BUSINESS ANALYTICS & TEXT MINING MODELING USING PYTHON

PROF. GAURAV DIXIT
Department of Management
IIT Roorkee

INTENDED AUDIENCE : UG & PG engineering students: All branches, MBA students, Professionals working in or aspiring for Business Analyst, Data Analyst, Data Scientist, and Data Engineer roles.

PREREQUISITES: Relevant sessions from the courses Business Analytics & Data Mining Modelling Using R Parts I and II

INDUSTRIES APPLICABLE TO : Big Data companies, Analytics & Consultancy companies, Companies with Analytics Division.

COURSE OUTLINE :
Objective of this course is to impart knowledge on use of text mining techniques for deriving business intelligence to achieve organizational goals. Use of Python based software platform to build, assess, and compare models based on real datasets and cases with an easy-to-follow learning curve.

ABOUT INSTRUCTOR :
Prof. Gaurav Dixit is an Assistant Professor in the Department of Management Studies at the Indian Institute of Technology Roorkee. He earned his doctoral degree from the Indian Institute of Management Indore and an engineering degree from Indian Institute of Technology (BHU) Varanasi. Previously, he worked in Hewlett-Packard (HP) as software engineer, and Sharda Group of Institutions as project manager on deputation. Gaurav's research focuses on information technology (IT) strategy, electronic commerce, electronic waste, data mining, text mining, and big data analytics and provides insights on business and social value of IT. His research has appeared in quality journals & conferences, including Resources, Conservation and Recycling, Journal of Global Information Technology Management, Sustainable Production and Consumption, Journal of Information Technology Management, ICIS conference, DIGITS conference, India Finance Conference.Indian Academy of Management conference, and Academy of Management conference.

COURSE PLAN :

Week 1: Introductory overview of Text Mining
- Introductory Thoughts
- Data Mining vs. Text Mining
- Text Mining and Text Characteristics
- Predictive Text Analytics
- Text Mining Problems
- Prediction & Evaluation
- Python as a Data Science Platform Python for Analytics
- Introduction to Python Installation
- Jupyter Notebook Introduction

Week 2: Python Basics
- Python Programming Features
- Commands for common tasks and control
- Essential Python programming concepts & language mechanics Built in Capabilities of Python
- Data structures: tuples, lists, dicts, and sets

Week 3: Built in Capabilities of Python
- Functions, Namespaces, Scope, Local functions, Writing more reusable generic functions

Week 4: Built in Capabilities of Python
- Generators
- Errors & Exception Handling
- Working with file Numerical Python
- N-dimensional array objects

Week 5: Numerical Python
- Vectorized array operations
- File management using arrays
- Linear algebra operations
- Pseudo-random number generation
- Random walks Python pandas
- Data structures: Series and DataFrame

Week 6: Python pandas
- Applying functions and methods
- Descriptive Statistics
- Correlation and Covariance Working with Data in Python
- Working with CSV, EXCEL files
- Working with Web APIs

Week 7: Working with Data in Python
- Filtering out missing data, Filling in the missing data, removing duplicates
- Perform transformations based on mappings
- Binning continuous variables
- Random sampling and random reordering of rows
- Dummy variables
- String and text processing
- Regular expressions
- Categorical type Data Visualization using Python
- Matplotlib Library
- Plots & Subplots

Week 8: Text mining modeling using NLTK
- Text Corpus
- Sentence Tokenization
- Word Tokenization
- Removing special Characters
- Expanding contractions
- Removing Stopwords
- Correcting words: repeated characters
- Stemming & lemmatization
- Part of Speech Tagging
- Feature Extraction
- Bag of words model
- TF-IDF model
- Text classification problem
- Building a classifier using support vector machine