掌握了这24个顶级Python库，你就是大神！

读芯术
发布时间：08-01 12:30 | 优质原创作者

全文共**11815**字，预计学习时长**24**分钟

Python有以下三个特点：

· 易用性和灵活性

· 全行业高接受度：Python无疑是业界最流行的数据科学语言

· 用于数据科学的Python库的数量优势

事实上，由于Python库种类很多，要跟上其发展速度非常困难。因此，本文介绍了24种涵盖端到端数据科学生命周期的Python库。

文中提及了用于数据清理、数据操作、可视化、构建模型甚至模型部署(以及其他用途)的库。这是一个相当全面的列表，有助于你使用Python开启数据科学之旅。

用于不同数据科学任务的Python库

用于数据收集的Python库

· Beautiful Soup

· Scrapy

作者最新文章

AI医生的可信度有多高？这取决于用户对机器的态度……

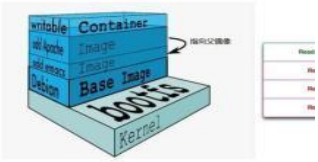
悄悄告诉你，未来的网站开发前景其实取决于物联网！

今日芯声 | 自己认输？美媒：美国将输掉对华科技战

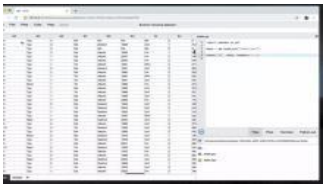
相关文章

21个必知的机器学习开源工具，涵盖5大领域

写给码农和运维的docker基础知识



再见Excel！我开源了一款与Python深度集成的神器级IDE



2019 必知的 10 大顶级 python

用于数据清理和数据操作的Python库

· Pandas

· PyOD

· NumPy

· Spacy

用于数据可视化的Python库

· Matplotlib

· Seaborn

· Bokeh

用于建模的Python库

· Scikit-learn

· TensorFlow

· PyTorch

用于模型解释的Python库

· Lime

· H2O

用于语音处理的Python库

· Librosa

· Madmom

· pyAudioAnalysis

用于图像处理的Python库

· OpenCV-Python

· Scikit-image

· Pillow

作为数据库的Python库

· Psycopg

· SQLAlchemy

用于模型部署的Python库

干货：如何正确地学习数据科学中的 python

用于数据收集的Python库

你是否曾遇到过这样的情况：缺少解决问题的数据？这是数据科学中一个永恒的问题。这也是为什么学习提取和收集数据对数据科学家来说是一项非常重要的技能。数据提取和收集开辟了前所未有的道路。

以下是三个用于提取和收集数据的Python库：

Beautiful Soup

收集数据的最佳方式之一就是抓取网站（当然是以合乎道德和法律的手段！）徒手做这件事需要耗费大量的劳动和时间。Beautiful Soup无疑是一大救星。

Beautiful Soup是一个HTML和XML解析器，可为被解析的页面创建解析树，从而用于从web页面中提取数据。从网页中提取数据的过程称为网页抓取。

使用以下代码可安装BeautifulSoup：

```
pip install beautifulsoup4
```

下面是一个可实现从HTML中提取所有锚标记的Beautiful Soup简单代码：

```
#!/usr/bin/python3

# Anchor extraction from html document

from bs4 import BeautifulSoup

from urllib.request import urlopen

with urlopen('LINK') as response:

soup = BeautifulSoup(response, 'html.parser')

for anchor in soup.find_all('a'):

print(anchor.get('href', '/'))
```

建议阅读下面的文章，学习如何在Python中使用Beautiful Soup：

《新手指南：在Python中使用BeautifulSoup进行网页抓取》

Scrapy

Scrapy是另一个可有效用于网页抓取的Python库。它是一个开源的协作框架，用于从网站中提取所需数据。使用起来快捷简单。

下面是用于安装Scrapy的代码：

```
pip install scrapy
```

Scrapy是一个用于大规模网页抓取的框架。可提供所有需要的工具有效地从网站中抓取数据，且依需要处理数据，并以使用者偏好的结构和格式存储数据。

下面是一个实现Scrapy的简单代码：

```
import scrapy

class Spider(scrapy.Spider):

name = 'NAME'

start_urls = ['LINK']

def parse(self, response):

for title in response.css('.post-header>h2'):

yield {'title': title.css('a ::text').get()}

for next_page in response.css('a.next-posts-link'):

yield response.follow(next_page, self.parse
```

