# GUJARAT TECHNOLOGICAL UNIVERSITY

**Syllabus for Master of Computer Applications, 5th Semester**
**Subject Name: Big Data Analytics (BDA)**
**Subject Code: 4659306**

**With effective from academic year 2018-19**

## 1. Learning Objectives:

- To be able to understand various techniques of Big Data Analytics
- To understand how to explore and communicate data using data visualization techniques

## 2. Prerequisites: Knowledge of R/Python and Database concepts

## 3. Recommended content knowledge

| Sr. No. | Course Content | Weightage |
|---|---|---|
| 1 | **Introduction**<br>Big Data Overview, BI versus Data Science, Current analytical architecture, Drivers of Big Data, Emerging Big Data Ecosystem and a New Approach to Analytics, Key Roles for the New Big Data Ecosystem, Examples of Big Data Analytics<br><br>Data Analytics Life Cycle Overview, Phases ( Discovery, Data Preparation, Model Planning, Mode Building, Communicate Results, Operationalize)<br><br>**Case Study: Global Innovation Network and Analysis ( GINA)** | 15% |
| 2 | **Mining Relationships Among Records**<br><br>**Association Rules :** Discovering Association rules in transaction Databases, Generating Candidate Rules, The apriori algorithm, Selecting strong rules, Data Formats, The process of Rule selection, Interpreting results, Rules and chance<br><br>**Collaborating Filtering:** Data and Format, User based collaborative filtering "People like you", Item-based Collaborative Filtering, Advantages and weaknesses of Collaborative filtering, Collaborative filtering vs Association Rules<br><br>**Cluster Analysis:** Introduction, measuring distance between two records, Measuring distances between two clusters**,** Hierarchical (Agglomerative) Clustering, Non-Hierarchical Clustering: The k-Means Algorithm | **25%** |
| 3 | **Forecasting Time Series**<br><br>**Handling Time Series:** Introduction, Descriptive vs. Predictive Modeling, Popular Forecasting Methods in Business, Time Series Components, Data-Partitioning and Performance Evaluation<br><br>**Regression-Based Forecasting :** A Model with Trend, A Model with Seasonality, A Model with Trend and Seasonality, Autocorrelation and ARIMA Models<br><br>**Smoothing Methods:** Introduction, Moving Average, Simple Exponential Smoothing, Advanced Exponential Smoothing | 25% |

| | | |
|---|---|---|
| 4 | **Social Network Analysis**<br><br>Introduction, Directed vs. Undirected Networks, Visualizing and Analyzing Networks, Social Data Metrics and Taxonomy, Using Network Metrics in Prediction and Classification, Collecting Social Network Data with R, Advantages and Disadvantages | 15% |
| 5 | **Text Mining**<br><br>Introduction, The Tabular Representation of Text: Term-Document Matrix and "Bag-of-Words" , Bag-of-Words vs. Meaning Extraction at Document Level, Preprocessing the Text, Implementing Data Mining Methods ,<br><br>Example: Online Discussions on Autos and Electronics | **20%** |

## 4. Main Reference Books:

1. EMC Education Services, Data Science and Big Data Analytics, WILEY
2. Galit Shmueli, Peter C Bruce, Inbal Yahav, Nitin R Patel, Kenneth C, Linchtendahl Jr, Data Mining for Business Analytics- concepts, techniques and Application in R, WILEY

## 5. Recommended Book(s):

1. Glenn J Myatt, Wayne P Johnson, Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining, Wiley, 2nd Edition
2. Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge University Press (Wiley India) , 2nd Edition
3. Mehmed Kantardzic, Data Mining: Concetps, Models, Methods and Algorithms, Wiley-IEEE, 2nd Edition
4. Field Cady, 'The Data Science Handbook The Data Science Handbook', Wiley Publication ISBN-13: 978-8126573332
5. Han, J., Kamber, M., Pei, J. Data mining concepts and techniques. Morgan Kaufmann, 2011
6. Michael Berthold, David J. Hand, Intelligent Data Analysis, Springer, 2007
7. Vincent Granville, Developing Analytic Talent: Becoming a Data Scientist, wiley, 2014
8. John W. Foreman (Author), Data Smart: Using Data Science to Transform Information into Insight, WILEY

## 6. Chapter wise Coverage from Main Book(s):

| Unit# | Book# | Topics |
|---|---|---|
| **1** | **1** | **Chapter 1,2** |
| **2** | **2** | **Chapter 14,15** |
| **3** | **2** | **Chapter 16, 17,18** |
| **4** | **1** | **Chapter 19** |
| **5** | **1** | **Chapter 20** |

