and normalize Instagram captions and comments (including hashtags, mentions, emojis, URLs), and perform tokenization, stop word removal, and lemmatization.
2. News Article Regex Extraction: Extract dates, names, locations, and email addresses from Instagram captions or comments using Python regex.
3. Multilingual Text Cleaning: Pre-process Hindi or Tamil Instagram captions, handling Sandhi, compound words, punctuation, and emojis.

Suggested Datasets - Instagram captions/comments, Twitter tweets

POS Tagging and Text Classification

1. POS Tagging on Instagram Posts: Build a POS tagging pipeline for English and Indian language Instagram captions and analyze syntactic patterns.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

2. Instagram Comment Classification: Classify comments into categories like positive, negative, or neutral using TF-IDF and Naïve Bayes.
3. Hashtag-based Topic Classification: Use hashtags from Instagram posts to classify posts into categories such as travel, food, fashion, or fitness.

Suggested Datasets - IMDb reviews, Amazon product reviews, BBC News

Feature Extraction

1. Topic Modeling on Instagram Captions: Apply LDA or NMF on captions from public Instagram pages to identify key topics or trends.
2. Text Summarization of Instagram Captions: Generate short summaries of multiple captions from a brand page or influencer account using TextRank or LSA.
3. Clustering Instagram Posts: Cluster Instagram captions or comments based on content similarity using Cosine similarity and K-Means clustering to find related post groups.

Suggested Datasets - Wikipedia articles, Reddit comments, Amazon reviews

Semantic and Sentiment Analysis

1. Semantic Analysis of Instagram Text: Analyze synonyms, antonyms, hypernyms, and word senses in Instagram captions using WordNet and handle multilingual text.
2. Sentiment Analysis of Instagram Comments: Compare lexicon-based (VADER, SentWordNet) and supervised ML approaches to classify comment sentiments.
3. Named Entity Recognition on Instagram Posts: Extract entities like brands, locations, or influencers from Instagram captions and comments using NER libraries.

Suggested Datasets - WordNet, VADER/SentWordNet, Instagram or IMDb comments

Mini-project on a case study of a chosen sample dataset (e.g., Instagram posts, comments, or any real-world text corpus).

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

P.V. Subbaiah

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

5. Model Question Paper – End Semester Exam Practical

NATURAL LANGUAGE PROCESSING

Programme : M.Sc. Data Science
Course Code : P26/CDS/DSC/202/P
Type of Course: DSC
No. of credits : 2

Duration: 2 Hours
Max. Marks: 50

Answer the following

1. Implement end-to-end NLP tasks using Python and relevant libraries (NLTK, spaCy, Scikit-learn, Gensim, VADER, etc.).
2. Demonstrate a mini-project case study using a sample dataset (e.g., Instagram posts/comments, IMDb reviews, or any real-world text corpus).

Prepared by	Checked & verified by	Approved by
Ms. Padmashree Teaching faculty	Dr. Sr. Sujatha Yeruva HoD	Prof. Uma Joseph Principal

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

SEMESTER - II

BIG DATA ANALYTICS AND DISTRIBUTED SYSTEMS

1. Course Description

Programme:	**M.Sc. Data Science**	**Max. Hours:**	**60**
Course Code:	**P26/CDS/DSC/203**	**Hours per week:**	**4**
Type of Course:	**DSC**	**Max. Marks:**	**100**
No. of Credits:	**4**		

2. Course Objectives

- Explaining the basic concepts, evolution, and applications of Big Data.
- Describe the Hadoop ecosystem components such as HDFS, MapReduce, YARN, Hive, and Pig.
- Develop simple MapReduce programs for Big Data processing.
- Examine NoSQL models and analytical tools used in Big Data analytics.

3. Course Outcomes

After the successful completion of the course, the student will be able to:

CO1: Define and recall the fundamental concepts, evolution, and key applications of Big Data.. (L II).

