Big Data Analytics

CSCE 587 Big Data Analytics [1]
Fall 2021
Lectures: In-Person
Sect 001 TTh 10:05-11:20am

John Rose
rose@cse.sc.edu/7-2405
Office: Innovation Center 2257
Office hours: TBA
TA: Qinyang Li
TA Office hours: TBA

Course Description[2]
CSCE/STAT 587 – Special Topics: Big Data Analytics (3) This course covers foundational techniques and tools required for data science and big data analytics. The course focuses on concepts, principles, and techniques applicable to any technology environment and industry and establishes a baseline that can be enhanced by further formal training and additional real-world experience.

Prerequisites[3]
This course introduces the student to concepts of big data management, data mining techniques and the underlying statistics that support bid data analytics. Since this is a 500-level course, relevant 500 level courses such as Database (CSCE 520) and Statistics for Engineers (STAT 509) cannot be listed as prerequisites. Consequently, lecture time will be devoted to addresses necessary topics from these courses. In this course we will use the programming language R as well as Hadoop/Map-Reduce/Pig/Hive as the primary tools for analysis.

Learning Outcomes[4]
By the end of the course the student will be able to:

1. deploy a structured lifecycle approach to data science and big data analytics projects
2. select visualization techniques and tools to analyze big data and create statistical models
3. use tools such as R and RStudio, and MapReduce/Hadoop/PIG/HIVE.

Required Texts, Other Materials, Suggested Readings[6]
This course does not have a required text. However, ad hoc readings from the field will be assigned. In addition, material from "Data Science and Big Data Analytics Student Guide" distributed by EMC Education Services will be provided to the students.

Course Overview[7]
Students will be expected to participate in lectures during designated lecture time. Lectures slides will be made available prior to the lecture. The instructor will reply to all feedback in a reasonable amount of time; the same is expected of the students. Specifically,
Communication: Responses to email communication and questions will be provided within 24 hours.
Assignment Grading: Grades for assignments will be returned within 72 hours of due date (assuming the assignment is turned in on time).
Test Grading: Grades for tests will be returned within 72 hours of due date.

Technology[8]
The software used for this course include R, and Hadoop/Map-Reduce/Pig/Hive. Each student will have their own virtual machine running on departmental hardware with this software pre-installed. Each student will be able to access their virtual machine via web browser and/or ssh/sftp.

Course Delivery Structure
The course will be delivered in a computer lab with 50% of the time devoted to lectures and the other 50% devoted to hands-on lab exercises.

Course Requirements
Homework: Students will complete assignments demonstrating mastery of material. These will be due at midnight on the due date (submitted via Microsoft Teams).

Final Exam
Sect001 Tuesday, December 7 at 9 am

Course Outline/Schedule
Lecture 1: Introduction to Big Data Analytics and VM Terminal Sessions
Lecture 2: Introduction to R and RStudio
Lecture 3: Introduction to R and RStudio continued
Lecture 4: Basic analysis in R
Lecture 5: Intermediate R
Lecture 6: Intermediate analysis in R
Lecture 7: Generic clustering consideration, K-means Clustering
Lecture 8: Hierarchical and model-based clustering, Association Analysis
Lecture 9: Basic Association Analysis, Association Rules Lab
Lecture 10: Association Rules speedup, Simple Linear regression
Lecture 11: Linear regression, Linear regression Lab
Lecture 12: Logistic regression
Lecture 13: Cont. of hypothesis testing and example of two sample test, Naïve Bayes
Lecture 14: Logistic regression
Lecture 15: Naïve Bayes, classification accuracy, holdout estimation and ROC analysis, NB Lab
Lecture 16: Decision trees part 1
Lecture 17: Decision trees part 2
Lecture 18: Review for Midterm Exam
Lecture 19: Midterm Exam
Lecture 20: Introduction to Hadoop and HDFS
Lecture 21: Using R with Hadoop
Lecture 22: First R/Hadoop program
Lecture 23: Intermediate R/Hadoop programming
Lecture 24: Pig
Lecture 25: Hive
Lecture 26: Intermediate Map-Reduce/Pig/Hive
Lecture 27: Discussion of Hadoop take home portion of final exam
Lecture 28: Review for Final Exam

Assignments[9]

Homework: Students will complete assignments demonstrating mastery of material. These will be due at midnight on the due date (submitted via dropbox).

HW 1: R Homework assignment

HW 2: K-means homework assignment

HW 3: Linear regression homework assignment

HW 4: Logistic regression homework assignment

HW 5: Naïve Bayes homework assignment

Project 1: Data set analysis (take home portion of midterm exam)

HW 6: Hadoop MapReduce/PIG/HIVE homework assignment.

GHW: Additional Hadoop MapReduce/PIG/HIVE homework assignment for graduate students.