## 7. Suggested Practical

**Tool: Python, Libraries of Python like Pandas, Sci-kit Learn etc., R ( R Studio and required packages)**

**Part 1: Data Pre-processing:**

Dataset:https://www.analyticsvidhya.com/blog/2016/07/practical-guide-data-preprocessing-python-scikit-learn/

1. Download loan data set (https://www.analyticsvidhya.com/blog/2016/07/practical-guide-data-preprocessing-python-scikit-learn/) and perform following operations
   i. Write program to read dataset ( Text,CSV,JSON,XML)
   ii. Performing Data Cleaning
      a. Handling Missing Data
      b. Removing Null data
      c. Rescaling Data
   iii. Dimensionality Reduction
   iv. Encoding Data
   v. Feature Selection
   vi. Implement Principle Component Analysis,
2. Use Loan data (above) and Fit KNN model to find out accuracy of model for prediction of loan.
3. Write a python code to predict profit of hotel chain given the population of the area (city) using the data at https://docs.google.com/spreadsheets/d/1Ks20skBgEefHFU36sFqVzozoFtz2EZE2rxB_IgXOrUg/edit?usp=sharing.
4. Write a python code to predict the price of house given square feet and number of bed rooms in the house for the dataset available at https://docs.google.com/spreadsheets/d/1DHVK7gKo4TSyj7mFLwofHamj1Sl4SOZma2q51w1ZvyE/edit?usp=sharing

**Part 2: Mining Relationships among Records**

5. Implement Apriori algorithm in python to find rules which explain association between different products for given transactions at a retail store. (The data is available at https://drive.google.com/file/d/1NUXoptUlHY8z4KcFKpFA6sQN5KnWzk3p/view?usp=sharing )
6. Implement text classification using neural network in python/R on Twenty Newsgroup dataset from UCI machine learning repository.
7. Generating Association rule mining e..g "Sythentic Data on Purchase of Phone faceplate"
   a. Recommender algorithms: Generating rules for Similar Book Purchases
8. Collaborative Filtering (use movielens dataset):
   b. Find similar items by using a similarity metric
   c. For a user, recommend the items most similar to the items (s)he already likes

**Part 3: Implement Clustering**

9. Implement clustering algorithm for grouping news articles.

10. Implement unsupervised machine learning algorithm (Clustering – K Means) in python on Titanic dataset to cluster data (use Titanic dataset) by removing the class label.
11. Implement unsupervised machine learning algorithm (Clustering – K Means) in python on Breast Tumour dataset to cluster data (use Breast Tumour dataset) by removing the class label.
12. Implement unsupervised machine learning algorithm (Clustering – Hierarchical) in python on Titanic dataset to cluster data (use Titanic dataset).

**Part 4: Various types of Text Analysis**

13. For the sentiment analysis dataset given in link https://drive.google.com/file/d/1x6H7_KJjkbDrpgZFS7I2wjsZsILeSJ4S/view?usp=sharing, implement the following in python,
    d. Clean and pre-process the dataset by removing URL, removing HTML tags, handling negation words which are split into two parts, converting the words to lower cases, removing all non-letter characters
    e. Split the dataset into training and testing set
    f. Implement feature extraction technique (to convert textual data to the numeric form)
    g. Build the classification model using Logistic Regression that classifies if a given sentiment text is positive or negative
    h. Obtain the accuracy score of the built model.

14. Implement a content based recommender system in python that recommends movies that are similar to a particular movie using movielens-20m-dataset available at https://kaggle.com.

**Part 5: Advanced Data Visualization**

15. Write a program to plot Chi square distribution
16. Write a program to plot Normal distribution
17. Write a program to plot Poisson distribution
18. Write a program to plot T distribution
19. Write a program to plot Binomial Distribution
20. Write a program to plot Central limit theorem
21. Write a program to plot Uniform distribution

**Part 6: Text pre-preprocessing using Python**

**Tools: NLTK (** http://www.nltk.org/) sci-kitlearn etc.

22. Removing stop words (the most common words in a language like “the”, “a”, “on” etc.)
23. 
24. Write a python code to perform spell check (edit distance algorithm)
25. Write a python code for finding the root words ( Stemming algorithm : A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish)
26. Write a python code to implement Tokenized algorithm for text processing
27. Write a python code Part of speech (PoS) tagging

**Part 7. Desirable:**

Abstractive/ Extractive text Summarization (single document, multi document)
Time series algorithm