CO2: Analyze the Hadoop ecosystem by differentiating the roles and interactions of HDFS, MapReduce, YARN, Hive, and Pig in Big Data processing. (LIV)

CO3: Apply Hadoop frameworks to develop, execute and analyze the performance and workflow of simple MapReduce programs for processing large datasets and related tools (Hive and Pig). (L III, IV)

CO4: Analyze and evaluate the suitability of NoSQL databases and analytical tools for specific Big Data applications, data models and Big Data analytical tools based on their structure and use cases. (L IV, V)

P.V. Sudha
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

4. Course Content

MODULE I: **15 Hours**

Introduction to Big Data

Overview of Big Data: Introduction to Big Data, Evolution of Big Data, Structuring Big Data, Elements of Big Data, Big Data Analytics. Exploring the Use of Big Data in Business Context: Use of Big Data in Social Networking, Use of Big Data in Preventing Fraudulent Activities, Use of Big Data in Detecting Fraudulent Activities in Insurance Sector, Use of Big Data in Retail Industry. Introducing Technologies for Handling Big Data: Distributed and Parallel Computing for Big Data, Introducing Hadoop. Understanding Hadoop Ecosystem: Hadoop Ecosystem, HDFS, Map Reduce, Hadoop YARN, HBase, Hive, Pig and Pig Latin, Sqoop, Zoo Keeper, Flume, Oozie.

MODULE II: **15 Hours**

Map Reduce Fundamentals and HBase

Understanding Map Reduce Fundamentals and HBase: The Map Reduce Framework, Techniques to Optimize Map Reduce Jobs, Role of HBase in Big Data Processing. Storing Data in Databases and Data Warehouses: RDBMS and Big Data, Non- Relational Database, Integrating Big Data with Traditional Data Warehouses, Big Data Analysis and Data Warehouse, Changing Deployment Models in Big Data Era. Processing Your Data with Map Reduce: Developing Simple Map Reduce Application, Points to Consider while Designing Map Reduce. Customizing Map Reduce Execution: Controlling Map Reduce Execution with Input Format, Reading Data with Custom Record Reader, Organizing Output Data with Output Formats, Customizing Data with Record Writer, Optimizing Map Reduce Execution with Combiner.

MODULE III: **15 Hours**

YARN, Hive and Pig

Understanding Hadoop YARN Architecture: Introduction YARN, Advantages of YARN, YARN Architecture, Working of YARN. Exploring Hive: Introducing Hive, Getting Started with Hive, Hive Services, Data Types in Hive, Built-In Functions in Hive, Hive DDL, Data Manipulation in Hive, Data Retrieval Queries, Using JOINS in Hive. Analyzing Data with Pig: Introducing Pig, Running Pig, Getting Started with Pig Latin, Working with Operators in Pig, Working with Functions in Pig, Debugging Pig, Error Handling in Pig.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

MODULE IV: **15 Hours**

Oozie, Analytical Approaches and Tools to Analyze Data

Using Oozie: Introducing Oozie, Installing and Configuring Oozie, Understanding the Oozie Workflow, Simple Application. NoSQL Data Management: Introduction to NoSQL, Types of NoSQL Data Models, Schema-Less Databases, Materialized Views, Distributed Models, Sharding, Map Reduce Partitioning and Combining, Composing Map Reduce Calculations. Understanding Analytics and Big Data: Comparing Reporting and Analysis, Types of Analytics, Developing an Analytic Team. Analytical Approaches and Tools to Analyze Data: Analytical Approaches, History of Analytical Tools, Introducing Analytical Tools, Comparing Various Analytical Tools.

5. Reference Books

1. DT Editorial Services. (2016). Big data: Black book. Dreamtech Press.
2. White, T. (2015). Hadoop: The definitive guide (4th ed.). O'Reilly Media.
3. S, R., & Vijaya Lakshmi, M. (2017). Big data analytics. McGraw Hill Education.
4. Ohlhorst, F. J. (2013). Big data analytics: Turning big data into big money. Wiley.
5. Li, K.-C., Jiang, H., Yang, L. T., & Cuzzocrea, A. (2015). Big data: Algorithms, analytics, and applications. CRC Press.

P. V. Sudha
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

M. Lijaltha Reena
HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

6. Syllabus Focus

a) Relevance to Local, Regional, National and Global Development Needs

S. No	Student Centric Methods Adopted	Type/Description of Activity
1.	National	Understand Bid data analytics to use in exams like **GATE/UGC-NET.**
2.	Global	Prepares professionals to address global challenges using Big Data technologies and advanced analytics.

b) Components on Skill Development/Entrepreneurship Development/Employability

SD/ED/EMP	Syllabus Content	Description of Activity
Skill Development	Module I & II	Practical Hands-on.
Employability	Module I, II, III & IV	Develops skills in Big Data tools, Hadoop ecosystem, and data analytics to enhance career opportunities in data-driven industries.

