Week2

Introduction to Python and Scientific Packages

Programming in Python

- Python is an **interpreted language**.
- **Code portability**. Runs on hardware/software platforms different from which used to develop the code.
- **Variables**: A variable stores a piece of data and gives it a name.
- **Lists**: What if we want to store many integers? We need a list!
- **Loops**: Repeat code until a conditional statement ends the loop.
- **Conditionals**: Sometimes you want to execute code only in certain circumstances.
- **Functions**: We can separate off code into functions, that can take input and can give output. They serve as black boxes from the perspective of the rest of our code.

Python Libraries for Data Science I

- Extensive first and third party libraries. Top Python Libraries for Data Science.
  - **NumPy** (aka **Num**erical **Py**thon) is the core numeric and scientific computation library in Python. General-purpose array-processing package.
  - **SciPy** (aka **Sci**entific **Py**thon) is extensively used for scientific and technical computations (extends NumPy).
  - **Matplotlib** is an essential library in Python for data visualization in data science. A plotting library.
  - **Pandas** (Python data analysis) is a foundational Python library for data analysis in data science. Data cleaning, data handling, manipulation, and modeling.
  - **Seaborn** is another library in Python for data visualization. Extension of Matplotlib. Statistical and graphical analysis in data science.

Python Libraries for Data Science II

- Top Python Libraries for Data Science.
  - **SciKit-Learn** is a robust machine learning library in Python. Data mining, feature engineering, training and deploying machine learning models.
  - **Statsmodels** - provides functionalities for descriptive and inferential statistics for statistical models.
  - **TensorFlow** - a framework for defining and running computations that involve tensors. Machine learning and deep learning framework.
  - **Keras** is a neural network Python library for deep learning model development, training, and deployment.
  - **PyTorch** - scientific computing package that uses the power of graphics processing units

Python Libraries for Data Science III

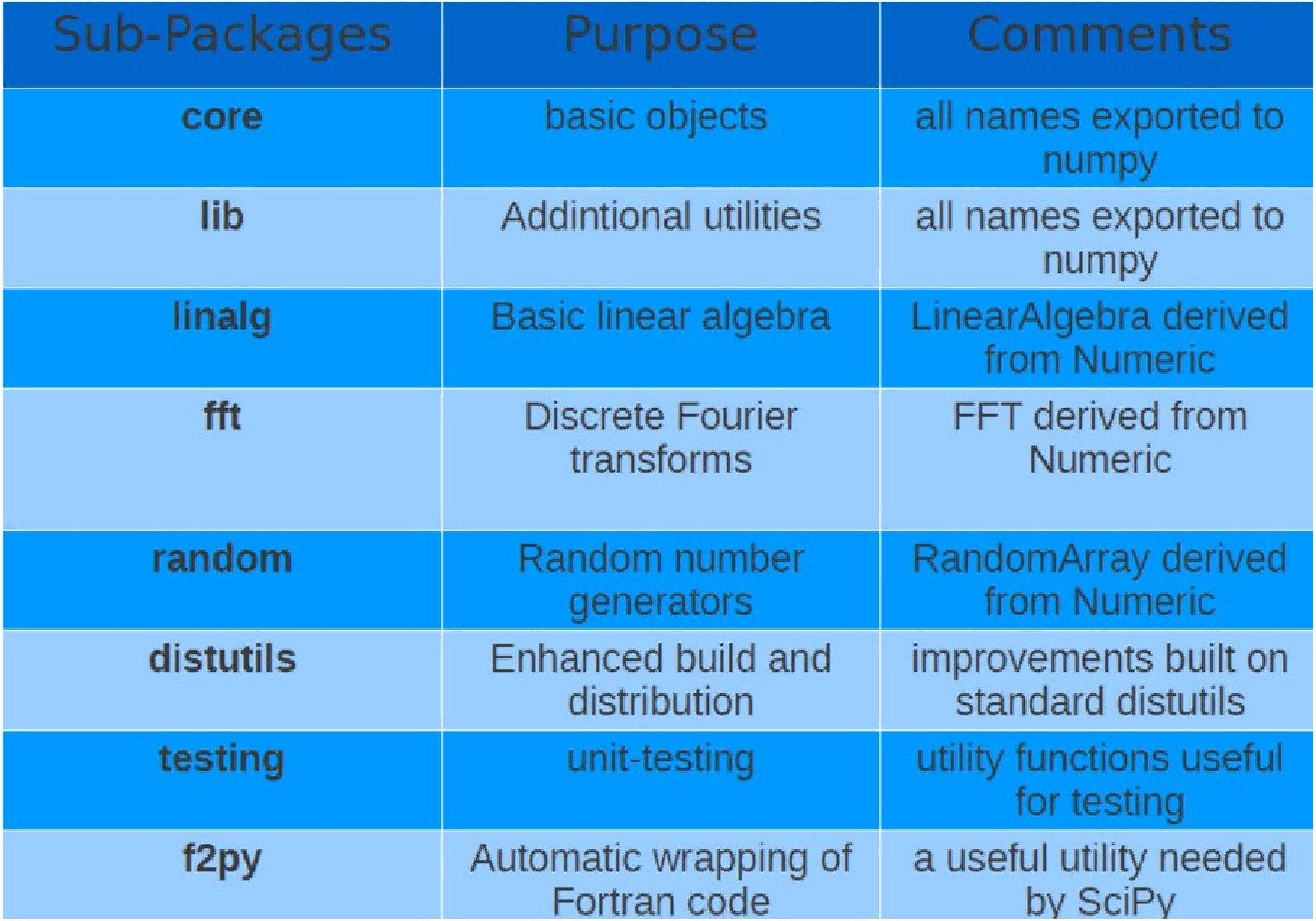
- Top Python Libraries for Data Science.
  - **Scrapy** - for web crawling frameworks
  - **BeautifulSoup** - for web crawling and data scraping
  - **NLTK** (**N**atural **L**anguage **T**ool **K**it) is a Python package essentially for natural language processing.
- One of the major reasons for its immense attraction is libraries and packages it has to offer.

