地址

下载地址

Python Tutorial

浏览器推荐Chrome，其他也可以。

背景

目前对于数据分析，Python为首选。除了内置的库外，Python还有大量的第三方库，也就是别人开发的，供你直接使用的东西。Python具有很多优势，比如，完成同一个任务，C语言要写1000行代码，Java只需要写100行，而Python可能只要20行。TIOBE，GIthub排名增长最快，因为方便，更新贡献的人也越来越多。Python是一个工具，会利用即可。它可以编写"间断而粗糙"的小程序（也就是脚本，scripting，一系列命令）。

Python是解释型编程语音，Java、C++是编译型语言。

1.由来

Python这个名字，来自Guido所挚爱的电视剧Monty Python’s Flying Circus。

2.应用领域

Python可以应用于众多领域，如：数据分析、组件集成、网络服务、图像处理、数值计算和科学计算等众多领域。目前业内几乎所有大中型互联网企业都在使用Python，如：Youtube、Dropbox、BT、Quora（中国知乎）、豆瓣、知乎、Google、Yahoo!、Facebook、NASA、百度、腾讯、汽车之家、美团等。

目的

了解Python，数据结构、语法规范、如何深入学习等。解决使用中遇到的一些环境配置问题。

如何学好?

多 练。

Python库

基础

NumPy（Numerical Python）是Python科学计算基础包，提供快速数组处理能力。

- 直接对数组进行数学运算。
- 线性代数运算、傅立叶变换、随机数生成。

Pandas（源于panel data）提供了快速便捷处理结构化数据的大量数据结构和函数。

- 处理表格型数据。
- 对表格各种操作，快速处理大量数据。

Matplotlib最流行的用于绘制表和其它二维数据可视化的Python库。

- 定制各种矢量图。

进阶

SciPy 是专门解决科学计算的库，包括数值积分，微分方程，矩阵分解，信号处理等。

Scikit-learn 目前为Python通用机器学习工具包。

- 包括分类
- 回归
- 聚类
- 降维
- 模型选用
- 预处理等

seaborn 除了Matplotlib，基于其开发的Seaborn在数据可视化方面功能也非常强大。相比于Matplotlib来说，Seaborn提供更高层次的API，可以让你在不需要了解那么多底层参数的情况下，同样能够画出比较有吸引力的图表。Examples Tutorial.

Statsmodels 是一个统计分析包，来源R语音分析的丰富性，包括回归模型、方差分析(ANOVA)、时间序列分析、非参数方法（核密度估计、核回归）、统计模型结果可视化。

深度学习库，Tensorflow、Pytorch、Kears等。

Install

1. Mac
   - 下载地址，下载Python 3.7 version for Mac
   - 安装，一直下一步即可。
   - 打开及检测

     进入Terminal，输入`jupyter lab`。 或者打开应用Anaconda Navigator后，点击jupyter lab。

     推荐前者，后者打开速度很慢。

2. Windows
   - 下载地址，下载Python 3.7 version for Windows
   - 安装
     - 在安装中，在安装路径一页，需**复制路径，以备后面粘贴 配置环境变量用**。
     - 在安装中，出现一个有两个选项页面，按照默认(第一个不勾选，第二个勾选)，最后一直下一步到安装完毕。

       ```
       □ Add Anaconda to my PATH environment variable
       ☑ Register Anaconda as my default Python 3.7
       ```

     - 安装完毕后，需要配置环境变量，打开 我的电脑 — 右键属性 — 高级系统设置 — 环境变量 — 选中Path变量一栏 — 点击编辑 — 新建，加入下面三行：

       ```
       粘贴内容\Anaconda3
       粘贴内容\Anaconda3\Scripts
       粘贴内容\Anaconda3\Library\bin
       ```

     - 进入cmd，需更新修复一个bug，按照下面步骤操作即可（Windows进入cmd方法：Windows+R，输入：cmd）：

       ```
       Windows 进入cmd，输入：

       conda update conda

       之后，输入需要确定是否更新，输入：y
       ```

   - 打开及检测

     进入cmd，输入`jupyter lab`。或者打开应用Anaconda Navigator后，点击jupyter lab。

推荐前者，后者打开速度很慢。

执行环境

执行环境有很多种，建议Anaconda — jupyter notebook（jupyter lab），可以消除入门的各种障碍。

- 2014年推出Jupyter notebook，一个支持多种语言的交互式网络代码"笔记本"，还支持Markdown和HTML内容。
- 2018年2月推出**Jupyter lab**。

也可以选用Pycharm，Sublime。

推荐用**Jupyter lab**。

Python包的升级

安装Anaconda中没有的Python包，可以在Terminal（Mac），cmd（Windows）输入：

```
conda install package_name
```

优先建议conda安装，会自动适配版本优化，如果conda中没有这个命令，可以用pip安装，输入：

```
pip install package_name
```

如果是很久之前就安装了，现在需要更新，输入：

```
conda update package_name
```

pip 可以用 --upgrade升级：

```
pip install --upgrade package_name
```

预备知识

1. 注释符号 `#`
2. 查询命令用法 `help（function_name）` 或 `function_name?`

3. `<Tab>`可以补全命令。
4. 和C，C++，JAVA一样，序列是从0开始的。

数据类型（Data Structure）

1. int float (数值) 、string （字符串）、tuple list （列表）、dictionary (字典)
2. Series、 DataFrame （Pandas）

实践

读取文件

函数	说明
read_csv	从文件、URL、文件型对象中加载带分隔符的数据。默认分隔符为逗号
read_excel	从Excel XLS 或 Xlsx file读取表格数据
read_hdf	读取pandas写的HDF5文件
read_json	读取JSON（JavaScript Object Notation）字符串中的数据
read_pickle	读取Python pickle 格式中存储的任意对象
read_sas	读取存储于SAS系统自定义存储格式的SAS数据集
read_sql	（使用SQLAlchemy）读取SQL查询结果
read_stata	读取Stata文件格式的数据集

```
data = pd.read_csv('file_path',sep=',')
```

参数：sep、header、encoding等。

数据清洗

处理缺失数据、数据整合、可视化统计、数据分组统计（类似透视表）group_by。

回顾

- 了解Python
- 数据结构
- 语法规范
- 如何深入学习
- 解决使用中遇到的一些环境配置问题(繁琐)。

Reference

1. **书籍 Python for Data Analysis网译书籍， 2nd codes。第一版 中文版， 密码:m4v2。**
2. Google's Python Class 里面含有video可以帮助了解数据的类型。
3. 视频教程 里面还有video可以帮助基本操作。来源官方文档The Python Tutorial。
4. 廖雪峰教程
5. Stata_Tutorial from Princeton.
6. jupyterlab 添加目录