下面是一个学习Scrapy并在Python中实现Scrapy的绝佳教程：

《使用Scrapy在Python中进行网页抓取（含多个示例）》

Selenium

Selenium是一个倍受欢迎的自动化浏览器工具。在业界常用于测试，但对于网页抓取也非常方便。Selenium在IT领域非常流行。

编写Python脚本来自动化使用Selenium的web浏览器是很容易的。它允许免费高效地提取数据，并将其存储在首选格式中以备后用。

关于使用Python和Selenium抓取YouTube视频数据的文章：

用于数据清理和数据操作的Python库

收集了数据之后，接下来要清理可能面临的任何混乱数据，并学习如何操作数据，方便数据为建模做好准备。

下面是四个可用于数据清理和数据操作的Python库。请记住，文中仅指明在现实世界中处理结构化（数值）数据和文本数据（非结构化）——而该库列表涵盖了所有内容。

Pandas

在数据操作和数据分析方面，Pandas绝无敌手。Pandas一度是最流行的Python库。Pandas是用Python语言编写的，主要用于数据操作和数据分析。

这个名称来源于术语“面板数据”，“面板数据”是一个计量经济学术语，指的是包含同一个人在多个时间段内的观察结果的数据集。

Pandas在Python or Anaconda中已完成预安装，但以防需要，安装代码如下：

pip install pandas

Pandas有以下特点：

· 数据集连接和合并

· 删除和插入数据结构列

· 数据过滤

· 重塑数据集

· 使用DataFrame对象来操作数据等

下面是一篇文章以及一份很棒的Cheatsheet，有助于使Pandas技能达标：

《Python中用于数据操作的12种有用的Pandas技术》

《CheatSheet:在Python中使用Pandas进行数据探索》

PyOD

难以发现异常值？这绝非个例。别担心，PyOD库就在这里。

以下代码可用于下载pyOD：

pip install pyod

PyOD是如何工作的？如何实现PyOD？下面一则指南将回答所有关于PyOD的问题：

《学习在Python中使用PyOD库检测异常值的绝佳教程》

NumPy

与Pandas一样，NumPy也是一个非常受欢迎的Python库。NumPy引入了支持大型多维数组和矩阵的函数，同时还引入了高级数学函数来处理这些数组和矩阵。

NumPy是一个开源库，有多方贡献者。在 Anaconda和Python中已预安装Numpy，但以防需要，下面是安装代码：

$ pip install numpy

下面是使用NumPy可执行的一些基本功能：

创建数组

import numpy as np

x = np.array([1, 2, 3])

print(x)

y = np.arange(10)

print(y)

output - [1 2 3]

[0 1 2 3 4 5 6 7 8 9]

基本运算

a = np.array([1, 2, 3, 6])

b = np.linspace(0, 2, 4)

