Data Management for Analytics
MSA 8040

Course Syllabus

Instructor: Dr. Houping Xiao
Email: hxiao@gsu.edu
Office: Buckhead 329

Class Meetings: Fall 2018
Buckhead 306
Thursday, 2:00pm – 4:30pm (*CRN: 94471*)
OR 7:15pm – 9:45pm (*CRN: 94059*)

Course Description:

The course is intended to introduce concepts in both Database and Data Mining areas for unstructured data analytics. Topics related to Database include 1) Concepts about both basic and advanced database; 2) Database design concepts, including the Relational Database Model, Entity Relationship (ER) Modeling and Normalization; 4) Advanced design and Implementation, including both introduction and advanced Structured Query Language (SQL). For unstructured data mining, we will cover the following topics: 1) Unstructured data generation, especially hands-on with web scraping, data storing and querying via MongoDB; 2) Clustering; 3) Topic Modeling; and 4) Classification. More details of each topic, please refer to the tentative course schedule.

Course Objectives:

By the end of the semester, students will be able to:

- Understand database
- Be familiar with relational database concepts
- Be proficient in manipulating data using SQL
- Understand structured and unstructured data
- Be familiar with MongoDB
- Be able to extract, store and query unstructured data
- Be able to recall and discuss algorithms for analysis of unstructured data
- Be familiar with Python
- Apply unstructured data analytics techniques to solve real problems

Textbooks:

- [1] Carlos Coronel and Steven Morris. Database systems: Design, Implementation, & Management. 13th Edition. Cengage Learning. ISBN-13: 978-1337627900

- [2] Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, and Fred Damerau. Text Mining: Predictive Methods for Analyzing Unstructured Information. First Edition. Springer. ISBN-10: 0387954333. ISBN-13: 9780387954332

Recommended References:

- [3] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. Database Systems: The Complete Book. 2nd Edition. Pearson. ISBN-13: 978-0131873254 ISBN-10: 0131873253
- [4] Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques. 3rd Edition. Elsevier. ISBN-10: 0123814790. ISBN-13: 9780123814791
- [5] NoSQL for Mere Mortals. 1st Edition, by Dan Sullivan (Addison-Wesley Professional: 2015).

IN-CLASS Course Structure

The course is designed as a combination of in class lecture with a lab session with hands-on work. Please refer to the tentative course schedule below for the lecture time. The goal of this pedagogical approach is to introduce the theoretical concepts and reinforce practical implementation in MongoDB or Python such that students can indeed master skills on their own after class. *Effort outside of reviewing and practicing the class materials are highly recommended.*

OUT-OF-CLASS Assistance

Office Hours: Thursday 4:30pm—6:00pm
Location: Buckhead 329

If you have questions about course concepts, please come to the Office Hours. If you could not come to the office hour, please email me to set up a meeting. To maximize the effectiveness of the meeting to help you better understand content you are struggling with, **in your email please include the specific questions you have and would like to discuss about the course content.**

Coursework and Grading

Coursework

- Class participation & Quizzes
- Mid-term Exam
- Homework
- Project

Grade Calculation

- 20% In-Class Quizzes (5% class participation + 15% Quizzes)
  The attendance of this class is required and regularly, we will administer quizzes at the beginning of class to assess the basic concepts of the material in the previous week. Totally, there are 13 quizzes and the best 10 out of them will be used for the final grades.
- 20% In-Class Mid-term Exam
  There will be one in-class mid-term exam. The mid-term exam will cover the material discussed in the course, homework and exercises. Exam missed due to an unexcused absence may not be made up (please refer to the Make-up policy in page 5).

You are not allowed to collaborate on the exam. Both you and your collaborator will receive a score of **ZERO** if any infraction is noticed and established. In addition, other actions may also be taken.

- 30% Homework

  A hard copy of solutions to all your home work should be handed in before the class **on the day that they are due**!

  1. **Please check Guidelines on Homework Assignment!**
  2. **Late homework or emailed homework will not be accepted.**

- 30% Project (20% Project Chapter + 40% Report + 40% Presentation)

  A group-based final project (**at most 4 persons per group, usually it is 3 or 4**) will be released on Week 10 to evaluate the students' understanding on the topics covered in the course. Each project includes a project chapter, a final report of their findings and a presentation to present their work. The grade will be based on the evaluation on their reports and demos from the instructor.

