CS 696 Intro to Big Data: Tools and Methods
Fall Semester, 2020 Syllabus

San Diego State University -- This page last updated 1/22/20

CS 696 Intro to Big Data: Tools and Methods Fall 2020

Instructor	Roger Whitney
Office	GMCS-561
Phone	619-594-3535
Email	whitney at sdsu.edu
Office Hours	3:15-5:15 pm Tuesday, Thursday

Course WWW Site: http://www.eli.sdsu.edu/courses/spring20/cs696/index.html
The course lecture notes, assignments, course videos and course wiki available at above course web site.

Texts:

Python Data Science Handbook, Jake VanderPlas, O'Reilly Media, Inc, 2017

Spark: The Definitive Guide, Matei Zaharia, Bill Chambers, February 2018, 9781491912218

The books are available on-line via SDSU library. See course website for links. In general I use on-line documentation and multiple texts to understand the different technologies we will use in this course. As the course progresses I will add to this list, but will restrict the books to those that are available for free on-line. Python Data Science Handbook is also available on-line at: https://jakevdp.github.io/PythonDataScienceHandbook/

Prerequisites: CS310 (Data Structures), Math 254, Math 245

Spark & Amazon's Cloud:
The class will use Amazon's cloud to run Spark on a cluster. Students will create their own accounts on Amazon cloud. This will incur a cost to the student which should be less than the cost of a text book. You can apply for a AWS Education account which gives you $100 of compute time for free. See course web site for link to Amazon's AWS Education program.

Short Description:
Tools, methods and practices to analysis, curate, search, query, visualize big data, ie data characterized by volume, variety, velocity, variability and veracity. These tools, methods include statistics, machine learning, visualization, no-sql databases, MapReduce, Hadoop**.**

Detailed Description:
This course assumes no prior experience with Data Science or Big Data. The goal of Big Data is to gain insight by analyzing data. The course will look at some statistical tech-

niques, machine learning and visualization to help analyze data. However most of the course will focus on software tools (Spark, Cassandra, Kafka, Jupyter Notebooks) to process data on clusters. We will use Jupyter notebooks in assignments to combine graphs, code, and text. Before one can analyze data it needs to be stored. While CSV and DataFrames are fine for small datasets we will look at Hierarchical Data Format (HDF5) and the Hadoop Distributed File System (HDFS). The latter needed when datasets are so large we need multiple machines to process the data. Also for storing data we will use a NoSQL database. NoSQL databases are common in Big Data. Multiple machines are required to analysis larger datasets. We will use Spark to scale to multiple machines. We will be using a cloud provider (Amazon) to run Spark on multiple machines. You will have to create an account on Amazon AWS. Using multiple machines to solve one problem introduces a number of issues. The problem has to be divided into separate pieces that can be solved independently. Data has to be distributed and results need to be communicated between machines. If not done correctly we will not see the benefit of using multiple machines.

While this is the second time the course has been offered it is being changed a fair amount. The course uses a lot of different technologies is a rapidly changing area. As a result there will be some rough edges to the course. You might have problems installing some software on your machines, or have connection issues. This is why the course is offered at the graduate level as graduate students should have the ability to deal with these issues.

Topics Covered

Python
Hadoop HDFS
Spark
    RDD, Spark SQL, Spark ML
Visualization - Matplotlib
Basic Statistics
    Correlation, conditional probability
No-SQL database
    Cassandra
Machine Learning
    Supervised, Unsupervised
    Linear Regression, K-nearest neighbor
Distributed Streaming
    Kafka

Learning Outcomes:

- Students will be able to implement programs to process and analyze big data
- Students will be able to use utilize Spark and no-sql databases to process and analyze big data
- Students will be able to set up a data pipeline to ingest data from databases (SQL or NO-SQL) and process the data on a cluster
- Use Notebooks to organize & run code, display and explain results

Information & Downloads:

Jupyter - http://jupyter.org/
Spark - http://spark.apache.org/
Kafka https://kafka.apache.org/
Cassandra http://cassandra.apache.org

Grading: Your grade in this course will be determined given below. There will be about 5 assignments.

Homework, Programs	55%
Exams (1)	25%
Project	20%

The exam will be on March 24. The project is due May 14.

Crash Policy: After registration closes adding students to the class will be handled via the waitlist system. If you are not able to register for this class you need to add your self to the class waitlist. The instructor has no ability to add any student to the course nor can the instructor see who is on the wait list.

