**­LIBRARY AND INFORMATION SCIENCE**

**Morgridge College of Education—University of Denver**

**LIS 4235 | Scripting for Large Databases (4 Credit Hours)   
Spring 2018**

**COURSE SYLLABUS**

**Instructor:** Peter Organisciak, Assistant Professor  
**E-mail:** Peter.Organisciak@du.edu

**Phone:** 303-871-3587

**Office Hours:** Mondays 12:30pm-2:30pm by appointment (<https://organisciak.youcanbook.me>)

Meetings during office hours are in my office, or online by arrangement.

Let me know if you need to meet at another time.

**Office:** Ruffatto Hall Room 247

**Class Meeting Times:** Tuesday 3:30 to 6:50 p.m. in Ruffatto Hall Room 305

## COURSE DESCRIPTION

Scripting for Large Databases is a bridge between structured data storage and data analysis. It will equip students to learn more from databases in aggregate, building foundational data science skills in the process. This course will introduce students to the basics of data storage and acquisition as part of a multi-step data gathering, processing, analysis, and visualization effort. The logic and structure of relational databases will be reviewed as well as non-relational (NoSQL) alternatives. Techniques and methods for automation and scalable data processing will be introduced using the Python programming language, with a focus on the SciPy Stack of data science libraries. Additionally, big data workflows will be covered. Skills will be integrated and applied by the student through the use of prepared data sources, along with use of APIs and web scraping techniques to acquire data through internet sources.

## LEARNING OUTCOMES

Upon satisfactory completion of the course, students will be able to:

1. Demonstrate knowledge of relational database fundamentals.
2. Use SQL and Pandas for structured data querying and analysis.
3. Understand the contexts of use for NoSQL, and apply them with MongoDB.
4. Apply fundamental data mining concepts in information science contexts.
5. Collect and organize data from the web.
6. Understand aggregations and transformations for structured and semi-structured data.
7. Understand and apply large data patterns, such as Map-Reduce and Split-Apply-Combine.

## SCHEDULE OVERVIEW

|  |  |  |
| --- | --- | --- |
| Week 1 | Mar 29 | Relational Databases, Data types, Introduction to the SciPy Stack |
| Week 2 | Apr 5 | Relational Databases continued, Data collection, Jupyter and Python |
| Week 3 | Apr 12 | SQLite, Basic Querying, Introduction to Pandas |
| Week 4 | Apr 19 | Split-Apply-Combine, Deletions and Updates, Indexing |
| Week 5 | Apr 26 | Summation, aggregation, and transformation |
| Week 6 | May 3 | Visualization |
| Week 7 | May 10 | MongoDB - Ingesting Data, document aggregations |
| Week 8 | May 17 | Document aggregations continued; Map-Reduce |
| Week 9 | May 24 | Regular Expressions |
| Week 10 | May 31 | Advanced techniques, review |

## Grading Overview

### Assignments

#### *Labs* – 625 pts

7 weekly labs – 4x100pts, 3x75pts

The labs are small exercises of specific skills and knowledge.

#### Analysis Project – 260 pts

Perform a data analysis on a dataset of interest. This assignment has three parts: a problem statement (week 5, 30pts) and a data collection progress report (week 8, 30pts), both shared on Canvas with colleagues for comment, and a final report (week 10, 200pts). In another college, there'd be an exam for this course. However, in our LIS context, it’s important to apply your skills with an engaging, creative problem.

#### Participation - 115pts

Your participation grade is based on engagement (online and in class), as well as attendance. Remember that poor attendance may indirectly affect your grade on the assignment and labs: missing class will make this course very difficult.

*Total: 1000 pts*

### Details

#### Time Due and Late Policy

* Assignments are due 2 hours before your section's lecture starts.
* Late assignments will lose 10% day, up to 50%. Late is better than never.
* 1 lab can be late without penalty, because sometimes life gets in the way.
* Last day for late assignments: June 8th.

#### Grading Scale

The Morgridge College of Education uses a standard grading scale. Grades/scores are based on points accumulated according to the following scale. (LIS students must make a "B" or better in all classes):

|  |  |  |
| --- | --- | --- |
|  | A 93%—100% | A- 90%—92% |
| B+ 87%—89% | B 83%—86% | B- 80%—82% |
| C+ 77%—79% | C 73%—76% | C- 70%—72% |
| D+ 67%—69% | D 63%—66% | D- 60%—62% |
|  | F 59% & below |  |

The highest official grade is an A, though I'd informally consider a 96% and up as an A+ :)

## General Notes and Accommodations

We are here for you to learn. My goal is for you to succeed, and I try to be fair, flexible, and sensitive to the realities of student life. Hopefully, this is reflected in the lessons, late-policy, grading, communications, workload etc.

Various policy accommodations are available, including for:

* observance of holy days,
* students with disabilities or medical issues, and
* student athlete conflicts.

The details of these accommodations are outlined more thoroughly below. In many of these cases, it is important to let me know early if you will need an accommodation.

