
python
import turtle
turtle.setup(650,350,200,200)
turtle.penup()
turtle.fd(-250)
turtle.pendown()
turtle.pensize(25)
turtle.pencolor("purple")
turtle.seth(-40)
for 1 in range(4):
turtle.circle(40, 80)
turtle.circle(-40, 80)
turtle.fd(40)
turtle.circle(40, 80/2)
turtle.fd(40)
turtle.circle(16, 180)
turtle.fd(40 * 2/3)

Python语言程序设计

从数据处理到人工智能

嵩 天
北京理工大学

单元开篇

从数据处理到人工智能

数据表示->数据清洗->数据统计->数据可视化->数据挖掘->人工智能

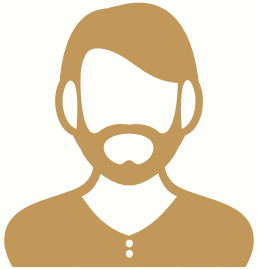
- **数据表示：采用合适方式用程序表达数据**
- **数据清理：数据归一化、数据转换、异常值处理**
- **数据统计：数据的概要理解，数量、分布、中位数等**

从数据处理到人工智能

数据表示->数据清洗->数据统计->数据可视化->数据挖掘->人工智能

- **数据可视化：直观展示数据内涵的方式**
- **数据挖掘：从数据分析获得知识，产生数据外的价值**
- **人工智能：数据/语言/图像/视觉等方面深度分析与决策**

从数据处理到人工智能



- Python库之数据分析
- Python库之数据可视化
- Python库之文本处理
- Python库之机器学习

python
语言程序设计

Python库之数据分析

Python库之数据分析

Numpy: 表达N维数组的最基础库

- Python接口使用，C语言实现，计算速度优异
- Python数据分析及科学计算的基础库，支撑Pandas等
- 提供直接的矩阵运算、广播函数、线性代数等功能

Python库之数据分析

Numpy: 表达N维数组的最基础库

```python
def pySum():
    a = [0, 1, 2, 3, 4]
    b = [9, 8, 7, 6, 5]
    c = []

    for i in range(len(a)):
        c.append(a[i]**2 + b[i]**3)

    return c

print(pySum())
```

```python
import numpy as np

def npSum():
    a = np.array([0, 1, 2, 3, 4])
    b = np.array([9, 8, 7, 6, 5])

    c = a**2 + b**3

    return c

print(npSum())
```

http://www.numpy.org

Python库之数据分析

Pandas: Python数据分析高层次应用库

- 提供了简单易用的数据结构和数据分析工具
- 理解数据类型与索引的关系，操作索引即操作数据
- Python最主要的数据分析功能库，基于Numpy开发

Python库之数据分析

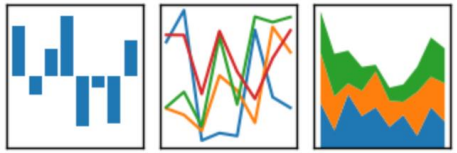
Pandas: Python数据分析高层次应用库

Series = 索引 + 一维数据

DataFrame = 行列索引 + 二维数据

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



http://pandas.pydata.org

Python库之数据分析

SciPy: 数学、科学和工程计算功能库

- **提供了一批数学算法及工程数据运算功能**
- **类似Matlab，可用于如傅里叶变换、信号处理等应用**
- **Python最主要的科学计算功能库，基于Numpy开发**

