# 第11章Python第三方库纵览

## 11.1网络爬虫方向

网络爬虫是自动进行HTTP访问并捕获HTML页面的程序。Python语言提供了多个具备网络爬虫功能的第三方库。

### 11.1.1requests

requests库是一个简洁且简单的处理HTTP请求的第三方库，其最大优点是程序编写过程更接近正常URL访问过程。这个库建立在Python语言的urllib3库基础上，类似这种在其他函数库之上再封装功能提供更友好函数的方式在Python语言中十分常见。在Python生态圈里，任何人都有通过技术创新或体验创建新发表意见和展示才华的机会。

Requests库支持非常丰富的链接访问功能，包括国际域名和URL获取、HTTP长连接和连接缓存、HTTP会话和cookie保持、浏览器使用风格的SSL验证、基本的摘要认证、有效的键值对cookie记录、自动解压缩、自动内容解码、文件分块上传、HTTP（S）代理功能、连接超时处理、流数据下载等。

有关requests库的更多可访问 <http://www.python-requests.org>

安装语法为：

|  |
| --- |
| python35 -m pip install requests -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

requests库提供了7个函数，用来表示对一个网页的HTTP请求

|  |  |  |
| --- | --- | --- |
| **requests库的常用函数** | | |
| **序号** | **函数** | **描述** |
| 1 | requests.request() | 构造一个请求，支撑以下各方法的基础方法 |
| 2 | requests.get() | 获取HTML网页的主要方法，对应于HTTP的GET |
| 3 | requests.head() | 获取HTML网页头的主要方法，对应于HTTP的HEAD |
| 4 | requests.post() | 向HTML网页提交POST请求的方法，对应于HTTP的POST |
| 5 | requests.put() | 向HTML网页提交PUT请求的方法，对应于HTTP的PUT |
| 6 | requests.patch() | 向HTML网页提交局部修改请求，对应于HTTP的PATCH |
| 7 | requests.delete() | 向HTML页面提交删除请求，对应于HTTP的DELETE |

|  |
| --- |
| import requests  resp=requests.get('http://www.baidu.com')  print(resp.text) |

### 11.1.2scrapy

Scrapy是Python开发的一个快速的、高层次的Web获取框架。不同于简单的网络爬虫功能，scrapy框架本身包含了成熟网络爬虫系统所应该具有的部分共用功能，它是一个半成品，任何人都可以根据需求方便地利用框架已有功能经过简单扩展实现专业的网络爬虫系统。

Scrapy用途广泛，可以应用于专业爬虫系统的构建、数据挖掘、网络监控和自动化测试等领域。Scrapy提供URL队列、异步多线程访问、定时访问、数据库集成等众多功能，基于它构建的爬虫系统可以7\*24运行，具备产品级运行能力。

有关scrapy库的更多介绍可以访问<https://scrapy.org/>

|  |
| --- |
| python35 -m pip install scrapy -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

Scrapy框架结构，通过扩展函数完善功能后，可以利用这套框架实现对Web页面不间断的获取。由于scrapy框架质量很高，可以用于产品设计和开发。

|  |
| --- |
| C:\Users\ADMINI~1\AppData\Local\Temp\__nyf7_clip_images\image_60320483_396.png |

## 11.2数据分析方向

数据分析是Python的一个优势方向，具有大批高质量的第三方库。

### 11.2.1numpy

Numpy是Python的一种开源数值计算扩展第三方库，用于处理数据类型相同的多维数组(ndarray)简称“数组”。这个库可用来存储和处理大型矩阵，比Python语言提供的列表结构要高效很多。Numpy提供了许多高级的数值编程工具，如矩阵运算、矢量运算、N维数据变换等。

Numpy内部是C语言编写，对外采用Python语言进行封装。因此，在进行数据运算时，基于numpy的Python程序可以达到接近C语言的处理速度。Numpy也成为Python数据分析方向各其他库的基础依赖库，已经成了科学计算事实上的“标准库”。