7. Pedagogy

S. No	Student Centric Methods Adopted	Type / Description of Activity
1.	Participative Learning	Presentations
2.	Experiential Learning	Quiz
3.	Problem solving	Group discussions and skill activities

P.V. Sudha
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

M. Lijaltha Reena
HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

8. Course Assessment Plan

a) Weightage of Marks in Continuous Internal Assessments and End Semester Examination

CO	Continuous Internal Assessments CIA -40%	End Semester Examination-60%
CO1	CIA 2 – Test 1: MCQ's, Quiz test or subjective	Written Exam
CO2	CIA 1 – Subjective	
CO3		
CO4	CIA 2 – Test 2: MCQ's or Presentation	

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University

CBCS 2026

b) Model Question Paper – End Semester Exam Theory

BIG DATA ANALYTICS AND DISTRIBUTED SYSTEMS

Course Code: P26/CDS/DSC/203 **MAX MARKS: 60**
Credits: 4 **TIME: 2 ½ Hours**

Note: This question paper consists of Section A and B. The answer to Section A & B must be written in the answer book given.

SECTION – A (Long Essay Type)

Answer ALL questions: **Marks: 4 x 10 =40**

1. Explain the concept of Big Data. Describe its elements. Summarize the role of Big Data analytics in modern organizations.

OR

2. Explain and illustrate the Hadoop ecosystem and its major components with suitable examples and with a diagram.
3. Analyze the MapReduce framework in detail. Discuss its working mechanism, phases and techniques used to optimize MapReduce jobs.

OR

4. Compare and analyze the relationship between Big Data technologies and traditional databases with suitable examples
5. Explain and examine the Hadoop YARN architecture. Evaluate its advantages in Big Data processing.

OR

6. Apply the concepts of Hive architecture to demonstrate how data is organized and processed in a Hive-based system. Use Hive Query Language (HQL) DDL operations to create and manage database structures in a practical example.
7. Describe and analyze the Oozie workflow scheduler. Explain its architecture and workflow management in Hadoop.

OR

8. Evaluate NoSQL data management. Compare different types of NoSQL analytical approaches.

P.V. Sudha
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

SECTION –B (Short Essay Type)

II. Write short notes on any **FIVE** of the following: **Marks: 5 x 4 = 20**

9. Define Big Data and explain its key characteristics (5Vs).
10. Explain the concepts of distributed and parallel computing and their role in processing Big data.
11. Analyze how input formats control MapReduce execution, and examine their impact on data splitting, processing efficiency, and overall performance in Big Data applications.
12. Explain the role of HBASE.
13. What is Hadoop YARN? Mention its main components.
14. Explain basic features of HIVE and uses in Hadoop.
15. Analyze the differences between reporting and analysis, and explain types of analytics.
16. Describe MapReduce partitioning and combining.

P. V. Sudeer
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

M. Lujaltha Reena
HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

CBCS 2026

Question Paper format – Blooms Taxonomy Level

SECTION A - INTERNAL CHOICE 4Q X 10 M = 40 M				
Question Number	**Module Covered**	**Question**	**CO**	**BTL (Blooms Taxonomy Level)**
1	Module 1	Explain the concept of Big Data. Describe its elements. Summarize the role of Big Data analytics in modern organizations.	CO 1	L II
2	Module 1	Explain and illustrate the Hadoop ecosystem and its major components with suitable examples and with a diagram.	CO 2	L II
3	Module 2	Analyze the MapReduce framework in detail. Discuss its working mechanism, phases and techniques used to optimize MapReduce jobs.	CO 2	L IV
4	Module 2	Compare and analyze the relationship between Big Data technologies and traditional databases with suitable examples	CO 2	L IV
5	Module 3	Explain and examine the Hadoop YARN architecture and its advantages in Big Data processing.	CO 3	L IV
6	Module 3	Apply the concepts of Hive architecture to demonstrate how data is organized and processed in a Hive-based system. Use Hive Query Language (HQL) DDL operations to create and manage database structures in a practical example.	CO 3	L III
7	Module 4	Describe and analyze the Oozie workflow scheduler. Explain its architecture and workflow management in Hadoop.	CO 4	L IV
8	Module 4	Evaluate NoSQL data management. Compare different types of NoSQL analytical approaches.	CO 4	L V

