DATA SCIENCE USING PYTHON

Deepashree Dwarakanath

Department of Masters of Computer Application

Dayananda Sagar College of Engineering

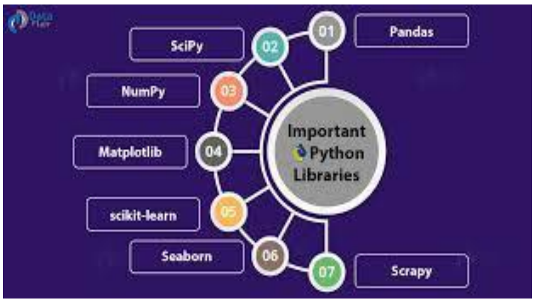
ABSTRACT

Python is an object oriented, scripted and interpreted language for both learning and real world programming. Python is a powerful high-level language created by Guido van Rossum. In this paper, we will provide an introduction to the main Python software tools used for Data science, Machine learning techniques and IOT. Briefly, this paper will first introduce Python as a language, and give introduction about Data science, Machine learning and IOT, and then describe packages that are popular in the Data science and Machine learning communities, such as NumPy, SciPy, TensorFlow, Keras ,Matplotlib etc. From there, we will move to show the importance of python for building IOT applications. We will use different code examples throughout. To aid the learning experience, execute following examples contained in this paper interactively using Jupiter notebooks .

Keywords: Machine learning · Data Science · IOT · Tools · Languages · Python

INTRODUCTION

1.1 Introduction to python



Python is a general-purpose, high-level programming language which became popular in the recent times .It allows programmer to write the code in fewer lines that is not In this paper we wish to give brief possible with other languages. The important feature in Python programming is it supports multiple programming paradigms. Python provides a large set of comprehensive standard library which is extensible. The main features of Python are Simple and easy to learn, Freeware and open source, High level , Simple and easy to learn, Freeware and open source, High level programming language, Platform independent, Portability, Dynamically typed, Both procedure oriented and Object oriented, Interpreted, Extensible, Embedded, Extensive Library.

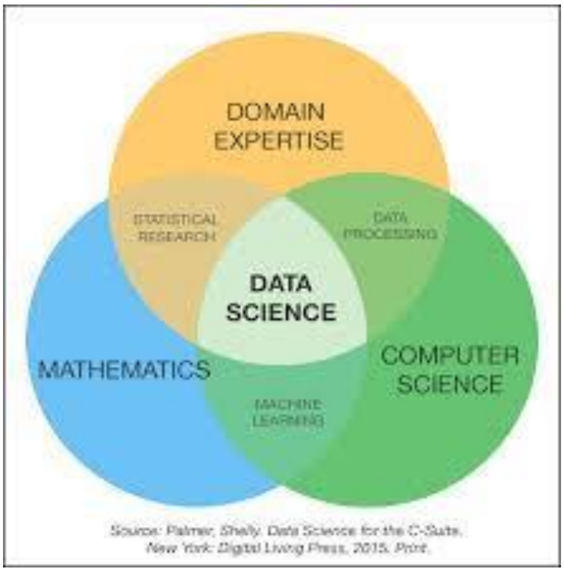
In this paper, we wish to give brief idea of python in the area of Data science, IOT and Machine learning. Python is known to have an abundance of libraries that assist with data analysis and scientific computing. For example, we can build python application which helps data analysts to analyse large amounts of data for scientific computing. The prerequisites for this paper are basic under-standing of statistics, as well as some experience in any C-style language. Some knowledge of Python is useful but not a must.

An accompanying GitHub repository is provided to aid the tutorial. It contains a number of notebooks of python code snippets for reference. It helps to go through number of examples related to different modules of Python.

Https://github.com/mdbloice/MLDS

1.2 Introduction to Data Science

Data science is a multi- disciplinary area that uses scientific methods, procedures, tools and systems to extract knowledge and get insights into structured and unstructured data. Data science is related to data analytics, data mining and big data. It understands the phenomenon of the data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science and information science.



