Intro to Python for Data Science
Arusha Tech

Anthony FAUSTINE

August 2017

Outline

1 Introduction

2 Python

3 Data Science

4 Python Packages for Data Science

Learning goal

- Understand python programming language and different python libraries for data science.
- Explore Python language fundamentals, including basic syntax, variables, control flow, data structure and functions.
- Build Numpy arrays, and perform basic and some linear algebra calculations.
- Create and customize plots using matplotlib.

Presenter Bio

- PhD student at Nelson Mandela African Institution of Science and Technology,
- **Research :** Applied machine learning and signal processing for computational sustainability.
  - Develop probabilistic-deep learning algorithm (Hybrid HMM-DNN) for energy dis-aggregation problem.
- Co-founder Pythontz
- Assistant Lecturer (UDOM), Researcher (Vicres, Hakikidawa).

Pythontz

python™
Tanzania
We aim to create a vibrant and diverse python community in Tanzania

Pythontz

About Pythontz

- A postive peer learning community for interested Python users in Tanzania.

Vision

- To create a vibrant and diverse python community in Tanzania.

Mission

- To foster the application of python programming across industries, learning centers, schools and community in Tanzania.

Outline

1 Introduction

2 Python

3 Data Science

4 Python Packages for Data Science

Introduction

What is Python ?

A very popular general-purpose programming language.

- Open source general-purpose language
- Dynamically semantics (rather than statically typed like Java or C/C++)
- Interpreted (rather than compiled like Java or C/C++)
- Object Oriented,

What can you use Python for ?

- Web development (Django)
- Web Scraping (Beautiful Soup)
- Scripting Language.
- Scientific programming and Numeric Computing.
- Automation and Embedded Sytstem.
- Desktop GUIs and 3D modelling.

But Why Python ?

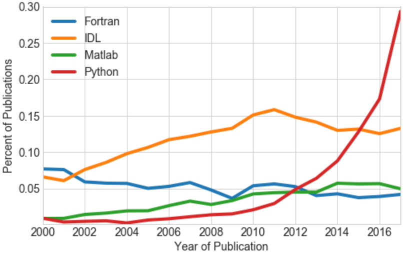


FIGURE – Jake VanderPlas PyCon 2017

- Python is a “teaching language”
- ....created to “bridge the gap between the shell and C
- “never intended. . . to be the primary language for programmers.”

Why is Python such an effective tool in science ?

1. Interoperability with Other Languages : You can use it in the shell on microtasks, or interactively, or in scripts, or build enterprise software with GUIs.
2. “Batteries Included” + Third-Party Modules : Python has built-in libraries and third-party liabraies for nearly everything.
3. Simplicity & Dynamic Nature : You can run your Python code on any architecture.
4. Open ethos well-fit to science : Easy to reproduce results with python
5. Python is the future of Machine Learning and AI.

Jake VanderPlas PyCon 2017

Why is Python such an effective tool in science ?

1. Interoperability with Other Languages : You can use it in the shell on microtasks, or interactively, or in scripts, or build enterprise software with GUIs.
2. “Batteries Included” + Third-Party Modules : Python has built-in libraries and third-party liabraies for nearly everything.
3. Simplicity & Dynamic Nature : You can run your Python code on any architecture.
4. Open ethos well-fit to science : Easy to reproduce results with python
5. Python is the future of Machine Learning and AI.

Jake VanderPlas PyCon 2017

Why is Python such an effective tool in science ?

1. Interoperability with Other Languages : You can use it in the shell on microtasks, or interactively, or in scripts, or build enterprise software with GUIs.
2. “Batteries Included” + Third-Party Modules : Python has built-in libraries and third-party liabraies for nearly everything.
3. Simplicity & Dynamic Nature : You can run your Python code on any architecture.
4. Open ethos well-fit to science : Easy to reproduce results with python
5. Python is the future of Machine Learning and AI.

Jake VanderPlas PyCon 2017

Why is Python such an effective tool in science ?

1. Interoperability with Other Languages : You can use it in the shell on microtasks, or interactively, or in scripts, or build enterprise software with GUIs.
2. “Batteries Included” + Third-Party Modules : Python has built-in libraries and third-party liabraies for nearly everything.
3. Simplicity & Dynamic Nature : You can run your Python code on any architecture.
4. Open ethos well-fit to science : Easy to reproduce results with python
5. Python is the future of Machine Learning and AI.

Jake VanderPlas PyCon 2017

Why is Python such an effective tool in science ?

