Lect 14 – Python for Data Analysis

Rob Capra

INLS 490-172

Some tips for Project 2

- After you split values on a line of data, you may need to explicitly convert some to integers such as: `uid = int(fields[0])`
- Don't forget that setdefault is your friend when building dictionaries.
- But dict.get is useful when counting things
- Tuples can be keys: ym2numtags[(y,m)]

Python for Data Analysis

- Data Analysis
  - Manipulating
  - Processing
  - Cleaning
  - Crunching
- Data (structured data)
  - Multidimensional arrays (matrices)
  - Tabular/spreadsheet-like data
    - Relational DBs
    - Tab/comma delimited text files
  - Multiple tables of data related by key columns

Alternatives

- Matlab
- R
- SAS
- Stata

Python Libraries for Analysis

- NumPy
  - Scientific computing
  - Fast, efficient multidim array object & functions
- Pandas
  - Data structures for structured data
  - Reshape, slice&dice, aggregate, select
  - DataFrame object similar to R data.frame
- Matplotlib
  - Plots & 2D visualizations
- Ipython
  - Shell for writing, testing, debugging
  - Notebook
  - GUI console with support for graphics

Installing Enthought Canopy Free
(PDA: Chapter 1, “Installation and Setup”)

- First, uninstall IDLE (good bye old friend!)
  - On Windows, use the “uninstall programs” control panel
- Download Enthought Canopy Free installer
  - www.enthought.com
- Verify the Python installation
  - C:\Users\rcapra\python
  - Look for “Canopy” in the version name
- Verify the ipython installation
  - C:\Users\rcapra\ipython –pylab
  - In[1]: import pandas
  - In[2]: plot(arange(10))
- Also can check: ipython notebook –pylab=inline

Installing Enthought Canopy Free

- First, uninstall IDLE (bye bye old friend!)
  - On Windows, use the “uninstall programs” control panel
- Download Enthought Canopy Free installer
- Verify the Python installation
  - C:\Users\rcapra\python
  - Look for “Canopy” in the version name
- Verify the ipython installation
  - C:\Users\rcapra\ipython --pylab