CIS 4730: Unstructured Data Management

2021- 05

Class Time: 04:45 PM - 07:15 PM (T, TH) Class Location: Zoom

Instructor: Rongen "Sophia" Zhang Email: rzhang6@gsu.edu

Office Hours: Optional Zoom drop-in hours / by appointment

[June 16th, 30th, July 14th 5: 00pm-6:00pm]

Prerequisites: CIS 3260 Database Management, CIS 3730 Introduction to Programming

Course Description

Data is the new oil; we need to find it, extract it, refine it, distribute it and monetize it.

– David Buckingham

Over 90 percent of digital data is unstructured – much of which is locked away across a variety of different data stores, in different locations and in varying formats. This course will discuss various issues and challenges in unstructured data management. At the same time, this course will introduce the best practices, underlying principles, and emerging technologies in storing, retrieving, and analyzing unstructured data.

Course Objectives

This course aims to prepare students with fundamental knowledge and skills in unstructured data management. After successful completion of this course, students will be able to:

1. Articulate the similarities and differences in managing structured and unstructured data
2. Collect and integrate unstructured data from multiple sources
3. Apply techniques to manage and store unstructured data
4. Prepare unstructured data for analysis
5. Use unstructured data to answer managerial questions and support decision-making
6. Develop and apply R programs for unstructured data management

Course Components

In general, each class meeting consists of two parts, separated by a 10-minutes break:

- Lecture: The first half of the class (before the break) in which we introduce and discuss theoretical knowledge on the weekly subject.
- Lab: The second half of the class (after the break) in which we develop and practice hands-on skills for unstructured data management. Students are expected to follow the instructor to code and submit the related lab assignments before next week of the class.

All course materials (lecture slides, lab notes, readings, project instructions, etc.) will be distributed electronically through the course website on iCollege.

Recommended Textbooks

(Note: We will mostly use class slides, notes, and reading materials, the following textbooks are recommended)

1. *Unstructured Data Analytics: How to Improve Customer Acquisition, Customer Retention, and Fraud Detection and Prevention*, by Jean Paul Isson (John Wiley & Sons, 2018). Free full text available from GSU library.

2. *The Definitive Guide to MongoDB: A Complete Guide to Dealing with Big Data Using MongoDB*, Second Edition, by David Hows, Peter Membrey, Eelco Plugge and Tim Hawkins (Apress, 2013). Free from GSU library.

Technology and Software Requirements

Because of the lab components of this course, students are required to either use the computers in the classroom or bring their own laptop. The computer/laptop should be able to install and execute the following software (all free):

1. **R**: R is a popular programming language for data analytics. We will learn basic R programming in our lab meetings. The final project and all lab exercises will be based on R.
2. **RStudio**: RStudio is an environment for R programming with many user-friendly features.
3. **MongoDB**: MongoDB is the most popular NoSQL database. We will learn about how to operate and query MongoDB databases.

COURSE SECTION POLICIES

Zoom Policies

• Students are expected to be punctual for class –- being tardy means missing important course announcements and disrupting the learning process for others. Students who arrive late are expected to enter and take a seat quietly.
• There will be 10-minute breaks during most classes. It will occur at regular intervals or based on the flow of the material.
• You are expected to be professional and respectful when attending class on Zoom.

- LOG in with your **full first name** and **last name** as listed on the class roster. Do not use a nickname or other pseudonym when you log in
- **Stay focused** for your own benefit. Please stay engaged in class activities. Close any apps on your device that are not relevant and turn off notifications.
- **Turn on your video when possible**. We would like to see your smiles. It is helpful to be able to see each other to create a pleasant and positive vibe for your classmates ☺
    - Exceptions: having limited internet bandwidth or being unable to find an environment without a lot of visual distractions
- **Mute your microphone when you are not talking**. This helps eliminate background noise.
- Be in a quiet place when possible. Find a quiet, distraction-free spot to log in. Turn off any music, videos, etc. in the background.
- Use the chat window for questions and comments that are **relevant to class**

Late or Make-up Work
No late or make-up projects or assignments will be allowed. Please submit your projects/assignments before the due date.

iCollege
Course materials, including slides, reading materials, problem sets, and solutions to quizzes or exams will be posted on the iCollege website of the course. The students are expected to check the iCollege website regularly and download the requisite materials.

Communication
The students' GSU e-mail addresses (as shown in GoSOLAR) will be used as the primary means of communication. Students should therefore check their GSU e-mail accounts regularly. The instructor cannot reply to iCollege emails from their GSU email account.

Student Evaluation

Gradable items are listed below (see the following sections for more details on each of these items). There will be occasional opportunities for extra credits.

