**Online Resources & Reference Material**

**General Python Resources**

* [Python Programming Language](http://www.python.org), the official Python website   
  (**Note:** We will primarily use Python 2.7 in this class).
* [Python 2.7 vs. Python 3.x](http://sebastianraschka.com/Articles/2014_python_2_3_key_diff.html) - Differences between the two versions with examples.
* [The Python Tutorial](http://docs.python.org/tutorial/)
* [The Python Standard Library](http://docs.python.org/library/index.html)
* [How to Think Like a Computer Scientist - The Python Version](http://openbookproject.net/thinkcs/python/english2e/)
* [Think Python](http://www.greenteapress.com/thinkpython), on-line book by Allen Downey
* [Learn Python the Hard Way](http://learnpythonthehardway.org/book/), another online book on Python

**Important Tools and Libraries**

* [IPython](http://ipython.org/): A[REPL](http://en.wikipedia.org/wiki/Read–eval–print_loop) for easy interactive python development. Extremely useful for testing ideas out one line of code at a time. We will use IPython Notebook (a Web based interactive shell for Python) extensively in this class.
* [Jupyter Notebook](http://jupyter.org/) (formerly IPython Notebook)
* [Jupyter Notebook Tutorial](https://www.youtube.com/watch?v=HW29067qVWk) - Nice  tutorial video by Corey Schafer.
* [matplotlib](http://matplotlib.sourceforge.net/):  A very nice plotting library, capable of generating production-level visualizations programmatically. Matlab-like syntax makes plotting very easy.
* [NumPy](http://numpy.scipy.org/): The fundamental package for scientific computing with Python.
* [SciPy](http://www.scipy.org/): the open source library for mathematics, science and engineering
* [scikit-learn](http://scikit-learn.org/stable/): a robust machine learning library building on top of NumPy, SciPy and matplotlib. Includes of a wide variety of modeling techniques.
* [Pandas (python data analysis library)](http://pandas.pydata.org/): data structures and tools for common data analysis tasks, including an efficient data frame implementation (similar to R).
* [BeautifulSoup](http://www.crummy.com/software/BeautifulSoup/): A general parsing library particularly useful for parsing html and xml.
* [NLTK:](http://nltk.org/) Natural Language Toolkit for Python, including tools for text preprocessing, tokenization, and vectorization (you may  also be interested in an [online book](http://nltk.org/book/) that shows how NLTK is used).
* [NetworkX](http://networkx.github.io/): Python language library for the creation, manipulation, and analysis of graphs and networks.

**Installation of Python and Scientific Libraries**

* [Anaconda](http://continuum.io/downloads) - (Mac, Windows, Linux) Python distribution for large-scale data processing and scientific computing (includes scientific and data analysis libraries such as Numpy, Pandas, and scikit-learn, as well as IPython). **This is the recommended package for this class.** Anaconda provides distributions both for Python 3.4 as well as Python 2.7 (for class examples, we will use Python 2.7).
* [Managing Python Environments in Anaconda](https://conda.io/docs/user-guide/tasks/manage-environments.html) (e.g., to set up Python 2 and Python 3 kernels on the same machine)
* [Notepad++](http://notepad-plus-plus.org/): Excellent Python-friendly text editor
* [Installing NumPy and SciPy](http://www.scipy.org/Installing_SciPy)
* [Installing scikit-learn](http://scikit-learn.org/stable/install.html)
* [Installing Pandas](http://pandas.pydata.org/pandas-docs/stable/install.html)
* [Standalone Python Distributions](http://www.python.org/getit/)