  1. **Please check Guidelines on each project carefully!**
  2. **Late submission policy:** deliverables submitted after due date will only be eligible for 60% of total points of the project the first week after.
  3. **No submission is accepted after one week.**
  4. **Peer-Evaluation**:

     To better achieve fairness in the class, at the end of the course you will be asked to evaluate yourself and the other members of your group on completing the project. These ratings are used for gauging team members' contributions. The grade you and your group members receive will depend in part on these peer evaluations. Rate each member based on the following criteria: (1) participation in group activities, (2) quality of work, (3) quantity of work, (4) finishing assigned work on time, and (5) ability to work as a team member. Please use the following scale to assign scores:

5	Exceptional effort, above and beyond the call of duty
4	Above average effort
3	Normal effort (this is the expected score!)
2	Below average effort
1	Unacceptable effort

     Then, submit the following note to the instructor:

     Your Name: _____ Score: _____

     Team Member #2: _____ Score: _____

     Team Member #3: _____ Score: _____

     Team Member #4: _____ Score: _____

Note: Please include a brief reason for any group member scoring either a "1" or a "5." I expect everyone to be thoughtful and diligent in completing this evaluation. ***You may get ZERO for the project if you receive "1"s from all other group members.***

Grading

Class participation	5%
Quizzes	15%
Mid-term Exam	20%
Project	30%
Homework (3)	30%

Grading Policy

The following grading scale will be used to translate number grades to letter grades:

96.5 – 100	A+
92.5 – 96.4	A
89.5 – 92.4	A-
85.5 – 89.4	B+
82.5 – 85.4	B
79.5 – 82.4	B-
69.5 – 79.4	C+
59.5 – 69.5	C
00.0 – 59.4	C-
92.5 – 100	D
89.5 – 92.4	F

Responsibility for Learning and The Honor Code

Being responsible for your own learning does not mean that you must always work in isolation. However, when working in groups we encourage you to be mindful of how much effort and Learning you are experiencing. Below, we outline our expectations for work in this course.

For HW problems, I encourage students to work together to solve and understand the problems. Nevertheless, each student is responsible for demonstrating he or she has good grasp of the material. Ultimately, each student's HW solution should reflect his or her own learning and be written in the students' own words. While students may work together to figure out how to solve the problems, each student must run his or her own analyses and turn in their own output. For the in-class quizzes, each should work independently, no discussion is allowed. Under no circumstance should a student email his or her HW solutions, project reports, and codes to a classmate. Working together (for the HW) is for the purpose of collaborating, not copying.

"As members of the academic community, students are expected to recognize and uphold standards of intellectual and academic integrity." As listed on https://deanofstudents.gsu.edu/student-conductpolicy-on-academic-honesty/.

Make-up Exam Policy

There is one mid-term exam in this course. Date for the exam is already set on the Tentative Course Schedule below. If there is an excusable reason for being unable to be present during the exam dates, please let me know as soon as possible to schedule a make-up exam. The make-up exam if at all possible will take place before the scheduled exam date. Students with unexcused absences for an exam will earn a 0 on the exam.

Special Needs

Students who wish to request accommodation for a disability may do so by registering with the Office of Disability Services. Students may only be accommodated upon issuance by the Office of Disability Services of a signed Accommodation Plan and are responsible for providing a copy of that plan to instructors of all classes in which accommodations are sought.

Tentative Course Schedule (Topics)

The course syllabus provides a general plan for the course; deviations may be necessary.

Week	Date	Topic	Corresponding Readings	Homework (HW) & Project
1	08/23/2018	Course Introduction, Introduction to Database	[1] Chapter 1 & 2	
2	08/30/2018	The Relational Database Model	[1] Chapter 3	
3	09/06/2018	Data Modeling & the Entity Relationship (ER) Modeling	[1] Chapter 4 & 5	HW1 out
4	09/13/2018	The Structured Query (SQL)	[1] Chapter 7	
5	09/20/2018	Advanced SQL	[1] Chapter 8	HW1 Due HW2 out
6	09/27/2018	Database Design, Data Index & Data Administration	[1] Chapter 9, 11 & 15	
7	10/04/2018	Business Intelligence and Data Warehouse	[1] Chapter 13	
8	10/11/2018	Mid-term Exam		HW2 Due
9	10/18/2018	Introduction to Unstructured Data Analytics	[2] Chapter 1 & 2	
10	10/25/2018	Unstructured data generation: Web Scraping, and introduction to MongoDB	Recommended reading [5] Chapter 6, 7, & 8	Final Project Out
11	11/01/2018	Hands-on MongoDB	Recommended reading [5] Chapter 6, 7, & 8	
12	11/08/2018	Clustering: k-means, Gaussian Mixture Models, Hierarchical clustering	[2] Chapter 5	HW3 Out

13	11/15/2018	Topic Modeling: Latent Dirichlet Allocation (LDA)	http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf	
14	11/22/2018	Thanksgiving Break		
15	11/29/2018	Classification: Nearest Neighbors Naïve Bayes, SVMs, Decision Trees, Random Forests,	[2] Chapter 4	HW3 Due
16	12/06/2018	Final Presentation and Report Due		

Your constructive assessment of this course plays an indispensable role in shaping education at Georgia State. Upon completing the course, please take the time to fill out the online course evaluation.