Python库之数据分析

SciPy: 数学、科学和工程相关功能库

http://www.scipy.org

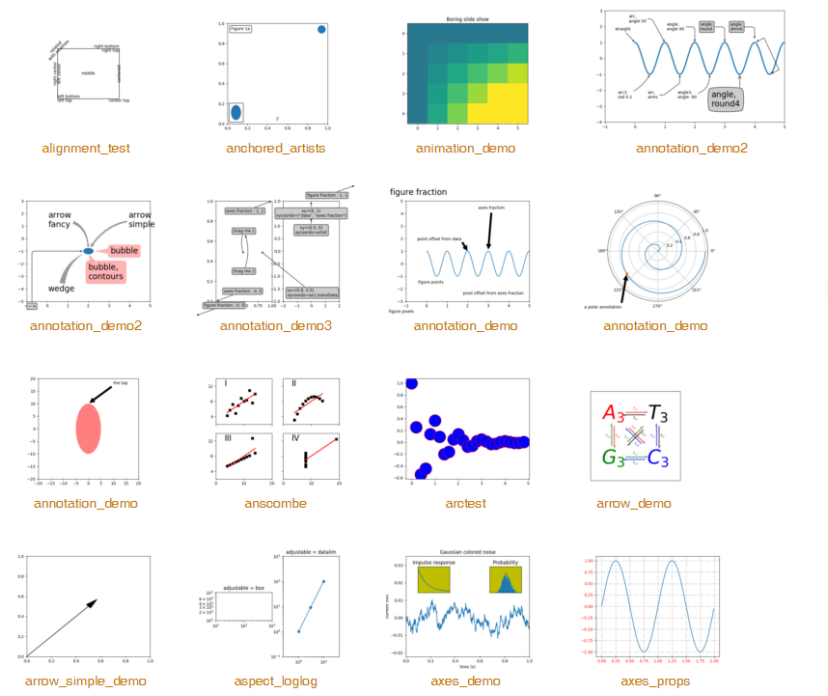
Python库之数据可视化

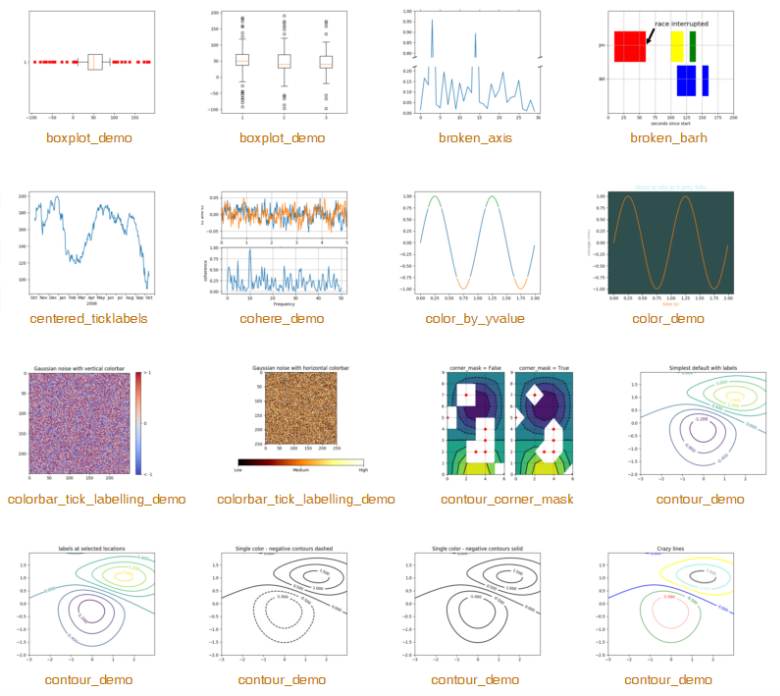
Python库之数据可视化

Matplotlib: 高质量的二维数据可视化功能库

- **提供了超过100种数据可视化展示效果**
- **通过matplotlib.pyplot子库调用各可视化效果**
- **Python最主要的数据可视化功能库，基于Numpy开发**