P.V. Sudha
PROFESSOR
Department of Computer Science & Engineering (A)
University College of Engineering
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

SECTION B - ANSWER ANY 5 OUT OF 8 (To compulsorily have ONE question from each module)		5Q X 4 M = 20 M		
9	Module 1	Define Big Data and explain its key characteristics	CO 1	L II
10	Module 1	Explain the concepts of distributed and parallel computing and their role in processing Big Data.	CO 1	L II
11	Module 2	Analyze how input formats control MapReduce execution, and examine their impact on data splitting, processing efficiency, and overall performance in Big Data applications.	CO 2	L IV
12	Module 2	Explain the role of HBASE.	CO 2	L II
13	Module 3	What is Hadoop YARN? Mention its main components.	CO 3	L II
14	Module 3	Explain basic features of HIVE and uses in Hadoop.	CO 3	L II
15	Module 4	Analyze the differences between reporting and analysis, and explain various types of analytics.	CO 4	L IV
16	Module 4	Describe MapReduce partitioning and combining.	CO 4	L IV

c) Question Paper Blueprint

Modules	Hours Allotted in the Syllabus	CO Addressed	Section A (No. of Questions)	Total Marks	Section B (No. of Questions)	Total Marks
1	15	CO-1	2	4x10=40	8 (By taking two questions from each Module)	5x4=20
2	15	CO-2	2			
3	15	CO-3	2			
4	15	CO-4	2			

9. CO-PO Mapping

CO	PO	Cognitive Level	Classroom sessions (hrs)
1	1, 2	Understand	15
2	1,2	Analyze	15
3	1, 2	Apply	15
4	1, 4	Analyze	15

P. V. Sudha
PROFESSOR
Department of Comp. Sc. & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN
Begumpet, Hyderabad-500 016.

BIG DATA ANALYTICS AND DISTRIBUTED SYSTEMS

Practical Syllabus

1. Course Description

Programme: M.Sc. Data Science
Course Code: P26/CDS/DSC/203/P
Type of Course: DSC
No. of Credits: 2

Max. Hours: 60
Hours per week: 2
Max. Marks: 50

2. Course Objectives

- To understand the installation, configuration, and operation of the Hadoop ecosystem including HDFS, MapReduce, Hive, and Pig.
- To develop skills in processing and analyzing large datasets using MapReduce programs, Hive queries, and Pig Latin scripts.

3. Course Outcomes

After the successful completion of the course, the student will be able to:

CO1: Apply Hadoop ecosystem tools (HDFS to manage and process large-scale data
CO 2: Analyze datasets by developing MapReduce programs and writing programs.

4. Course Content

1. Installing Hadoop in different modes: standalone, Pseudo distributed.
2. Perform some tasks by using web-based tools of Hadoop system.
3. Implement the following file management tasks in Hadoop:
   Adding file and directories & Creating file, retrieving file and deleting files
4. Write a Map Reduce program for basic word count.
5. Write a Map Reduce program for sorting text data.
6. Write a Map Reduce program for analyzing student report.
7. Write a MapReduce program for mining weather data.
8. Installing and running Hive, practice some Hive commands.
9. Using Hive; create, insert, update, alter, delete, and drop the tables
10. Using Hive; query the data from the database tables.
11. Using Hive; create views, use functions, create indexes for the database tables.
12. Installing and running Pig, practice some Pig commands.
13. Write Pig Latin scripts using evaluate functions to analyze your data.
14. Write Pig Latin scripts using math functions to analyze your data.
15. Write Pig Latin scripts using string functions to analyze your data.

P. V. Sridevi
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

5. Model Question Paper – End Semester Exam Practical

BIG DATA ANALYTICS AND DISTRIBUTED SYSTEMS

Programme: M.Sc. Data Science
Course Code: P26/CDS/DSC/203/P **Duration: 2 Hours**
Type of Course: DSC **Max. Marks: 50**
No. of credits: 2

Answer any ONE of the Following

1. Write a Map Reduce program for basic word count.
2. Write a Map Reduce program for sorting text data.
3. Using Hive create, insert, update the table.
4. Write a Map Reduce program for analyzing student report.
5. Using Hive; create, insert, update, alter, delete, and drop the tables.

