CSCI E-63 Big Data Analytics (24038) 2017 Spring term (4 credits)

Zoran B. Djordjević, PhD, Senior Enterprise Architect, NTT Data, Inc.

Lectures: Fridays starting on January 27th, 2017, from 5:30 to 7:30 PM (EST), 1 Story Street, Room 306, Cambridge, MA

Optional Online Sections: Saturdays, starting January 28th, 2017 at 10-11:30 AM (EST).

The explosion of social media and the computerization of every aspect of social and economic activity resulted in creation of large volumes of mostly unstructured data: web logs, videos, speech recordings, photographs, e-mails, Tweets, and similar. In a parallel development, computers keep getting ever more powerful and storage ever cheaper. Today, we have the ability to reliably and cheaply store huge volumes of data, efficiently analyze them, and extract business and socially relevant information. The key objective of this course is to familiarize the students with most important information technologies used in manipulating, storing, and analyzing big data. We will examine the basic tools for statistical analysis, R and Python, and several machine learning algorithms. The emphasis of the course will be on mastering Spark 2.0 which emerged as the most important big data processing framework. We will examine Spark ML (Machine Learning) API and Spark Streaming which allows analysis of data in flight, i.e. in near real time. We will learn about so-called NoSQL storage solutions exemplified by Cassandra for their critical features: speed of reads and writes, and ability to scale to extreme volumes. We will learn about memory resident databases (VoltDB, SciDB) and graph databases (Ne4J). Students will gain the ability to initiate and design highly scalable systems that can accept, store, and analyze large volumes of unstructured data in batch mode and/or real time. Most lectures will be presented using Python examples. Some lectures will use Java and R.

Prerequisites: Familiarity with Intermediate Python or Java is advised. Most assignments could easily be done in Python, Scala, Java or R. We will assume no familiarity with Linux and will introduce you to all essential Linux commands. Students need access to a computer with a 64 bit operating system and at least 4 GB of RAM. Note: 8 GB or more of RAM is strongly advised.

Lectures: Lectures will be delivered live and made available after lectures for online viewing through Zoom Web Conferencing tool. Streaming recording will also be available. Links to Zoom recorded lectures will be accessible on the course Web site within a few hours after the end of the lecture. Streaming video recorded lectures will become available with a delay of one to two days.

References: Detailed handouts with references to material on the Web will be handed out every week. There is no required text book.

Grading: Practically every class will be followed by a homework assignment. Grades on the solutions for class assignments constitute approximately 85% of the final grade. 15% of the grade will be earned through the final project. Final projects will be assigned a few weeks before the end of the class. For the final project you will produce a paper (10+ pages of MS Word text, 10+ PowerPoint Slides, a working demo, 15 minute YouTube video of your presentation and a brief 2 minute YouTube video that might be presented to the class on the day of final presentations. Several students will be invited to present their final projects live to the entire class. **Grades:** 95% or higher cumulative grade on all assignments and the final project gives you an A as the final grade in the course, 90-94.9% gives you an A-, 85-89.9% a B+, 80-84.9% a B, etc.

Communications: zdjordj@fas.harvard.edu, Canvas class site and Piazza, once class starts.

Tentative List of Class Topics:

	Date	Topic
1	01/27/2017	Basic Statistics
2	02/03/2017	Relationships and Representations, Graph Databases
3	02/10/2017	Introduction to Spark 2.0
4	02/17/2017	Language processing with Spark 2.0
5	02/24/2017	Analysis of Streaming Data with Spark 2.0
6	03/03/2017	Applications of Spark ML Library
7	03/10/2017	Basic Neural Network and TensorFlow
	03/17/2017	Spring Break
8	03/24/2017	Advance TensorFlow
9	03/31/2017	Assessing Quality of Machine Learning Algorithms
10	04/07/2017	Analysis of Images, OCR Applications
11	04/14/2017	Analysis of Speech Signal
12	04/21/2017	Question Answer Systems
13	04/28/2017	Page Rank like Search systems with TensorFlow
14	05/05/2017	Analysis of Streaming Data with TensorFlow
15	05/12/2017	Final Project Presentations