Late Policy: Late homework will be accepted, but with a penalty. An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 80%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.

Email & Assignments: Unless indicated by the assignment all assignments are to be submitted to the course repository. No assignments will be accepted via email.

No Extra Credit: There will not be any extra credit assignments. There will not be any extra credit problems in the assignments.

Cheating: Any one caught cheating will receive an F in the course.

UNIVERSITY POLICIES

Accommodations: If you are a student with a disability and are in need of accommodations for this class, please contact Student Ability Success Center at (619) 594-6473 as soon as possible. Please know accommodations are not retroactive, and I cannot provide accommodations based upon disability until I have received an accommodation letter from Student Ability Success Center.

Student Privacy and Intellectual Property: The Family Educational Rights and Privacy Act (FERPA) mandates the protection of student information, including contact information, grades, and graded assignments. I will use Blackboard to communicate with you, and I will not post grades or leave graded assignments in public places. Students will be notified at the time of an assignment if copies of student work will be retained beyond the end of the semester or used as examples for future students or the wider public. Students maintain intellectual property rights to work products they create as part of this course unless they are formally notified otherwise.

Religious observances: According to the University Policy File, students should notify the instructors of affected courses of planned absences for religious observances by the end of the second week of classes.

Academic Honesty: The University adheres to a strict policy prohibiting cheating and plagiarism. Examples of academic dishonesty include but are not limited to:

- copying, in part or in whole, from another's test or other examination;
- obtaining copies of a test, an examination, or other course material without the permission of the instructor;
- collaborating with another or others in work to be presented without the permission of the instructor;
- falsifying records, laboratory work, or other course data;
- submitting work previously presented in another course, if contrary to the rules of the course;
- altering or interfering with grading procedures;
- assisting another student in any of the above;
- using sources verbatim or paraphrasing without giving proper attribution (this can include phrases, sentences, paragraphs and/or pages of work);
- copying and pasting work from an online or offline source directly and calling it your own;
- using information you find from an online or offline source without giving the author credit;
- replacing words or phrases from another source and inserting your own words or phrases.

The California State University system requires instructors to report all instances of academic misconduct to the Center for Student Rights and Responsibilities. Academic dishonesty will result in disciplinary review by the University and may lead to probation, suspension, or expulsion. Instructors may also, at their discretion, penalize student grades on any assignment or assessment discovered to have been produced in an academically dishonest manner.

Resources for students: A complete list of all academic support services--including the Writing Center and Math Learning Center--is available on the Student Affairs' Academic Success website. Counseling and Psychological Services (619-594-5220) offers confidential counseling services by licensed therapists; you can Live Chat with a counselor at http://go.sdsu.edu/student_affairs/cps/therapist-consultation.aspx between 4:00pm and 10:00pm, or call San Diego Access and Crisis 24-hour Hotline at (888) 724-7240.

Classroom Conduct Standards: SDSU students are expected to abide by the terms of the Student Conduct Code in classrooms and other instructional settings. Prohibited conduct includes:

- Willful, material and substantial disruption or obstruction of a University-related activity, or any on-campus activity.
- Participating in an activity that substantially and materially disrupts the normal operations of the University, or infringes on the rights of members of the University community.
- Unauthorized recording, dissemination, or publication (including on websites or social media) of lectures or other course materials.
- Conduct that threatens or endangers the health or safety of any person within or related to the University community, including
    1. physical abuse, threats, intimidation, or harassment.
    2. sexual misconduct.

Violation of these standards will result in referral to appropriate campus authorities.

Medical-related absences: Students are instructed to contact their professor/instructor/coach in the event they need to miss class, etc. due to an illness, injury or emergency. All decisions about the impact of an absence, as well as any arrangements for making up work, rest with the instructors. Student Health Services (SHS) does not provide medical excuses for short-term absences due to illness or injury. When a medical-related absence persists beyond five days, SHS will work with students to provide appropriate documentation. When a student is hospitalized or has a serious, ongoing illness or injury, SHS will, at the student's request and with the student's consent, communicate with the student's instructors via the Vice President for Student Affairs and may communicate with the student's Assistant Dean and/or the Student Ability Success Center.

SDSU Economic Crisis Response Team: If you or a friend are experiencing food or housing insecurity, or any unforeseen financial crisis, visit sdsu.edu/ecrt, email ecrt@sdsu.edu, or walk-in to Well-being & Health Promotion on the 3rd floor of Calpulli Center.