Data Management for Data Science

Computer Science 210

Spring 2021

Prerequisites: CS 142 (Data 101: Data Literacy) OR CS 111 (Introduction to Computer Science)
CS students may use this course as an elective for the major.

Course Description

This course is designed to provide students with the knowledge and skills needed to acquire and curate real word data, to explore the data to discover patterns and distributions, and to manage large datasets with databases.

Students will learn the minimal aspects of Python as needed to acquire and curate datasets. Much of their work will be done using Python libraries that deliver maximum benefit with minimal programming effort: to get data from various online data sources online, detect which aspects of data are uncurated or unreliable and understand why it is so, learn various domain independent and domain dependent ways to curate the data, and get the curated data into a form that can be explored, managed and analyzed. Students will also learn how to get datasets into database-ready form, and do basic analysis of such datasets using relational databases and SQL, and NoSQL databases.

The course content is designed to be accessible to all SAS students regardless of their major. Although the course has CS 111 as one of the pre-requisites (for Computer Science students), it does not require students to have any programming experience, since in the other pre-requisite course (CS 142), students are only nominally exposed to R.

Learning Objectives

At the end of the course, students will be able to use basic Python and a small set of libraries to acquire data from various online sources including CSV files and JSON formatted files. They will be able to store data from these sources using simple, easy to use Python storage structures. Students will be able to perform curation using basic Python programming and regular expressions. They will be able to understand and further curate various aspects of dataset attributes using the Python libraries NumPy and Pandas. They will be able to explore and find patterns and distributions in data through visualization using the Python library matplotlib. Students will be able to get structured data into a relational database and manage it using SQL, and get unstructured data into a NoSQL database and manage it with NoSQL management features.

Text/Resources

No required text.

Content for the Python part of the course will be made available via Jupyter notebooks, online documentation (usage samples and APIs accessible from within Jupyter notebooks, and other external sites), and various online data sources.

Content for the database management part of the course will be made available via lecture slides, and accompanying online material.

Topics

Data Management Overview
Setting up Python. Getting started with basic Python elements.
Building Simple Program Logic with Python
Storing and Managing Data in with Python lists and dictionaries
Curating and Extracting Data with Python Regular Expressions
File Handling in Python
Working with CSV and JSON data formats
Managing Numeric Data with Numpy
Storing and Analyzing Data with Pandas
Data Curating and Management with Numpy and Pandas
Visualizing and Exploring Data with Python matplotlib
Creating Relational Databases
Relational Data Management with SQL
NoSQL Data Management

Coursework Requirements

To assess that students have acquired basic literacy in all the concepts, tools, and techniques they are taught, they will be given 6 quizzes periodically through the semester (typically bi-weekly), of which the lowest scoring quiz will be dropped for grading.

Students are expected to work on 4-5 homework assignments through the semester. These assignments are intended for the student to learn by doing—data management is very much a hands-on experience. By doing the assignments, students will learn firsthand how to seek out various data sources; get data from such sources; curate, structure, visualize/explore/discover, and manage such data. Each assignment will typically have a 2 to 3-week window for submission. The assignments will be done either individually or in groups.

Students will be required to do one course project, in groups. They will start working on the project midway through the semester. Each group will elect to pick one rich data source from which they will acquire data, curate/structure/explore the data, and present the results of their analysis and discovery. Each group will be required to use a particular subset of concepts and tools they have learned in the course that most effectively apply to their choice of data source and dataset. Student groups will submit their project in incremental stages, each of which will be graded with feedback to keep them on track for successful completion.

The final exam will assess the student's ability to put together the concepts and tools they have learned in the course in solving a range of very particular and localized data management problems arising in various real-world scenarios.