1. Interoperability with Other Languages : You can use it in the shell on microtasks, or interactively, or in scripts, or build enterprise software with GUIs.
2. “Batteries Included” + Third-Party Modules : Python has built-in libraries and third-party liabraies for nearly everything.
3. Simplicity & Dynamic Nature : You can run your Python code on any architecture.
4. Open ethos well-fit to science : Easy to reproduce results with python
5. Python is the future of Machine Learning and AI.

Jake VanderPlas PyCon 2017

Why is Python such an effective tool for Data Science

1. Very rich scientific computing libraries
2. All DS tasks can be performed with Python :
   - accessing, collecting, cleaning, analysing, visualising data
   - modelling, evaluating models, integrating in prod, scaling

http ://slides.com/utstikkar/introtopython-pythonproglanguage#/3

Why is Python such an effective tool for Data Science

1. Very rich scientific computing libraries
2. All DS tasks can be performed with Python :
    - accessing, collecting, cleaning, analysing, visualising data
    - modelling, evaluating models, integrating in prod, scaling

http ://slides.com/utstikkar/introtopython-pythonproglanguage#/3

Why is Python such an effective tool for Data Science

1. Very rich scientific computing libraries
2. All DS tasks can be performed with Python :
    - accessing, collecting, cleaning, analysing, visualising data
    - modelling, evaluating models, integrating in prod, scaling

http ://slides.com/utstikkar/introtopython-pythonproglanguage#/3

Why is Python such an effective tool for Data Science

1. Very rich scientific computing libraries
2. All DS tasks can be performed with Python :
    - accessing, collecting, cleaning, analysing, visualising data
    - modelling, evaluating models, integrating in prod, scaling

http ://slides.com/utstikkar/introtopython-pythonproglanguage#/3

PYTHON 2 VS. PYTHON 3

- 2 major versions of Python in widespread use : Python 2.x and Python 3.x
- Some features in Python 3 are not backward compatible with Python 2
- Some Python 2 libraries have not been updated to work with Python 3
- Bottom-line : there is no wrong choice, as long as all the libraries you need are supported by the version you choose.
- In this workshop : Python3

Resource to learn Python

10 Resources to Get Started Learning Python

Outline

1 Introduction

2 Python

3 Data Science

4 Python Packages for Data Science

What is Data science

`The future belongs to the companies and people that turn data into products. By Mike Loukides June 2, 2010`

Data science : deals with analyzing and manipulating data to derive insights and build data products.

- The end goal of DS ⇒ data product :

  Data product : any tool created with the help of data to make a more informed decision.

What is Data science

`The future belongs to the companies and people that turn data into products. By Mike Loukides June 2, 2010`

Data science : deals with analyzing and manipulating data to derive insights and build data products.

- The end goal of DS ⇒ data product :

  Data product : any tool created with the help of data to make a more informed decision.

What is Data science

`The future belongs to the companies and people that turn data into products. By Mike Loukides June 2, 2010`

Data science : deals with analyzing and manipulating data to derive insights and build data products.

- The end goal of DS ⇒ data product :

  Data product : any tool created with the help of data to make a more informed decision.

Data science vs Machine learning

Machine learning : a set of algorithms that learn from data in order to make predictions or inference.

- Data Science is the real-world application of machine learning, with the goal of creating data products.

Outline

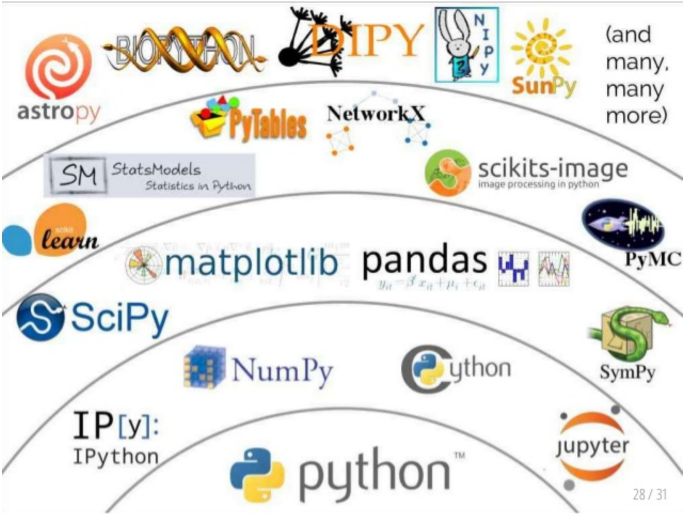
1 Introduction

2 Python

3 Data Science

4 Python Packages for Data Science

Python’s Scientific Stack



Jupyter

Jupyter : Open-source web application for interactive and exploratory computing.

