Data Analytics in Python
Fall 2018

Hardeep Johar (hj2203@columbia.edu)
Wednesday 9:00 am - 12:15 pm

Course Description
Description
The goal of this course is to give you a working knowledge of how to use Python to extract knowledge and information from data. At the end of this course you will be competent in using Python libraries to work with and analyze offline as well as online data. We will see how to get data from files (csv, html, json, xml) and relational databases (mysql), cover the rudiments of data cleaning, and examine data analysis, machine learning (regression, decision trees, clustering), deep learning (tensorflow) and data visualization packages (numpy, Pandas, Scikit-learn) available in Python. We will examine how we can use Python libraries for extracting value from text (text mining) and understanding the properties of networks (network analysis). Finally, we will cover the basics of functional programming, an essential tool for dealing with big data, by looking at the map/reduce option and on the pyspark API.

The course contains a group project component that will require you to gather, store, and analyze a data set of your choosing. Individual evaluation components include quizzes, assignments, class participation, and an in-class written exam.

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 analysis skills - will be substantial!

Prerequisites
Prior exposure to some programming language is helpful and you should have taken B8136 (Introduction to Programming) or an equivalent course or cleared the waiver exam. I encourage you to explore online Python programming courses before the start of the semester (bearing in mind that we’ll be working with Python 3.6 and not 2.7).
Evaluation components

1. Assignments

Assignment due dates will be posted on the calendar and all assignments will require submission online in the assignments folder. It is your responsibility to check the calendar and keep track of due dates. Assignments turned in within a week after the due date will be automatically penalized 25% and anything turned in more than a week late will be automatically penalized 50%. Assignments turned in more than two weeks late will receive no credit.

2. Project

There is no better way to learn something than to go out and use it so start thinking about a data set you’d like to analyze. Final submission will include a (brief) design report, Python code, and an in-class presentation and demonstration. A significant part of your project grade will come from how other students rate your work so your focus should not be only on the analysis but also on the presentation of your work and the results in an easy to understand manner.

3. Quizzes

There will be several short quizzes - in class and online. The purpose of the quizzes is to quickly check recall and, generally, if you’ve been paying attention in class and doing in your home assignments in a timely fashion you shouldn’t have to worry too much about them.

Computers in class

Computers are a requirement for this course and you are expected to bring one for every class. We’ll do a lot of programming - the best way to learn is to see something in action and Python is an especially good language for making things happen. Make sure that your laptops have sufficient charge for the class!

**Mac vs Windows**: Either is fine but, if you have the choice, then please use a Mac. It is much easier to install needed libraries on a Mac than it is on a Windows
machine. In particular, if you have a Mac and are using some sort of Windows emulator then please use Mac OS-X and not the Windows emulator. The double redirection will make everything a lot slower. But, either Mac or Windows will work so don’t worry if you’re a Windows user.

**Python2 versus Python3.** Unfortunately, there are two versions of Python and they aren’t compatible. We’ll be using Python3 (current version: 3.6, though this may change before the start of the semester) mainly because Python2 is slated to go away (though, in the open source world, that doesn’t mean a whole lot!) but also because it is a better language.

## Course schedule

(Note: While we will definitely cover all these topics, I might tweak the time spent on each and the ordering a little!)

- **Week 1:** Python review
- **Week 2:** Web scraping
- **Week 3:** SQL
- **Week 4:** Data analysis with numpy and pandas
- **Week 5:** Data analysis and visualization
- **Week 6:** Text mining
- **Week 7:** Analysis of networks
- **Week 8:** Machine learning
- **Week 9:** Machine learning
- **Week 10:** Neural networks and Deep learning
- **Week 11:** Very basic introduction to big data programming (pyspark and map/reduce)
- **Week 12:** Project presentations

## Texts

There is no text for this class. The following will be helpful if you want to go above and beyond the material covered in the course:


**Online resources**

Python documentation: [http://docs.python.org/3.6/index.html](http://docs.python.org/3.6/index.html)
Python tutorial: [https://docs.python.org/3.6/tutorial/](https://docs.python.org/3.6/tutorial/)
Python Regular Expressions [https://docs.python.org/3.6/library/re.html](https://docs.python.org/3.6/library/re.html)
SQL: [http://www.w3schools.com/sql/](http://www.w3schools.com/sql/)
MySQL: [http://www.plumcreek.us/mysql/](http://www.plumcreek.us/mysql/)