python 数据处理 基础教程

作者：齐显东
联系：xqiad@connect.ust.hk

Python 必备

1. The Jupyter Notebook (交互式python+ 文档编辑（markwown） +画图 一体化)
2. Pandas （Python Data Analysis Library）
3. Numpy （计算 / 矩阵）
4. Scikit-Learn （机器学习库，一行解决 KMeans, Decision Tress）
5. Matplotlib （画图）
6. Seaborn （高端，高效画图）

工作必备：

1. Markdown 文档编辑。（10倍速度与latex / word，且整洁美观）
2. Windows 下命令行。
    1. Xshell
    2. Cygwin
    3. Cmder

Python 推荐：

1. 用 virtualenv 管理自己的python 环境。（就像容器，隔离各个python 环境）
2. 用 Anaconda 管理自己的python 环境.(可选)

Python 进阶：

1. 如何在服务器上（没有屏幕显示）使用 “交互的Jupyter Notebook”
    1. 教程-1
    2. 教程-2

学习建议：

1. pandas, numpy 等库随时用，随时查。
2. 在 kaggle 比赛网站学习, 非常好的“数据处理”例子，习的最新技巧。
    1. 优秀kaggle例子-1
    2. 优秀kaggle例子-2

The Jupyter Notebook (一次运行, 多次阅读,保存运行结果)

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code

网址：http://jupyter.org
初级教程： http://python.jobbole.com/87527/?repeat=w3tc
推荐教程： http://nbviewer.jupyter.org/github/phelps-sg/python-bigdata/blob/master/src/main/ipynb/pandas.ipynb （顺便学习 pandas 数据处理库）

Pandas （Python Data Analysis Library）

推荐教程：（十分钟） http://pandas.pydata.org/pandas-docs/stable/10min.html

Numpy

Numpy是Python的一个科学计算的库，提供了矩阵运算的功能，其一般与Scipy、matplotlib一起使用。其实，list已经提供了类似于矩阵的表示形式，不过numpy为我们提供了更多的函数。

推荐教程：https://zhuanlan.zhihu.com/p/24309547
一般教程：http://www.cnblogs.com/smallpi/p/4550361.html

（pandas， numpy 都有很多很棒的 readtext, loadcsv 的方法，自动解析“逗号，空格，tab”并根据 column 分成python-list，建议采用）

Scikit-Learn

Scikit-Learn是一个基于python的用于数据挖掘和数据分析的简单且有效的工具，它的基本功能主要被分为六个部分：分类(Classification)、回归(Regression)、聚类(Clustering)、数据降维(Dimensionality Reduction)、模型选择(Model Selection)、数据预处理(Preprocessing)，前面写的很多文章算法都是出自该扩展包。

Python机器学习库SKLearn包含的内容（目录）
链接：http://blog.csdn.net/cheng9981/article/details/61649552

python seaborn 绘图函数库

Seaborn其实是在matplotlib的基础上进行了更高级的API封装，从而使得作图更加容易，在大多数情况下使用seaborn就能做出很具有吸引力的图，而使用matplotlib能制作具有更多特色的图。

Seaborn推荐教程（建议收藏，随时查用）.

1. http://blog.csdn.net/suzyu12345/article/details/69029106