Project 2: Hadoop MapReduce/PIG/HIVE Project (take home portion of final exam)

Midterm exam: Covers lectures 1 – 17. It is comprised of an in-class written exam as well as a take-home applied-exam. The overall midterm grade will be a weighted average of the in-class and take-home

MidtermScore = ¾ in-class written exam + ¼ take-home applied project

Final exam: Covers entire semester: It is comprised of an in-class written exam as well as take-home applied-exam. The overall midterm grade will be a weighted average of the in-class and take-home

FinalExamScore = ¾ in-class written exam + ¼ take-home applied project

Grading Scheme[10]

Final grade: 90 <= A, 87 <= B+ <90, 80<= B < 87, 77 <= C+ < 80, 70<= C < 77, 65 <= D+ <70, 60 <= D < 65, F < 60

Grades will be calculated from homework (30%), in-class labs (20%), midterm (20%), and Final Exam (30%).

Difference between Undergraduate and Graduate Work:

While letter grades for both undergraduate and graduate students are based on the same percentage boundaries, the denominators used to calculate those percentages are different. Graduate students are assigned additional problems in both homework and exams as well as one additional homework assignment.

Course Policies

Attendance[11]: Attendance is mandatory. The in-class labs (20% of overall grade) are graded on participation. You must be in class to receive credit for in-class labs. If you weren't in-class, then you didn't participate. In-class labs can not be made up except in the case of a documented excused absence.

Tardiness, late assignments: homework is due at midnight on the due date (submitted via Teams). Late assignments will be charged 10% per day. No projects or assignments will be accepted after 5 days.

Policy on disabilities or special needs[12]: Reasonable accommodations are available for students with a documented disability. If you have a disability and may need accommodations to fully participate in this class, contact the Student Disability Resource Center: 803-777-6142, TDD 803-777-6744, email sadrc@mailbox.sc.edu, or stop by LeConte College Room 112A. All accommodations must be approved through the Student Disability Resource Center. See https://www.sa.sc.edu/sds/.

Student-to-Instructor (S2I) Interaction[14]: Students will listen/view lectures online via videos and interact with the professor through email, MS Teams and (possibly) discussion boards. The professor will post announcements, provide individual feedback to students, and hold online office hours via MS Teams.

Students-to-Student (S2S) Interaction[14]: Students will engage in discussions through email and possibly MS Teams.

Student-to-Content (S2C) Interaction[14]: Students will engage with course content by completing in-class labs, homework assignments, and exam projects.

Violations of academic honesty[13]: Assignments and examination work are expected to be the sole effort of the student submitting the work. Students are expected to follow the University of South Carolina Honor Code and should expect that every instance of a suspected violation will be reported. Students found responsible for violations of the Code will be subject to academic penalty under the Code in addition to whatever disciplinary sanctions are applied.

Honor code violations include any of the actions described below(excerpted from http://www.sc.edu/policies/ppm/staf625.pdf):

A. Plagiarism: Use of work or ideas without proper acknowledgment of source. Prohibited behaviors include:
1. Partial or incomplete citation of work or ideas.
2. Improperly paraphrasing by acknowledging the source but failing to present the material in one's own words.
3. Paraphrasing without acknowledgment of the source.
4. Multiple submissions of the same or substantially the same academic work for academic credit.
5. Copying, partially or entirely, any material without acknowledgement of the source.

B. Cheating: Improper collaboration or unauthorized assistance in connection with any academic work. Prohibited behaviors include:

1. Requesting unauthorized assistance
2. Copying another individual's or group's academic work.
3. Receiving and utilizing academic work for purposes of fulfilling an academic requirement.
4. Completing any academic work for someone else or permitting someone else to complete academic work on your behalf.
5. Using any bribe, coercion or unauthorized aid (e.g., outside source, cell phone, calculator, notes, previous testing materials) for an unfair academic advantage.
6. Using, possessing or distributing the contents of any examination (e.g., unauthorized access to test/quiz information, unauthorized duplication of test/quiz materials) without authorization.
7. Taking, misplacing, or damaging property if the student knows or reasonably should know that an unfair academic advantage would be gained.

C. Falsification: Misrepresenting or misleading others with respect to academic work or misrepresenting facts for an academic advantage. Prohibited behaviors include:
1. Signing in for another student who is not in attendance, or requesting this action.
2. Interfering with an instructor's ability to evaluate accurately a student's competency or performance on any academic work.
3. Fabrication of documents submitted in connection with academic work.

D. Complicity: Assisting or attempting to assist another in any violation of the Honor Code. Prohibited behaviors include:
1. Sharing academic work with another student (either in person or electronically) without the permission of the instructor.
2. Communicating (either in person or electronically) with another student(s) or other individual(s) during an examination without the permission of the instructor.

PROCEDURES

A. All allegations must be referred to the Office of Academic Integrity for investigation. The instructor should notify students that they are being referred to the Office of Academic Integrity.

B. Non-academic sanctions are determined by the Office of Academic Integrity in conjunction with the college liaison.

C. Academic sanctions are determined at the discretion of the instructor of record and occur following the case resolution by the Office of Academic Integrity. The academic sanction in CSCE/STAT 587 will be a grade of zero on the assignment/test/project for which the honor code was violated.