CS 6010: DATA SCIENCE PROGRAMMING

Semester Hours:	3.0	*Contact Hours:* 3
Coordinator:	Shuteng Niu	
Text:	Various	
Author(s):	VARIOUS	
Year:	Various	

SPECIFIC COURSE INFORMATION

Catalog Description:

This course introduces modern programming languages, platforms, and software packages as well as their application to analyze data of various volumes, velocities, and varieties in the field of data science. Students will be required to leverage modern software libraries at all level of data preparation and analysis. Prerequisites: Admission to MS in CS program, admission to MS/Ph.D. in DS program, or permission of instructor.

Course type: **ELECTIVE**

SPECIFIC COURSE GOALS

- I am able to perform data cleaning and analysis using modern programming languages and computational platforms.
- I am able to test statistical hypotheses using programmatic techniques using modern software libraries.
- I am able to visualize data with various complexities for analysis using modern software libraries.
- I am able to use high performance libraries to improve computation for data science problems.
- I am able to analyze data of significant size independently using software tools.
- I am able to follow best practices for developing, writing, and documenting code.

LIST OF TOPICS COVERED

- (~ 10%) Programming Fundamentals
    - Syntax, Data types, Variables, Functions, Control Flow
    - Array/List creation and manipulation
    - Functions, Packages, and Objects

    - Library use
    - Version Control
- (~ 15%) Libraries and Methods for Data Ingestion and Preparation
    - Data structures for storing, analyzing, and querying data
    - Accessing, cleaning, integrating, transforming, analyzing, and testing data
- (~ 15%) Libraries and Methods for Data Exploration and Visualization
    - Libraries for Plotting/Visualization
    - Create and interpret various plots
- (~ 15%) Libraries and Methods for Probability and Statistics
    - Libraries for Hypothesis Testing, Confidence Intervals, and Statistical Validation
- (~15%) Libraries and Methods for Machine Learning
    - Developing basic Machine Learning models for application
    - Dealing with imbalanced data
- (~15%) High Performance Libraries and Their use (e.g. Theano, Keras, TensorFlow, etc.)
- (~15%) Advanced computational platforms (e.g. AWS, Azure, etc.)
    - Deploying virtual environments for Data Science (Deep Learning, GPU, etc.)
    - Executing projects and analysis at a large scale

EXAMPLE PROJECTS

- **Introductory:**
    - Storytelling with Data (https://www.dataquest.io/blog/data-science-portfolio-project/)
    - Pandas Dataframes (https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python#gs.tggqks0)
- **Intermediate:**
    - Mine and Analyze Twitter Streaming Data
    - What can you learn from visualizing Hubway Data? (http://hubwaydatachallenge.org/trip-history-data/)
- **Advanced:**
    - Build a recommendation system based on the MoveLens Data
    - Perform an Analysis of the Yelp Dataset. What can you learn?