Dealing with data in Python
Fall 2015

Instructor: Hardeep Johar
Teaching Assistant:
Class hours: Friday 10:10 - 12:40
Location: 
Office hours:
Contact: hj2203@columbia.edu

Course Description
The goal of this course is to introduce you to the basics of programming in Python and to give you a working knowledge of how to use programs to deal with data. Data is the lifeblood of an organization competency in programming is an essential skill for successfully extracting information and knowledge from data. In this course we will first cover the basics of programming and then focus on using Python on the entire data management process - from data acquisition to analysis of data - big data and small data. At the end of this course you can expect to be familiar with working with relational databases, using SQL based languages such as MySql, dealing with formatted data (XML, JSON, etc.), using Python to work with and analyze data from databases as well as from the web, and using Big Data processing frameworks like Hadoop and MapReduce.

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.

Prerequisites
There is no prerequisite for this class. Prior exposure to some programming language will be helpful so students are encouraged to take the Python Programming course on Codeacademy or Google prior to the start of this course.

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
Web crawling (urllib): https://docs.python.org/3.4/howto/urllib2.html
BeautifulSoup: http://www.crummy.com/software/BeautifulSoup/bs4/doc/
NLTK: http://www.nltk.org/
Big Data: http://www.mmds.org/

Grading components (tentative)

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mini Quizzes</td>
<td>20%</td>
</tr>
<tr>
<td>Home assignments</td>
<td>15%</td>
</tr>
<tr>
<td>Midterm exam</td>
<td>15%</td>
</tr>
<tr>
<td>Project</td>
<td>30%</td>
</tr>
<tr>
<td>Final exam</td>
<td>15%</td>
</tr>
<tr>
<td>Participation</td>
<td>5%</td>
</tr>
</tbody>
</table>

Topics covered:

1. Python programming (first 4 weeks)
2. Regular expressions
3. Web scraping and web crawling
4. xml, json, html basics
5. Relational databases: Entity relationship model
6. Relational databases: Normalization
7. Relational databases: SQL
8. Data analysis and data visualization (Python libraries)
9. Natural language processing
10. Big data (MapReduce and Hadoop)