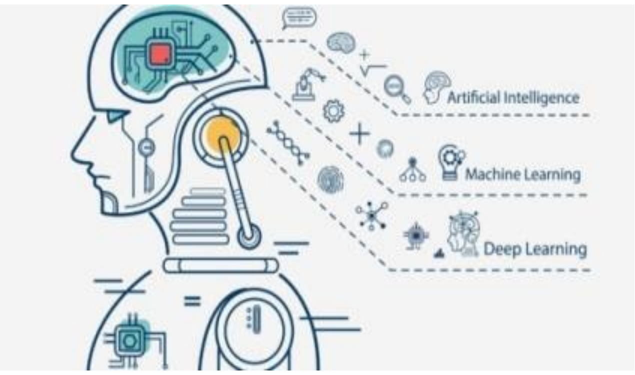
Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Statistics is one of the most important disciplines to provide tools and methods to find structure in and to give deeper insight into data, and the most important discipline to analyse and quantify uncertainty. Python provides various predefined modules to work on Data science projects.

1.3 Introduction to Machine Learning

The term machine learning refers to the automated detection of meaningful patterns in data. In the past couple of decades it has become a common tool in almost any task that requires information extraction from large data sets. We are surrounded by a machine learning based technology: search engines learn how to bring us the best results (while placing profitable ads), anti-spam software learns to filter our email messages, and credit card transactions are secured by software that learns how to detect frauds. Digital cameras learn to detect faces and intelligent personal assistance applications on smartphones learn to recognise voice commands.

Cars are equipped with accident prevention systems that are built using machine learning algorithms. Machine learning is also widely used in scientific applications such as bioinformatics, medicine, and astronomy. One common feature of all of these applications is that, in contrast to more traditional uses of computers, in these cases, due to the complexity of the patterns that need to be detected, a human programmer cannot provide an explicit, fine- detailed specification of how such tasks should be executed. Taking example from intelligent beings, many of our skills are acquired or refined through learning from our experience (rather than following explicit instructions given to us). Machine learning tools are concerned with endowing programs with the ability to “learn “and adapt.



Because machine learning is typically used to process large volumes of data, you may want to choose a powerful low-level language. However, if you’re only just beginning to explore this field, it might be better to start with Python. Python is beginner-friendly, and can do the same thing that other coding languages can, but in fewer lines of code. If you are interested in exploring machine learning with Python, this paper will serve as your guide. This is paper gives overview of programming machine learning using Python.

2. OBJECTIVES OF STUDY

1. To conceptualize the features of Python

2. To investigate python modules for Data Science like Numpy which is used for matrix and vector manipulation, SciPy, the 2D plotting library Matplotlib etc

3. To focus on python modules for Machine learning like Tensor flow numerical computations for machines learning, Keras for neural networks and deep learning.

3. RELATED WORKS

3.1 Basic Features of Python

Python is a general-purpose interpreted, interactive, object- oriented, and high-level programming language. It was created by Guido van Rossum during 1985-19990. Like Perl, Python source code is also available under the GNU General Public License (GPL). This paper gives enough understanding on Python programming language.

• Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

• Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

• Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

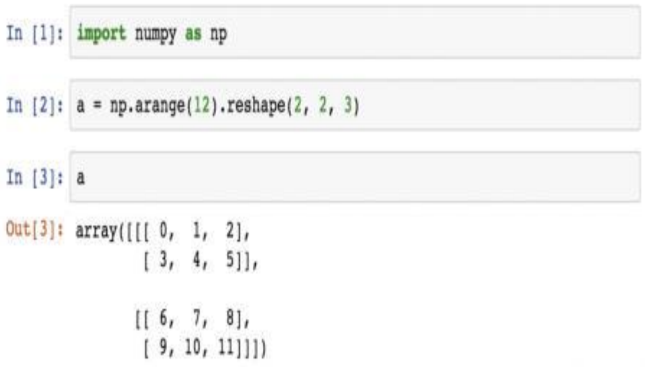
• Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

3.2 Python for Data Science

These are the most essential Data Science libraries you have to know:

- Numpy
- Matplotlib
- Scipy

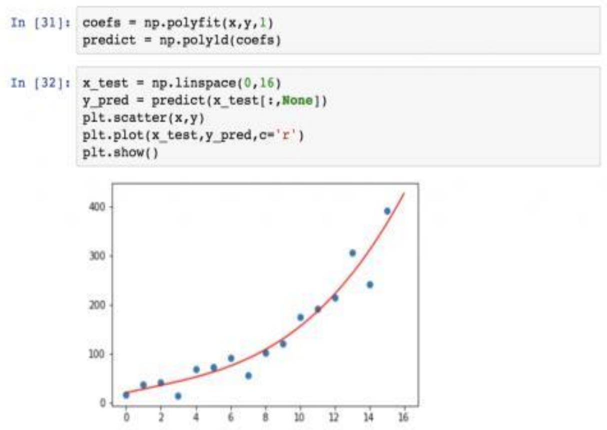
Numpy: Numpy will help us to manage multi-dimensional arrays very efficiently. Maybe it is difficult to do that directly, but since the concept is a crucial part of data science, many other libraries (well, almost all of them) are built on Numpy. Simply to say, without Numpy it is difficult to use Pandas, Matplotlib, Scipy or Scikit-Learn.



```
In [1]: import numpy as np

In [2]: a = np.arange(12).reshape(2, 2, 3)

In [3]: a

Out[3]: array([[[ 0,  1,  2],
                [ 3,  4,  5]],

               [[ 6,  7,  8],
                [ 9, 10, 11]]])
```

3-dimensional Numpy array:

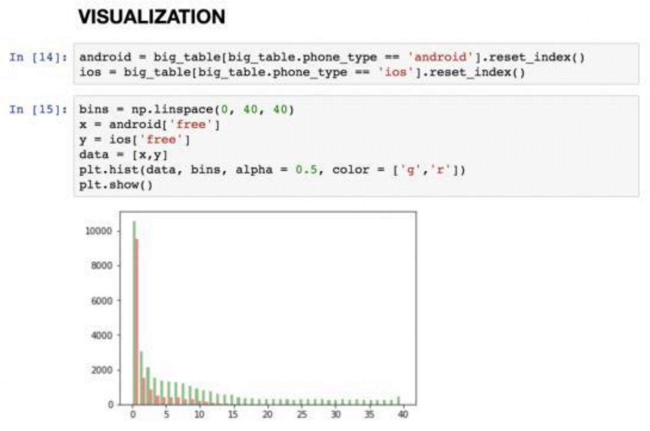
But on the other hand, it also has a few well-implemented methods. It's quite to use Numpy random function, which is found slightly better than the random module of the standard library. And when it comes to simple predictive analytics tasks like linear or polynomial regression, Numpy polyfit function will be favourite.



```
In [31]: coefs = np.polyfit(x,y,1)
         predict = np.poly1d(coefs)

In [32]: x_test = np.linspace(0,16)
         y_pred = predict(x_test[:,None])
         plt.scatter(x,y)
         plt.plot(x_test,y_pred,c='r')
         plt.show()
```

Matplotlib:

Data visualization is very important. Data visualization helps us to better understand the data, discover things that wouldn't discover in raw format and communicate findings more efficiently to others. The best and most well-known Python data visualization library is Matplotlib. It is not easy to use, but usually it provides many functions like bar chart, scatterplot, pie chart, histogram am etc which are useful for projecting many dimensions of data.

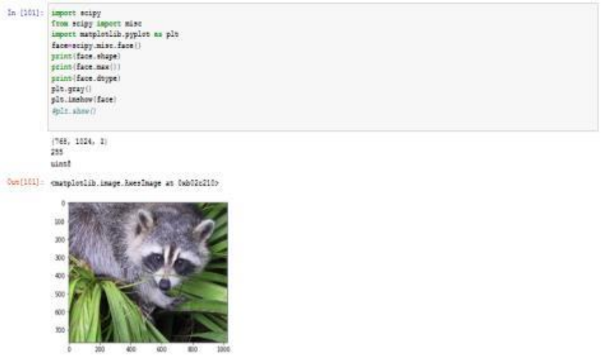


VISUALIZATION

```
In [14]: android = big_table[big_table.phone_type == 'android'].reset_index()
         ios = big_table[big_table.phone_type == 'ios'].reset_index()

In [15]: bins = np.linspace(0, 40, 40)
         x = android['free']
         y = ios['free']
         data = [x,y]
         plt.hist(data, bins, alpha = 0.5, color = ['g','r'])
         plt.show()
```