Grading Component	Percentage
Mid-term exam	15%
Final Exam	15%
Lab assignments	30%
Lecture assignments	10%
Course project	30%
- Demo presentation (25%) - Peer evaluation (5%)	
Total	**100%**

Final Grade
The following grading scale will be used to calculate final grades:

Points	Grade
97.0 – 100.0	A+
93.0 – 96.9	A
90.0 – 92.9	A-
87.0 – 89.9	B+
83.0 – 86.9	B
80.0 – 82.9	B-
77.0 – 79.9	C+
73.0 – 76.9	C
70.0 – 72.9	C-

60.0 – 69.9	D
< 59.9	F

Exam

Exam questions will be a mix of multiple choice, true/false, and short-answer questions. The exam will cover only materials from lectures, including the slides and the required readings. In other words, the exam will not include lab related materials. The exam will only be open in iCollege on the day and time listed in the syllabus. Students missing an exam will receive a zero on that exam. Make-up exam will not be given. Exam may be taken on a different date only if the instructor is given a legitimate reason (jury duties, religious holiday, scheduled surgeries, pregnancy, etc.) Proof of reasons must be scanned and sent to the instructor ahead of time (unless absence was due to a legitimate emergency, in which case legal proof must be sent afterwards).

Lecture Assignments

We have two lecture assignments for this course. In each lecture assignment, you will be given a short video on a specific topic. The two topics (and their respective video) for this semester are the following:

1. **What is a vector?**
   - Video URL: https://www.youtube.com/watch?v=fNk_zzaMoSs
   - Due: 11:59pm, 6/29/2021
   - Note: Do not confuse this with the vectors in R. They are very different. In this assignment, you should summarize the idea of vectors solely based on content in the video. You won't earn any points if your summary is mainly about the vectors in R.

2. **How does a neural network recognize handwritten digits?**
   - Video URL: https://www.youtube.com/watch?v=aircAruvnKk
   - Due: 11:59pm, 7/13/2021

For each assignment, your task is to write a report to summarize the topic. The report must contain at least 500 words and you need to include at least 2 diagrams (or tables or screenshots) in your summary to illustrate the ideas.

Submissions should be made through the iCollege course website by their deadlines. No email/late submission will be accepted. It is student's responsibility to ensure the submissions went through in time. Network or technical issues (either on your machine or iCollege) cannot be used as an excuse.

Lab Assignments

Lab sessions will have several hands-on exercises and assignments. Requirements and details of these lab assignments will be provided in class handouts/slides. Each of the assignments is usually just a few lines of R script which reinforce topics that we have learned during the class. In three of these lab sessions, you will be asked to upload your work to iCollege. You are encouraged to discuss with peers if you have questions, but each person needs to turn in their work individually. If students are not able to submit their work by the end of the lab session, submissions are open for a week till the beginning of the next class (that is, by **next Monday 11:59pm**).

Submissions should be made through the iCollege course website by their deadlines. No email/late submission will be accepted. It is student's responsibility to ensure the submissions went through in time. Network or technical issues (either on your machine or iCollege) cannot be used as an excuse.

Course Project

The final project requires group work, and each group should consist of 4 students. For the group project, you need to identify a business problem that can be solved by acquiring, filtering, extracting, validating, cleansing, and analyzing unstructured data. The business problem can be from your organization (or a team member's organization). You need to identify a source through which you can collect the unstructured data you need to solve the business problem. You need to write the proper code to collect, analyze, visualize, and interpret the data. Based on your findings you need to provide some recommendations to solve the business problem you identified.

Detailed information about the project will be given in a separate handout. Project codes are due by 11:59 pm, Aug 1, 2021. To encourage teamwork spirit and minimize the free rider phenomenon, team members will evaluate each other's contribution right after the project presentation.

Submissions should be made through the iCollege course website by their deadlines. Email/late submission will have 40% discount of the score of the assignment. It is the student's responsibility to ensure the submissions went through in time. Network or technical issues (either on your machine or iCollege) cannot be used as an excuse.

Attendance and Participation

Class attendance, while not mandatory, is expected. The instructor encourages everyone to participate in class activities, discussions, and respond to questions from other students. In evaluating your class participation in discussions, both the quantity and quality of participation is taken into account. Principles for class participation include:

1. Show a respectful and positive attitude towards him/herself, classmates, and teacher
2. Work with others in paired/group-based exercises
3. Contribute to classroom discussion
4. Attend classes and focus on class work (that is, not using social media, sending emails/texts, or doing anything irrelevant to class activities)

Grading Policy

It is expected that no more than 35 percent of the students in a given class section will receive a grade of A+, A, or A-. The majority of the remaining students are expected to receive grades of B+, B, or B-. Those students demonstrating significantly lagging performance shall earn grades at the C-level or lower as appropriate.