Python库之数据可视化





http://matplotlib.org

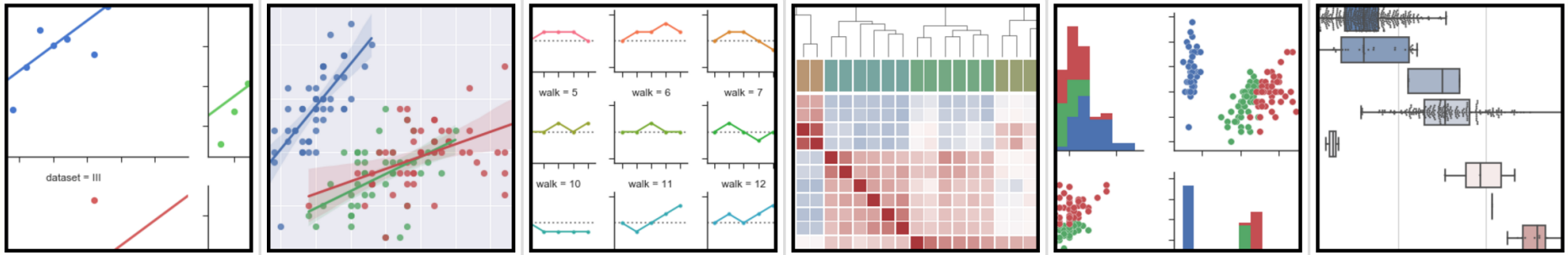
Python库之数据可视化

Seaborn: 统计类数据可视化功能库

- 提供了一批高层次的统计类数据可视化展示效果
- 主要展示数据间分布、分类和线性关系等内容
- 基于Matplotlib开发，支持Numpy和Pandas

Python库之数据可视化

Seaborn: 统计类数据可视化功能库



http://seaborn.pydata.org/

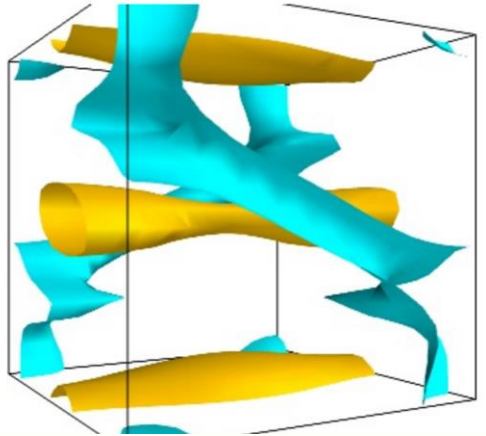
Python之数据可视化

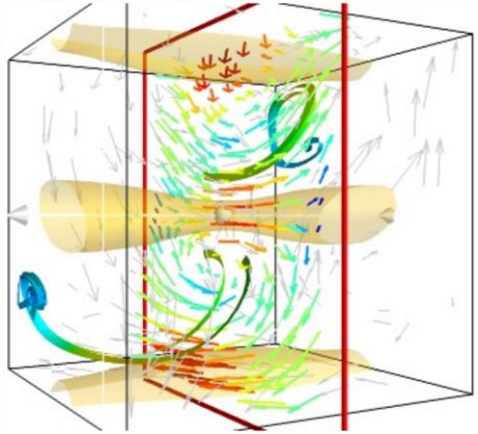
Mayavi：三维科学数据可视化功能库

- 提供了一批简单易用的3D科学计算数据可视化展示效果
- 目前版本是Mayavi2，三维可视化最主要的第三方库
- 支持Numpy、TVTK、Traits、Envisage等第三方库

