SI 601 Syllabus Winter 2014 - Data Manipulation

Tuesdays 8:30am-11:30am, NQ1255

Instructor: Kevyn Collins-Thompson (kevynct@umich.edu)  http://www.umich.edu/~kevynct/
GSI: Danny Wu (tzuyu@umich.edu)

Instructor Office Hours: NQ4360 Tuesdays 2:00pm – 4:00pm, or by appointment. (For any questions or issues related to the course.)

GSI Office Hours: NQ4352  Fridays 10am – noon, for technical support on assignments, projects.

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'm not generally 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:

Python Software Foundation (2013). "Python v2.7.3 Documentation." (http://docs.python.org/2/)

Recommended:

(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 lab.

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

**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 other work and must indicate on your turned-in assignment who you worked with. 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 2.7.5 for the in-class lab assignments. For instruction on installing Python, please refer to the first class Welcome email.

**Grading:**

**Homework (60 %)** - There will be 5 x 100 point homework assignments during the term. Assignments will be posted on CTools.

**Lab (20 %)** - There will be 5 x 20 point in-class lab assignments during the term. I will drop the lowest score of these labs. Labs will be posted on CTools.

**Project (20 %)** - There will be an individual project worth 100 points on a topic of your choice. This will involve selecting a data source, retrieving data and manipulating it in some interesting way. You will put together a project proposal at the halfway point in the term (1 page, 20 points), give a short presentation during the last class (15 points) and submit a final report (typically 3-4 pages, 65 points).

**Late Penalty:** You have 3 free late days to use during the course. One late day equals one 24 hour period after the due date of the assignment. No fractional late days: they are all or nothing. Once you have used up your late days, 20% penalty for each day an assignment is late. You don't need to explain or get permission to use late days. Late days can't be used for the final project.

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

**Schedule:**

<table>
<thead>
<tr>
<th>Class Date</th>
<th>Topic</th>
<th>Assignments Due (Wed am after class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 14</td>
<td>Course introduction Basics of Programming with Python</td>
<td>Install software as described in 601 welcome email.</td>
</tr>
<tr>
<td>Jan 21</td>
<td>Extracting Patterns from Text with Regular Expressions</td>
<td>Homework 1, Lab 1</td>
</tr>
<tr>
<td>Jan 28</td>
<td>Fetching and Parsing Web content: HTML,</td>
<td>Homework 2, Lab 2</td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
<td>Project Proposal Due</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Feb 4</td>
<td>Fetching data from Large Online Services Querying data in a SQL Database</td>
<td>Homework 3, Lab 3</td>
</tr>
<tr>
<td>Feb 11</td>
<td>Large-scale data manipulation with MapReduce and Hadoop</td>
<td>Homework 4, Lab 4</td>
</tr>
<tr>
<td>Feb 18</td>
<td>Advanced topics: data mining, machine learning.</td>
<td>Homework 5, Lab 5</td>
</tr>
<tr>
<td>Feb 25</td>
<td>Course review, final project presentations</td>
<td>Project report and presentation</td>
</tr>
</tbody>
</table>