INSTITUTE OF COMPUTER SCIENCES AND INFORMATION TECHNOLOGY (ICS/IT)
FACULTY OF MANAGEMENT AND COMPUTER SCIENCES
THE UNIVERSITY OF AGRICULTURE, PESHAWAR
KHYBER PAKHTUNKHWA PAKISTAN

Phone: +92-91-9221323, Ext: 3214, Email: dicsit@aup.edu.pk, Web: www.aup.edu.pk

Program:	**MS-Data Science**
Course Title:	**Text Mining and Analytics**
Course Code:	**DS-727**
Course Hours:	**03**
Total Weeks:	**16**
Total Credit Hours:	**48**

Course Objective

Given the dominance of text information over the Internet, mining high-quality information from text has become increasingly critical. The actionable knowledge extracted from text data facilitates our life in a broad spectrum of areas, including business intelligence, information acquisition, social behaviour analysis and decision making. This course will cover essential topics in text mining, including basic natural language processing techniques, document representation, text categorization and Clustering, document summarization, sentiment analysis, social network and social media analysis, probabilistic topic models, and text visualization.

Learning Outcome

In this course, we will introduce a variety of fundamental principles, techniques and modern advances in text mining. At the end of the course, the student will be able to solve some novel text mining problems.

Course Prerequisites:

The prerequisites for this course are:

- Programing Language (Java or Python)
- Calculus
- Basic Probability Theory and Statistics

Week Wise Course Content Distribution:

Week-1

- Introduction to Text Mining
  - Difference between Data and Text Mining
  - Introduction to Major topics

Week-2

- Natural Language Processing (NLP)
  - Methods in NLP
    - Tokenization, Part-of-speech tagging chunking, syntax parsing and named entity recognition.
    - StopWords, Stemming

Week-3

- Document Representation
    - Unstructured Text Document
    - Structured Text Document
    - Matrix Representation

Week-4

- Document Similarity/Dissimilarity Algorithms
    - TF/IDF
    - Jaccard Similarity
    - Cosine Similarity
    - Euclidean Distance

Week-5

- Text Classification
    - Vector Space Model
    - Classification Algorithms
        - Naïve Bayes, k Nearest Neighbor (kNN).

Week-6

- Text Clustering
    - Connectivity-based Clustering (Hierarchical Clustering)
    - Centroid-based Clustering (k-Means Clustering)

Week-7

- Topic Modeling
    - Probabilistic Latent Semantic Indexing (pLSI)
    - Latent Dirichlet Allocation (LDA)

Week-8

- Document Summarization

Week-9

- Social media and network analysis
    - Page Rank

Week-10

- Sentiment Analysis
    - Polarity Prediction, Aspect Identification
    - Review Mining
    - Sarcasm Detection

Week-11

- Sentiment Analysis Models
    - Classification Model
    - Lexicon based Model
    - Boolean Rule-based Model

Week-12

- Text Mining in Local and Resource-Poor Languages
    - Urdu Text

    - Pashto Text

Week-13

- Sentiment Analysis in Urdu Text
    - Polarity Prediction Urdu Text
    - Review Mining Urdu Text
    - Aspect Identification Urdu Text
    - Sarcasm Detection Urdu Text

Week-14

- Document Categorization

Week-15

- Text Visualization

Week-16

- Tools for Text Analysis
    - Introduction to Python
    - Libraries for Text Mining and Analysis

Total Marks: 100

Recommended Books and Materials:

1. Mining Text Data. Charu C. Aggarwal and ChengXiang Zhai, Springer, 2015.
2. Speech & Language Processing. Dan Jurafsky and James H Martin, Pearson Education India, 2000.
3. Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.

Course Proposed and Designed by:

Dr Rafi Ullah Khan
Institute of Computer Science and Information Technology
The University of Agriculture Peshawar
rafiyz@aup.edu.pk
rafyz@gmail.com