Dealing with Data
INFO-GB.2346.01: Spring 2017

Class hours: Thursdays 6:00-9:15pm, February 9, 2017 – May 4, 2017
Class Location: TBD
Instructors: Konstantin Bauman
   Email: kbauman@stern.nyu.edu
   Office: KMC 8-178
   Office hours: Thursdays 5:00-6:00pm and by appointment.
Teaching Assistant: TBD

Course Description
In this course you will get a working knowledge of how to deal with data. Data is the lifeblood of an organization and successfully extracting information and knowledge from data is an important driver of organizational success. This course focuses on the entire data management process - from data acquisition and processing to analysis of data and presenting results.

This is an intensive hands on course so be prepared for a lot of work and a significant time commitment. But your reward - proficiency in data management skills - will be substantial.

The list of topics that we plan to cover includes:

- Relational databases and SQL
- Basics of Python
- Crawling and parsing HTML
- Accessing data sources through API's
- Regular expressions
- Basics of Data Visualization
- Textual data and natural language processing tools
- Basics of Big Data processing

Course Objectives
At the completion of this course, students will be able to:

- Retrieve and manage data coming in a variety of formats and from different sources
- Store and query data in relational databases
- Handle unstructured and textual data
- Visualize and effectively present data
- Build simple predictive models
Prerequisites

Having some basic familiarity with programming will help, but no prior knowledge of programming is required. However, to ensure that we get the maximum out of our class time, I expect all students to complete a small pre-module assignment that will bring them to the appropriate familiarity level with the basics of programming. This will allow us to use the class time to learn how to leverage the Python ecosystem and libraries to accomplish various data management tasks (as opposed to learning Python). Since this is a hands-on class, you are expected to bring your laptop to every class (and remember to charge it, so that it lasts for the duration of the class).

Reference material

- Python documentation: http://docs.python.org/3.4/index.html
- Python tutorial: https://docs.python.org/3.4/tutorial/
- Python Regular Expressions https://docs.python.org/2/library/re.html
- Unix basics tutorial: http://www.tutorialspoint.com/unix/
- Web crawling (urllib): https://docs.python.org/3.4/howto/urllib2.html
- BeautifulSoup: http://www.crummy.com/software/BeautifulSoup/bs4/doc/
- Pandas: http://pandas.pydata.org/pandas-docs/stable/
- NLTK: http://www.nltk.org/
- Big Data http://www.mmds.org/

Textbooks

While we do not have any required textbook for the course, the following books will be useful references for the material that we will be covering in class.

- Learn Python the Hard Way: This book is also available for free on the web. We will not be really using this book but it is a useful textbook for those that have never been exposed to programming before.
- Python for Data Analysis: This is a useful book, discussing in details the “Python pandas” library, which provided many useful functionality for data management in Python.
- Natural Language Processing with Python: This book describes how to use NLTK, a commonly used Python library for processing textual data.
- Mining the Social Web: This book covers topics that may be a little bit too advanced for this class, such as building data mining models. However, it is a useful companion and illustrates some nice applications.
• Doing Data Science: Similar to the Mining the Social Web, this book deals with topics that may be a little bit too advanced for the material that we cover in this class. However, there are many good examples on how to collect and process data, so it may be worth adding it to your collection, especially if you plan on taking more classes in that front (e.g., the data mining class).
• Python Data Visualization Cookbook: This is a book for those interested in studying deeper visualization in Python.

Useful Material for Learning Python
• Learn Python the Hard Way
• Code Academy, Python track
• Google’s Python class
• Useful IPython notebooks

Grading components

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mini Quizzes</td>
<td>20%</td>
</tr>
<tr>
<td>Homework assignments</td>
<td>20%</td>
</tr>
<tr>
<td>Term Project</td>
<td>30%</td>
</tr>
<tr>
<td>Final exam</td>
<td>30%</td>
</tr>
</tbody>
</table>

**Quizzes:** We’ll have several, very short, online quizzes mainly to reinforce points made in class and also to help you get your hands dirty. All quizzes are open book and you’re welcome to check your solutions on your computers. Quizzes will be lightly graded so make a good faith effort and you’ll do fine. Two quiz scores will be dropped in computing your quiz grade so no worries if you miss one or two.

**Homework assignments:** We’ll have a few home assignments as well. Like the quizzes, assignments are not meant to be diagnostic but rather to help you practice and learn so they will be lightly graded. You can consult with others, ask me questions, use google or stackoverflow for help, but do try them on your own first. One home assignment score will be dropped in computing your home assignment score.

**Late Assignment Submission Policy:** You are free to submit late, but there is a 3% grade penalty for every additional day after the deadline for a maximum of 7 days. (After that, no credit will be given.) Given the generous late submission policy, penalties are strictly enforced, and no extensions are granted. Please plan accordingly, and do not leave submission for the last minute.
**Term Project:** The most important learning experience in the class is the project. There is no better way to learn something than to go out and use it, so start thinking about an application that you think you’d like to build. It is also a great opportunity to learn how to work with a team on data science problems.

