SI 330: Data Manipulation

Monday and Wednesdays 8:30-10:00 NW 1255
All syllabus details subject to change

Instructor: Christopher Brooks (brooksch@umich.edu)
Instructional Aide: Jaik Prasad (jaiki@umich.edu)

Office hours:

- Brooks by appointment (see calendar), in NQ 4439 or online via hangouts
- Prasad, Wednesday's 10:00-12:00 in NQ 1282

Communication

The best way to contact any member of the teaching team is by using Slack. We try to answer questions that are sent via Slack (si330datamani-fwa5497.slack.com) within about 48 hours. Responses on weekends and holidays may be slower.

If you have questions about course material, homework, or labs, please feel free to come and talk with us during office hours. Jaik should be your first choice for technical questions; conceptual questions are best directed to Chris. Personal matters should be communicated via email to Chris.

About this Course

This is a programming course. This course offers more advanced material for BSI students that deepens knowledge of programming and software development beyond the 106/206 sequence, focusing especially on current computing methods and data structures for obtaining, transforming, and manipulating data. The skills covered in SI 330 will be critical for for future Information Analytics and Data Science courses. This course is a programming course -- you will need to be a programmer to succeed.

Much of the work involved in data science (50-80% is the figure often quoted by data scientists, closer to 80-90% in my own professional experience) is not just in exploration or visualization, but in the initial data manipulation/engineering stages, e.g. gathering, transforming, cleaning, and aggregating. This is the "sewer work" of data science: it's not glamorous but it's absolutely critical. Thus, an important goal of the course is not only to cover specific Python-based tools for data manipulation, but also to teach useful high-level, step-by-step ways of thinking about and solving data computing problems, breaking down a complex task into the right kinds of intermediate stages. For example, the operations provided by many powerful data computing platforms, from SQL databases to Spark large-scale computing, can be thought of in terms of 'split-apply-combine' operations applied to data. Given the fast-moving nature of data science and computing, you'll find this high-level understanding very useful not only for using today's computing tools, but for learning new technologies in the future.

My Teaching Philosophy

I have two jobs in this course: (a) to teach you and help you learn and (b) to evaluate whether you have learned. Sometimes these roles conflict. I set a high bar on expectations, but my goal with the assessments I give is to lower the stakes at any one time, and give you ample opportunity to demonstrate to me that you are capable with this material. You will have some opportunity to make mistakes, and in most cases you will be given an opportunity to correct mistakes with reflection if you so choose.

Textbook & Technology

Python for Data Analysis: Data Wrangling with Pandas, NumPy, and iPython 2nd edition by Wes McKinney (O'Reilly). Copyright 2017 Wes McKinney. The library has digital copies.

It is expected that you will bring a capable computer to each and every class.

Evaluation & Late Policy

Assignment 1: Python Toolkits for Data Manipulation	15%
Assignment 2: Scalable Structured Data	15%
Assignment 3: Big Data	15%
In-class Midterm: Cumulative to Databases IV	20%
Project:	25% (maybe split on presentation/project code)
Quiz 1-4:	2.5% each for a total of 10%, released at least 1 week in advance of the due date

I want everyone to be able to succeed. Assignments must be handed in on time (no exceptions). However, once feedback is returned to you, you will have 168 hours (7 days from when it was returned to you!) to make modifications to your assignment and submit the final assignment for grading.

Tentative Schedule of Topics

Date	Topic	Reading (before class)
Introduction		
9/4	Course overview, slack, and jupyter	
9/9	Python Programming Assessment (ungraded) & Python Review	
Python Toolkits for Data Manipulation		
9/11	Regular Expressions	Python docs
9/16	Pandas I: Numpy, Series and DataFrames	Chapters 3, 4, 5
9/18	Pandas II: Loading data & Data Cleaning	Chapters 6, 7
9/23	Pandas III: Joining & Combining	Chapter 8
9/25	Pandas IV: Aggregation & Grouping	Chapter 10
9/30	Pandas V: In-class problem set	
	Other structured data formats: HTML, XML, JSON	
10/1	Assignment 1 due: Pandas (10%)	
Scalable Structured Data		
	Databases I: Introduction	
	Databases II: DDL	

	Databases III: DML 1	
	Databases IV: DML2	
	Databases V: ETL and ORM	
	Assignment 2 due: Databases (10%)	
10/28	Midterm (Cumulative up to and including Databases IV)	
10/11	Project Pitches (2 min lightning + 800 words)	
Big Data		
	Big Data I:	
	Big Data II:	
	Big Data III:	
	Big Data IV:	
	Assignment 3 due: Big Data	
Other Forms of Data		
	Time Series, Natural Language	
	Networks, Images	
Extended Topics		
	Project Presentations I	
	Project Presentations II	
	Project Due	
	Advanced python/pandas topics: performance, cython	
	Advanced databases topics: triggers, ufuncs, data modelling	
	Advanced scalability topics: service oriented computing with aws	

Original Work

Unless otherwise specified in an assignment, all submitted work must be your own, original work. You may discuss general approaches with others on individual assignments, but may not copy code or answers wholesale and must indicate on your turned-in assignment who you worked with and how. Any excerpts from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the School's policy on Academic and Professional Integrity will result in severe penalties, which might range from failing an assignment, to failing a course, to being expelled from the program, at the discretion of the instructor and the Associate Dean for Academic Affairs.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let us know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress. As soon as you make us aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/sswd/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. We will treat any information you provide as private and confidential.

Student Mental Health and Wellbeing

The University of Michigan is committed to advancing the mental health and wellbeing of its students, while acknowledging that a variety of issues, such as strained relationships, increased anxiety, alcohol/drug problems, and depression, directly impacts students' academic performance.

If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact the Counseling and Psychological Services (CAPS) at (734) 764-8312 and https://caps.umich.edu/ during and after hours, on weekends and holidays or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (732) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources.

For a more comprehensive listing of the broad range of mental health services available on campus, please visit: http://umich.edu/~mhealth/