**References for Data Analysis in Python**

* [Python Scientific Lecture Notes](http://scipy-lectures.github.io/index.html): including detailed notes on NumPy, Matplotlib, and Scipy
* [NumPy and SciPy Documentation](http://docs.scipy.org/doc/): including NumPy User Guide and Cookbook
* [Tentative NumPy Tutorial](http://wiki.scipy.org/Tentative_NumPy_Tutorial): from SciPy.org
* [Pandas Tutorials:](http://pandas.pydata.org/pandas-docs/stable/tutorials.html) This is a guide to many pandas tutorials, geared mainly for new users.
* [10 Minutes to Pandas](http://pandas.pydata.org/pandas-docs/stable/10min.html)
* [10 Minute Tour of Pandas](https://vimeo.com/59324550) - A short video introduction by Wes Mackinny.
* [Getting Started with Python for Data Science](http://www.kaggle.com/wiki/GettingStartedWithPythonForDataScience): from Kaggle.com - includes information about installing Python and relevant libraries.
* [Introduction to NumPy and Matplotlib](http://www.youtube.com/watch?v=3Fp1zn5ao2M) - YouTube video
* [Matplotlib User's Guide](http://matplotlib.org/contents.html)
* [Matplotlib Pyplot Tutorial](http://matplotlib.org/users/pyplot_tutorial.html)
* [IPython Documentation](http://ipython.org/documentation.html)
* [IPython Notebook Tutorials and Examples](http://nbviewer.ipython.org/github/ipython/ipython/blob/2.x/examples/Notebook/Index.ipynb)
* [Natural Language Processing with Python](http://nltk.org/book/): Online book on text processing and analysis using NLTK.

**Other Relevant Tools & Resources**

* [Weka - Open source data mining package written in Java](http://www.cs.waikato.ac.nz/ml/weka/)
* [Weka Content Presentations on SlideShare](http://www.slideshare.net/wekacontent/presentations)
* [The Weka Manual](http://prdownloads.sourceforge.net/weka/WekaManual-3-6-10.pdf?download), as included in the distribution
* [On-line stats book](http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html)
* [OpenIntro's textbook on basic statistics](http://www.openintro.org/stat/textbook.php) *(openintro.org) -* Basic statistics skills to get you started for analysis tasks.
* [Statistical formulas for programmers](http://www.evanmiller.org/statistical-formulas-for-programmers.html)
* [**Knowledge Discovery Nuggets (KDnuggets)**](http://www.kdnuggets.com/) - Great starting point: tools, links, companies, etc.

**Data Sets**

* [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/) - A repository of more than 200 data sets for machine learning and data mining
* [mldata.org](http://mldata.org/) - A large machine learning data repository
* [**Movie Ratings Data**](http://facweb.cs.depaul.edu/mobasher/classes/ect584/data/movielens.zip) - Real movie ratings data from [www.movielens.org](http://www.movielens.org) Web site. Contains ratings on 1600+ movies by 1000 users
* [Kaggle.com Competition Data Sets](http://www.kaggle.com/competitions) - Data sets from a variety of competitions. Also a good source for class project ideas.
* [Stanford Large Network Dataset Collection](http://snap.stanford.edu/data/) - A variety of network data sets, including data from social networks, product reviews, online communities, etc.
* [Yelp Data Set Challenge](http://www.yelp.com/dataset_challenge/) - Reviews and check-in data on thousands of businesses.
* [Online Grocery Shopping Data from Instacart.](https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2)
* [Million Song Dataset](http://labrosa.ee.columbia.edu/millionsong/) - Freely-available collection of audio features and metadata for a million contemporary popular music tracks.
* [All the News](https://www.kaggle.com/snapcrack/all-the-news) - 143,000 articles from 15 American publications
* [Public Data sets on Amazon Web Services](http://aws.amazon.com/publicdatasets/) - Large public data sets (including data sets for US Census, Wikipedia, Freebase, human genome project), ready for big data analytics on the cloud.
* [Data.gov](http://catalog.data.gov/dataset) - Publically available data sets from Federal, State, and local government, including economic, geological, demographic and many other types of data sources. This site also includes a list of other [Open Data Sites](http://www.data.gov/opendatasites) with similar publicly available data sources from various cities, states, and countries.
* [KDnugget's list of data sets for data mining](http://www.kdnuggets.com/datasets/)
* [Infochimps Data Market](http://www.infochimps.com/datasets) - Thousands of data sets, including data from various social networks and collaborative tagging sites such as Twitter, Delicious, Last.fm, MusicBrainz, as well as data sets from many other domains.