5460. Big Data Scaling (2022 Spring)

Maizie (Xin) Zhou, PhD

Syllabus

Class Information

Class Hours:	Monday and Wednesday	2:45 - 4:00 pm
Room:	Sony Building 2001-A	
Office Hours:	By appointment (maizie.zhou@vanderbilt.edu)	
Graders:	Tabitha Lee (tabitha.see.ya.lee@vanderbilt.edu)	
	Teppei Kotake (teppei.kotake@vanderbilt.edu)	
	Jeerthi M Kannan (jeerthi.m.kannan@vanderbilt.edu)	
Grades:	Grading will be based on the following elements:	
	Homework Assignments, In-class live coding participation, Midterm	
	Final project (writeup and presentation).	

Assessment:

• **Homework Assignments 30%:** There will be homework assignments nearly for every topic. They are intended primarily to help you prepare for the exam and project. You are allowed to work in groups on the homework, but you must write up your own solutions in your own words. ASSIGNMENTS ARE DUE AT **11:59 midnight** OF THE DUE DATE THROUGH BRIGHTSPACE.

• **In class live coding participation 10%:** There will be living coding participation nearly for every topic on Thursday.

• **Midterm 30%:** Midterm will be a take-home exam.

• **Final Project 30%:** Projects are required to be related in a substantive way to at least one of the central topics of the course. Final projects can be done in groups of 1 - 3 people. We encourage you to form a group of 3 members, since groups of 3 usually lead to the best outcomes. We will talk about more details in class.

Late days Policy

Each student will have a total of 5 free late (calendar) days applicable to any assignment except the midterm and final project. Free late days can be used at any time, no questions asked. Each 24 hours or part thereof that a homework is late uses up one full late day. Once these late days are exhausted, any homework turned in late will be penalized 10% per late day.

Late days are never transferable between students, even students in the same group.

Class Announcements: All students are held responsible for all announcements made in the class and Campuswire.

Campuswire: We will use Campuswire for all homework/midterm/project announcements, questions and public communications, holding office hours. You should have already received the invite by email!

Course Materials: The course relies on Jupyter notebooks, Powerpoint slides, and online resources (https://spark.apache.org/docs/latest/). The lecture slides and Jupyter notebooks for each week will be found in the **box** as the course progresses. The textbook (Mining of Massive Dataset: http://www.mmds.org/) is useful and recommended, but not required. You can download it for free or purchase the hardcopy from Cambridge University Press.

All academic work at Vanderbilt is done under the Honor System.

The course will discuss data mining and machine learning algorithms, and the emphasis will be on MapReduce and Spark as tools for creating parallel algorithms that can process very large amounts of data.

Topics include Cloud Computing, Distributed File Systems, MapReduce, Apache Spark, SparkSQL, Regression, Tree Methods, Locality Sensitive Hashing (LSH), Clustering, Recommendation Systems, Link Analysis, Mining Data Streams.

Schedule:

Week	Topic / Contents
1	Cloud Computing Intro to Clouds, Cloud Computing, Popular Commercial Cloud Architectures Parallel processing in python
2	Big Data Processing Distributed File Systems, HDFS, Hadoop, MapReduce Hands-on tutorial for Google Clouds (Dataproc and cloud storage)
3	Apache Spark Spark Architecture, Resilient Distributed DataSets (RDDs), Transformations and actions (slides) SparkContext and RDD Basics (notebook)
4	Spark DataFrames ● Spark DataFrames Section Introduction (slides) ● Spark DataFrame Basics (notebook) ● Spark DataFrame Basic Operations (notebook) ● Groupby and Aggregate Functions (notebook) ● Missing Data (notebook) ● Dates and Timestamps (notebook)
5	Spark SQL More Tutorial: DataFrame, SparkSQL, and RDD (notebook)
6	Review (slides) Midterm
7	Regression ● Regression (slides)

	- Introduce Linear Regression in PySpark (notebook) - Data Transformations, VectorAssembler (notebook) - Linear Regression Example (notebook) - Introduce Logistic Regression in PySpark (notebook) - Data Transformations and Pipeline (notebook) - Logistic Regression Example (notebook)
8	Tree Methods - Large Scale Machine Learning: Decision Tree (slides) - Introduce Tree Methods in PySpark (notebook) - Three Tree Models Comparison Example (notebook) - Random Forest Classification Project (notebook)
9	Locality-Sensitivity Hashing, Clustering - Locality-Sensitivity Hashing; Clustering (slides) - Introduce Clustering in PySpark (notebook) - Clustering Example (notebook)
10	Recommendation System - Introduction to Recommender Systems and Collaborative Filtering (slides) - Recommendation System Example (noteboook)
11	PageRank, Link Analysis - PageRank (slides) - Introduce NetworkX for PageRank (notebook)
12	Mining Data Streams
13	Final Project Presentations
14	Project Writeup Due