Introduction to Python for Text Analysis: Basic Setup

Jennifer Pan

Institute for Quantitative Social Science
Harvard University

(Political Science Methods Workshop, February 21 2014)

*Much credit to Andy Hall and *Learning to Think Like a Computer Scientist*

Installation

Mac

- Python is pre-installed on Mac OS X
- If you want to install a different version, go to http://www.python.org/downloads/mac-osx/
- For trouble-shooting see http://docs.python.org/2/using/mac.html

Windows

- Python must be installed on windows
- Go to http://www.python.org/downloads/windows/
- For trouble-shooting see http://docs.python.org/2/using/windows.html

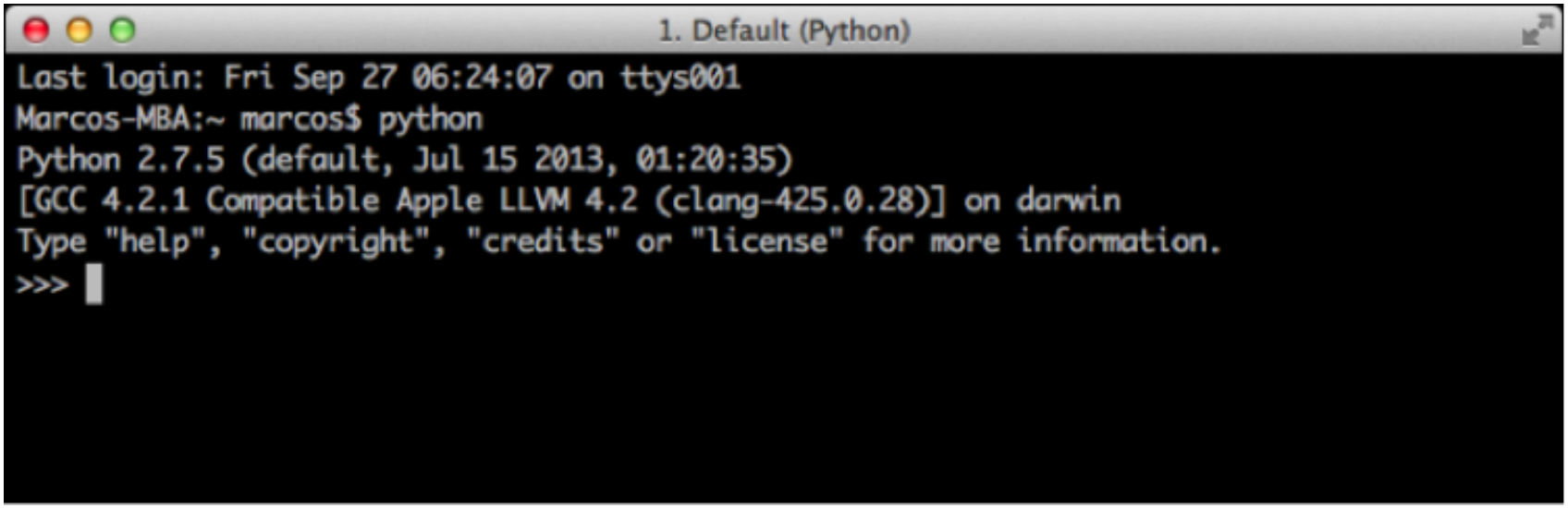
Terminal / Command Prompt

If you're programming with Python, it's a good idea to take a minute to familiarize yourself with Terminal or some kind of command line interface, because you're going to be running scripts.

- Here's an intro from lifehacker
- Here's a more in-depth guide from Learn Code the Hard Way