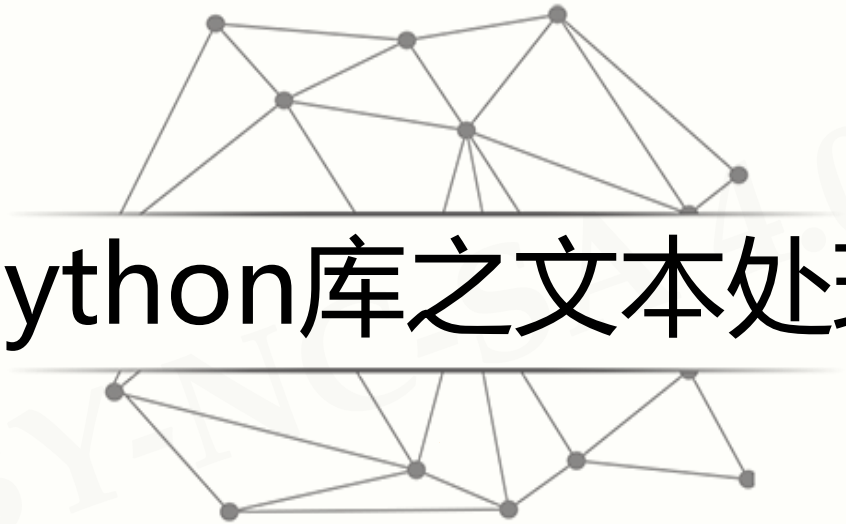
Python之数据可视化

Mayavi：三维科学数据可视化功能库





http://docs.enthought.com/mayavi/mayavi/



Python库之文本处理

Python之文本处理

PyPDF2：用来处理pdf文件的工具集

- 提供了一批处理PDF文件的计算功能
- 支持获取信息、分隔/整合文件、加密解密等
- 完全Python语言实现，不需要额外依赖，功能稳定

Python之文本处理

PyPDF2：用来处理pdf文件的工具集

```
from PyPDF2 import PdfFileReader, PdfFileMerger
merger = PdfFileMerger()
input1 = open("document1.pdf", "rb")
input2 = open("document2.pdf", "rb")
merger.append(fileobj = input1, pages = (0,3))
merger.merge(position = 2, fileobj = input2, pages = (0,1))
output = open("document-output.pdf", "wb")
merger.write(output)
```

http://mstamy2.github.io/PyPDF2

Python之文本处理

NLTK：自然语言文本处理第三方库

- 提供了一批简单易用的自然语言文本处理功能
- 支持语言文本分类、标记、语法句法、语义分析等
- 最优秀的Python自然语言处理库

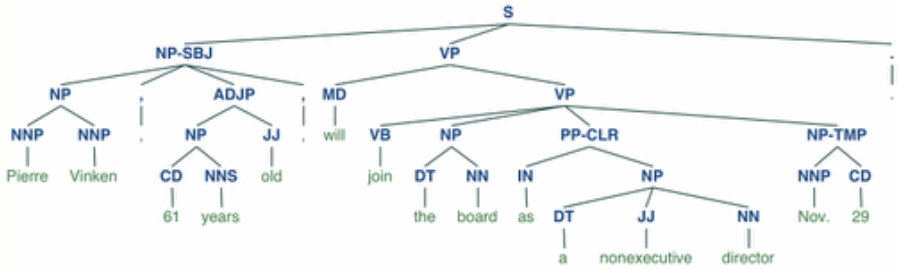
Python之文本处理

NLTK：自然语言文本处理第三方库

```
from nltk.corpus import treebank

t = treebank.parsed_sents('wsj_0001.mrg')[0]

t.draw()
```



http://www.nltk.org/

Python之文本处理

Python-docx：创建或更新Microsoft Word文件的第三方库

- **提供创建或更新.doc .docx等文件的计算功能**
- **增加并配置段落、图片、表格、文字等，功能全面**

Python之文本处理

Python-docx：创建或更新Microsoft Word文件的第三方库

```
from docx import Document

document = Document()

document.add_heading('Document Title', 0)

p = document.add_paragraph('A plain paragraph having some ')

document.add_page_break()

document.save('demo.docx')
```

