CAS CS 505
Data Science in Python
Fall 2016

Meeting Place: ??

Meeting Time: ??

Instructor: Prof. Mark Crovella

- **Office:** MCS-140E
- **Office Hours:** ??
- **Email:** crovella@bu.edu

Teaching Fellow:??

- **Office Hours:** ??
- **Office Hours Location:** ??
- **Lab Tutoring Hours:** ??.
- **Email:** ??

Overview of the Course

This course is targeted at students who require a basic level of proficiency in working with and analyzing data. The course emphasizes practical skills in working with data, while introducing students to a wide range of techniques that are commonly used in the analysis of data, such as clustering, classification, regression, and network analysis. The goal of the class is to provide to students a hands-on understanding of classical data analysis techniques and to develop proficiency in applying these techniques in a modern programming language (Python).

Broadly speaking, the course breaks down into three main components, which we will take in order of increasing complication: (a) unsupervised methods; (b) supervised methods; and (c) methods for structured data.

Lectures will present the fundamentals of each technique; focus is not on the theoretical underpinnings of the methods, but rather on helping students understand the practical settings in which these methods are useful. Class discussion will study use cases and will go over relevant Python packages that will enable the students to perform hands-on experiments with their data.

Prerequisites: Students taking this class must have some prior familiarity with programming, at the level of CS 105, 108, or 111, or equivalent. CS 132 or equivalent (MA 242, MA 442) is required. CS 112 is also helpful.

Readings

There is no text. Lecture notes will be posted online.

Some recommended texts are:

1. Python for Data Analysis (http://shop.oreilly.com/product/0636920023784.do)
2. Programming Collective Intelligence (http://shop.oreilly.com/product/9780596529321.do)

Web Resources

The slides I use are actually executable python scripts, using the `ipython notebook`. If you have ipython notebook, you can download and execute the examples on your own computer, and you can modify them any way you'd like, play around with them, experiment, etc.

The slides I use in lecture are published on `github`. The repository is `https://github.com/mcrovella/CS505-Data-Science-in-Python`. If you want to access the repository using `git`, please feel free but you can simply download directly from the web site if you prefer.

Homeworks and Project

1. There will be six programming assignments. In a typical assignment you will analyze one or more datasets using the tools and techniques presented in class.
2. In addition, there will be a final project. For the project you will extract some knowledge or conclusions from the analysis of dataset of your choice. The analysis will be done using a subset of the methods we described in class.

   The project will have three essential components: 1) a data collection piece (which may involve crawling or calls to an API, combining data from different sources etc), 2) a data analysis piece (which will involve applying different techniques we described in class for the analysis) and 3) a conclusion component (where the results of the data analysis will be drawn). The students will submit a 5-page report explaining clearly all the three components of their project. Finally a poster presentation will be required where the students will be prepare to present their effort and results in front of their poster.

   As an example, you may choose to collect data from Twitter related to a specific topic (e.g., Ebola virus) and then measure the intensity of posts about a topic in different areas of the world etc. Other examples of projects may include (but are not limited to): analysis of MBTA data, analysis of NYC data, crawling of YouTube (or other social media data) and analysis of social behavior like trolling, bullying etc.

   The project is due by the end of the exam week. The project presentations will be given in the form of a final poster explaining components 1, 2 and 3 of the project.

   You are expected to work individually on homeworks and on the final project. There will be no final exam.

   CHECK THIS WITH KATHERINE

3. Homeworks will be submitted via *websubmit*. The URL for submitting an assignment is `http://cs-websubmit.bu.edu/main.py`. If you have questions about homework submission or grading, please start by checking with the TF, and if you still have questions, then feel free to direct them to me.

Submitting Homework

For showing your analytical / mathematical work, there are three options, in increasing order of quality:

1. You can scan handwritten notes into PDF. Note that these must be **clear** and **neat** because the grader will simply read them as best s/he can.
2. You can write up your work in Word, using the built-in equation editor for the mathematics. Then save as PDF.
3. You can learn and use LaTeX. This is the tool that produces a truly professional, publishable document. It is what serious computer scientists use. You can learn to use it in a few hours, starting from `http://www.latex-tutorial.com/`. Eventually you will find it useful for lots of your coursework, so it makes sense to learn it now.

