Syllabus: CIS 4930/6930 Spring 2014

Introduction to Data Science/Data Intensive Computing

Catalog Description: See UF course catalog.

Pre-requisites and Co-requisites:

Data Structures and Algorithms (COP3530) or equivalent.

Optional but preferred: Information and Database Systems I (CIS 4301) and an introductory course to probabilities and statistics are pre-requisites.

Course Objectives:

In order to address the growing need from both industry and academia (e.g., medical and bio informatics, financial, law enforcement, economics, decision support, social networks) for big data analytic skills including, data management, data mining, machine learning and data visualization, this is the first of the three-course series in the Data Science curriculum. The aim is to bring student with basic programming and data structure background to be abreast with common tools used for Data Science application development. This course serves as an introduction to the basics of data science including programming for data analytics, file management, relational databases, classification, clustering and regression. The foundation is laid for big data applications ranging from social networks to medical and business informatics.

Instructor:

Daisy Zhe Wang, Office: E456, E-mail: daisyw@cise.ufl.edu, Office Hours: Mon 1:40-3pm/Wed 1:40-2:30pm or by appointment. Covers Part 1 (Data Manipulation, at Scale) and Part 3 (Graph/Text Data Analysis & Communicating Results) in the first 7 weeks of class (Jan-Feb).

Sanjay Ranka, Office: E532 Email: ranka@cise.ufl.edu Office Hours: Wed/Fri 3:00-4:00pm or by appointment. Covers Part 2 (Statistical Analytics) and Part 4 (Parallel Computing) in the second 7 weeks of class (March-April).

Course Information:

- Credits: 3
- Section: 111H/11C3/11F9/1A04
- Meeting Times: MWF: 6th (12:50 to 1:40pm)
- Where: CSE Building, E121
- Teaching Assistant: Kun Li and Yupeng Yan
- Laboratory: N/A
- Material and Supply Fees: None
- **Class web page:** [http://www.cise.ufl.edu/class/cis6930sp14IDS/](http://www.cise.ufl.edu/class/cis6930sp14IDS/)

**Textbooks and Software Required:** None. We will be using Amazon Web Services (AWS) and software supported on top of AWS. AWS credits will be given to each student.

**Recommended Reading:**


**Course Outline and Topics:**

This course will give an introduction to the basic data science techniques including programming in SQL, Map-Reduce, R, and Python. We also cover topics including relational databases, data visualization, classification, clustering, regression and parallel computing platforms. The topics are as follows:

**Part 0: Introduction**

**Part 1: Data Manipulation, at Scale**

- MapReduce, Hadoop, relationship to databases, algorithms, extensions, languages
- Databases, SQL and the relational algebra
- Parallel databases, parallel query processing, in-database analytics
- Key-value stores and NoSQL; tradeoffs of SQL and NoSQL

**Part 2: Statistical Analytics**

- Programming in Python and R
- Basic Data Mining
  - Basic statistical modeling, introduction to machine learning, overfitting
  - Supervised learning: Linear and Logistic Regression, Classification
  - Unsupervised learning: Clustering, Association Rule mining

**Part 3: Graph/Text Data Analysis & Communicating Results**

- Graph Analytics: PageRank, community detection, recursive queries, iterative processing
- Text Analytics: TF/IDF, conditional random fields
- Visualization, data products, visual data analytics

**Part 4: Parallel Computing**
Concurrency and Data Decomposition
- Message Based Parallelism – MPI
- Thread Based Parallelism – OpenMP

Attendance and Expectations:
- We strongly encourage class attendance and participation.
- Please return your projects in time. Late returns will cause 20% deduction in your grade for that project/presentation for each late day.
- If I postpone or cancel the office hour, I will post it in the announcements.
- Please avoid any activities that will disturb the flow of the lectures: Silence your cell phones, pagers, etc.

Grading Policy – Methods of Evaluation

Projects (100 %)

Grading Scale:
Roughly the boundaries will be:

- 90 -- 100 A
- 85 -- 89 B+
- 80 -- 84 B
- 75 -- 79 C+
- 70 -- 74 C
- 65 -- 69 D+
- 60 -- 64 D
- 0 -- 59 E

The boundary for A-, B-, C- will be decided at the end of the semester.

Makeup Exam Policy: N/A

Honesty Policy:
All students admitted to the University of Florida have signed a statement of academic honesty committing themselves to be honest in all academic work and understanding that failure to comply with this commitment will result in disciplinary action. This statement is a reminder to uphold your obligation as a UF student and to be honest in all work submitted and exams taken in this course and all others.

Accommodation for Students with Disabilities:
Students requesting classroom accommodation must first register with the Dean of Students Office. That office will provide the student with documentation that he/she must provide to the course instructor when requesting accommodation.

**UF Counseling Services:**

Resources are available on campus for students having personal problems or lacking clear career and academic goals. The resources include:

- University Counseling Center, 301 Peabody Hall, 3921575, Personal and Career Counseling.
- SHCC mental Health, Student Health Care Center, 3921171, Personal and Counseling.
- Center for Sexual Assault/Abuse Recovery and Education (CARE), Student Health Care Center, 3921161, sexual assault counseling.
- Career Resource Center, Reitz Union, 3921601, career development assistance and counseling.

**Software Use:**

All faculty, staff and student of the University are required and expected to obey the laws and legal agreements governing software use. Failure to do so can lead to monetary damages and/or criminal penalties for the individual violator. Because such violations are also against University policies and rules, disciplinary action will be taken as appropriate. We, the members of the University of Florida community, pledge to uphold ourselves and our peers to the highest standards of honesty and integrity.