- Allows to create and share documents that contain live code, equations, visualizations and explanatory text.

- It is a platform for Data Science at scale.
- Covers all the life-cycle of scientific ideas :ideas to publications.
- Demo

Numpy and Sci-py

Numpy : the fundamental Python package for scientific computing.

- Provide high-performance vector, matrix and higher-dimensional data structures.
- Offers Matlab-ish capabilities within Python.

Sci-py : Collections of high level mathematical operations

- linear algebra.
- Optimization
- Integration etc.

statsmodels

statsmodels : statistical modelling toolbox

Matplotlib

Matplotlib is an excellent 2D and 3D graphics library for generating scientific figures.

- It provides both a very quick way to visualize data from Python and publication-quality figures in many formats.

Other data visualization packages : Seaborn and Bokeh.

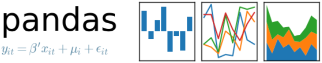
Other Python Library for Visualization

87.62
88.905
56
Ba
Barium
137.327
71
Lu
1.0
10.0
100.0
Pd
Tc
Ag
Cd

Pandas

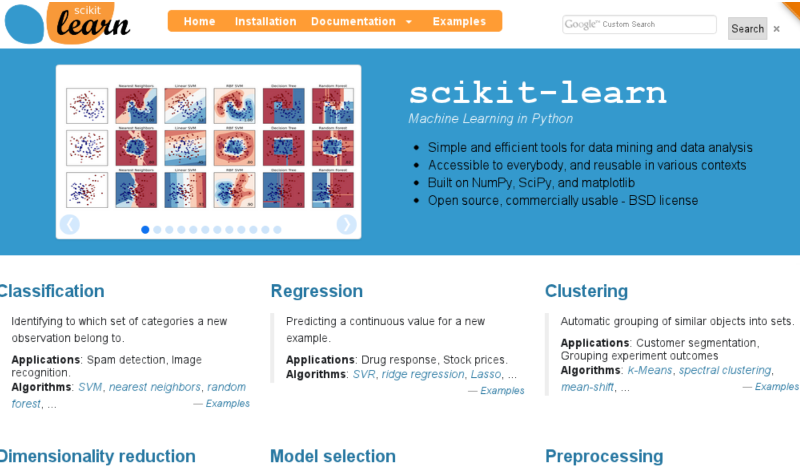
Panda : a python package providing fast, flexible, and expressive data structures for data analysis.

- A fundamental high-level building block for doing practical, real world data analysis in Python.
- Designed to work with relational or labeled data or both.



Scikit-Learn for ML

Scikit-Learn (sklearn) is Python's premier general-purpose machine learning library.



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which set of categories a new observation belong to.

Applications: Spam detection, Image recognition.
Algorithms: *SVM*, *nearest neighbors*, *random forest*, ... — *Examples*

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.
Algorithms: *SVR*, *ridge regression*, *Lasso*, ... — *Examples*

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: *k-Means*, *spectral clustering*, *mean-shift*, ... — *Examples*

Dimensionality reduction

Model selection

Preprocessing

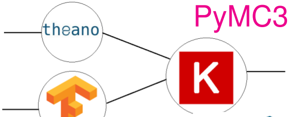
Python ML and AI libraries

Tensorflow

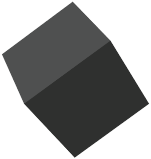
Theano

Pytorch

Keras



Edward

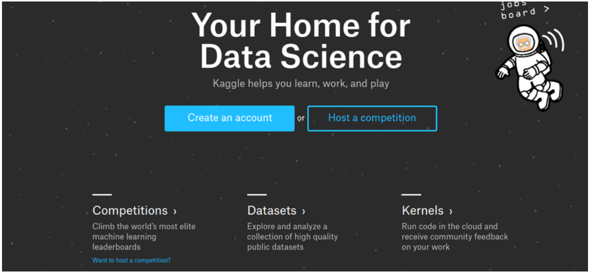


PyMC3

NLTK

Data Science Platform

Kaggle : helps you learn, work, and play.



Your Home for Data Science

Kaggle helps you learn, work, and play

Competitions ›
Climb the world's most elite machine learning leaderboards
Want to host a competition?

Datasets ›
Explore and analyze a collection of high quality public datasets

Kernels ›
Run code in the cloud and receive community feedback on your work

Data set :

- Academic Torrents
- UCI Machine learning repository

THANK YOU

Practical Session

Practical Session