c = a - b

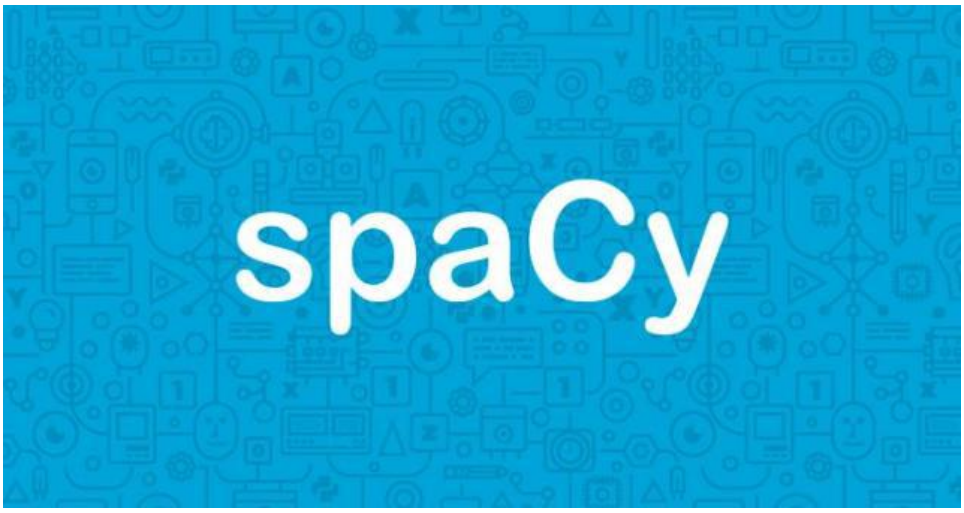
print(a**2)

output - [1. 1.33333333 1.66666667 4.]

[1 4 9 36]

以及更多其他功能！

SpaCy



目前已经讨论了如何清理数据和处理数值数据。但是如果正在处理文本数据呢?到目前为止，现有的库都无法解决该问题。

Spacy是一个非常有用且灵活的自然语言处理(NLP)库和框架，用于清理创建模型的文本文档。与类似用途的其他库相比，SpaCy速度更快。

在Linux中安装Spacy：

pip install -U spacy

python -m spacy download en

以下是学习spaCy的课程：

《简化自然语言处理——使用SpaCy（在Python中）》

用于数据可视化的Python库

下一步是什么呢？数据可视化！此处假设已得到验证，并且发掘了隐藏的观点和模式。

下面是三个用于数据可视化的绝佳Python库。

Matplotlib

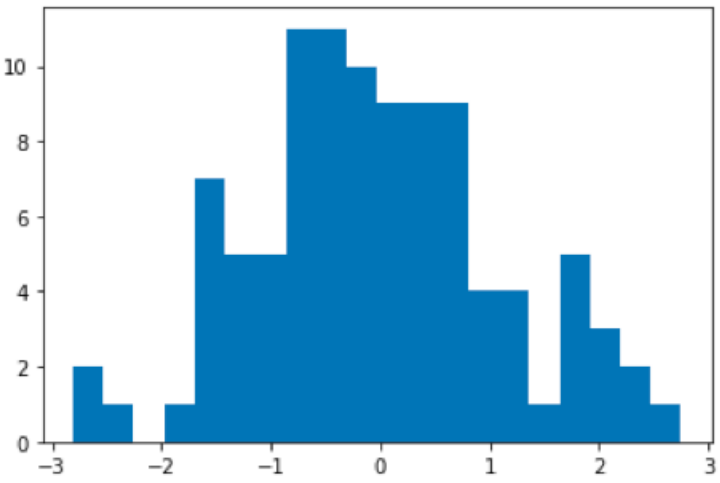
以下是安装Matplotli的代码：

$ pip install matplotlib

下面是使用Matplotlib构建的不同类型图示的部分例子：

柱状图

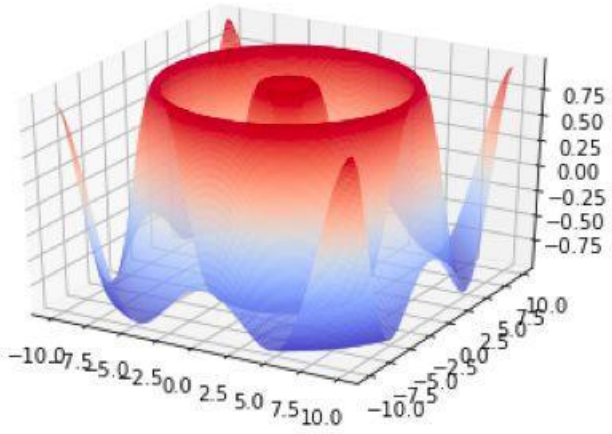
```
%matplotlib inline

import matplotlib.pyplot as plt

from numpy.random import normal

x = normal(size=100)

plt.hist(x, bins=20)

plt.show()
```