Prepared by	Checked &Verified by	Approved by
Ms. Jyothi Reddy Teaching faculty	Dr. Sr. Sujatha Yeruva HoD	Prof. Uma Joseph Principal

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

SEMESTER –II

DEEP LEARNING

1. Course Description

Programme: M.Sc. Data Science
Course Code: P26/CDS/DSC/204
Course Type: DSC
No. of credits: 4

Max. Hours: 60
Hours per week: 4
Max. Marks: 100

2. Course Objectives

- To understand the fundamentals of neural networks and deep learning
- To develop deep learning models for image processing and solving complex tasks.
- To explore advanced deep learning techniques for text analysis.
- To design, train, evaluate, and optimize deep learning models using various Keras APIs and tools.

3. Course Outcomes

On completion of the course, the student will be able to:

CO1: Understand and implement neural network models using Keras including perceptron, multilayer perceptron, and optimization techniques. (L II)
CO2: Apply CNNs and GANs for image recognition and generation tasks. (LIII)
CO3: Apply word embeddings and recurrent neural networks for NLP applications. (LIII)
CO4: Analyze advanced deep learning systems using Keras functional API. (LIV)

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

4. Course Content

MODULE I: **15 Hours**

Foundations of Neural Networks, Working with Keras

Perceptron, The first example of Keras code, Multilayer perceptron, Problems in training the perceptron and a solution, Activation functions, sigmoid, ReLU, One-hot encoding, Defining a simple neural net in Keras, Running a simple Keras net and establishing a baseline, Improving the simple net in Keras with hidden layers and with dropout, Testing different optimizers in Keras, Increasing the number of epochs, Controlling the optimizer learning rate, Increasing the number of internal hidden neurons, Increasing the size of batch computation, Summarizing the experiments run for recognizing handwritten charts, Adopting regularization for avoiding overfitting Hyperparameters tuning, Predicting output, A practical overview of backpropagation, Towards a deep learning approach.Installing and Configuring Keras, Keras API, Keras architecture, tensor, Composing models in Keras, Overview of predefined neural network layers, predefined activation functions, metrics, optimizers. Saving and loading the weights and the architecture of a model, Checkpointing, Using TensorBoard and Keras, Quiver and Keras.

MODULE II: **15 Hours**

Deep Learning with Convolutional Networks, Generative Adversarial Networks

Deep convolutional neural network, Local receptive fields, Shared weights and bias, Pooling layers, Max and Average pooling, ConvNets, LeNet and LeNet code in Keras, Understanding the power of deep learning, Recognizing and Improving the CIFAR-10 images with deep learning, Very deep convolutional networks for large-scale image recognition, Recognizing cats with a VGG-16 net, Utilizing Keras built-in VGG-16 net module. Overview of Generative Adversarial Networks and GAN applications, Deep convolutional generative adversarial networks, Keras adversarial GANs for forging MNIST and CIFAR, WaveNet.

MODULE III: **15 Hours**

Word Embeddings and Recurrent Neural Network

Distributed representations, word2vec, skip-gram and CBOW word2vec models, Extracting word2vec embeddings, Using third-party implementations of word2vec, Exploring GloVe, Using pre-trained embeddings and learning from scratch, Fine-tuning learned embeddings from word2vec and GloVe, Look up embeddings. Recurrent Neural Network: SimpleRNN cells, SimpleRNN with Keras, RNN topologies, Vanishing and exploding gradients, Long short term memory, LSTM with Keras, sentiment analysis, Gated recurrent unit, GRU with Keras, POS tagging, Bidirectional and Stateful RNNs.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

MODULE IV: **15 Hours**

Additional Deep Learning Models, AI Game_Playing

Keras functional API, Regression networks, Keras regression example, Unsupervised learning: autoencoders, composing deep networks, Customizing Keras, Keras example using the lambda layer and building a custom normalization layer, Generative models. Reinforcement learning, Maximizing future rewards, Q-learning, The deep Q-network as a Q-function, Balancing exploration with exploitation, Keras deep Q-network for catch.