有关numpy库的更多介绍可访问：<http://www.numpy.org/>

安装语法：

|  |
| --- |
| python35 -m pip install numpy -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.2.2scipy

Scipy是一款方便、易用、专为科学和工程设计的Python工具包，它在是numpy库的基础上增加了众多的数学、科学以及工程计算中常用的库函数。它包括统计、优化、整合、线性代数、傅立叶变换、信号分析、图像处理、常微分方程求解等众多模块。顾名思义，scipy主要用于科学和工程计算。

有关scipy库的更多介绍可访问: htts://scipy.org

安装语法：

|  |
| --- |
| python35 -m pip install scipy -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.2.3pandas

pandas是基于numpy扩展的一个重要第三方库，它是为了解决数据分析任务而创建的。pandas提供了一批标准的数据模型和大量快速便捷处理数据的函数和方法，提供了高效地操作大型数据集所需要的工具。

Pandas最初被作为金融数据分析工具而开发，因此，pandas为时间序列分析提供了很好的支持。Pandas的名称来自面板数据（panel data）和Python数据分析（data analysis）。Panel data是经济学中关于多维数据集的一个术语。Pandas提供两种最基本的数据类型：Series和DataFrame ，分别代表一维数组和二维数组类型。

有关pandas库的更多介绍要访问: <http://pandas.pydata.org/>

**安装语法：**

|  |
| --- |
| python35 -m pip install pandas -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

## 11.3文本处理方向

Python语言非常适合处理文本，因此，在这个方向也形成了大量有价值的第三方库。常用的有pdfminer、penpyxl、python-docx、beautifulsoup4。

### 11.3.1pdfminer

pdfminer是一个可以从PDF文档中提取各类信息的第三方库。与其他PDF相关的工具不同，它能够完全获取并分析PDF的文本数据。pdfminer能够获取PDF中文本的准确位置、字体、行数等信息，能够将PDF文个把转为HTML及文本格式。pdfminer包含两个重要的工具：pdf2txt.py和dumppdf.py。

pdf2txt.py能够从PDF文件中提取所有文本内容。dumppd.py能够把PDF文件内容变成XML格式，并进一步提取其中的图片。

在关pdfminer库的更多介绍可访问 <https://euske.github.io/pdfminer/>

安装语法：

|  |
| --- |
| python35 -m pip install pdfminer -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.3.2openpyxl

Openpyxl是一个处理Microsoft Excel文档的Python第三方库，它支持读写Excel的xls、xlsx、xlsm、xltx、xltm等格式文件，并进一步能处理Excel文件中的工作表、表单和数据单元。

有关openpyxl库的更多介绍可访问 <http://openpyxl.readthedocs.io/>

安装语法：

|  |
| --- |
| python35 -m pip install openpyxl -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.3.3python-docx

Python-docx是一个处理Microsoft Word文档的Python第三方库，它支持读取、查询以及修改doc、docx等格式文件，并能够对Word常见样式进行编程设置，包括字符样式、段落样式、表格样式、页面样式等，进一步可以使用这个库实现添加和修改文本、图像、样式和文档等功能。

有关python-docx库的更多介绍可访问 <https://pypi.python.org/pypi/python-docx>

安装语法：

|  |
| --- |
| python35 -m pip install python-docx -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.3.4beautifulsoup4

beautifulsoup4库，也称为Beautiful Soup库或bs4库，用于解析和处理HTML和XML。需要注意，它不是BeautifulSoup库。它的最大优点是能根据HTML和XML语法建立解析树，进而高效解析其中的内容。

HTML建立的Web页面一般非常复杂，除了有用的内容信息处，还包括大量用于页面格式的元素。直接解析一个Web网页需要深入了解HTML语法，而且比较复杂。Beautifulsoup4库将数据从HTML和XML文件中解析出来，它能够提供一种符合习惯的方法去遍历搜索和修改解析树，将专业的Web页面格式解析部分封装成函数，提供了若干有用且便捷的处理的函数。

