PVP SIDDHARTHA INSTITUTE OF TEHNOLOGY, KANURU, VIJAYAWADA
(AUTONOMOUS)

INFORMATION TECHNOLOGY

MINNING MASSIVE DATASETS

Course Code	19IT4602F	**Year**	III	**Semester**	II
Course Category	PE	**Branch**	IT	**Course Type**	Theory
Credits	3	**L-T-P**	3-0-0	**Prerequisites**	DBMS, DS
Continuous Internal Evaluation :	30	**Semester End Evaluation:**	70	**Total Marks:**	100

Course Outcomes		**Blooms Taxonomy Level**
Upon successful completion of the course, the student will be able to		
CO1	Understating the fundamentals of concepts Distributed file systems, Data Streams and Social Networks	L2
CO2	Determine the Concepts of Data Streams and Link Analysis	L3
CO3	Compare and Contrast the Concepts Link Analysis	L4
CO4	Deduce Graph concepts for Social Networks	L4

Contribution of Course Outcomes towards achievement of Program Outcomes & Strength of correlations (3:Substantial, 2: Moderate, 1:Slight)

	PO 1	**PO 2**	**PO 3**	**PO 4**	**PO 5**	**PO 6**	**PO 7**	**PO 8**	**PO 9**	**PO 10**	**PO 11**	**PO 12**	**PSO 1**	**PSO 2**
CO 1	2												1	
CO 2	2		2	2									1	
CO 3	2			2									1	
CO 4	2	2											1	

Syllabus

Unit No	**Contents**	**Mapped CO**
I	**Data Mining**: What is data Mining? Statistical Limits on Data Mining, Things Useful to Know MapReduce and the new software stack: Distributed file systems, MapReduce, Algorithms usingMap Reduce, and complexity theory for Map Reduce	**CO1**

II	**Finding similar items:** Application for near-neighbor search, shingling of documents, Similarity- preserving summaries of sets, locality-sensitive hashing for documents and distance measures	**CO1**
III	**Mining Data Streams**: The Stream Data Model, Sampling data in a stream, filtering streams and counting distinct elements in a stream	**CO1, CO2**
IV	**Link Analysis**: Page Rank, Efficient computation of Page Rank, Topic-sensitive Page Rank, Link Spam,Hubs and Authorities	**CO1, CO2, CO3**
V	**Mining Social Network Graphs**: Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities, Partitioning of graphs, Simrank and Neighborhood properties of graphs	**CO1,CO4**

Learning Recourses

Text Books

1. Jure Leskovec, Anand Rajaraman and Jeffrey D. Ullman, Mining of Massive Datasets, Third edition, 2020.

References

1. H. Garcia-Molina, J. D. Ullman, and J. Widom. Database Systems: The Complete Book. Second Edition. Pearson Prentice Hall, 2009
2. J.Lin and Ch. Dyer. Data-Intensive Text Processing with MapReduce. Morgan and Claypool Publishers, 2010 http://lintool.github.com/MapReduceAlgorithms/
3. T. Hastie, R. Tibshirani, and J. Friedman. Elements of Statistical Learning: Second Edition. Springer, 2009

e-Resources & other digital material

http://www-stat.stanford.edu/~tibs/ElemStatLearn/