5. References

1. Deep Learning with Keras, by Antonio Gulli, Sujit Pal, Packt Publishing, 2017.
2. Deep Learning by Ian Goodfellow, Yoshua Bengio and Aaron Courville published by MIT Press.
3. Neural Networks and Deep Learning: A Textbook, by Charu C. Aggarwal, Springer.
4. Deep Learning (The MIT Press Essential Knowledge series) by John D. Kelleher
5. Deep Learning From Scratch: Building with Python from First Principles by Seth Weidman, O`Reilley.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

5.Syllabus Focus

a) Relevance to Local, Regional, National and Global Development Needs

S.No	Local /Regional/National /Global Development Needs	Relevance
1	National Development	Deep Learning enable intelligent automation, data analysis, and innovative solutions that support technological advancement and national development
2	Global Development	Deep Learning provide global solutions for sectors like healthcare, agriculture, finance, and communication through advanced data-driven technologies.

b) Components on Skill Development/Entrepreneurship Development/Employability

SD/ED/EMP	Syllabus Content	Description of Activity
SD	Modules I & II	Hands- on Practical's
EMP	Modules III & IV	Mini Project

6. Pedagogy

S. No	Student Centric Methods Adopted	Type / Description of Activity
1.	Participative	Seminars
2.	Experimental	Practical demonstrations on Deep Learning techniques
3.	Problem solving	Programming assignments

P.V. Subbaiah
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

CBCS 2026

7. Course Assessment Plan

a) Weightage of Marks in Continuous Internal Assessments and End Semester Examination

CO	Continuous Internal Assessments CIA - 40%	End Semester Examination- 60%
CO1	CIA 2 – Test 1: MCQ's, Quiz test or subjective	Written Exam
CO2	CIA 1 - Subjective	
CO3		
CO4	CIA 2 – Test 2: MCQ's or Presentation	

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

P. V. Rao
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

b) Model Question Paper- End Semester Exam

Deep Learning

MODEL QUESTION PAPER THEORY

Course Code: P26/CDS/DSC/204 **Max Marks: 60**
Credits: 4 **Time: 2 ½ Hrs.**

I: Answer the following: **4x 10 = 40**

1. Explain the Perceptron model and discuss the problems in training a perceptron. How are these problems solved using multilayer perceptrons?

OR

2. Explain the Keras architecture and describe the steps involved in defining, training, and saving a neural network model in Keras.
3. Explain the architecture of Convolutional Neural Networks including local receptive fields, shared weights, and pooling layers.

OR

4. Describe Generative Adversarial Networks (GANs) and explain their working with suitable examples.
5. Explain Word2Vec models (CBOW and Skip-gram) and discuss how word embeddings are used in NLP tasks.

OR

6. Demonstrate the architecture and working of Recurrent Neural Networks (RNNs). Discuss the role of LSTM networks in solving the vanishing gradient problem.
7. Explain the Keras Functional API and discuss how it is used to build complex deep learning models.

OR

8. Describe Reinforcement Learning and Q-learning. Describe how a Deep Q-Network (DQN) works in game playing.

II: Answer any Five: **5 x 4 = 20**

9. Define Perceptron and mention its limitations.
10. Discuss One-Hot Encoding. Explain its purpose in neural networks.
11. Describe Pooling Layer in Convolutional Neural Networks?
12. Explain applications of Generative Adversarial Network (GAN).
13. Discuss about genism library
14. Differentiate between LSTM and GRU.
15. Explain in detail Autoencoders.
16. Explain any two examples on Reinforcement Learning.

P.V. Sridevi
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

DEPARTMENT OF COMPUTER SCIENCE, ST. FRANCIS COLLEGE FOR WOMEN, HYDERABAD

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

CBCS 2026

Question Paper format – Blooms Taxonomy Level

SECTION A - INTERNAL CHOICE				**4Q X 10 M = 40 M**
Question Number	**Module Covered**	**Question**	**CO**	**BTL (Blooms Taxonomy Level)**
1	Module 1	Explain the Perceptron model and discuss the problems in training a perceptron. How are these problems solved using multilayer perceptrons?	CO 1	L II
2	Module 1	Explain the Keras architecture and describe the steps involved in defining, training, and saving a neural network model in Keras.	CO 1	L II
3	Module 2	Explain the architecture of Convolutional Neural Networks including local receptive fields, shared weights, and pooling layers.	CO 2	L II
4	Module 2	Describe Generative Adversarial Networks and explain their working with examples.	CO 2	L II
5	Module 3	Explain Word2Vec models and discuss how word embeddings are used in NLP tasks.	CO 3	L II
6	Module 3	Demonstrate the architecture and working of Recurrent Neural Networks. Discuss the role of LSTM networks in solving the vanishing gradient problem.	CO 3	LIII
7	Module 4	Explain the Keras Functional API and discuss how it is used to build complex deep learning models.	CO 4	L II
8	Module 4	Describe Reinforcement Learning and Q-learning. Describe how a Deep Q-Network (DQN) works in game playing.	CO 4	L II
SECTION B - ANSWER ANY 5 OUT OF 8 (To compulsorily have ONE question from each module)				**5Q X 4 M = 20 M**
9	Module 1	Define Perceptron and mention its limitations.	CO 1	L I
10	Module 1	Discuss One-Hot Encoding.Explain its purpose in neural networks.	CO 1	L II
11	Module 2	Describe Pooling Layer in CNN'S	CO 2	L II
12	Module 2	Explain applications of GAN.	CO 2	L II
13	Module 3	Discuss about genism library	CO 3	L II
14	Module 3	Differentiate between LSTM and GRU.	CO 3	L IV
15	Module 4	Explain in detail Autoencoders.	CO 4	L II
16	Module 4	Explain any two examples on Reinforcement Learning.	CO 4	L II