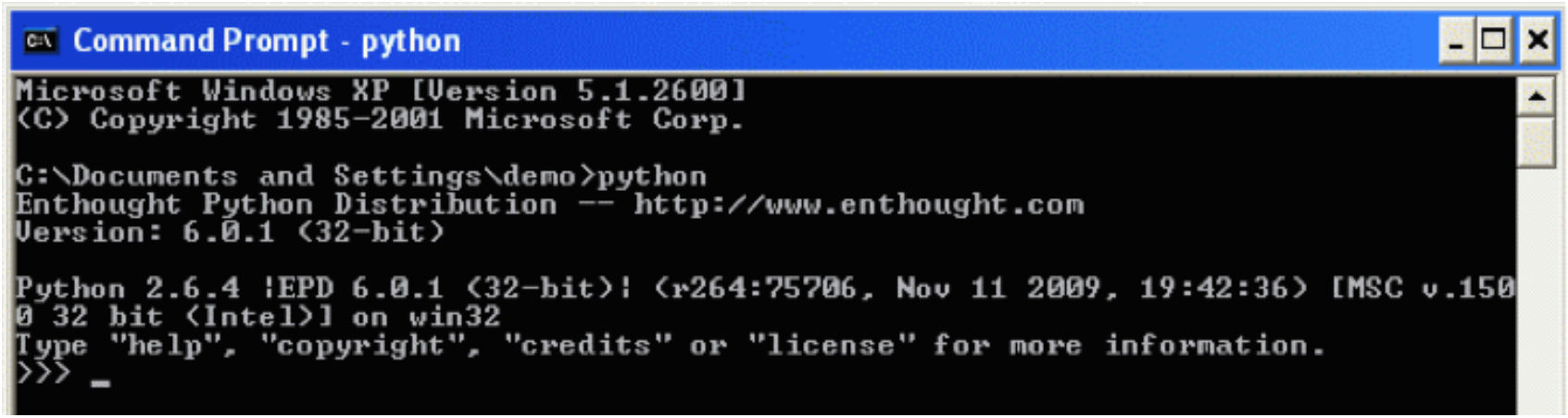
Terminal in Mac

- Type `terminal` in Spotlight to open terminal (alternatively, go to Applications > Utilities > Terminal)
- Test things are working by opening terminal and type `python`
- Success if:



Command Prompt in Windows

- Type `cmd` in program search bar to open the command prompt
- Set up python by adding `C:\Python27` to your PATH variable. See exact steps at stackoverflow
- Test things are working by opening command prompt and type `python`
- Success if:



Choosing an Integrated Development Environment

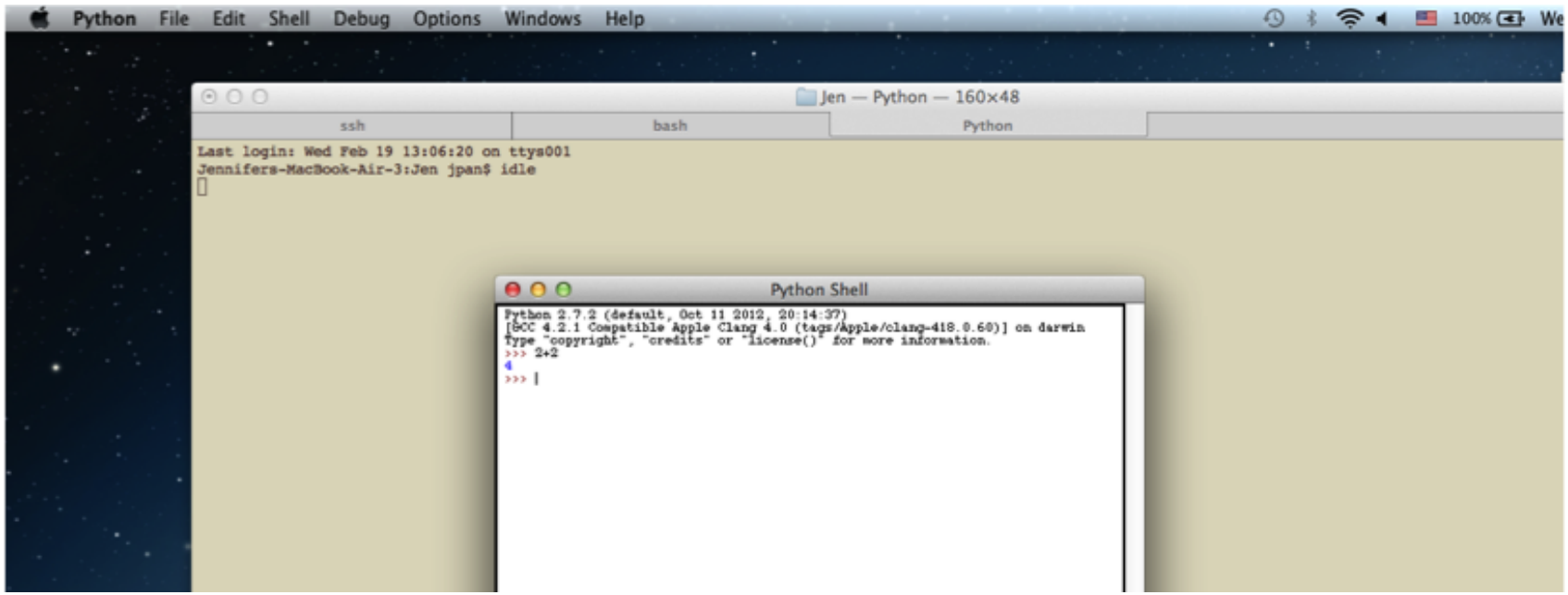
You can write python code in any text editor (e.g., TextEdit (mac) Notepad (windows), just save the file with `.py`. It's helpful to have an IDE because the editor will mark things up (i.e., color code, automatically indent) to avoid mistakes, and you can run bits of code (i.e., for testing).

