Data Science with Python

Phil Mui, Summer 2016

philmuiscu@gmail.com (@philmui)

1. Description

Data science is one of the most in-demand skills in the world today. Harvard Business Review recently published an article with the title: "Data Scientist: The Sexiest Job of the 21st Century" (https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/). This class will prepare you well to become a data scientist.

Data science involves the application of scientic methodologies to extract understanding from and make predictions based on data sets from a broad range of sources. Data science involves knowledge and skills from three areas: 1) computer science, 2) math/statistics and 3) domain specific expertise.

The objective of this course is to teach the analytical mindset & programming skills relevant to data science. Students will learn the Python programming language, along with a set tools for data science in Python, including the Jupyter (IPython) Notebook, NumPy, Pandas, matplotlib, PySpark, and Tableau data visualization. Students will learn skills that cover the various phases of exploratory data analysis: importing data (web, JSON, CSV, relational data), cleaning and transforming data, algorithmic thinking, grouping and aggregation, visualization, time series, statistical modeling/prediction and communication of results. The course will utilize data from a wide range of sources and will culminate with a final project and presentation.

As one of the faster growing fields today, data science has new tools coming faster than publishers can create books. To supplement the traditional readings on data science, we will be sharing state of the art tools and environments that are used by the best practitioners today: including matplotlib and Tableau for data visualization, Spark and PySpark for big data analytics. We aim to equip you with the latest industry thinking & toolsets to become a productive data scientist.

2. Learning Objectives

In this course you will learn:

- To read, write, run and debug programs in the Python programming language.
- To use the core software packages for data science in Python.
- To import data sets into Python using text files, HTTP, CSV, JSON files
- To wrangle data sets by cleaning and transforming them.
- To perform exploratory data analysis to extract insight out of data.
- To visualize data sets for exploration and presentation purposes.
- To build statistical models to understand and make predictions based on data.
- To use the IPython Notebook to create and present reproducible stories involving code and data.

3. Course Materials

We will be using a number of resources throughout the class.

- Camino. We will use SCU's Camino for all course discussions and to distribute course materials.
- Online Python Tutorial (https://www.codecademy.com/learn/python) on codecademy.com. This is an excellent, self-guided tutorial on the Python programming language. We will not be spending any time in class covering the basics of Python; instead you will learn it by completing this online tutorial. You will be required to complete this course by the beginning of the 3rd class meeting.
- Python for Data Analysis (P4DS) (https://www.amazon.com/Python-Data-Analysis-Wrangling-IPython/dp/1449319793), Wes McKinney, O'Reilly (2012). We will be doing weekly readings out of this book. you should have a hard or soft copy and bring it with you to class starting the second week.

The following resource are optional, but highly recommended:

- Tableau Training Videos (http://www.tableau.com/learn/training)
- Spark Online Documentation (http://spark.apache.org/docs/1.6.2/)
- Learning the Pandas Library (https://www.amazon.com/Learning-Pandas-Library-Munging-Analysis/dp/153359824X), Matt Harrison, Hairysun (2016)

4. Grading

Your grade for this course will be determined as follows:

- Online Python Tutorial (https://www.codecademy.com/learn/python) on codecademy.com: 20%
- Weekly homework: 20%
- Midterm: 30%
- Final project and presentation: 30%

5. Course Honor Code

All work submitted for grading must be the original product of the signatory individual or team members. Unless explicitly stated otherwise, sharing of work between teams is not allowed. You should not make use of materials connected to previous offerings of this course or related courses at other institutions. When referencing the words or ideas of others is appropriate, you should make proper citation, but that will not excuse the lack of individual contribution. When in doubt, check with me.

Confirmed violations will result in a failing grade for the entire course, a notation in the student's academic record, and possibly suspension or even expulsion from the program. The penalty will be non-negotiable, and will certainly not allow any recourse actions (e.g., a "do-over" or makeup assignment). For group work, all members will share the consequence of any tainted submissions.

Specifics of the official LSB Graduate Student Honor Code and the protocol for handling academic integrity violations are published at:

https://www.scu.edu/business/graduates/academics/policies/
(https://www.scu.edu/business/graduates/academics/policies/)

6. Software and computers

This course will involve extensive programming and computer work, both in and out of class. You are required to bring a laptop to class each time and and have the following software installed by the first class meeting:

- Enthought Canopy Python Distribution (https://store.enthought.com/downloads/): a scientific-oriented Python distribution from Enthought. This includes EPDFree, a free base scientific distribution (with NumPy, SciPy, matplotlib, Chaco, and Jupyter) and EPD Full, a comprehensive suite of more than 100 scientific packages across many domains. EPD Full is free for academic use but has an annual subscription for non-academic users.
- A text editor for your platform:
  - For Windows, I recommend Atom.io or Sublime Text
  - For Mac, I recommend Atom.io, nano, Text Wrangler, TextMate or Sublime Text
  - For Linux, I recommend nano, emacs or Sublime Text
- Chrome or Firefox browser

7. Tools

In addition to the core Python language, we will cover some or all of the following open source packages for data science in Python.

Core:

- python 2.7
- canopy python distribution
- juypter / ipython
- numpy
- pandas

Visualization:

- matplotlib
- tableau
- ggplot

Modeling and prediction:

- pyspark
- spark 1.6.2

Development:

- git
- virtualenv (venv)

Web related technologies:

- HTTP Requests
- XPath
- CSS Selector

8. Schedule

We will be meeting every Saturday starting July 16th to August 27th, 2016. Please note that we are NOT meeting in-person on July 09 nor on September 03.

Here is our latest guide to the topics in the class -- subject to changes based on feedback throughout the class:

- 07/09 (No in-person class) Python
  - Download & install Enthought Canopy Python Distribution (https://store.enthought.com/downloads/).
  - Register & try to complete Online Python Tutorial (https://www.codecademy.com/learn/python) on codecademy.com
- 07/16 Introduction to Data Science & data ingestion
  - (Bring your laptop with Enthought Python 2.7 environment installed)
  - Introducing Data Science (P4DS Chapter 2)
  - git, virtual environments
  - Data Ingestions (CSV / text, Excel, PDF, API)
  - Interactive Data Science: Jupyter / IPython (P4DS Chapter 3)
- 07/23 Web, Numpy, Pandas
  - web scraping
  - numpy, pandas (P4DS Chapter 4-5)
  - Data Loading, Storage, File Formats (P4DS Chapter 6)
- 07/30 Data Wrangling
  - Data Loading, Storage, File Formats (P4DS Chapter 6)
  - Data Clean, Transform, Merge, Reshape (P4DS Chapter 7)
- 08/06 Data Visualization
  - matplotlib, pandas, maps (P4DS Chapter 8)
  - Tableau visualization & SQL databases
  - Midterm Exam
- 08/13 Descriptive Statistics with Python
  - matplotlib, pandas, maps (P4DS Chapter 8)
  - Tableau visualization & SQL databases
  - DataFrame (P4DS Chapter 5 review)
- 08/20 More Statistics with Python
  - Data summarization (P4DS Chapter 9)
  - if time permits: Series & Time (P4DS Chapter 10)
- 08/27 Final Project Presentations
  - Spark & PySpark (http://spark.apache.org/docs/1.6.2/)
  - Map & Reduce with Wikipedia data
  - Web Traffic Analysis with NASA data
- 09/03 (No in-person class)
  - Submit summary report

Very excited to be spending this summer with you in this class!