CHECK WITH KATHERINE.

For showing your computational work, you will often submit code and/or scripts showing your code runs. For the code, simply submit them as .py files. For the scripts showing your code executing, you can use the built-in logging system of ipython. I recommend:

1. `%logstart -ort hwk-file.txt backup`
2. run your code in the interpreter showing the output
3. `%logstop`.

At which point the file `hwk-file.txt` will contain a record of the inputs and outputs of your code. (Note that output from "print" statements will not show up however). v

Piazza

We will be using Piazza for class discussion. The system is really well tuned to getting you help fast and efficiently from classmates, Ms. Zhao, and myself. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza. Our class Piazza page is at: `https://piazza.com/bu/???`. We will also use Piazza for distributing materials such as homeworks and solutions.

When someone posts a question on Piazza, if you know the answer, please go ahead and post it. However pleased *don't* provide answers to homework questions on Piazza. It's OK to tell people *where to look* to get answers, or to correct mistakes; just don't provide actual solutions to homeworks.

Programming Environment

We will use `python` as the language for teaching and for assignments that require coding. Instructions for installing and using Python are on Piazza.

Course and Grading Administration

Assignments will be submitted using `websubmit`. Ms. Zhao will explain how to submit assignments.

Final grades will be computed based on the following:

50% Homework assignments.

50% Final Project

The exact cutoffs for final grades will be determined after the class is complete.

You need to consistently work the problem sets each week. Plan to set aside a regular time each week to do them.

Academic Honesty

You may discuss homework assignments with classmates, but you are solely responsible for what you turn in. Collaboration in the form of discussion is allowed, but all forms of cheating (copying parts of a classmate's assignment, plagiarism from books or old posted solutions) are NOT allowed. We – both teaching staff and students – are expected to abide by the guidelines and rules of the Academic Code of Conduct (which is at `http://www.bu.edu/dos/policies/student-responsibilities/`).

You can probably, if you try hard enough, find solutions for homework problems online. Given the nature of the Internet, this is inevitable. Let me make a couple of comments about that:

1. If you are looking online for an answer because you don't know how to start thinking about a problem, talk to Ms. Zhao or myself, who may be able to give you pointers to get you started. Piazza is great for this – you can usually get an answer in an hour if not a few minutes.

2. If you are looking online for an answer because you want to see if your solution is correct, ask yourself if there is some way to verify the solution yourself. Usually, there is. You will understand what you have done *much* better if you do that. So ... it would be better to simply submit what you have at the deadline (without going online to cheat) and plan to allocate more time for homeworks in the future.

Course Schedule

Date	Topics	Reading	Assigned	Due
9/6 9/8	Introduction to Python Git, Github, Python notebooks, Pandas		HW 0	
9/13 9/15	Probability and Statistics Refresher Linear Algebra Refresher			HW 0
9/20 9/22	Numpy, Scikit-learn, Distance and Similarity Functions Intro to Timeseries		 HW 1	
9/27 9/29	Clustering, k-means Clustering II			
10/4 10/6	Hierarchical Clustering Expectation Maximization and GMM		 HW 2	HW 1
10/11 10/13	NO CLASS; Monday Schedule DB Clustering and Comparing Clustering Algorithms			
10/18 10/20	Dimensionality Reduction - SVD I SVD II and Web Scraping			
10/25 10/27	Open Classification: Decision Trees			 Proj Proposals
11/1 11/3	Classification: SVM, Naive Bayes Regression: Linear Regression			 Proj Proposals Finalized
11/8 11/10	Logistic Regression Linear Regression II		HW 3	
11/15 11/17	Recommendation Systems Network Analysis I		 HW 4	 HW 3
11/22 11/24	Network Analysis II NO CLASS; Thanksgiving Break			Proj Proposals prog report
11/29 12/1	Graph Clustering Text Analysis and Topic Modeling		HW 5	HW 4
12/6 12/8	Wrapup Poster Session			