CS 696 Intro to Big Data: Tools and Methods Fall 2016

Instructor	Roger Whitney
Office	GMCS-561
Phone	619-594-3535
Email	whitney at sdsu.edu
Office Hours	3:15-5:15 pm Tuesday, Thursday
	10 am to Noon Friday
Credits	3 Units

Course WWW Site: http://www.eli.sdsu.edu/courses/fall16/cs696/index.html
The course lecture notes, assignments, course videos and course wiki available at above course web site.

Texts:
G**etting Started with Julia,**   Ivo Balbaert, Packt Publishing 2015, 178328479X
Julia for Data Science, Voulgaris, Technics Publications, 2016, 1634621301
Hadoop: The Definitive Guide 4th Edition, O'Reilly Media, 2015,1491901632

The books are available on-line via SDSU library. See course website for links.

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

Hadoop & Amazon's Cloud:
The class will use Amazon's cloud to run Hadoop 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.

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 techniques, machine learning and visualization to help analyze data. To perform the analysis we will use the new programming language Julia, as it is interactive, fast, fairly simple and interacts with existing languages. Common data science languages are either slow (Python, Matlab, R), not interactive (Java) or complex (Scala). To support insights into datasets one needs visualization, code and text explaining the process used and insights

gained. 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. Time permitting the course might use a NewSQL database. Multiple machines are required to analysis larger datasets. We will use Hadoop and Spark to scale to multiple machines. Hadoop to use MapReduce in batch mode and Spark to handle stream processing. We will be using a cloud provider to run Hadoop & Spark on multiple machines, likely Amazon or Cloudera, on which you will have to create an account. 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.

This is the first time that this course has been offered. 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 should expect that there will be some dealing with some of the technologies. 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
  Julia
  Visualization
  Basic Statistics
      Correlation, conditional probability, Bayes, Hypotheses testing
  Time Series Analysis
  No-SQL databases
  Machine Learning
      Supervised, Unsupervised
      Linear Regression
      K-nearest neighbor
      K-means
  Data Cleaning
  Hadoop & Spark
      Hadoop architecture, HDFS, Mahout
      MapReduce, Pig

Learning Outcomes:
  • Students will be able to implement programs to process, visualize and analyze big data
  • Students will be able to use utilize Hadoop and no-sql databases to process and analyze big data
  • Students will be able to implement programs to process, visualize and analyze big data

Information & Downloads:
    Julia - http://julialang.org/downloads/
    Atom/Juno (Julia IDE) - http://junolab.org/
    Jupyter - http://jupyter.org/
    Hadoop - http://hadoop.apache.org/
    Spark - http://spark.apache.org/

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

Homework, Programs	1/2
Exams (2)	1/2

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.

Disabled Students: If you are a student with a disability and believe you will need accommodations for this class, it is your responsibility to contact Student Disability Services at (619) 594-6473. To avoid any delay in the receipt of your accommodations, you should contact Student Disability Services as soon as possible. Please note that accommodations are not retroactive, and that accommodations based upon disability cannot be provided until you have presented your instructor with an accommodation letter from Student Disability Services. Your cooperation is appreciated.