COEN242: Big Data

The COEN: 242 Big Data (Spring 2021) covers the fundamentals of parallel computing algorithms, specifically designed for learning tasks on large-scale datasets. The course reviews methods for dealing with both large and high-dimensional datasets, emphasizing distributed implementations. We will offer a hands-on project using the Spark development platform and introduce the theory behind statistical data analysis.

Course Information

- **Instructor**: Zhiqiang Tao
  - Office: Bergin 110
  - Contact: ztao@scu.edu
- **Class and Office Hours**:
  - Class hours: **07:10 pm - 09:00 pm [TR]**
  - Location: video records
  - Office hours: **05:10 pm - 07:00 pm [MW]**
  - Communication: Camino, Piazza, and Email
- **Prerequisite:** AMTH 108 or 210, and COEN 178 or 280

Online Asynchronous Mode

For all the students enrolled in COEN-242 Spring 2021, *lecture video records* will be provided in classes on **Tuesday / Thursday**, and *online office hours* will be held in classes on **Monday / Wednesday**.

Course Objective

The learning outcomes are: 1) Understanding the internals of important BigData technologies; 2) Developing data-intensive applications using Spark and MapReduce.

Course Content

- Apache Spark fundamentals, multi-threaded/cluster execution.
- Resilient distributed datasets (RDD) and map-reduce operations,
- Working with Key-value pairs, joins.
- Persistence and iterative algorithms, lazy evaluation, PageRank
- Convex optimization, stochastic gradient descent, matrix and tensor factorization.
- Parallelizing computation for machine learning problems
- Big data with machine learning
  - Linear regression, generalized linear models, ridge and lasso regularization.
  - Classification, logistic regression, loss functions. ROC curves and AUC.
  - Feature selection and dimensionality reduction
- Big data with data mining
  - Graph embedding and link prediction

Textbook

This course will not require any specific textbook. Lecture slides, tutorials, and papers will be provided to cover the topics in this class. Students will also be asked to find more self-learning content from online resources. Some recommended reference books are listed (in no particular order) as follows.

- Karau, H., Konwinski, A., Wendell, P. and Zaharia, M., 2015. Learning Spark: Lightning-Fast Big Data Analysis. O'Reilly Media, ISBN:978-1449358624
- Boyd, S., and Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press, ISBN: 978-05218337803.
- Christopher Bishop. Pattern Recognition and Machine Learning, Springer, ISBN: 978-0387310732.
- Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman. Mining of Massive Datasets Cambridge University Press, ISBN: 978-1108476348.

Programming

All homework assignments are required by Apache Spark. However, the knowledge of Python is not strictly required, as our course will cover Python to the extent necessary to proceed with assignments.

Grading

Students will be graded on random quizzes, three assignments, a (*take-home*) mid-term examination, and a final project. The final grade will be composed as follows.

- Random Quizzes (10%)
- Homework (30%)
- Mid-Term Exam (30%)
- Final Project (30%)

Tentative Schedule

Week	Video Lectures [TR]	Online [MW]	HW / Exam
1	Course overview & Introduction to Python	N/A on M	
2	Spark fundamentals & Map-Reduce	Office hours	
3	Map-Reduce Operations in Spark	Office hours	HW1
4	Key-Value Pairs & Partitioning	Office hours	
5	Lazy Evaluation, Resilience & Persistence	Midterm review	HW2
6	Optimization & Parallelizing Computations	Office hours	Midterm Exam
7	Regression & Statistical Learning	Office hours	

Week	Video Lectures [TR]	Online [MW]	HW / Exam
8	Feature Selection & Dimensionality Reduction	Office hours	HW3
9	Graph Embedding & Link Prediction	Office hours	
10	Final project presentation	Final project presentation	