3D 图表

```
from matplotlib import cm

from mpl_toolkits.mplot3d import Axes3D

import matplotlib.pyplot as plt

import numpy as np
```

```
ax = fig.gca(projection='3d')

X = np.arange(-10, 10, 0.1)

Y = np.arange(-10, 10, 0.1)

X, Y = np.meshgrid(X, Y)

R = np.sqrt(X**2 + Y**2)

Z = np.sin(R)

surf = ax.plot_surface(X, Y, Z, rstride=1,cstride=1, cmap=cm.coolwarm)

plt.show()
```



目前已经介绍了Pandas、NumPy和Matplotlib，那么请查看下面的教程，该教程结合了以上三个库进行讲解：

《使用NumPy、Matplotlib和Pandas在Python中进行数据探索的终极指南》

Seaborn

Seaborn是另一个基于matplotlib的绘图库。它是一个为绘制有吸引力的图像而提供高级接口的python库。matplotlib能实现功能，Seaborn只是以另一种更吸引人的视觉方式来实现。

Seaborn 的一些特点：

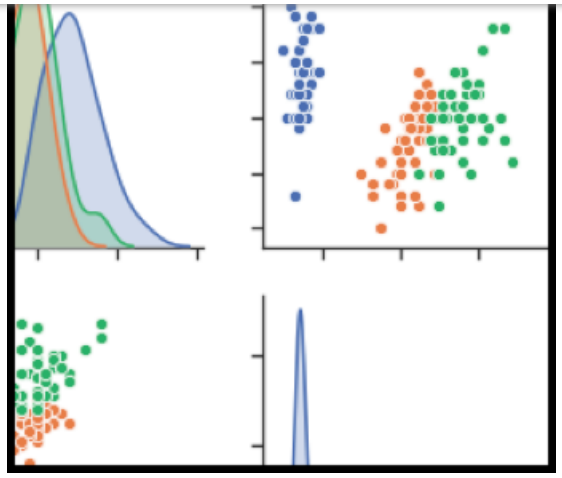
· 作为一个面向数据集的API，可用于查验多个变量之间的关系

· 便于查看复杂数据集的整体结构

· 用于选择显示数据中模式的调色板的工具

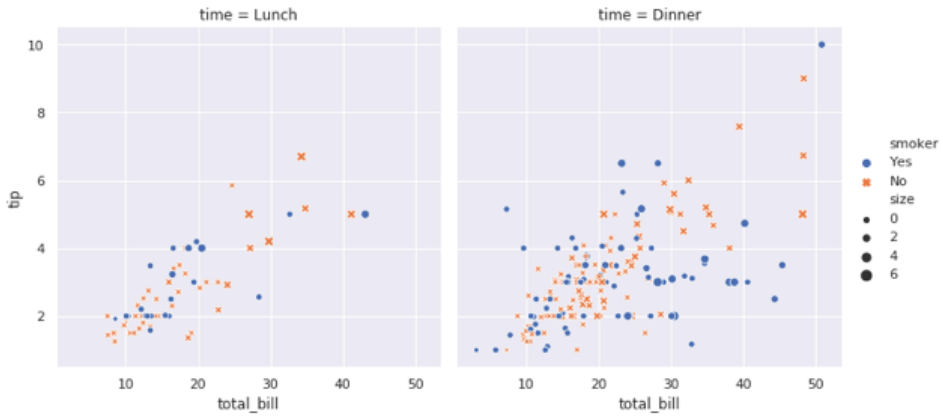
下面一行代码可用于安装Seaborn：

```
pip install seaborn
```