The expectation is that you use the material we’ll cover in this class to plan, design, and implement a small software application. Your project grade will depend on how well your work illustrates your understanding of the course material.

The project will be executed by student teams of 3-4 students. Teams are encouraged to interact with the instructor and TA electronically or face-to-face in developing their project reports. Each team will present its project at the end of the semester. We will discuss the project requirements and presentations in class. Deadlines for project deliverables will be in the class calendar on NYClasses and the syllabus.

I strongly recommend you to start working on your project early – you will have to find data as well as formulate the problem.

**Final Exam:** The final exam will be a take-home to be completed during the week following the last class. The subject matter covered and the exact dates will be discussed in class. The exam will test your understanding of the subject matter and your ability to express the learned concepts in a concise way. You will be able to access all course material during the final exam.

**Academic Integrity**

Integrity is critical to the learning process and to all that we do here at NYU Stern. As members of our community, all students agree to abide by the NYU Stern Student Code of Conduct, which includes a commitment to:

- Exercise integrity in all aspects of one's academic work including, but not limited to, the preparation and completion of exams, papers and all other course requirements by not engaging in any method or means that provides an unfair advantage.
- Clearly acknowledge the work and efforts of others when submitting written work as one's own. Ideas, data, direct quotations (which should be designated with quotation marks), paraphrasing, creative expression, or any other incorporation of the work of others should be fully referenced.
- Refrain from behaving in ways that knowingly support, assist, or in any way attempt to enable another person to engage in any violation of the Code of Conduct. Our support also includes reporting any observed violations of this Code of Conduct or other School and University policies that are deemed to adversely affect the NYU Stern community.

The entire Stern Student Code of Conduct applies to all students enrolled in Stern courses and can be found here: [http://www.stern.nyu.edu/uc/codeofconduct](http://www.stern.nyu.edu/uc/codeofconduct)
**General conduct and behavior**

Students are also expected to maintain and abide by the highest standards of professional conduct and behavior. Please familiarize yourself with Stern’s Policy in Regard to In-Class Behavior & Expectations (http://www.stern.nyu.edu/portal-partners/current-students/undergraduate/resources-policies/academic-policies/) and the NYU Disruptive Behavior Policy (http://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/bullying--threatening--and-other-disruptive-behavior-guidelines.html).

**Students with disabilities**

If you have a qualified disability and will require academic accommodation of any kind during this course, you must notify me at the beginning of the course and provide a letter from the Moses Center for Students with Disabilities (CSD, 998-4980, www.nyu.edu/csd) verifying your registration and outlining the accommodations they recommend. If you will need to take an exam at the CSD, you must submit a completed Exam Accommodations Form to them at least one week prior to the scheduled exam time to be guaranteed accommodation.

**Course Schedule**

<table>
<thead>
<tr>
<th>Class</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>February 9</td>
<td>1. Introduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Administrative setup</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Setting up EC2 and MySQL Workbench</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Basic UNIX commands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Entity Relationship Models and Diagrams</td>
</tr>
<tr>
<td>2</td>
<td>February 16</td>
<td>1. Entity Relationship Diagrams</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Relational Databases</td>
</tr>
<tr>
<td>3</td>
<td>February 23</td>
<td>1. SQL: Selection queries</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. SQL: Aggregate Queries</td>
</tr>
<tr>
<td>4</td>
<td>March 2</td>
<td>1. SQL: Joins</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Guest Speaker - case study</td>
</tr>
<tr>
<td>Week</td>
<td>Date</td>
<td>Lectures</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| 5    | March 9    | 1. Python: Variables; Basic data types;  
|      |            | 2. Python: Lists, Tuples, Dictionaries                                    |
| -    | March 16   | Spring Break (No Class)                                                  |
| 6    | March 23   | 1. Python: Control Flow statements; Loops;  
|      |            | 2. Python: Functions                                                     |
| 7    | March 30   | 1. Working with the Web: XML, HTML, JSON, Web Crawling  
|      |            | 2. Interacting with Web APIs using Python and CURL;                      |
| 8    | April 6    | Regular Expressions                                                      |
| 9    | April 13   | 1. Python and SQL: Interacting with Relational Databases using Python  
|      |            | 2. Python Pandas and Visualization                                        |
| 10   | April 20   | Natural Language Processing                                              |
| 11   | April 27   | 1. Big Data: MapReduce, Hadoop Distributed File System (HDFS)          
|      |            | 2. Guest Speaker – Big Data in Finance                                   |
| 12   | May 4      | 1. Review Section                                                       
|      |            | 2. Project presentations                                                 |
|      |            | Final Exam (online)                                                     |