CS 7311: Data-Driven Computational Methods and Infrastructure

Fall 2018

Syllabus updated Aug 30 2018

Instructor	Dr. Jelena Tešić (pronounced as Yeh-LE-nah TE-shich)
Office Number	CMAL 307B
Email Address	**jtesic@txstate.edu**
Faculty Profile	https://cs.txstate.edu/accounts/profiles/j_t463/
Office Hours	Tue Thu 3:30 – 5 p.m. or by appt

Section Information	CS7311 DERR
Class Meetings	Thursday 6:30 p.m. – 9:20 p.m.
Open Labs	DERR 231 (Linux Lab) MCS 590 (Windows Lab)

Course Description

This course covers computational and statistical methods for using large-scale data sets ('big data') to answer scientific and business questions. It focuses on framing research questions, understanding how data can answer them, and using modern software tools such for scalable data storage, processing, and analysis.

Course Objectives

The students will be able to:

- Formulate concrete research questions to address business or scientific objectives
- Identify or collect data to answer research questions
- Design tools to process, clean, and organize data for subsequent analysis
- Create and run data processing and analysis pipelines to compute statistical results over large-scale data sets using modern high-performance computing infrastructure
- Present results clearly using data visualizations and written prose
- Interpret analysis results and identify their implications for business concerns or scientific interest
- Determine appropriate data processing technology to support a desired analysis method

Course Material

There is no official textbook
Instructor will provide set of online references and tutorials for specific data science problems

What is expected of student registered for CS 7311 in Fall 2018?

Students are expected to:

(1) Attend instructional meetings, in person or via Zoom.
(2) **Read announcements** from the instructor posted on TRACs course site.
(3) Be informed and prepared for the class
(4) Schedule presentations on time
(5) Clearly communicate with the instructor regarding and issues, delays or unforeseen circumstances in timely manner. Emailing is the fastest way to reach the instructor.

Assessment

Survey and Tutorial Presentation in Class 30pt **- schedule it with the instructor, FCFS, announced in TRACS**

- Data Science paper – selection will be provided by instructor on TRACS – you are welcome to propose a paper you want to present
- Walk through tutorial on data science tools and/or walk-through examples in Jupyter Notebook
- In depth analysis of chosen pipeline – Team Presentation and Discussion

Research Project 50 pt

- Formulate the business or scientific objective where data science pipeline
  - Identify or collect data to answer research questions
- Design tools to process, clean, and organize data for subsequent analysis
  - Determine appropriate data processing technology to support a desired analysis method
- Setup team codebase (github.txstate.edu) and collaboration tools. – due **Thu Oct 25th at 11:55 p.m.**
- Create and run data processing and analysis pipelines on modern high-performance computing infrastructure (if possible)
- Interpret analysis results and identify their implications for business concerns or scientific interest
- Clean, package, share, and deliver the code using github.txstate.edu repository**- due Thu Dec 13th at 11:55 p.m.**

Project Presentation in Class 40pt

- Present your research idea, related work, data source, and what tools and framework you plan to use
- Write a paper/report draft and submit it to instructor for review – **due Thu Oct 25th at 11:55 p.m.**
- Present the final project result in class
- Write a scientific research paper or business proposal - **due Thu Dec 13th at 11:55 p.m.**

Extra Credit: 100 pt range, you can earn up to 120 pt

Communication

Best way to contact the instructor is to send her an email.

All announcements, resources, and updated will be posted on TRACS.
https://tracs.txstate.edu/
We will use the TRACS website for the following:

- Announcements (Announcement Tool)
- Grades (Gradebook tool)
- Programming assignment submissions will be using assignments tool or git.txstate.edu
- Resources (Resources tool)
  - lectures – pdf copies of lecture notes posted prior to the lecture
  - papers – main repository for all papers presented and discussed in the class

Course Schedule

Following is a tentative schedule for the class. **Exact topics and dates may be updated as the course progresses.**

Class	Date	Lecture	Assignments/In Class Presentations
1	Aug 30	Introduction	Define Research Problem
2	Sep 6	Fundamentals of Data Science	Present Research/Tool/Paper
3	Sep 13	Data Science Tools	Present Research/Tool/Paper
4	Sep 20	Frameworks	Present Research/Tool/Paper
5	Sep 27	Methodology	Present Research/Tool/Paper
6	Oct 4	Pipeline	Present Research/Tool/Paper
7	Oct 11	Analysis and Improvement	Present Research/Tool/Paper
8	Oct 18	…	Project Checkpoint Presentation
9	Oct 25	…	Project Checkpoint Presentation
10	Nov 1	Data Visualization	Analysis of proposed pipelines
11	Nov 8	Data Science Pipelines	Analysis of proposed pipelines

12	Nov 15	Data Science Pipelines	Analysis of proposed pipelines
13	Nov 29	Demos and Final Project Presentations	
14	Dec 6	Demos and Final Project Presentations	
15	**Dec 13**	**Demos and Final Project Presentations (FINAL EXAM DATE)**	

Policies

Grade Grievance Policy: If a student believes a mistake has been made in grading an assignment, the student has one week after an assignment is returned to resubmit an assignment for re-grading if they believe there is an error.

Drop Policy Students will not be automatically dropped for non-attendance: if you are planning to drop the class or withdraw from the class, follow the instructions listed on registrar's web site:
http://www.registrar.txstate.edu/registration/drop-a-class.html
It is your responsibility to be familiar with the University Policy on dropping classes as described in the catalog and the TXSTATE website (see), to observe relevant deadlines, and to follow proper procedures for dropping classes.

Incomplete Policy CS department has a strict policy regarding 'Incomplete grade'. It has to be approved by the chairman and thus an 'Incomplete grade' will only be granted under unexpected and truly severe situations, which must be supported by some official documents.

E-mail Policy: During the work week, instructor will respond to personal emails within 24 hours. Instructor will review communication over the weekend but will respond on Monday to most situations. If you need to reach me by email, please use the subject line: Your Name, Course Name/Number, Topic. Please allow a full 24 hours before emailing me again about the same question or issue, and on Monday for inquiries sent over the weekend.

Plagiarism Policy: Except where explicitly and specially allowed (such as group project), all work submitted in class is expected to be your individual work. Plagiarism will not be tolerated and if detected will result in an automatic 'F' grade. Please refer to http://www.txstate.edu/effective/upps/upps-07-10-01.html for Texas State's Honor Code.

Extra Credit Policy: See Assesment

Accommodations for students with disability

Any student requiring special accommodations, should inform me during the first two weeks of classes. The student should also contact the office of disability services at the LBJ student center. Students who qualify for extra time for exams must take their test with ATSD and must schedule their test at the same time the test is given in class.

Academic Honor Code and Conduct

You are expected to adhere to

- the University's Academic Honor Code http://www.txstate.edu/honorcodecouncil/Academic-Integrity.html
- Code of Student Conduct - http://www.dos.txstate.edu/handbook/rules/cosc.html
- Texas State Mission and Shared Values: http://universityplan2023.avpie.txstate.edu/overview/Texas-State-Mission-and-Goals.html.