DEPARTMENT OF COMPUTER SCIENCE, ST. FRANCIS COLLEGE FOR WOMEN, HYDERABAD

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

c)Question Paper Blueprint

Modules	Hours Allotted in the Syllabus	CO Addressed	Section A (No. of Questions)	Total Marks	Section B (No. of Questions)	Total Marks
1	15	CO-1	2	4x10=40	8 (By taking two questions from each Module)	5x4=20
2	15	CO-2	2			
3	15	CO-3	2			
4	15	CO-4	2			

NOTE: From
Section-A any 4 questions can be answered (INTERNAL CHOICE).
Section-B any 5 questions can be answered. (EXTERNAL CHOICE)

9. CO-PO Mapping

CO	PO	Cognitive Level	Classroom sessions (hrs)
1	1	Understand	15
2	1,2	Apply	15
3	1, 2	Apply	15
4	1, 2	Analyze	15

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

P.V. Sridevi
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

CBCS 2026

Deep Learning Practical

1. Course Description

Programme: M.Sc. Data Science
Course Code: P26/CDS/DSC/204/P
Course Type: DSC
No. of credits: 2

Max. Hours: 60
Hours per week: 4
Max. Marks: 50

2. Course Objective

1. To understand the fundamentals of Neural Networks and Deep Learning and implement models using the Keras framework.
2. To develop and apply deep learning techniques such as CNNs, RNNs, word embeddings, GANs, and reinforcement learning for solving real-world problems.

3. Course Outcomes

CO1: Analyzing deep learning models using Keras, including neural networks, convolutional networks, and recurrent networks.
CO2: Applying deep learning techniques for image recognition, natural language processing, and AI-based game playing.

4. Course Content

1. Neural Networks, Working with Keras:
   a. Running a simple Keras net and establishing a baseline.
   b. Installing and Configuring Keras.
2. Deep Learning with ConvNets, Generative Adversarial Networks:
   a. Recognizing and improving the images with deep learning.
   b. Working with Keras adversarial Generative Adversarial Networks.
3. Word Embeddings and Recurrent Neural Networks:
   a. Extracting word2vec embeddings.
   b. Implementing Simple Recurrent Neural Networks with Keras.
4. Additional Deep Learning Models, AI Game Playing:
   a. Regression using Keras.
   b. Keras using the lamda layer.
5. Mini Project.

P.V. Sridhar
PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

CBCS 2026

5. Model Question Paper

Deep Learning Practical

COURSE CODE: P26/CDS/DSC/204/P **Max.Marks: 50**
Duration: 2 Hrs.

Answer any one of the following.

1.Implement Handwritten digit recognition with keras using MNIST dataset.
2.Implement Image Classification Using CNN by considering CIFAR-10 Dataset.
3.Using Keras implement Generative Adversarial Networks (GAN)
4.Implement word2vec using gensim library.
5.Write a program to implement Simple Recurrent Neural Network.

Prepared by	Checked & verified by	Approved by
Ms. Khalida Tabassum	**Dr. Sr. Sujatha Yeruva**	**Dr. Uma Joseph**
Teaching Faculty	**Head of the Department**	**Principal**

HOD Computer Science
ST FRANCIS COLLEGE FOR WOMEN,
Begumpet, Hyderabad-500 016.

PROFESSOR
Department of Computer Science & Engineering
University College of Engineering (A)
Osmania University,
Hyderabad-500 007.