Data Science & Programming in Python

# Learning Objectives

* Develop an understanding of Python syntax
* Explore Python libraries which are of value to data scientists
* Garner a deeper understanding of programming fundamentals
* Become familiarized with various technical topics
* Gain experience in working on a substantial project

# Class Expectations

Learning a programming language is difficult, so to truly get value out of this course will require your effort outside of class meeting times. We have a lot of ground to cover, and just 14 hours with which to cover it. I welcome you to ask questions during class if they pertain to the current topics being discussed; however, if you have questions unrelated to the topic at hand, please save them until after class. I will be happy to stay after class to answer questions about topics from previous class periods. If the demand exists, I would be willing to hold a “review session” immediately following our scheduled class time where we discuss the homework, work on additional examples, go over previous topics, etc.

# Grading

Your grade in this course will consist of 100% participation. The final grading scale is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A++** | **A+** | **A** | **A-** | **A?** |
| >= 94.5 | 89.5 – 94.49 | 84.5 – 89.49 | 79.5 – 84.49 | < 79.5 |

Unfortunately, since this course consists of 0 semester hours, your final grade will not affect your GPA.

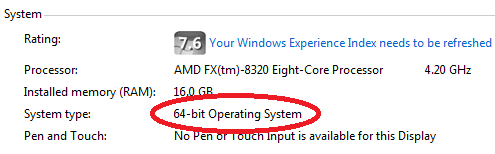
# Course Structure

Class will meet twice a week, with Mondays focusing more on Data Science topics and Tuesdays focusing on Computer Science ones. Therefore, students who are less interested in the CS portions may skip all but the first CS session at their discretion, though I would personally recommend attending them all.

The first four weeks are designed to help you get your bearings in the world of Python, with discussions of syntax and programming fundamentals as well as brief exercises in data collection, cleaning, visualization and modeling. In the final three weeks, we will be taking a deep dive into the more commonly-used modeling methodologies as well as working on a capstone project.

# Required Items

* Anaconda: “Data Science Platform” for Python (<https://www.continuum.io/downloads>)
  + Contains over 100 packages (libraries) for data science, including all the ones we’ll need
  + Contains several different IDEs (think RStudio), including Spyder, which is the one I recommend you to use (feel free to use another if you have a preference)
  + Please choose the Python 3.5 version. If you are unsure whether you use 32-bit or 64-bit, right-click “Computer” and go to “Properties” and find this:



* GitHub: Online code repository
  + You must create a GitHub account (<https://github.com/>)
    - Make sure you keep track of your username and password!
  + You must download the git client (<https://git-scm.com/downloads>)
  + Using these tools, you will be able to synchronize your code across devices, or work more effectively in a team. More importantly, you will be able to build a portfolio of code which you can show off to future employers.

# Additional Resources

* The Python Foundation has a tutorial on their website which is pretty good: <https://docs.python.org/3/tutorial/>. Sections 3-5 are a handy reference for data types, control statements and data structures.
* Google’s short but great Python tutorial at <https://developers.google.com/edu/python/>. This tutorial is designed for Python 2.7, but the vast majority of the content is still relevant.
* <https://www.edx.org/course/learn-program-using-python-utarlingtonx-cse1309x>: edX’s “Intro to Python” course.
* edX has a course called “Programming with Python for Data Science” at <https://courses.edx.org/courses/course-v1:Microsoft+DAT210x+4T2016/info>. This course assumes you are already familiar with Python basics, and covers topics such as data manipulation, exploration, transformation and modeling utilizing the same libraries we will cover in our course.
* <http://www-bcf.usc.edu/~gareth/ISL/> *An Introduction to Statistical Learning*. Amazing and free book with great insights on modeling. Contains many R examples as well, if that’s your fancy!

# Final Project

This seven-week course will cap off with presentations of an involved project. Projects can be undertaken alone, or in a group. Your project can be on whatever topic you would like, but it must involve at least two of the following:

* Data Harvesting & Cleaning
* Visualization
* Model Building

Try to find something you are passionate about, or at least interested in, and you will be amazed at what you can create! There are many APIs available to provide you with data, and if there’s no API for your topic, you can always scrape the web. In the end, we’ll have done some learning, and have a great project you can put into your GitHub repository.

Please feel free to consult me if you get stuck anywhere on the journey. I am here to help.

# Course Schedule\*

|  |  |  |  |
| --- | --- | --- | --- |
| **Mon** | **Day** | **Topic** | **Homework** |
| Oct | 24 | Intro, Python v. R, Algorithms, Pseudocode, Demo |  |
|  | 25 | Syntax, Data Types, Control Statements, Data Structures | HW1 |
|  | 31 | Pandas, JSON, File Input, SQL, Basic Data Visualization | HW2 |
| Nov | 1 | Recursion, Scope, Exceptions, Algorithm Implementation |  |
|  | 7 | Strings, RegEx, Text Mining, Twitter API | HW3 |
|  | 8 | Lambda, Map, Reduce, Filter, List Comprehensions |  |
|  | 14 | Dimensionality, Linear Algebra, PCA, Numpy | HW4 |
|  | 15 | Object-Oriented Programming |  |
|  | 21 | HTML, Web Scraping |  |
|  | 22 | Cluster Analysis |  |
|  | 28 | Regression Modeling, Logistic Regression |  |
|  | 29 | Partition Modeling |  |
| Dec | 5 | Kaggle Competition |  |
|  | 6 | Project Presentations |  |

\*Subject to change