Python入门

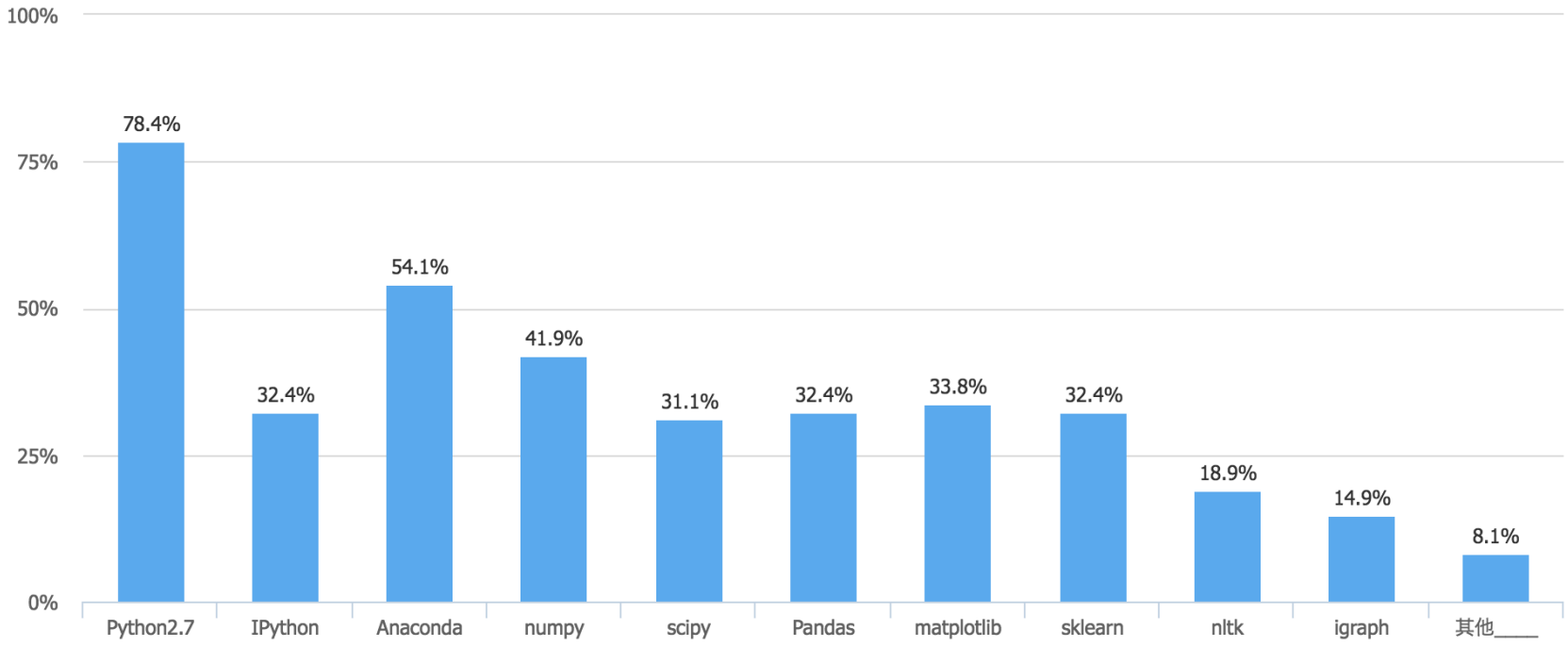
“数据分析是Python的杀手锏。”

七月在线 王博士

2016年8月27日

调查问卷

- Python环境安装调查（74份回收/383人）



- Python基础教程 http://www.runoob.com/python/python-tutorial.html

主要内容

- 安装Python与环境配置
- Anaconda安装和使用
- 常用数据分析库Numpy、Scipy、Pandas和matplotlib安装和简介
- 常用高级数据分析库nltk、igraph和scikit-learn介绍
- Python2和Python3区别简介

主要内容

- 安装Python与环境配置
- Anaconda安装和使用
- 常用数据分析库Numpy、Scipy、Pandas和matplotlib安装和简介
- 常用高级数据分析库nltk、igraph和scikit-learn介绍
- Python2和Python3区别简介

安装Python与环境配置

- 安装Python 2.7.12
  - https://www.python.org/downloads/
  - Mac OS X：自带python 2.7或者brew install python
- 环境变量配置
  - Windows（cmd输入）：path=%path%;C:\Python 或右键计算机->属性->高级系统设置->系统属性->环境变量->双击path ->添加“; C:\Python”安装路径
  - Linux：export PATH="$PATH:/usr/local/bin/python”

