CIS 8045 – Unstructured Data Management

Class Time: Thursday 5:30-9:45pm
Class Modality: Hybrid (F2F: GSU Buckhead Executive Ed Ctr 306; Online: WebEx)
Instructor: Ling Xue Email: lxue5@gsu.edu
Office: RCB 907 Office Hours: By Appointment

Prerequisites: CIS 8040

Covid-19 Policy: Masks, social distancing, and vaccination are not mandatory for the F2F sessions. However, for your own health and the safety of others, **I strongly recommend you to mask up and practice social distancing all the time, and get vaccinated for taking this course**.

Catalog Description:
This course addresses the *unstructured* data management skills needed for modern data analysis, including those salient to big data and real-time data environments. The focus is on unstructured data and its environment. Unstructured data includes various types of free-form data (such as text and image), user generated content, social media, location-aware data, and digital media among others. Topics covered include the unconventional techniques used to manage and represent unstructured data (such as various NoSQL approaches) for the analytics purpose and methods used to analyze unstructured data (such as text analytics).

Course Description:
This course addresses the *unstructured* data management skills needed for modern data analysis including those salient to big data and real-time data environments. Unstructured data includes various types of free-form data (such as text and image), user generated content, social media, location-aware data, and digital media among others. The course focuses on the unconventional techniques used to manage and represent unstructured data (such as various NoSQL approaches) for the analytics purpose and methods used to analyze unstructured data (such as text analytics). Students will learn about various software and databases as well as methods for storing, manipulating, and mining unstructured data for better decision-making.

Course Credit: 3.0 Credit Hours

Course Objectives:
Upon completion of the course, students should be able to:
- Articulate similarities & differences between managing structured and unstructured data.
- Apply various NoSQL techniques to manage and represent unstructured data
- Integrate data from multiple sources (such as online sites and social media)
- Prepare unstructured data for analysis
- Apply techniques for text analytics and image processing

Recommended Readings:
MongoDB Fundamentals. 2020. By Amit Phaltankar, Juned Ahsan, Michael Harrison, Liviu Nedov. Published by Packt Publishing. ISBN (e-book): 978-1839213045

Hands-On Graph Analytics with Neo4j. By Estelle Scifo. Published by Packt Publishing. ISBN (e-book): 978-1839215667

The Deep Learning Workshop. 2020. By Mirza Rahim Baig, Thomas V. Joseph, Nipun Sadvilkar, Mohan Kumar Silaparasetty, Anthony So. Published by Packt Publishing. ISBN (e-book): 978-1839210563

Required Software:

NoSQL systems: MongoDB and RoboMongo (GUI client), Neo4j and Neo4j Desktop,
Development tools: Python (in Google Colaboratory, no installation required);

Homework Assignments:

A series of hands-on homework assignments will be provided to you to further explore the topic/technique covered in class. Each is an individual activity. With these hands-on assignments, students gain proficiency in the various software and databases assigned for this class.

Term Project:

A group-based term project is for students to integrate the NoSQL skillsets with those of text analytics.

Typical class session:

Class sessions will comprise (1) lectures/discussions of relevant techniques, concepts and features, (2) instructor demonstrations, and (3) student lab sessions with hands-on work. The purpose of this pedagogical approach is to introduce and reinforce ideas and skill sets so that you can master these on your own after class hours. To bring this knowledge to a highly proficient, professional level, you will have to spend time and effort outside of class reviewing and practicing the class material. To ensure that you have the basic knowledge that will allow you to function on your own after class, be sure to ask the instructor questions during class, either during the lecture/discussion, demo, or lab.

Class Attendance

All students are required to attend all classes and complete in-class exercises, except when precluded by emergencies, religious holidays or bona fide extenuating circumstances. If one or more class is missed, it is the student's responsibility to determine the specific material covered during their absence and make the necessary arrangements for making up what is missed.

Course Grading

Grading Component	**Percentage**
Homework Assignments	25%
Term Project	30%
Final Exam	35%
Class Participation (In-Class Exericise)	10%

Total	**100%**

A+ = >=98	A = 93 – 97.9	A- = 90 – 92.9
B+ = 87 – 89.9	B = 82 – 86.9	B- = 80 – 81.9
C+ = 76 – 79.9	C = 71 – 75.9	C- = 68 – 70.9
D+ = 65 – 67.9	D = 60 – 64.9	F <= 59.9

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. If the class cannot be held at the scheduled time or place, it may be held via an online forum.

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 and at-home assignments 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.

Tentative Class Schedule			
Class Date	**Topic**	**Modality**	**Due**
1: 1/6	Introduction to NoSQL databases. MongoDB basics: data representation, and basic query, etc.	Online	
2: 1/13	MongoDB advanced queries and data manipulation (e.g., aggregation).	F2F	
3: 1/20	MongoDB advanced use cases: e.g., index, PyMongo, etc.	Online	
4: 1/27	Neo4j basics: data representation, manipulation and query, etc.	F2F	HW1
5: 2/3	Neo4j advanced use cases.	Online	
6: 2/10	Processing unstructured text data using deep learning	F2F	HW2
7: 2/17	Processing unstructured image data using deep learning	Online	
8: 2/24	Other topics **Final Exam**	F2F	HW3

Note: HW-Homework; Online sessions would be delivered through WebEx in iCollege.