There are **many** to choose from (`https://wiki.python.org/moin/PythonEditors`) You may want to try python before deciding on an IDE

To keep things simple for now, use the IDE that comes with your python installation: IDLE

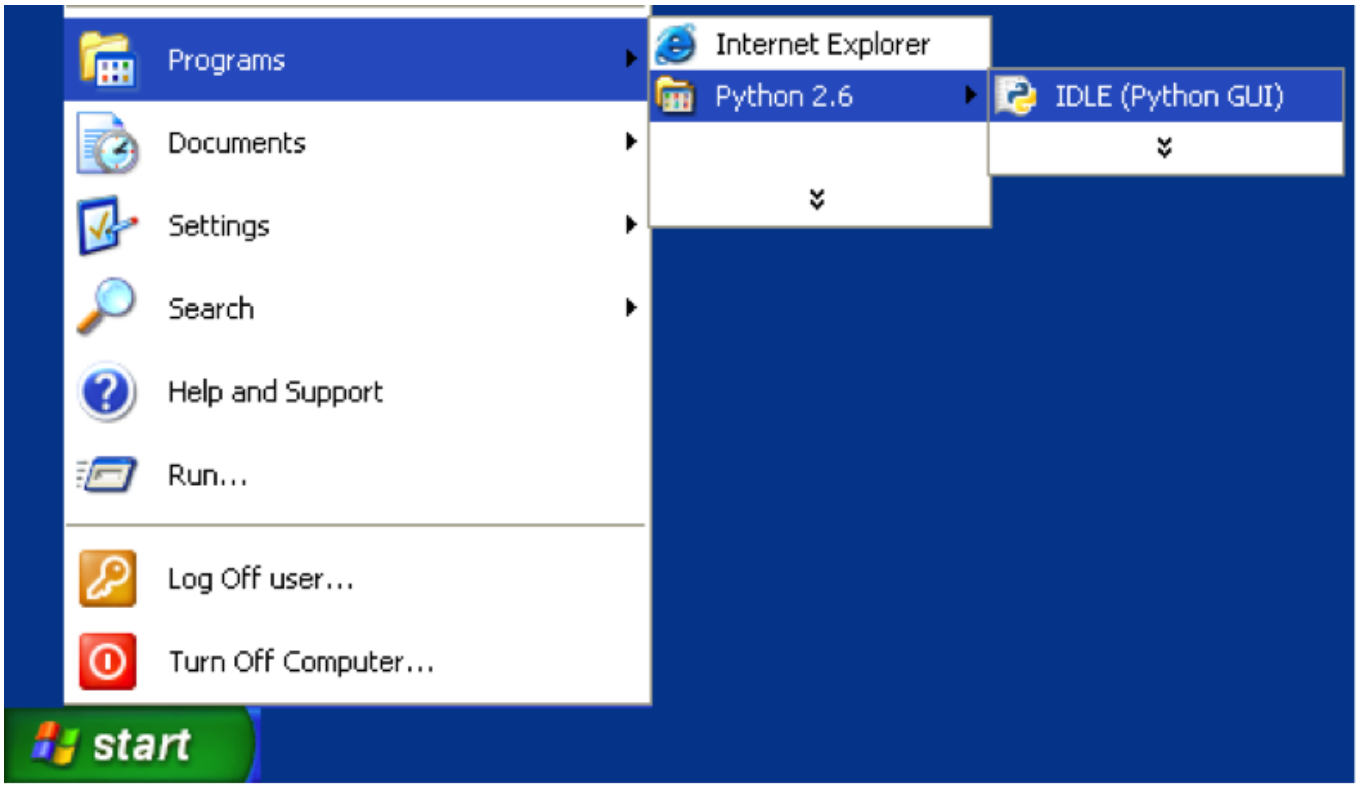
IDLE for Mac

- Use the pre-installed IDLE IDE
- Go to terminal and type `idle`
- Success:



IDLE for Windows

- Use the pre-installed IDLE IDE
- Go to program search bar and type `idle`, select `IDLE (Python GUI)`
- Success:



Modules for Scraping and Text Analysis

Modules (like packages in R) are pre-packaged sets of functions
Some modules come pre-install with Python

- `urllib2`: module for opening URLs
- `string`: module for working with strings
- `re`: module for regular expression matching
- `csv`: module for saving output into `.csv` format

Other helpful modules that you need to install:

- `BeautifulSoup`: module for HTML parsing, can parse anything, but can be slow (`lxml` is faster alternative)
- `nltk`: lots of tools for working with human language data
- `scikit-learn`: machine learning module, e.g., naive bayes, svm (requires `NumPy`, `SciPy`, `matplotlib`)
- `gensim`: module for topic models

How to Install of Module: BeautifulSoup

- Find `BeautifulSoup` version compatible with your Python version
- Download a file that ends with `.tar.gz`
- In terminal, cd into directory with `.tar.gz` file, type:
  `$ tar -xzf [filename]`
  (tar is the Unix command to unpack the file. The letters after the dash are Unix options. The x tells Unix to extract from the tar ball, the z tells Unix its a zip file, and the f tells Unix to name the unpacked directory according to the name of the tar file itself.)
- cd into new directory with name that looks like the name of the tar file, type ls and you should see a file called setup.py (every Python module comes with this file, which tells the computer how to install the module). Type:
  `$ python setup.py install`
- To test if it installed properly, launch Python in terminal and type:
  `$ import bs4`
- If it imports without error, you've succeeded!

For more information

bit.ly/JenPan_Python

Please send corrections to jjpan@fas.harvard.edu