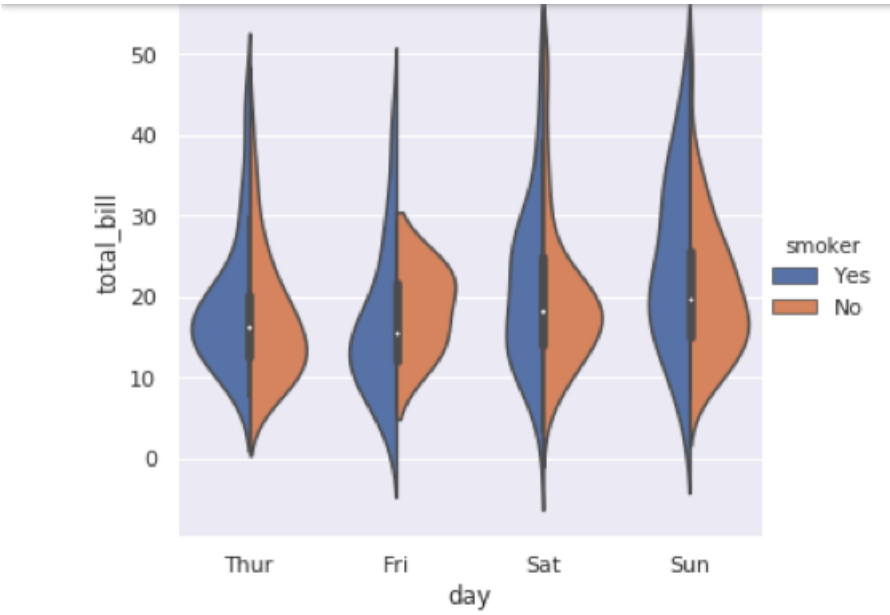
浏览下面这些很酷的图表，看看seaborn能做些什么：

```
import seaborn as sns

sns.set()

tips =sns.load_dataset("tips")

sns.relplot(x="total_bill",y="tip", col="time",

hue="smoker",style="smoker", size="size",

data=tips);
```



下面是另外一个例子：

```
import seaborn as sns

sns.catplot(x="day",y="total_bill", hue="smoker",

kind="violin",split=True, data=tips);
```



Bokeh

Bokeh是一个面向现代网页浏览器的交互式可视化库，为大量数据集提供优美的通用图形结构。

Bokeh可用于创建交互式绘图、仪表板和数据应用程序。

安装：

pip install bokeh

了解更多关于Bokeh的知识及其实际应用：

《使用Bokeh的交互式数据可视化（在Python中）》

用于建模的Python库

现在到了本文最令人期待的部分——建模！这也是大多数人一开始接触数据科学的原因。

Scikit-learn

就像用于数据操作的Pandas和用于可视化的matplotlib一样，scikit-learn是Python构建模型中的佼佼者。没有什么能与之媲美。

事实上，scikit-learn建立在NumPy，SciPy和matplotlib之上。它是开放源码的，每个人都可以访问，并且可以在各种环境中重用。

Scikit-learn支持在机器学习中执行的不同操作，如分类、回归、聚类和模型选择等。命名它——那么scikit-learn会有一个模块。

建议浏览以下链接以了解有关scikit-learn的更多信息：

《Python中的Scikit-learn——笔者去年学到的最重要的机器学习工具！》

TensorFlow

TensorFlow由谷歌开发，是一个流行的深度学习库，可帮助构建、培训不同模型。是一个开放源码的端到端平台。TensorFlow提供简单的模型构建，强大的机器学习生产，以及强大的实验工具和库。

TensorFlow提供多个抽象级别，可根据需要进行选择。TensorFlow通过使用高级Keras API来构建和训练模型，这使TensorFlow入门和机器学习变得容易。