Scipy:

Mathematics deals with a huge number of concepts that are very important but at the same time, complex and time-consuming. However, Python provides the full-fledged scipy library that resolves this issue for us. In this scipy, we will be learning how to make use of this library along with a few functions and their examples.



```
In [101]: import scipy
          from scipy import misc
          import matplotlib.pyplot as plt
          face=scipy.misc.face()
          print(face.shape)
          print(face.max())
          print(face.dtype)
          plt.gray()
          plt.imshow(face)
          #plt.show()

          (768, 1024, 3)
          255
          uint8

Out[101]: <matplotlib.image.AxesImage at 0xb02c210>
```

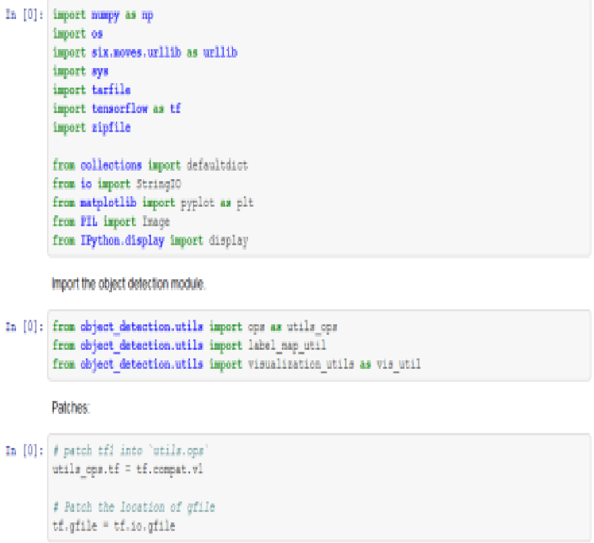
3.3 Python for Machine learning

The following libraries are general-purpose libraries for anything involving advanced data manipulation. This means they can all be used in implementing machine learning, and many of the higher level machine learning libraries makes use of some or all of these libraries. Getting acquainted with them is highly recommended if you plan on getting anywhere with scientific Python programming.

This list is by no means exhaustive; it is meant to be a starting point for you as you explore machine learning through Python! Already we discussed Numpy, Matplotlib and Scipy which are used for machine learning too. But we will see other modules which are used in machine learning.

Tensor flow:

Tensor flow is almost certainly the most well-known open source machine learning library available for Python, and for good reason. It was developed by Google, and is used in nearly every Google application that utilizes machine learning. If you've used Google Photos or voice search, then you've been using tensor flow. Tensorflow is extremely well documented and supported, and is optimised for speed. It is more difficult to learn, however, because it is actually a Python front-end coded on top of C or C++.



```
In [0]: import numpy as np
        import os
        import six.moves.urllib as urllib
        import sys
        import tarfile
        import tensorflow as tf
        import zipfile

        from collections import defaultdict
        from io import StringIO
        from matplotlib import pyplot as plt
        from PIL import Image
        from IPython.display import display
```

Import the object detection module.

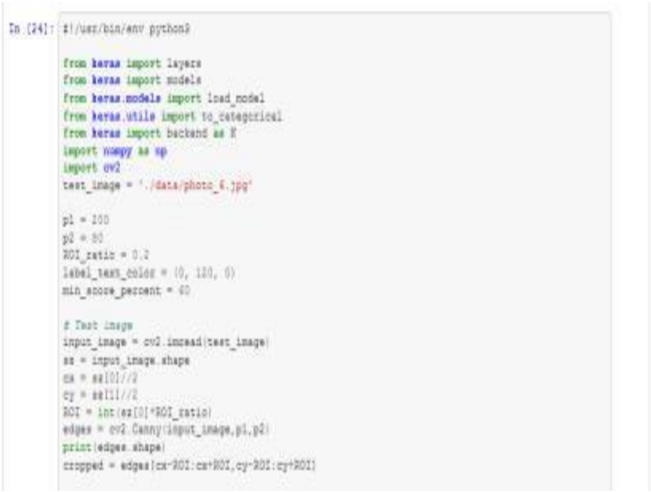
```
In [0]: from object_detection.utils import ops as utils_ops
        from object_detection.utils import label_map_util
        from object_detection.utils import visualization_utils as vis_util
```