http://python-docx.readthedocs.io/en/latest/index.html

Python库之机器学习

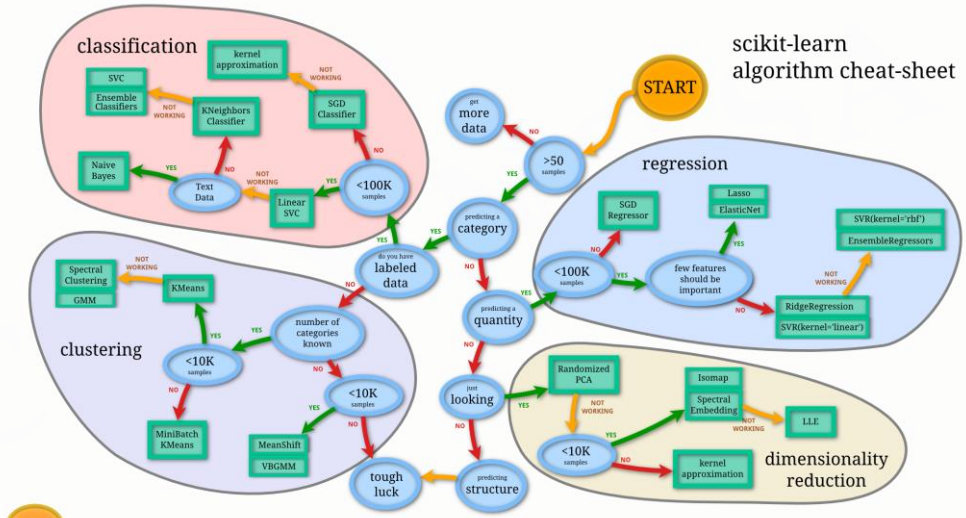
Python之机器学习

Scikit-learn：机器学习方法工具集

- 提供一批统一化的机器学习方法功能接口
- 提供聚类、分类、回归、强化学习等计算功能
- 机器学习最基本且最优秀的Python第三方库

Python之机器学习

Scikit-learn：与数据处理相关的第三方库



http://scikit-learn.org/

Python之机器学习

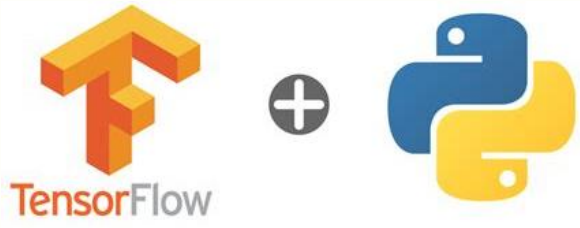
TensorFlow：AlphaGo背后的机器学习计算框架

- 谷歌公司推动的开源机器学习框架
- 将数据流图作为基础，图节点代表运算，边代表张量
- 应用机器学习方法的一种方式，支撑谷歌人工智能应用

Python之机器学习

TensorFlow：AlphaGo背后的机器学习计算框架

```
import tensorflow as tf
init = tf.global_variables_initializer()
sess = tf.Session()
sess.run(init)
res = sess.run(result)
print('result:', res)
```



https://www.tensorflow.org/

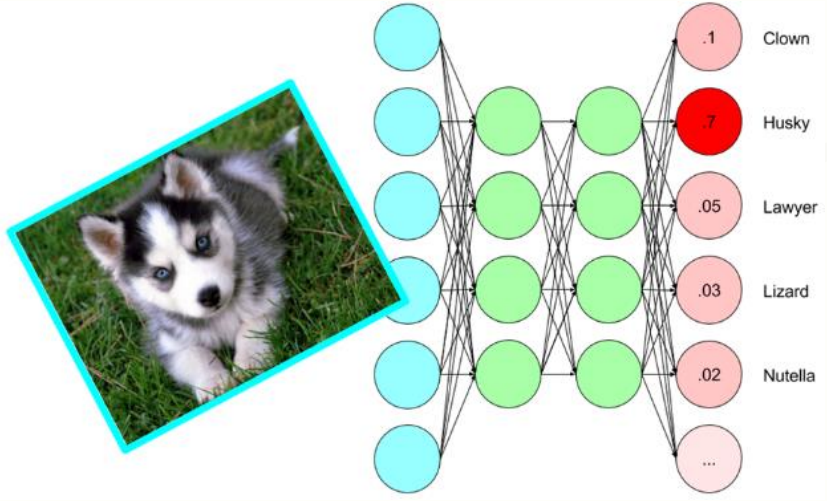
Python之机器学习

MXNet：基于神经网络的深度学习计算框架

- 提供可扩展的神经网络及深度学习计算功能
- 可用于自动驾驶、机器翻译、语音识别等众多领域
- Python最重要的深度学习计算框架

Python之机器学习

MXNet：基于神经网络的深度学习计算框架



mxnet

https://mxnet.incubator.apache.org/

单元小结

从数据处理到人工智能

- Numpy、Pandas、SciPy
- Matplotlib、Seaborn、Mayavi
- PyPDF2、NLTK、python-docx
- Scikit-learn、TensorFlow、MXNet