安装pip

- Pip已经在Python 2 >=2.7.9或Python 3 >=3.4中自带，但需要更新。 https://pip.pypa.io/en/stable/installing/
  - Linux或者OS X：
    - pip install -U pip
  - Windows（cmd输入）：
    - python -m pip install -U pip
- 安装大部分python库：
  - pip install <some software>
  - pip uninstall <some software>

主要内容

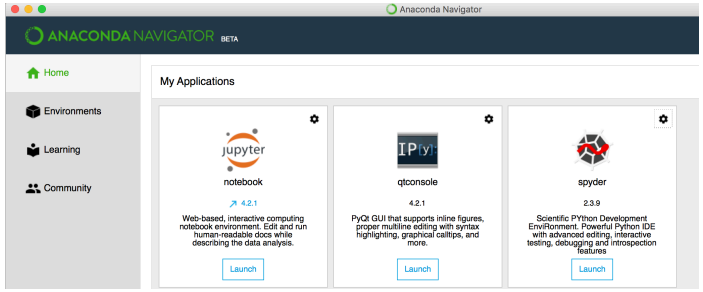
- 安装Python与环境配置
- Anaconda安装和使用
- 常用数据分析库Numpy、Scipy、Pandas和matplotlib安装和简介
- 常用高级数据分析库nltk、igraph和scikit-learn介绍
- Python2和Python3区别简介

Anaconda安装

- Anaconda
  - 下载：https://www.continuum.io/downloads
  - 命令行创建和启动环境：
    - conda create --name py27 python=2.7
    - activate py27
  - 列出安装packages：conda list
  - 安装package：conda install numpy（conda install会安装或更新库所依赖的各种库，pip install不会更新）

Anaconda使用

- Jupyter QtConsole or IPython（命令行式）
  - conda update ipython 或如需本地安装pip install ipython
- Jupyter Notebook（Web）：
  - 如需本地安装：pip install jupyter
- spyder（IDE，如需本地安装推荐Pycharm）



主要内容

- 安装Python与环境配置
- Anaconda安装和使用
- 常用数据分析库Numpy、Scipy、Pandas和matplotlib安装和简介
- 常用高级数据分析库nltk、igraph和scikit-learn介绍
- Python2和Python3区别简介

安装数据分析库

- 安装命令pip install/conda install
  - pip install numpy
  - pip install scipy
  - pip install pandas
  - pip install matplotlib

Numpy

- 提供常用的数值数组、矩阵等函数
- 优点：
  - 是基于向量化的运算
  - 进行数值运算时Numpy数组比list效率高

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

In [2]: np.arange(10)
Out[2]: array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])

In [3]: a = np.arange(10)
        a ** 2
Out[3]: array([ 0,  1,  4,  9, 16, 25, 36, 49, 64, 81])
```

Scipy

- 是一种使用NumPy来做高等数学、信号处理、优化、统计的扩展包 http://docs.scipy.org/doc/
  - Linear Algebra (scipy.linalg)
  - Statistics (scipy.stats)
  - Spatial data structures and algorithms (scipy.spatial)

```
In [1]: import numpy as np
        from scipy import linalg
        A = np.array([[1,2],[3,4]])
        A

Out[1]: array([[1, 2],
               [3, 4]])

In [2]: linalg.det(A)

