An Introduction to Using Python with Data

Patrick Hall

jpatrickhall@gmail.com

Course Goals

- To communicate a general understanding of software languages and their uses in processing and analyzing data.

- To convey specific knowledge regarding the Python language through examples related to processing and analyzing data.

Course Overview

- Software languages and programming for processing and analyzing data
- Python
  - Introduction
  - Basic operations and strings
  - Controlling the flow of a program
  - Reading and writing files
  - Data structures
  - Defining your own functions
  - Scraping data from the web
  - Numerical Python (NumPy) and data analysis
  - Plotting Results and IPython

Introductions

- About me …
  - Research Statistician Developer for SAS Enterprise Miner

    http://www.sas.com/en_us/software/analytics/enterprise-miner.html

  - Cloudera Certified Data Scientist

    http://www.cloudera.com/content/cloudera/en/training/certification/ccp-ds.html

Introductions

- About you …
  - Your education and experience in processing and analyzing data.
  - Your education and experience with programming and Python.
  - Your goals for this class.

Preliminary Course Instructions

As a group …

1. Download course materials
   https://github.com/jphall663/bellarmine_py_intro/archive/master.zip
2. Download Anaconda Python version 2.0.1
   http://repo.continuum.io/archive/index.html
3. Install Anaconda Python version 2.0.1
4. Set working directory in Spyder IDE

Course Logistics

- Schedule
- Course Materials
- Python Documentation

  https://docs.python.org/2/tutorial/
- Questions and Discussions
- Hands-on Examples

Break time.

Software Languages and Programming for Data Processing and Analysis.

Break time.

Python: Basic Operations and Strings.

Section Goals

https://docs.python.org/2/tutorial/introduction.html

- The interactive shell
- Operations for assignment and comparison
- Strings
- Escape Characters
- Slicing

https://docs.python.org/2/library/stdtypes.html#string-methods

- String functions

Exercise 1

- Working With Strings

Controlling the Flow of Your Python Program.

Section Goals

https://docs.python.org/2/tutorial/controlflow.html

  - `if` statements
  - `for` statements
  - `break` and `continue` statements
  - `pass` statements

- `enumerate` statements

Reading and Writing Files with Python.

Section Goals

https://docs.python.org/2/tutorial/inputoutput.html#reading-and-writing-files

- Opening and closing files
- File modes

https://docs.python.org/2/reference/compound_stmts.html

- `with` statements

- Combining `for` loops, `if` statements and file operations to read and write files.

Exercise 2

- Loops and File I/O

Basic Data Structures in Python.

Section Goals

https://docs.python.org/2/tutorial/datastructures.html

- Lists
- List Comprehensions
- Sets
- Dictionaries
- Looping Techniques
- Conditions

Section Goals

https://docs.python.org/2/library/collections.html

  - Counters

Exercise 3

- Lists, Dictionaries and Sets

Defining Your Own functions.

Section Goals

https://docs.python.org/2/tutorial/controlflow.html#defining-functions

  - Defining functions

Scraping Data from the Web.

Section Goals

https://docs.python.org/2/howto/urllib2.html

- `urllib2` fetches HTML and other data from websites
- Fetching URLs using the `urlopen` function
- Reading information from an URL using the `read` function

http://www.crummy.com/software/BeautifulSoup/bs4/doc/

- `BeautifulSoup` parses HTML into more meaningful data
- `prettify` function
- `get_text` function
- `find_all` function

Section Goals

  - Data Sources on the Web

Exercise 4

- Scraping Data from the Web

Numerical Python (NumPy) and Data Analysis.

Section Goals

http://wiki.scipy.org/Tentative_NumPy_Tutorial

- What is NumPy?
- The Basics:
  - NumPy Arrays
  - Basic Array Operations
  - Indexing, Slicing and Iterating
- Iteration vs. vector operations

Section Goals

https://docs.python.org/2/library/csv.html

- CSV and delimited data
- Reading CSV data using the `csv` module
- Potential problems with CSV data

Section Goals

http://www.kaggle.com/c/titanic-gettingStarted

- What is Kaggle?
- What is predictive modeling?
- The famous Titanic data set

Section Goals

- Numpy data types
- Masking arrays

Exercise 5

- Numpy: Kaggle Titanic Competition

Plotting Results and IPython.

Section Goals

http://matplotlib.org/

- solution_6.py
- Adding values to a plot
- Decorating a plot
- Magic Numbers
- `matplotlib` examples

Section Goals

- Starting an IPython session
- Creating an IPython notebook
- Sharing an IPython notebook using GitHub
  - https://gist.github.com/
  - http://nbviewer.ipython.org/
  - http://nbviewer.ipython.org/github/jphall663/bellarmine_py_intro/blob/master/Titanic.ipynb

Exercise 6

- IPython: Graphing Results

The end.