SD 212: Data Science & Programming II

Course Policy, Spring AY2023

Instructors:

- Prof. Daniel S. Roche, 438 Hopper Hall, x36814, roche@usna.edu (Coordinator).
- Dr. Alex Timcenko, 444 Hopper Hall, x36823, timcenko@usna.edu.

MGSP Leaders:

- 1/C Greg Smith. 27th company, m235952@usna.edu
- 1/C Dale Sturdifen 9th company, m236120@usna.edu

Course Description: This course builds on the programming skills developed in the prerequisite course and moves the focus towards a wider software ecosystem in order to solve more complex data science tasks. Students will learn and apply foundational principles of program organization including classes and objects, interfaces, inheritance, abstraction, and decoupling. In addition, important command-line skills will be developed for data gathering and cleaning, as well as library and software acquisition and use. These principles will be utilized through high-level programming in Python to analyze and manipulate real-world data sets.

Credits: 3-2-4

Prerequisites: SD211 Intro to Data Science and Programming

Learning Objectives:

1. Observe the structure of new datasets and perform basic data cleaning and manipulation using command-line tools. (supports outcome 2)
2. Understand how regular expressions can be used to describe tokens and their use in programs to manipulate plain-text inputs. (supports outcome 1)
3. Write programs to analyze real datasets using popular data science libraries. (supports outcome DS-6)
4. Utilize basic object-oriented principles such as inheritance and operator overloading to develop and structure complex programs. (supports outcome 1)
5. Understand how libraries are packaged, distributed, downloaded, and installed using standard tools.
6. Examine how data science has been and can be used to impact society at large. (supports outcome 4)

Student Outcomes: Graduates of the program will have an ability to:

1. **Analysis.** Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions.
2. **Implementation.** Design, implement, and evaluate a computing-based solution to meet a given set of computing requirements in the context of the program's discipline.
3. **Communication.** Communicate effectively in a variety of professional contexts.
4. **Ethics.** Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.
5. **Teamwork.** Function effectively as a member or leader of a team engaged in activities appropriate to the program's discipline.

DS-6. Data. Apply theory, techniques, and tools throughout the data analysis lifecycle and employ the resulting knowledge to satisfy stakeholders' needs

Syllabus:

- **Unit 1: Welcome back** (Classes 1–3)
  Course overview, Data science pipeline, Python review
- **Unit 2: Command line** (Classes 4–7)
  Files and directories, bash commands, Piping and redirection
- **Unit 3: Regular expressions** (Classes 8–12)
  Regex syntax, Python re, Command-line tools
- **Unit 4: Error handling** (Classes 13–15)
  try/except, return codes in bash
- **Unit 5: Data cleaning** (Classes 16–18)
  Missing data, Outliers, Preprocessing
- **Unit 6: Info Challenge** (Classes 19–20)

- **Unit 7: Hardware and OS** (Classes 21–23)
  CPU, Memory hierarchy, Filesystems, Role of the operating system
- **Unit 8: Concurrency** (Classes 24–28)
  Multithreading, Python GIL, Multiprocessing, pickle, shell job control
- **Unit 9: Numpy** (Classes 29–30)
  ndarray, dtype, ufuncs, Native library performance
- **Unit 10: OOP in Python** (Classes 31–34)
  Operator overloading, Inheritance, Naming conventions
- **Unit 11: Typing** (Classes 35–36)
  Type hints, Linters, Static vs run-time checks
- **Unit 12: sklearn** (Classes 37–39)
  Reading documentation, Classification, Regression
- **Unit 13: Versions and packaging** (Classes 40–41)
  git, pip, conda

Updates to the course policy: In case this course policy needs to be changed during the semester, students will be notified by email and verbally during class. The current version will always be posted on the course website.