Out[2]: -2.0
```

Pandas

- 是一种构建于Numpy的高级数据结构和精巧工具，快速简单的处理数据。
  - 支持自动或明确的数据对齐的带有标签轴的数据结构。
  - 整合的时间序列功能。
  - 以相同的数据结构来处理时间序列和非时间序列。
  - 支持传递元数据（坐标轴标签）的算术运算和缩减。
  - 灵活处理丢失数据。
  - 在常用的基于数据的数据库（例如基于SQL）中的合并和其它关系操作。
- 数据结构：Series和DataFrame

Pandas

In [1]:
```
import pandas as pd
import numpy as np
s = pd.Series([1,3,5,np.nan,6,8])
s
```

Out[1]:
```
0    1.0
1    3.0
2    5.0
3    NaN
4    6.0
5    8.0
dtype: float64
```

In [2]:
```
dates = pd.date_range('20130101',periods=6)
dates
```

Out[2]:
```
DatetimeIndex(['2013-01-01', '2013-01-02', '2013-01-03', '2013-01-04',
               '2013-01-05', '2013-01-06'],
              dtype='datetime64[ns]', freq='D')
```

Pandas

In [3]:
```
df = pd.DataFrame(np.random.randn(6,4),index=dates,columns=list('ABCD'))
df
```

Out[3]:

	A	B	C	D
2013-01-01	0.117033	1.618955	0.628805	-0.522834
2013-01-02	-2.611320	0.050968	-0.186566	-0.693047
2013-01-03	0.062438	1.363331	-1.235562	-0.054896
2013-01-04	2.296841	-0.005361	0.803157	-0.384893
2013-01-05	-0.361252	-0.699216	0.194242	-0.931266
2013-01-06	0.894726	0.751127	0.685620	0.346222

In [4]:
```
# df.head()
# df.tail()
# df.describe()
# df.T
df.sort_values(by='B')
```

Out[4]:

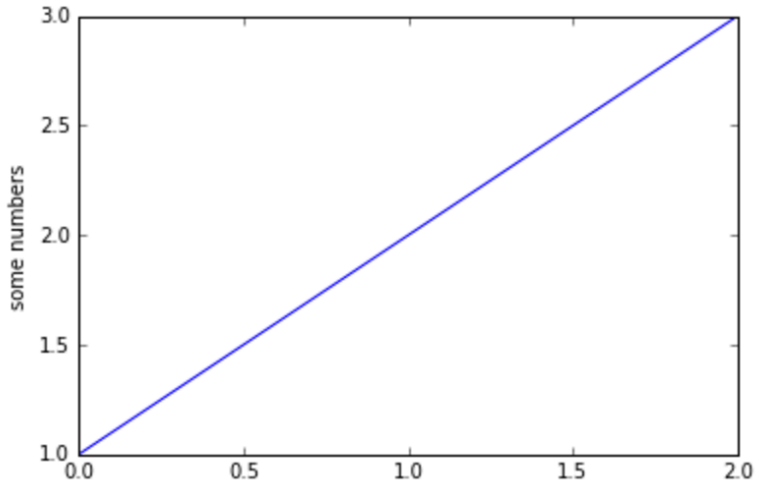
	A	B	C	D
2013-01-02	1.478557	-1.702842	-1.129036	0.209293
2013-01-01	-1.728759	-1.271772	-1.199566	-0.358619
2013-01-03	-0.626600	-0.643887	-1.009176	1.704757
2013-01-04	-0.645785	0.008506	-0.405815	-0.093509
2013-01-05	-0.075609	1.308982	0.050332	1.949689
2013-01-06	-1.014709	1.739562	1.048480	-1.149290

matplotlib

Python绘图库

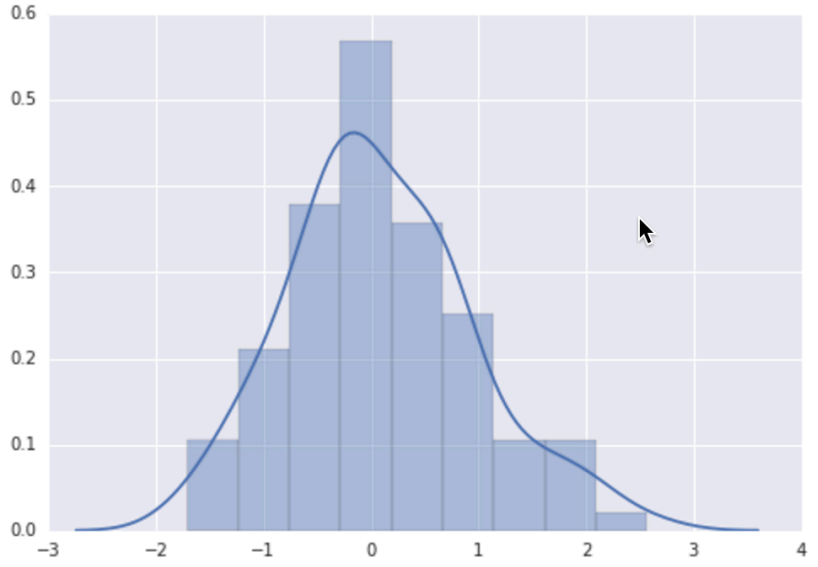
In [1]:
```
%matplotlib inline
```

In [2]:
```
import matplotlib.pyplot as plt
plt.plot([1,2,3])
plt.ylabel('some numbers')
plt.show()
```



In [3]:
```
import seaborn as sns
sns.set(color_codes=True)
```

In [5]:
```
import numpy as np
x = np.random.normal(size=100)
sns.distplot(x);
```



主要内容

- 安装Python与环境配置
- Anaconda安装和使用
- 常用数据分析库Numpy、Scipy、Pandas和matplotlib安装和简介
- 常用高级数据分析库nltk、igraph和scikit-learn介绍
- Python2和Python3区别简介

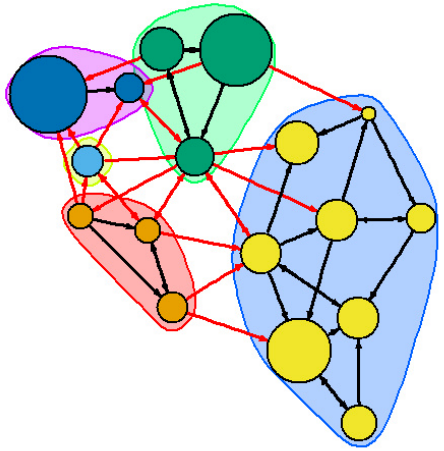
nltk

- 自然语言处理工具包（Natural Language Toolkit）
  - 安装：pip install -U nltk
  - 引入：import nltk
  - 下载预料库：nltk.download()
- 应用：
  - 文本提取
  - 词汇切分
  - 词频分析
  - 词袋模型
  - 情感分析

igraph

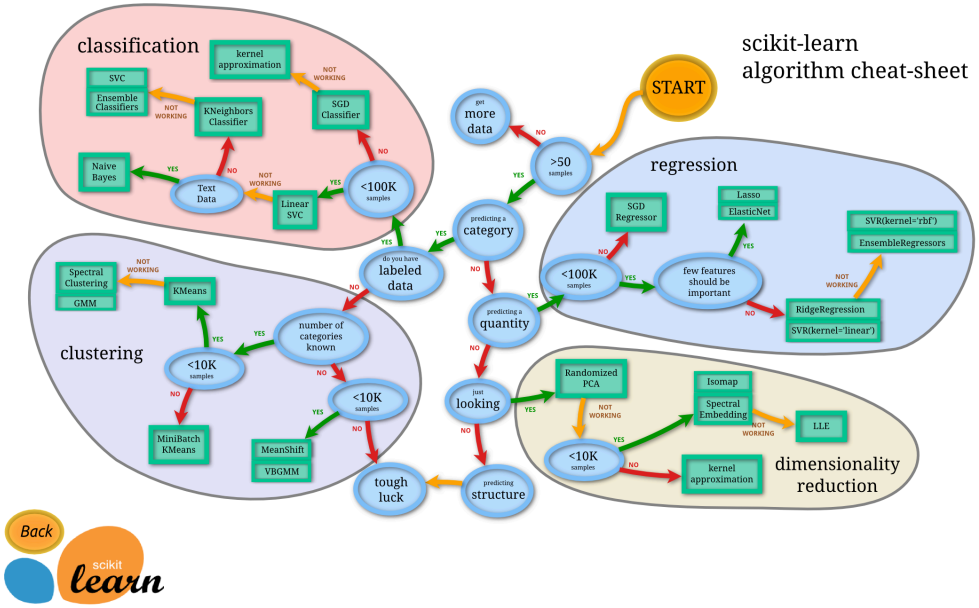
- 图计算和社交网络分析 http://igraph.org/python/
- 安装：
  - pip install -U python-igraph
  - conda install -c marufr python-igraph=0.7.1.post6

