**Course number and name:** **Data Mining I – CS 02505**

**Credits and contact hours:** 3 credits / 3 contact hours

**Instructor's or course coordinator's name:** Anthony Breitzman

**Instructional materials:** None required. If you feel you want a book, here are a couple of good ones. Larose – Discovering Knowledge in Data; Han, Kamber, Pei – Data Mining Concepts and Techniques

Specific course information

**Catalog description:** This is a first graduate level course in Data Mining, which is designed to teach students the key steps in data mining, along with the primary algorithms related to data acquisition, cleansing, and supervised and unsupervised learning.

**About this course:** This hands-on course will teach students how to harness massive data sets to find interesting results or to solve real world problems. Some of the objectives include:

- understand the complexity of mining massive datasets with high dimensions.
- use state-of-the art techniques to reduce the dimension of a problem without losing the intelligence hidden in the data.
- recognize which algorithms for extracting knowledge from a given set of data are most appropriate for a given problem.
- interpret results so that customers or companies can make intelligent business and operations decisions.
- Gain a working knowledge of R and Python.

**Prerequisites:** Being accepted in the MS-DS or MS-CS programs or related COGS. (It's assumed that students will be familiar with Linear Algebra and Data Structures)

Specific Topics:

- Introduction to Data Mining and Knowledge Discovery
- Overview of Python and Jupyter Notebooks
- Data Mining Lifecycle: Six Phases
- Obtaining Data; Web Crawlers (etiquette, and spider traps); Reddit and Google APIs
- Data Quality, Data Cleansing, Handling Missing Data and Identifying Misclassifications
- Graphical Methods for Identifying Outliers
- Data Transformation: Min-Max Normalization; Z-Score Standardization

- Overview of Supervised versus Unsupervised Learning approaches
- Hierarchical Clustering; k-Nearest Neighbor Algorithm; distance functions and database considerations
- Decision Trees; Classification and Regression Trees; C4.5 and CART Algorithm
- Naïve Bayes
- Artificial Neural Networks; Backpropagation
- Logistic Regression
- Association Rules; Market Basket Analysis
- Model Evaluation Techniques
- Principal Component Analysis