## **Big Data Technology “IS 6733”**

### Tools

The following cloud-based environment will be provided to each student: Jupyter, Spark, Python, Python libraries for big data analytics. Github will be used for course content, projects, and assignments.

### Recommended Background

This course is for those new to data science. No prior programming experience is needed, although the ability to install applications and utilize a virtual machine is necessary to complete the hands-on assignments.

### About the Course

According to McKinsey’s report, *The Age of Analytics: Competing in a Data Driven World*, December 2016, “Data and analytics capabilities have made a leap forward in recent years. The volume of available data has grown exponentially, more sophisticated algorithms have been developed, and computational power and storage have steadily improved. The convergence of these trends is fueling rapid technology advances and business disruptions”. This course is designed for students new to data science and interested in understanding why the Big Data Era has come to be. In this course, you will gain an understanding of what insights big data can provide through hands-on experience with the tools and frameworks used by big data scientists and engineers. This course provides an introduction to one of the most common frameworks, Spark, that has made big data analysis easier and more accessible. By following along with provided lessons, you will experience with data processing and leverage graph analytics to model problems. In the final course project, you’ll apply the skills you learned to do basic analyses of big data, to develop and implement a big data application.

### Syllabus

* Introduction to Big Data
* Big Data Characteristics
* Hadoop Ecosystems
* Big Data Integration
* Big Data Processing
* Graph Analytics for Big Data
* Machine Learning for Big Data
* Course Big Data Project
* Big Data Visualization (If time permits)

### Grading Criteria

* Quizzes and Exams (30%)
* Assignments (30%)
* Course project (40%)
* Extra grade (5%)

### What is the course project?

The course project is designed to help you practice, apply, and showcase the skills you’ve learned. In some Specializations, the Capstone Project is spread out across the courses. In others, it is the last course in the Specialization.

### Textbooks:

[1] Martella, Claudio, et al. Practical graph analytics with apache giraph. Apress, 2015.

[2] Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale by Tom White

[3] Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems by Aurélien Géron

[4] Goodfellow, Ian, et al. Deep learning. Vol. 1. Cambridge: MIT press, 2016

We will not follow the books chapter-by-chapter, but rather use them as a reference. Slides will be uploaded in GitHub.

### Prerequisite: None

This course is for those new to data science. No prior programming experience is needed, although the ability to install applications and utilize a virtual machine is necessary to complete the hands-on assignments.

### Class Schedule

### Class Time: Tuesday and Thursday 10:00 am – 12:45 pm

### Office Hours:

* Tuesday 1:00 pm – 2:00 pm
* Thursday 1:00 pm – 2:00 pm

### Office Location:

* NPB 3.138E08 @ Open Cloud Institute (OCI) – Cell (210) 872.7259

|  |  |  |
| --- | --- | --- |
| **No.** | **Date** | **Topic** |
| 1 | Aug 21st, 23rd | **Introduction to Cloud Computing and Big Data** At the end of this module you will be able to:   * Describe the Big Data landscape * The three key sources of Big Data * Data at Scale – Working with Big Data |
| 2 | Aug 28th, 30th | **Big Data Characteristics** At the end of these modules you will be able to:   * Recognize different data types * Characteristics of Big Data (Vs) * Data models (vector Space vs. Graph Data Model) * Streaming data vs. static data |
| 3 | Sept 4th, 6th  **Assignment #1** | **Hadoop Ecosystem (Part I, II)** At the end of these modules you will be able to:   * HDFS, YARN and MapReduce |
| 4 | Sept 11th, 13th | **Big Data Processing -**  At the end of these modules you will be able to:   * SQL vs. NoSQL * Retrieve data from databases and big data management systems * Introduction to Apache Spark * SparkSQL and Spark DataFrames |
| 5 | Sept 18th, 20th  Quiz.#1  **Assignment #2** | **Big Data Processing- Natural Language Processing (Part I)** At the end of these modules you will be able to:   * + Tokenization   + Lemmatization   + Bigram Collocation Detection |
| 6 | Sept 25th, 27th  Submit  Project  Definition | **Big Data Processing- Natural Language Processing (Part II)** At the end of these modules you will be able to:   * + Topic Modeling   + Text Clustering and Classification |
| 7 | Oct. 2nd, 4th  Quiz.#2  **Assignment #3** | **Machine Learning with Big Data (Part II)** At the end of these modules you will be able to:   * + Deep Learning |
| 8 | Oct. 9th , 11th | **Machine Learning with Big Data (Part II)** At the end of these modules you will be able to:   * + Convolution Neural Network |
| 9 | Oct. 16th , 18th  **Assignment #4** | **Machine Learning with Big Data (Part III)** At the end of these modules you will be able to:  Deep Learning and Convolution Neural Network |
| 10 | Oct.23th,25th | **Machine Learning with Big Data**  At the end of these modules you will be able to:  LSTM and GAN |
| 11 | Oct.30th, Nov. 1st  Quiz #3  **Assignment #5** | **Graph Analytics for Big Data (part I)** At the end of these modules you will be able to:   * Introduction to Graphs * Examples (Smart Cities, Social Network, Etc.)   Path Analytics |
| 12 | Nov. 6th, 8th | **Graph Analytics for Big Data (Part II)**  * Connectivity Analytics * Connectedness: Indegree and Outdegree * Global Property: Modularity * Key Player, Eigenvector Centrality   Large Scale Graph Processing Giraph and GraphX |
| 13 | Nov. 13th, 15th  **Assignment #6**  Quiz #4 | **Graph Analytics for Big Data (Part III)**  * Graph Computation and Tensorflow * Deep Learning and Convolution Neural Network |
| 14 | Nov. 20th, | **Project Review** |
|  | Nov. 22nd- 23rd | **Thanksgiving Holidays.** |
| 15 | Nov. 27th, 29th |  |
|  | Dec. 4th | **Project Review** |
| 16 | Dec. 6th | **Student Study Day. Classes do not meet.** |
| 17 | Dec. 8th, 14th | **Final Exams**. No Final Exams on Sunday.  **Project Review** |
| 18 | Dec. 17th | **Final Grades Due; 2 p.m. Deadline for faculty to enter final grades.** |