Boston University Metropolitan College

Course Title: Big Data Analytics

MET CS 777

Course Format: On Campus/Online

Instructor Name: Kia Teymourian , kiat@bu.edu

Office hours: by appointment

Course Description

This course is an introduction to large-scale data analytics. Big Data analytics is the study of how to extract actionable, non-trivial knowledge from massive amount of data sets. This class will focus both on the cluster computing software tools and programming techniques used by data scientists, as well as the important mathematical and statistical models that are used in learning from large-scale data processing. On the tools side, we will cover the basics systems and techniques to store large-volumes of data, as well as modern systems for cluster computing based on Map-Reduce pattern such as Hadoop MapReduce, Apache Spark and Flink.

Students will implement data mining algorithms and execute them on real cloud systems like Amazon AWS, Google Cloud or Microsoft Azure by using educational accounts. On the data mining models side, this course will cover the main standard supervised and unsupervised models and will introduce improvement techniques on the model side.

This course can be taken by students with not exclusively computer science backgrounds who have basic knowledge of programming.

Books

There is no textbook for the class. All class material will be conveyed during lecture.

Recommended Books**:**

- **"Data Mining: Concepts and Techniques",** Third Edition. (The Morgan Kaufmann Series in Data Management Systems) by Jiawei Han, Micheline Kamber, Jian Pei. Morgan Kaufmann. ISBN-13: 978-9380931913
- **Learning PySpark**. Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0. by Tomasz Drabas

(Author), Denny Lee. 2017" Example pySpark Implementations available here https://github.com/drabastomek/learningPySpark

- **Mining of Massive Datasets** by Jure Leskovec, Anand Rajaraman, Jeff Ullman. published by Cambridge University Press. 2014. By agreement with the publisher, you can download the book for free from this page http://www.mmds.org/
- **Python for Data Analysis:** Data Wrangling with Pandas, Numpy and IPython by W. McKinney, O'reilly Publishing, 2013
- **Learning PySpark.** Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0. by Tomasz Drabas, Denny Lee. 2017. Example pySpark Implementations available here https://github.com/drabastomek/learningPySpark
- **Python Data Science** Handbook by Jake VanderPlas, O'Reilly Publishing, ISBN-13: 978-1491912058

Courseware

Blackboard
Course Notes

Class Policies

1) **Attendance & Absences** – clearly state your attendance policy, limit to absences, etc. List all unusual required meetings (e.g. exhibits, guest lectures, field trips, etc.)
2) **Assignment Completion & Late Work** – detail your policy regarding how students should submit completed assignments (in person, by email, on courseware site, etc.), as well as how you will address late work.

Weekly programming assignments submitted through blackboard on-line. Late homework is accepted with 50% penalty. Final projects are submitted through blackboard on-line. Students will present their projects on the last day of class. Both quiz and final are closed-book and are in-class

3) **Academic Conduct Code** – Please use the following wording, or an equivalent, in your syllabus: "Cheating and plagiarism will not be tolerated in any Metropolitan College course. They will result in no credit for the assignment or examination and may lead to disciplinary actions. Please take the time to review the Student Academic Conduct Code:

Academic conduct code as specified below:

http://www.bu.edu/met/metropolitan_college_people/student/resources/conduct/code.html.

NOTE: [This should not be understood as a discouragement for discussing the material or your particular approach to a problem with other students in the class. On the contrary – you should share your thoughts, questions and solutions. Naturally, if you choose to work in a group, you will be expected to come up with more than one and highly original solutions rather than the same mistakes.]

Grading Criteria

Give a detailed list of percentage weights for assignments, papers, class participation and examinations as applicable. If you have complex grading criteria, please spell this out here as clearly as possible. Remember: the syllabus is a contract between you and your students, and will be referred to as such in the event a dispute arises.

6 Homework Assignments: 30%
6 Quizzes: 20%
Term Project: 20%
Final Exam: 30%

Class Meetings, Lectures & Assignments

List in a legible format all of the class meetings, lectures, and assignments. One example, based on a computer science course:

Lectures, Readings, and Assignments subject to change, and will be announced in class as applicable within a reasonable time frame.

6 Quizzes and 6 Homework Assignments

Assignments are designed based on real-world public data sets.

The tentative lecture schedule is:

Module 1 - Map Reduce Data Processing Pattern

- Introduction to Big Data Analytics. What is Big Data? What are the challenges?
- Introduction to Apache Hadoop and MapReduce. Apache Spark.
- Spark programming. (Python and pySpark)
- Resilient Distributed Dataset (RDDs).

Module 2 - Large-Scale Data Processing and Storage

- RDDs, DataFrames, Spark SQL
- PySpark + NumPy + SciPy, Code Optimization, Cluster Configurations
- Recap – Relational Databases and SQL
- NoSQL, Column-oriented and Document-based Databases
- Distributed Object Storage Systems

Module 3 - Introduction to Modeling and Optimization Basics

- Introduction to modeling: numerical vs. probabilistic vs. Bayesian
- Optimization basics: Gradient descent (batch and stochastic),
- Newton's method,
- Expectation maximization,
- Markov Chain Monte Carlo (MCMC)

Module 4 – Supervised Learning on Large-Scale Data

- Introduction to supervised learning
- Linear regression and generalized linear models
- Regularization
- Support Vector Machine (SVM) and the kernel trick
- Outlier Detection

Module 5 – Unsupervised Learning on Large-Scale Data

- Introduction to unsupervised learning
- K-means / K-medoids
- Mixture of Gaussians and Gaussian EM
- Matrix factorization

Module 6 – Text and Pattern Mining

- Text Mining: Latent Semantic Indexing
- Text Mining: Topic models
- Pattern Mining: Association rule mining and the Apriori algorithm
- Pattern Mining: Maximal association rule mining, Random Forest