NumPy

Numpy (**Num**erical **Py**thon) - The Fundamental Package for Scientific Computing with Python. https://numpy.org/

- NumPy offers high-quality mathematical functions and supports logical operations on built-in *multi-dimensional array objects*.
- NumPy arrays are significantly faster than traditional Python lists and way more efficient in performance.
- Some of the features provided by NumPy
  - Basic array operations such as addition and multiplication
  - Mathematical, logical, shape manipulation operations
  - Indexing, slicing, flattening, and reshaping the arrays
  - Stacking, splitting, and broadcasting arrays
  - I/O Operations
  - Fourier transform capabilities
  - Basic linear algebra
  - Basic statistical operations
  - Random number generation
  - ….

NumPy module organization



Sub-Packages	Purpose	Comments
core	basic objects	all names exported to numpy
lib	Addintional utilities	all names exported to numpy
linalg	Basic linear algebra	LinearAlgebra derived from Numeric
fft	Discrete Fourier transforms	FFT derived from Numeric
random	Random number generators	RandomArray derived from Numeric
distutils	Enhanced build and distribution	improvements built on standard distutils
testing	unit-testing	utility functions useful for testing
f2py	Automatic wrapping of Fortran code	a useful utility needed by SciPy

SciPy

- SciPy is a scientific computation library in Python. A collection of mathematical functions and algorithms built on Python's extension NumPy https://scipy.org/
- It provides the user with high-level commands and classes for manipulating and visualizing data.
- It is widely used in machine learning and scientific programming and comes with integrated support for linear algebra and statistics.
- Some of the features provided by SciPy
    - Search for minima and maxima of functions
    - Calculation of function integrals
    - Support for special functions
    - Signal processing
    - Multi-dimensional image processing
    - Work with genetic algorithms
    - Fourier transforms
    - Solving ordinary differential equations
    - ….
- The scipy namespace itself only contains functions imported from numpy. Therefore, importing only the scipy base package does only provide numpy content, which could be imported from numpy directly (NOT USED as *import scipy*).

SciPy modules

Subpackage	Description
cluster	Clustering algorithms
constants	Physical and mathematical constants
fftpack	Fast Fourier Transform routines
integrate	Integration and ordinary differential equation solvers
interpolate	Interpolation and smoothing splines
io	Input and Output
linalg	Linear algebra
ndimage	N-dimensional image processing
odr	Orthogonal distance regression
optimize	Optimization and root-finding routines
signal	Signal processing
sparse	Sparse matrices and associated routines
spatial	Spatial data structures and algorithms
special	Special functions
stats	Statistical distribution and function

i.e., from scipy import linalg, io

Matplotlib

- Matplotlib is the core plotting and data visualization package in Python. https://matplotlib.org/
- A 2D graphical Python library which produces publication quality figures. However, it also supports 3D graphics (mplot3d toolkit), but this is very limited.
- Matplotlib is capable of producing high-quality figures in various formats. It offers interactive cross-platform environments for plotting.
- It provides a MATLAB/Mathematica-like interface for simple plotting *pyplot submodule* with secondary x-y axis support, and facilitates the creation of subplots, labels, grids, legends, use a logarithmic scale or polar coordinates etc.
  - Matplotlib also allows full control of axes properties, font styles, line and marker styles, and some more formatting entities.
- You can generate line plots (Charts), bar charts, histograms, power spectra, pie charts, error charts, box plots, scatter plots, stem plots, contour plots, etc., with just a few lines of codes in Matplotlib.

Programming in Python

- https://python-course.eu/
- https://www.codecademy.com/catalog/language/python
- https://docs.python.org/
- https://scipy-lectures.org/
- https://www.naukri.com/learning/articles/top-10-powerful-python-libraries-for-data-science/
- https://computation.physics.utoronto.ca/tutorials/
- https://moodle2.units.it/course/view.php?id=6837
- https://jckantor.github.io/CBE30338/
- https://matplotlib.org/stable/tutorials/index.html