LDSCI7236 Theory and Applications of Data Analytics

Activity: Lab Session 1: Program Design in Python

Professor (assoc.) Ioannis Kypraios

Python

- **Why using Python:**
  - **Easy to learn**
  - **Fast prototyping and code development and management**
  - **Interpreted language and Object Oriented:** With Python you can write code in the form of a script which can be loaded to run from a command line. There is no need to do first compilation of the code. Additionally, Python is Object Oriented programming language which makes easy to reuse code you have previously written and extend it, you can create objects, and APIs and Libraries.
  - **Open source libraries**: Python has an extended programming community and forums where you can find reusable code from previous projects, load into your program a new library or use functions from an API.

Python and Data Analytics

- **Variety and plethora of libraries for developing advanced programs:**
  - **NumPy**
  - **Scikit-learn**
  - **Matplotlib and Seaborn**
  - **Pandas**
  - **Pefile**
  - **Volatility**

Python and Data Analytics: NumPy

- **NumPy is one of the most important libraries for data science and AI. You can use NumPy and its functions and APIs to develop advanced ML algorithms.**
  - https://numpy.org/
- **NumPy multidimensional arrays**: One of the most important strengths of NumPy is it offers the creation and handling of multidimensional array objects, called ndarrays
  - A multidimensional array is a matrix but that has **more than 2 dimensions**

    i.e. [m x n]
  - ndarrays can be used for solving advanced linear algebra and matrix calculations as well as allow complex operations on image data.

Python and Data Analytics: NumPy

❑ TIP – Matrix operations with NumPy

- **A vector is similar to a matrix but it has only one row and several columns. Vectors are represented in Pythion as type list.**
- **The usual linear algebra rules apply for matrices and vectors in Python i.e. the allowed operations are:**
  - **Addition**
  - **Subtraction**
  - **Scalar multiplication**

  N/B: The following rules hold:

  1. **Required condition for addition and subtraction are the matrices to be of the same size.**
  2. **The result of the addition and subtraction of two matrices is another matrix whose elements are the result of the sum or subtraction of corresponding elements in row and column order.**
  3. **The product of two matrices or two vectors does not follow the commutative property.**
  4. numpy library **provides the** dot() **function to calculate the product of two matrices**

Python and Data Analytics: Scikit-learn

- scikit-learn library **consists of a series of models and algorithms in ML that can be re-used for developing advanced predictive applications such as:**
  - Classification
  - Regression
  - Dimensionality reduction
  - Clustering
  - Data preprocessing
  - Feature extraction
  - Hyperparameter optimization
  - Model evaluation

- https://scikit-learn.org/stable/

Python and Data Analytics: Matplotlib and Seaborn

- **Both libraries are used for data visualisation, and graphical representation and plotting. By graphical representation we can perform an exploratory data analysis (EDA) which assists us in the selection of more accurate predictive models.**
- matplotlib **is a data plotting tool.**
- **Seaborn is an extension of** matplotlib **and uses features of** scikit-learn **for providing with various visualisation tools.**

- https://matplotlib.org/
- https://seaborn.pydata.org

Python and Data Analytics: Pandas

- pandas package **enables the data cleaning for proceeding with the data analysis phase.**

  That step is very important for the correct analysis to be achieved when applying afterwards an ML algorithm on the data.

- **https://pandas.pydata.org/**

Some Other Python Libraries

- **Pefile library is essential for analysing Windows OS executable files during the phases of static malware analysis, looking for possible indications of compromise or the presence of malicious code in executables. Thus, Pefile allows the analysis of the Portable Executable (PE) file format which is the standard for the object files on the Microsoft platform.**
- **Pefile through the PE file analysis can also support the analysis of .exe files, .dll libraries and .sys device drivers.**
- **To install Pefile:**

pip install pefile

Some Other Python Libraries

- **Volatility: It is a tool used for malware analysis which allows the analysis of the runtime memory of an executable process. It identifies the presence in our system of a malware code.**
- **Volatility is a Python-programmable utility which can be found with different distributions of Linux implementations such as Kali Linux for malware analysis or penetration testing.**
- **Volatility allows the extraction from the memory dumps of all the processes running on the system as well as information about injected Dynamic-Link Libraries (DLLs) together with any hidden processes within the runtime memory (not typically detected by anti-virus software).**

For Labs…

Anaconda 1of4

- **Anaconda is a collection of over 700 packages developed in Python containing amongst other ML and data analysis libraries:**
  - **NumPy**
  - **SciPy**
  - **Scikit-learn**
  - **Pandas**
  - **Matplotlib**
- **https://www.anaconda.com/products/individual**
- **Recommended >= 3.7; Latest stable: 3.11**
- **Anaconda allows you to configure custom environments within which you can select the Python libraries you want to install.**

Anaconda 2of4

- **Conda utility is used for installing or updating existing packages and libraries.**
- To access the help menu:

```
conda -h
```

- To install a new package:

```
conda install
```

- To create and activate a custom environment where we want to install, e.g. version Python 3.7:

```
conda create –n py37 python=3.7
activate py37
```

Anaconda 3of4

❑ **Useful** conda **commands:**

- activate py37
- conda install –n py37 PACKAGE-NAME

e.g. conda install –n py37 seaborn

- conda list –n py37
- conda update conda

  conda update -all

Anaconda 4of4
Jupyter Notebooks

- **Jupyter Notebook is a development environment which allows in a single document the integration of both the Python code and the result of its execution such as images or graphics. It assists the developer with the debugging of his code and its testing.**
- **Jupyter Notebook is a web-based utility. It comes pre-installed with Anaconda. So, it is not necessary to install it separately since it comes readily available. To run it:**

```
jupyter notebook
```

For specifying the listening port e.g. 9000 of the service:

```
jupyter notebook  --port 9000
```

- Once Jupyter has started to open an existing notebook inside the root directory:

  http://localhost:8888/tree

Anaconda
Deep Learning 1of3

- **Deep Learning Libraries:**
  - **TensorFlow**
  - **Keras**
  - **PyTorch**
- **TensorFlow has been specifically developed to program deep neural networks (DNNs). To install and test with Anaconda:**

1. Install TensorFlow with conda:

```
conda install –n py37 –c conda-forge tensorflow
```

2. Install a specific version of TensorFlow by using the following command:

```
conda install –n py37 –c conda-forge tensorflow=2.0.0
```

3. Test the installation of TensorFlow by the following test lines:

```
activate py37
python
>>> import tensorflow as tf
>>> hello = tf.constant('Hello, TensorFlow!')
>>> sess = tf.Session()
>> print(sess.run(hello))
```

- https://www.tensorflow.org/

Anaconda
Deep Learning 2of3

- **Keras can be installed on top of TensorFlow as a high-level interface for NNs development.**
  - **To install:**

    conda install –n py37 –c conda-forge keras

  - https://keras.io/

Anaconda
Deep Learning 3of3

❑ **PyTorch has been developed by Facebook for performing large-scale image analysis.**

- ➢ **To install:**

conda install –n py37 –c peterjc123 pytorch

- ➢ https://pytorch.org/

❑ Common applications for using the PyTorch:

- Natural Language Processing (NLP)
- Large-scale image processing
- Social media analysis

LDSCI7236 Theory and Applications of Data Analytics

Activity: Lab Session 1: Program Design in Python

Professor (assoc.) Ioannis Kypraios