A grade of "C-\" is considered a passing grade for this course and a "C-" is considered passing for prerequisite purposes for this course as well as for all electives. Refer to the University catalog for information concerning +/- grading and quality points for GPA calculations.

Course Schedule

The course outline below provides a general plan for the class. However, the plan is subject to change to accommodate students' learning progress and unexpected events. All changes to the outline will be updated and posted in iCollege.

Tentative Class Schedule		
Date	**Lecture**	**Lab**
June 8	Student Survey, Introduction to unstructured data	R: Introducing R and RStudio
June 10	XML and JSON	R: Data types
June 15	NoSQL databases	R: Data input & output & summary
June 17	MongoDB	MongoDB Atlas & Robo 3T
June 22	Information retrieval	R: Flow control
June 24	Scoring, weighting & vector space	R: Data manipulation
June 29	Regular expressions	R: Text processing
July 1	Mid-term Exam	Team Project Overview
July 6	Web crawling & web APIs	R: Web scraping
July 8	Text categorization & clustering	R: Reviewing R labs
July 13	Deep learning applications	Review and Course wrap-up
July 15	Final Exam	
July 20	Guest Lecture (Zaiyong Zhang - Siemens Gamesa Senior Data Analyst)	Working on Project
July 22	Emerging topics and applications	Working on Project
July 27	Project Presentation	

Important Note

This syllabus provides a general guideline for the conduct of this course; however, deviations may be necessary. Updates will be given during the semester and posted online through icollege

Some R Resources

In addition to the class handouts, there are a number of good resources available on the web for both learning R and seeking answers to questions about how to accomplish various tasks.

- The official intro, "An Introduction to R", available online in HTML and PDF
- Quick-R. A website with very clear and well-organized tutorials.
- Thomas Lumley, "R Fundamentals and Programming Techniques" (a large PDF slide deck)
- R Reference Card
- Online resources such as Stack Overflow and the R-help Mailing List
- RStudio Documentation

- R and Data Mining: Examples and Case Studies
- ... and so much more

Academic Honesty

Students may have general discussions about assignments with fellow classmates, but each student must develop his or her solution to each Mini-Project. It is each student's responsibility to keep his/her own work secure. DO NOT share computer files of Mini-Project Assignments with classmates. Failing to adequately protect one's work does not relieve the student from academic dishonesty charges.

University regulations will be enforced regarding dishonorable or unethical conduct (Cheating, Plagiarism, Falsification, Unauthorized Collaboration or Multiple Submissions). The penalties for incidents of academic dishonesty can lead to expulsion from the University (see General Catalogue p. 64, Student Handbook p. 130 or http://www2.gsu.edu/~wwwdos/codeofconduct_conpol.html).

In this class, there will be zero tolerance for dishonorable or unethical conduct. Electronic or physical sharing of answers will be considered cheating and will not be tolerated.

Cheating on examinations involves giving or receiving unauthorized help before, during, or after an examination. Examples of unauthorized help include sharing information with another student during an examination, intentionally allowing another student to view one's own examination, and collaboration before or after an examination which is specifically forbidden by the instructor.

Submission for academic credit of a work product, or a part thereof, represented as its being one's own effort, which has been developed in substantial collaboration with assistance from another person or source, or computer-based resource, is a violation of academic honesty. It is also a violation of academic honesty to knowingly provide such assistance. Collaborative work specifically authorized by an instructor is allowed. (*Collaboration on all individual assignments is forbidden. If your instructor discovers that you have had unauthorized assistance or collaboration, the instructor is obligated to file a report with the Dean's Office.)*

If a student is charged with Academic Dishonesty, for each charge, a zero (0) will be given for the assignment, a minimum of point equivalent of one final grade (i.e. B- to a C-) will be deducted from the final course total points and a written Notice of Academic Dishonesty will be given to the Dean's office. The student will also receive a copy of the notice.

Unless specifically stated by the instructor, all exams are to be completed by the student alone. Within-group collaboration is allowed on project work.

Collaboration between project groups will be considered cheating unless specifically allowed by an instructor. Copying work from the Internet without a proper reference will be considered plagiarism and subject to disciplinary action as delineated in the Student Handbook.

Student Assessment

Your constructive assessment of this course plays an indispensable role in shaping education at Georgia State and helping to improve this course for future students. You can send emails to instructor to provide feedback during the semester. Upon completing the course, please take time to fill out the online course evaluation. Your feedback is much appreciated. ☺