Even if it is not a stated policy, if you feel you are disadvantaged in your learning in some way, let me know. If the course content is not matching with certain cultural contexts, I can try to adapt it, within reason.

This syllabus provides an overview of the course, but things may change while it is underway. If that happens, it's in service of the learning outcomes. Sometimes I'll see in class that we need more time on a topic, or an event or speaker will make a topic pertinent sooner, or a new publication will arrive and displace an old one. Sometimes, I'll realize that there's a better way to learn something, and I'll want *you* to benefit, not just next year's students.

If you are having problems or concerns, always feel welcome to contact me at [Peter.Organisciak@du.edu](mailto:Peter.Organisciak@du.edu).

## Student Responsibilities

As a student in this course, you are expected to challenge yourself, to actively participate in your education, and to search both inside and outside of the classroom for answers to your questions.

It is important for students to perform their work honestly, and to contribute toward an inclusive space.

Don't plagiarize. Empathize with your classmates, and don't antagonize them. Speak about yourself, not others.

### Honor code

All work submitted in this course must be your own and produced exclusively for this course. The use of sources (ideas, quotations, paraphrases) must be properly acknowledged and documented. For the consequences of violating the Academic Misconduct policy, refer to the University of Denver website on the Honor Code ([www.du.edu/honorcode](http://www.du.edu/honorcode)). See also <http://www.du.edu/studentconduct> for general information about conduct expectations from the Office of Student Conduct.

**Inclusive Learning Environments**

In this class, we will work together to develop a learning community that is inclusive and respectful. Our diversity may be reflected by differences in race, culture, age, religion, sexual orientation, socioeconomic background, and myriad other social identities and life experiences. The goal of inclusiveness, in a diverse community, encourages and appreciates expressions of different ideas, opinions, and beliefs, so that conversations and interactions that could potentially be divisive turn instead into opportunities for intellectual and personal enrichment.

A dedication to inclusiveness requires respecting what others say, their right to say it, and the thoughtful consideration of others’ communication. Both speaking up and listening are valuable tools for furthering thoughtful, enlightening dialogue. Respecting one another’s individual differences is critical in transforming a collection of diverse individuals into an inclusive, collaborative and excellent learning community. Our core commitment shapes our core expectation for behavior inside and outside of the classroom.

### Class participation

Students are expected to complete reading assignments, attend all scheduled classes, actively participate in class discussions, and submit written assignments on due dates. Since concepts and techniques not stressed in the textbooks will be discussed in class, attendance and class participation are essential. You will be graded on your contribution to class discussions and activities. The class Canvas is for your use and convenience. It will be used to post information, discuss issues, and aid group communication.

## Texts

There are no required textbooks. Readings are linked from Canvas for each week.

## Schedule Detail

Note that the readings below are related to concepts. More instruction and tutorial-style content will be shared on Canvas. Also, given the technical nature of this course, I’ll be tracking the tenor of the course really closely and may adapt the schedule in response to perceived needs. Again, please consider Canvas as the schedule and reading list of record.

**Week 1**

*Topics*

Relational Databases, Data types, Introduction to the SciPy Stack

*Readings and Resources*

Adrienne Watt. (2014). Database Design (2nd ed.). Retrieved from <https://opentextbc.ca/dbdesign01>. Chapters 1, 2, 4, 7.

**Week 2**

*Topics*

Relational Databases continued, SQL, Data collection, Jupyter and Python

*Readings and Resources*

Semeler, A. R., Pinto, A. L., & Rozados, H. B. F. (2017). Data science in data librarianship: Core competencies of a data librarian. Journal of Librarianship and Information Science. <https://doi.org/10.1177/0961000617742465>

*Introduction to SQL*. (2018). Launch School. Retrieved from <https://launchschool.com/books/sql>

* Read the Introduction.

Hock-Chuan, Chua. (2014). MySQL by Examples for Beginners. Retrieved from <https://www.ntu.edu.sg/home/ehchua/programming/sql/MySQL_Beginner.html>.

*Activities*

* Lab #1

**Week 3**

*Topics*

SQL continued, SQL Lite, Joins, Basic Querying, Introduction to Pandas

*Readings and Resources*

*Introduction to SQL*. (2018). Launch School. Retrieved from <https://launchschool.com/books/sql>

* Read the section on SQL Joins.

*Activities*

* Lab #2

**Week 4**

*Topics*

Split-Apply-Combine, Deletions and Updates, Indexing

*Readings and Resources*

Wickham, H. (2011). The split-apply-combine strategy for data analysis. Journal of Statistical Software, 40(1), 1–29.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., … Pirahesh, H. (1997). Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. Data Mining and Knowledge Discovery, 1(1), 29–53. <http://www.cs.ucr.edu/~vagelis/classes/FIU-CLASSES/COP6727/publications/dataCube.pdf>

*Activities*

* Lab #3
* Due Next Week: Problem Statement

**Week 5**

*Topics*

Summation, aggregation, and transformation (SAC cont.)

