SI 601 W16 Syllabus

SI 601 Winter 2015 - Data Manipulation

Wednesday 5:30pm - 8:30pm, NQ 2255

Instructor: Yuhang Wang (yuhangw@umich.edu)

Grader: Shao-Chi Wang (shaochi@umich.edu)

Instructor Office Hours: NQ 1243, Monday 6pm - 8pm

If you have questions about course material, homework, lab, or projects, please feel free to come and talk with me during my office hours. You can also contact me via email: please put "**601**" in the subject line so I can be sure to attend to it. Please note that I may not be available on email over the weekend.

Note: Some syllabus details are subject to change.

Description:

SI 601 aims to help students get started with their own data harvesting, processing, and aggregation. Data analysis is crucial to evaluating and designing solutions and applications, as well as understanding users' information needs and how they may want to use it. In many cases, the data we need to access is distributed online among many Web pages, stored in a database, or available in a large text file. Often these data (e.g., Web server logs) are too large to obtain and/or process manually. Instead, we need an automated way to gather the data, parse it, and summarize it before we can do more advanced analysis. In this course, you will learn to use Python and its modules to accomplish these tasks in a quick and easy - yet useful and repeatable – way. The companion for this half-semester course, SI 618: "Exploratory Data Analysis," teaches how to further glean insights from the data through analysis and visualization.

Prerequisites:

At least some knowledge of programming is required.

Texts:

Required:
Charles Severance (2010). Python for Informatics: Exploring Information.
(http://open.umich.edu/education/si/resources/python-opentextbook/winter2010)

Python Software Foundation (2015). "Python v2.7.11 Documentation."
(http://docs.python.org/2/)

Recommended:
Wes McKinney (2012). Python for Data Analysis. O'Reilly Media. ISBN: 978-1-4493-1979-3, Ebook ISBN: 978-1-4493-1978-6

(There will be multiple other sources used throughout the course, but I will note them in the slides)

Classroom Policy:

Students are asked to attend class on time and remain through the entire class. Students will need to bring their laptops for the in-class labs as well as LectureTools participation.

I don't take attendance in this class, but we do have in-class labs that must be completed for credit. So it will be in your best interest to attend.

Original Work:

Unless explicitly specified, all submitted work must be your own, original work. You may discuss general approaches with others on individual assignments, but you should work on the code by yourself. It is a violation of the original work policy to copy code or other work. If you did have extended discussions with other students that helped you with any assignment, you must indicate on your turned-in assignment who you worked with. In written project reports, 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 (stated in the Master's and

Doctoral Student Handbooks) 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 me 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 me 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. I will treat any information you provide as private and confidential.

Course Requirements:

You are required to bring a laptop to class, and at your first opportunity, install Python 2.7.x for the in-class lab assignments. For instruction on installing Python, please refer to the SI 601 welcome email I sent out.

Grading:

Homework (50%) - There will be up 6 x 100 point weekly homework assignments during the term. I will drop the lowest score of these assignments. Assignments will be posted on CTools.

Lab (20 %) - There will be 6 x 20 point weekly in-class lab assignments during the term. I will drop the lowest score of these labs. Labs will be posted on CTools.

Project (30%) - There will be an individual project worth 100 points on a topic of your choice. This will involve selecting data sources, retrieving data and manipulating it in some interesting way. In particular, your project should show off your ability to take two datasets that have different formats and/or access methods and manipulate them to combine them and extract a useful byproduct. The manipulation could involve filtering, format conversion, handling missing or noisy data, matching records from one data source with corresponding records in the other, and so on. You will put together a project proposal at the halfway point in the term (1 page, 20 points), a short slide deck summarizing results toward the end of the course (3-5 slides, 15 points) and a final report (typically 4-5 pages, 65 points). You can also optionally give a short presentation during the last class. These are always a lot of fun!

Please see the course Wiki for a long list of resources for potential project ideas.

Late Homework Penalty: The lab and homework assignments are due at 6pm on Mondays when class begins. No late submissions will be accepted under normal circumstances. Extensions will only be granted to students with good, documented reasons (e.g. medical grounds or other extenuating circumstances beyond the student's control) at the instructor's discretion.

Letter Grades: Assignment of the final letter grade will be done in accordance with the School of Information Masters Student Handbook guidelines.

Schedule:

Class Date	Topic	What's Due Before Class?
Jan. 6, 2016	Course introduction Basics of Programming with Python	Install software as described in SI 601 welcome email.
Jan. 13, 2016	Text Encodings Extracting Patterns from Text with Regular Expressions	Due: Lab 1, Homework 1
Jan. 20, 2016	Fetching and Parsing Web content: HTML, XML, JSON. Web APIs.	Due: Lab 2, Homework 2
Jan. 27, 2016	Querying data in a SQL Database	Due: Lab 3, Homework 3

		1-page project idea proposal
Feb. 3, 2016	Big data processing on Hadoop: MapReduce and Hive	Due: Lab 4, Homework 4
Feb. 10, 2016	Big data processing on Hadoop: Spark and SparkSQL (1)	Due: Lab 5, Homework 5
Feb. 17, 2016	Big data processing on Hadoop: Spark and SparkSQL (2)	Due: Lab 6, Homework 6
Feb. 24, 2016	Course review, final project presentations	Due: Project slides Project report