使用TensorFlow从阅读这些文章开始：

《TensorFlow 101:理解张量和图像以便开始深入学习》

《开始使用Keras和TensorFlow在R中进行深度学习》

PyTorch

什么是PyTorch？其实，这是一个基于Python的科学计算包，其功能如下：

· NumPy的替代品，可使用GPU的强大功能

· 深度学习研究型平台，拥有最大灵活性和最快速度

PyTorch提供以下功能：

· 混合前端

· 工具和库：由研发人员组成的活跃社区已经建立了一个丰富的工具和库的生态系统，用于扩展PyTorch并支持计算机视觉和强化学习等领域的开发

· 云支持：PyTorch支持在主要的云平台上运行，通过预构建的映像、对GPU的大规模训练、以及在生产规模环境中运行模型的能力等，可提供无摩擦的开发和轻松拓展

以下是两篇有关PyTorch的十分详细且易于理解的文章：

《PyTorch简介——一个简单但功能强大的深度学习库》

《开始使用PyTorch——学习如何建立快速和准确的神经网络（以4个案例研究为例）》

用于数据解释性的Python库

你真的了解模型如何工作吗？能解释模型为什么能够得出结果吗？这些是每个数据科学家都能够回答的问题。构建黑盒模型在业界毫无用处。

所以，上文中已经提到的两个Python库可以帮助解释模型的性能。

LIME

LIME是一种算法（库），可以解释任何分类器或回归量的预测。LIME是如何做到的呢？通过可解释的模型在局部不断接近预测值，这个模型解释器可用于生成任何分类算法的解释。

安装LIME很简单：

pip install lime

下文将从总体上帮助开发LIME背后的直觉和模型可解释性：

《在机器学习模型中建立信任（在Python中使用LIME）》

H2O

相信很多人都听说过H2O.ai，自动化机器学习的市场领导者。但是你知道其在Python中也有一个模型可解释性库吗？

H2O的无人驾驶AI，提供简单的数据可视化技术，用于表示高度特征交互和非线性模型行为，通过可视化提供机器学习可解释性（MLI），说明建模结果和模型中特征的影响。

通过下文，阅读有关H2O的无人驾驶AI执行MLI的更多信息。

《机器学习可解释性》

用于音频处理的Python库

音频处理或音频分析是指从音频信号中提取信息和含义以进行分析、分类或任何其他任务。这正在成为深度学习中的一种流行功能，所以要留意这一点。

LibROSA

LibROSA是一个用于音乐和音频分析的Python库。它提供了创建音乐信息检索系统所需的构建块。

这是一篇关于音频处理及其工作原理的深度文章：

《利用深度学习开始音频数据分析（含案例研究）》

Madmom

Madmom是一个用于音频数据分析的很棒的Python库。它是一个用Python编写的音频信号处理库，主要用于音乐信息检索（MIR）任务。

以下是安装Madmom的必备条件：

· NumPy

· SciPy

· Cython

· Mido

· PyTest

· Fyaudio

· PyFftw

安装Madmom的代码：

pip install madmom

下文可用以了解Madmom如何用于音乐信息检索：

《学习音乐信息检索的音频节拍追踪（使用Python代码）》

pyAudioAnalysis

pyAudioAnalysis是一个用于音频特征提取、分类和分段的Python库，涵盖广泛的音频分析任务，例如：

· 对未知声音进行分类

· 检测音频故障并排除长时间录音中的静音时段

· 进行监督和非监督的分割

· 提取音频缩略图等等

可以使用以下代码进行安装：

pip install pyAudioAnalysis

用于图像处理的Python库

因此，请确保熟悉以下三个Python库中的至少一个。

OpenCV-Python

谈到图像处理，OpenCV首先浮现在脑海中。OpenCV-Python是用于图像处理的Python API，结合了OpenCV C ++ API和Python语言的最佳特性。主要用于解决计算机视觉问题。

OpenCV-Python使用了上文提到的NumPy。所有OpenCV阵列结构都与NumPy数组相互转换。这也使得与使用Numpy的其他库（如SciPy和Matplotlib）集成变得更加容易。

在系统中安装OpenCV-Python：

pip3 install opencv-python

以下是两个关于如何在Python中使用OpenCV的流行教程：

《基于深度学习的视频人脸检测模型建立（Python实现）》

《16个OpenCV函数启动计算机视觉之旅（使用Python代码）》

Scikit-image

Scikit-image是另一个用于图像处理的python库，是用于执行多个不同图像处理任务的算法集合。可用于图像分割、几何变换、色彩空间操作、分析、过滤，形态学、特征检测等等。