Textbooks:

- Charles Severance. *Python for Everybody: Exploring Data in Python 3*, online, 2023.
- Wes McKinney. *Python for Data Analysis*, O'Reilly Media, 2nd ed., 2017.
- William E. Shotts. *The Linux Command Line*, No Starch Press, 2nd ed., 2019.
- Jeroen Janssens. *Data Science at the Command Line*, O'Reilly Media, 2nd ed., 2021.
- Mendel Cooper. *Advanced Bash-Scripting Guide*, Public domain, rev. 10, 2014.

Course Website: https://www.usna.edu/Users/CS/roche/courses/s23sd212

Extra Instruction: Extra instruction (EI) is strongly encouraged and should be scheduled by email. (For Dr. Roche, first go here to check available times.) EI is not a substitute lecture; students should come prepared with specific questions or problems.

Collaboration: The guidance in the Honor Concept of the Brigade of Midshipmen and the Computer Science Department Honor Policy must be followed at all times. See https://www.usna.edu/CS/resources/honor.php. Specific instructions for this course:

- Collaboration or assistance from any human other than the instructors, MGSP leaders, and those enrolled in SD212 this semester is not permitted. This includes any written or electronic materials from previous semesters.
- Homework: Students may collaborate on homework with others in the same class, but must cite this collaboration clearly. Every student must actually complete their own assignment and understand anything they turn in.
- Labs: Most normal labs are treated like homeworks for purposes of collaboration (see above). Some labs may be clearly indicated as *individual effort*; these are treated like projects: No discussion or collaboration is permitted except with the instructors and MGSP leaders. Any online resources used must be clearly and specifically cited how they are used.
- Exams: No collaboration is allowed. Any group study guides should be shared with the instructor.

All collaboration and outside sources should always be cited. The same rules apply for giving and receiving assistance. If you are unsure whether a certain kind of assistance or collaboration is permitted, you should assume it is not, work individually, and seek clarification from your instructor.

Classroom Conduct:

Everyone in the classroom will show appropriate respect to each other at all times.

This class relies on active engagement and frequent interaction. Use of electronic devices during class time outside of note-taking apps is not permitted.

The section leader is responsible for recording attendance, bringing the class to attention, notifying the CS department office if the instructor is more than 5 minutes late, and directing the class in useful work in the instructor's absence.

Drinks are permitted, but they must be in closable containers. Food, alcohol, and tobacco (of all kinds) are prohibited. Electronic devices must be silent during class and should never serve as a distraction to other students.

Absences:

Students are responsible for all class material. Notes will be posted for each lecture, along with recommended readings. However, this material is not exhaustive and students missing class should arrange to copy notes from a classmate.

Late Policy: **Homework** solutions will generally be discussed immediately, and so no late submissions of homeworks will be accepted for credit. The same deadline applies even in the case of excused absences; students who will miss class should ensure that their work is still submitted on time (typically, electronically).

Labs: Each student has up to 3 *grace days* they may use at their discretion at any point during the semester for lab deadlines. By emailing the instructor *before the deadline*, each grace day used extends the deadline of 1 lab for 1 student by 1 day.

Grading:

The work of the class consists of:

- Homeworks (2-3 per week)
- Labs (weekly)
- Midterm exams (6-week and 12-week)
- Final exam

Any student that completes every homework assignment to a satisfactory level will have their two lowest homework grades dropped at the end of the semester. The definition of "satisfactory level" is based on effort and is at the sole discretion of the instructor. *Work submitted late may count for this requirement, even if it is late and gets zero credit.*

Plus/minus grades will be assigned based on the following numerical cutoffs:

	-		+
A	90–92	93–100	
B	80–82	83–86	87–89
C	70–72	73–76	77–79
D		60–66	67–69
F		0–59	

Here is a breakdown of percentages by grading period.

	6 weeks	12 weeks	16 weeks	Final
Homeworks	20%	20%	20%	15%
Labs	40%	40%	40%	40%
Midterms	40%	40%	40%	20%
Final				25%

Submitted:

____

Prof. Daniel S. Roche
Course Coordinator