Patches:

```
In [0]: # patch tf1 into `utils.ops`
        utils_ops.tf = tf.compat.v1

        # Patch the location of gfile
        tf.gfile = tf.io.gfile
```

Keras:

Built on top of Theano and Tensorflow is Keras, a high-level library for working with datasets. Keras is best known for being one of the easiest machine learning libraries out there because it is coded entirely in Python, while using either Theano or Tensorflow as a back- end. It is the most beginner- friendly library for machine learning, and includes functions for creating training datasets and more. Keras' neural networks API was developed for fast experimentation and is a good choice for any deep learning project that requires fast prototyping



```
In [24]: #!/usr/bin/env python3

         from keras import layers
         from keras import models
         from keras.models import load_model
         from keras.utils import to_categorical
         from keras import backend as K
         import numpy as np
         import cv2
         test_image = './data/photo_6.jpg'

         p1 = 200
         p2 = 50
         ROI_ratio = 0.2
         label_text_color = (0, 120, 0)
         min_score_percent = 60

         # Test image
         input_image = cv2.imread(test_image)
         ss = input_image.shape
         cx = ss[0]//2
         cy = ss[1]//2
         ROI = int(ss[0]*ROI_ratio)
         edges = cv2.Canny(input_image,p1,p2)
         print(edges.shape)
         cropped = edges[cx-ROI:cx+ROI,cy-ROI:cy+ROI]
```

CONCLUSION

In this paper we have presented usage of Python as a tool in various research areas like Data Science, Machine learning and IOT. Along with Python language, there are many other languages are used for Data Science, Machine learning and for developing iot devices like Java, C++ etc. But right now most of the developers use python scripting language than Java, C++. Because of its easy syntax, secure coding, and it's simplicity. When it comes to robust and performance, developers choose Python. Iot, when integrated with AI, will help developers to work with Python further.

With respect to the future work there is still huge space for this language to serve other upcoming research areas because of its features like simplicity, extensive library, inbuilt and extensible modules. In future we will propose python as a powerful tool which is used by many research communities.

ACNOWLEDGEMENT

We would like to thank Department of MCA-VTU, Dayananda Sagar college of Engineering for encouragement and support in writing this paper.

REFERENCES

1. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science 349(6245), 255–260 (2015)

2. Le Cun, Y., Bengio, Y., Hinton, G.: Deep Learning. Nature 521(7553), 436-444(2015).

3. Hilzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics – state-of-the-art, future challenges and research directions. BMC Bioinform. 15(s6), I1(2014).

4. Wolfram, S.: Mathematica: A System for Doing Mathematics by Computer. Addi-son Wesley Longman Publishing Co., Inc., Boston (1991)

5. Engblom, S., Lukarski, D.: Fast MATLAB compatible sparse assembly on multicore computers. Parallel Comput. 56, 1–17 (2016) https://www.researchgate.net/publication/330513589_Internet_of_Things_IOT_Using_Raspberry_Pi

6.Python Machine Learning: A Guide To Getting Started | Built In

7.https://www.researchgate.net/publication/330513589_Internet_of_Things_IOT_Using_Raspberry_Pi

8.https://www.techaheadcorp.com/blog/ top-6-programming-languages-for-iot- projects/
9.https://www.google.com/role-of- python-in-iot-development

10.Https://nbviewer.jupyter.org/github/ehmatthes/intro_programming/blob/master/notebooks/introducing_functions.i pynb

11. Https://data36.com/python-libraries-packages-data-scienTists/