*Readings and Resources*

Jake VanderPlas. (2016). Aggregations: Min, Max, and Everything In Between. In Python Data Science Handbook. Retrieved from <https://jakevdp.github.io/PythonDataScienceHandbook/02.04-computation-on-arrays-aggregates.html>

*Activities*

* Lab #4

**Week 6**

*Topics*

Visualization, JSON, Semi-structured data

*Readings and Resources*

Plotting with matplotlib. (2018). Retrieved March 25, 2018, from <http://pandas.pydata.org/pandas-docs/version/0.13/visualization.html#visualization>

Browse the visualization galleries for Matplotlib and Seaborn.

*Supplemental*

Jake VanderPlas. (2016). *Between. In Python Data Science Handbook*. Retrieved from <https://jakevdp.github.io/PythonDataScienceHandbook/>

* Chapter 4: Visualization with Matplotlib

*Activities*

* Lab #5

**Week 7**

*Topics*

MongoDB - Ingesting Data, document aggregations

*Readings and Resources*

Karl Seguin. (2011). *The Little MongoDB Book* (2.6). <http://openmymind.net/mongodb.pdf>

* Read Introduction, Getting Started, and Ch. 1, The Basics

*Activities*

* Lab #6
* Due next week: Data Collection Progress Report

**Week 8**

*Topics*

Document aggregations continued; Map-Reduce

*Readings and Resources*

Aggregation Pipeline. (2018). Retrieved March 25, 2018, from <https://docs.mongodb.com/manual/core/aggregation-pipeline>

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM, 51(1), 107–113. <http://users.cis.fiu.edu/~mrobi002/teaching/GoogleMapreduce-osdi04.pdf>.

*Supplemental Reading*

Kaushik Sathupadi. (2010). Map Reduce - A really simple introduction. Blog Post. Retrieved from <http://ksat.me/map-reduce-a-really-simple-introduction-kloudo/>.

**Week 9**

*Topics*

Regular Expressions, MongoDB wrap-up

*Readings and Resources*

Daniel Jurafsky, & James H. Martin. (2016). Speech and Language Processing (3rd ed.). Retrieved from <https://web.stanford.edu/~jurafsky/slp3/>.

* Read Chapter 2.1, Regular Expressions

Knox, D. (2013). Understanding Regular Expressions. Programming Historian. Retrieved from <https://programminghistorian.org/lessons/understanding-regular-expressions>

Rubular -<http://rubular.com/>

*Activities*

* Lab #7
* Due Next Week: Analysis Paper

**Week 10**

*Topics*

Advanced techniques, review

*Activities*

* Was there a colleague that was especially helpful or generous in helping you learn? Email Dr. O to let him know.

## Policy Details

**Students with Disabilities/Medical Issues**

If you qualify for academic accommodations because of a disability or medical issue please submit a Faculty Letter to me from Disability Services Program (DSP) in a timely manner so that your needs may be addressed. DSP is located on the 4th floor of Ruffatto Hall; 1999 E. Evans Ave.303.871. / 2372 / 2278/ 7432. Information is also available on line at [http://www.du.edu/disability/dsp.](http://www.du.edu/disability/dsp)

**Religious Accommodations Policy**

University policy grants students excused absences from class or other organized activities or observance of religious holy days, unless the accommodation would create an undue hardship.  Faculty are asked to be responsive to requests when students contact them IN ADVANCE to request such an excused absence. Students are responsible for completing assignments given during their absence but should be given an opportunity to make up work missed because of religious observance.

Once a student has registered for a class, the student is expected to examine the course syllabus for potential conflicts with holy days and to notify the instructor by the end of the first week of classes of any conflicts that may require an absence (including any required additional preparation/travel time). The student is also expected to remind the faculty member in advance of the missed class, and to make arrangements in advance (with the faculty member) to make up any missed work or in-class material within a reasonable amount of time.

**Student Athletes**

If you are a student-athlete, you should inform me of any class days to be missed due to DU sponsored varsity athletic events in which you are participating. Please provide me with an absence policy form by the end of the first week of class. You will need to make up any missed lectures, assignments, and/or exams.

**Mental Health & Wellness**

As part of the University’s Culture of Care & Support we provide campus resources to create access for you to maintain your safety, health, and well-being. We understand that as a student you may experience a range of issues that can cause barriers to learning, such as strained relationships, increased anxiety, alcohol/drug concerns depression, difficulty concentrating and/or lack of motivation. These stressful moments can impact academic performance or reduce your ability to engage. The University offers services to assist you with addressing these or ANY other concerns you may be experiencing. If you or someone you know are suffering from any challenges, you should reach out for support. You can seek confidential mental health services available on campus in the Health & Counseling Center (HCC). Another helpful resource is Student Outreach & Support (SOS), where staff work with you to connect to all the appropriate campus resources (there are many!), develop a plan of action, and guide you in navigating challenging situations. If you are concerned about one of your peers you can submit a report through our Pioneers Care System. More information about HCC, SOS, and Pioneers CARE can be found at:

Health & Counseling Services (<http://www.du.edu/health-and-counseling-center/>)

Student Outreach & Support and Pioneers Care reporting <http://www.du.edu/studentlife/studentsupport/>