```
In [1]: from igraph import *

In [2]: g = Graph([(0,1), (0,2), (2,3), (3,4), (4,2), (2,5), (5,0), (6,3), (5,6)])
        g
Out[2]: <igraph.Graph at 0x103e50810>

In [3]: summary(g)
        IGRAPH U--- 7 9 --

In [4]: g.degree()
Out[4]: [3, 1, 4, 3, 2, 3, 2]
```



Scikit-learn

- Scikit-learn是建立在Scipy之上的一个用于机器学习的Python模块。
  - 安装：pip install -U scikit-learn / conda install scikit-learn



主要内容

- 安装Python与环境配置
- Anaconda安装和使用
- 常用数据分析库Numpy、Scipy、Pandas和matplotlib安装和简介
- 常用高级数据分析库nltk、igraph和scikit-learn介绍
- Python2和Python3区别简介

Python2和Python3区别简介

- print()函数
- 整除：3/2=1.5(python3);3/2=1(python2)
- 支持Unicode (utf-8) 字符串
- xrange()函数被集成在range()函数中
- raising exceptions：raise IOError("file error")
- Handling exceptions：except NameError as err
- 取消.next()
- For 循环变量和全局命名空间泄漏: [... for var in (item1, item2, ...)]
- 比较不可排序内容抛出错误
- 通过input()解析用户输入：把用户的输入存储为一个 str 对象
- 返回可迭代对象，而不是列表
- 使用 __future__ 模块支持python3：division、print_function

```
In [1]: 3/2
Out[1]: 1

In [2]: from __future__ import division

In [3]: 3/2
Out[3]: 1.5
```

感谢大家！

恳请大家批评指正！