在安装scikit-image前，请先安装以下软件包：

· Python（> = 3.5）

· NumPy（> = 1.11.0）

· SciPy（> = 0.17.0）

· joblib（> = 0.11）

这就是在机器上安装scikit-image的方法：

pip install -U scikit-learn

Pillow

Pillow是PIL（Python Imaging Library）的新版本。它是从PIL派生出来的，在一些Linux发行版（如Ubuntu）中被用作原始PIL的替代。

Pillow提供了几种执行图像处理的标准程序：

· 逐像素操作

· 掩模和透明处理

· 图像过滤，例如模糊，轮廓，平滑或边缘监测

· 图像增强，例如锐化，调整亮度、对比度或颜色

· 在图像上添加文字等等

安装Pillow：

pip install Pillow

查看以下关于在计算机视觉中使用Pillow的AI漫画：

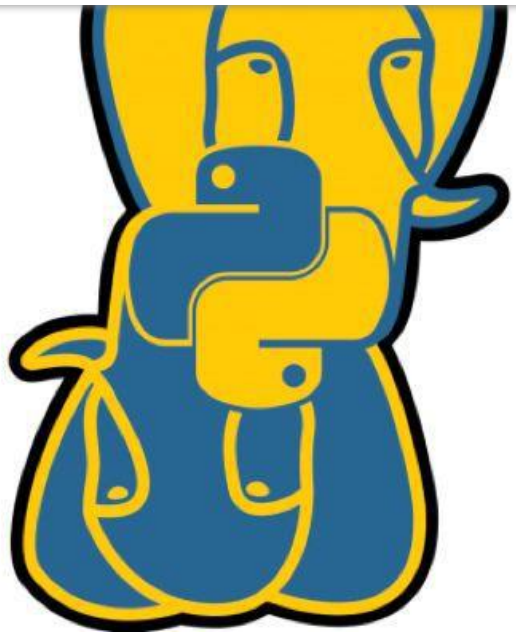
《AI漫画：Z.A.I.N —— 第二期：使用计算机视觉进行面部识别》

用于数据库的Python库

学习如何从数据库存储、访问和检索数据是数据科学家必备的技能。但是如何在不首先检索数据的情况下做到建模呢？

接下来介绍两个与SQL相关的Python库。

psycopg



Psycopg是Python编程语言中最流行的PostgreSQL（高级开源代码关系数据库）适配器。Psycopg的核心是完全实现Python DB API 2.0规范。

目前的psycopg2实现支持：

· Python版本2.7

· Python 3版本（3.4到3.7）

· PostgreSQL服务器版本（7.4到11）

· PostgreSQL客户端库版本（9.1以上）

以下是安装psycopg2的方法：

pip install psycopg2

SQLAlchemy

SQL是最流行的数据库语言。SQLAlchemy是pythonSQL工具包和对象关系映射器，它为应用程序开发人员提供了SQL的全部功能，且极具灵活性。

SQL旨在实现高效、高性能的数据库访问。SQLAlchemy将数据库视为关系代数引擎，而不仅仅是表的集合。

要安装SQLAlchemy，可以使用以下代码行：

pip install SQLAlchemy

用于部署的Python库

你知道哪些模型部署？部署模型意味着将最终模型放入最终应用程序（技术上称为生产环境）。

Flask

Flask是一个用Python编写的Web框架，广泛用于部署数据科学模型。Flask由两个部分组成：

· Werkzeug：Python编程语言的实用程序库

· Jinja：Python的模板引擎

查看下面的示例以打印“Hello world”：

```
from flask import Flask

app = Flask(__name__)

@app.route("/")

def hello():

return "HelloWorld!"

if __name__ == "__main__":

app.run()
```

以下文章是学习Flask的良好开端：

《在生产中将机器学习模型部署为API的教程（使用Flask）》

留言 点赞 关注