有关beautifulsoup4库的更多介绍可访问 <https://www.crummy.com/software/BeautifulSoup/>

安装语法：

|  |
| --- |
| python35 -m pip install beautifulsoup4 -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

## 11.4数据可视化方向

数据可视化是指根据数据特点将其展示为易于理解的图形的过程。Python语言在数据可视化方面具有较强的优势。常用的有matplotlib、TVTK、mayavi。

### 11.4.1matplotlib

Matplotlib是提供数据绘图功能的第三方库，主要进行二维图表数据展示，广泛用于科学计算的数据可视化。使用这个库可以利用Pyhton程序绘制超过100种数据可视化效果。

有关matplotlib库的更多介绍可访问 <http://matplotlib.org/>

安装语法：

|  |
| --- |
| python35 -m pip install matplotlib -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.4.2TVTK

TVTK库是在标准的VTK库之上用Traits库进行封装的Python第三方库。视觉工具函数库（Visualization Toolkit，VTK）是一个开源、跨平台、支持平行处理的图形应用函数库，它是专业可编程的三维可视化工具。TVTK在Python生态系统中被等同于VTK。

有关TVTK库更多的介绍可访问 <http://docs.enthought.com/mayavi/tvtk/>

TVTK在Windows平台上暂时不能使用pip命令安装，建议使用文件安装，在其它平台可使用pip命令安装。

### 11.4.3mayavi

虽然VTK 3D可视化软件包功能强大，Python的TVK封装使用方便简洁，但是要用这些工具快速编写实用的三维可视化程序仍然需要花费不少精力。Mayavi基于TVK开发，完全用Python编写，提供了一个更为方便实用的可视化软件，可以简洁地嵌入到用户编写的Python程序中，或者直接使用其面向脚本的API快速绘制三维可视化图形。值得注意的是mayavi也被称为mayavi2。

有关mayavi库的更多介绍可访问 <http://docs.enthought.com/mayavi/mayavi/>

mayavi库在Windows平台上暂时不能使用pip命令安装，建议使用文件安装，在其它平台可使用pip命令安装。

## 11.5用户图形界面方向

Python标准库内置了一个GUI库—Tkinter，这个库基于Tcl/Tk开发，然而这个库十分陈旧，提供的开发控制也很有限、编写出来的GUI风格与现代程序GUI风格相差甚远，从用户体验角度说，Tkinter库并不成熟。

常用的3个高质量的用户图形界面Python生态库：PyQt5、wxPython、PyGTK

### 11.5.1PyQt5

PyQt5是Q5应用框架的Python第三方库，它有超过620个类和近6000个函数和方法。它是Python中最为成熟的商业级GUI第三方库。这个库是Python语言当前最好的GUI第三方库，它可以Window、Linux和MacOS X等操作系统上跨平台使用。

PyQt5采用“信号-槽”机制将事件和对应的处理程序进行绑定。简单说，信号就是事件、槽是事件的处理程序。例如，鼠标点击操作是信号，处理鼠标点击的程序是槽。一个信号可以连接多个槽，也可以不连接槽。信号可以连接其他信息。PyQt5窗体有很内置信息，也可以自定义信号。

有关PyQt5库的更多介绍可访问 https://www.riverbank.computing.com/software/pyqt/

安装语法：

|  |
| --- |
| python35 -m pip install pyqt5 -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.5.2wxPython

wxPython是Python语言的一套优秀的GUI图形库，它是跨平台GUI库wxWidgets的Python封装，可以使用Python程序员能够轻松地创建健壮可靠、功能强大的图形用户界面的程序。其中，wxWidgets使用C++语言编写，C/C++语言编写的功能库进行二次封装是Python语言的重要特点之一。

有关wxPython库的更多介绍可访问 <https://wxpython.org/>

安装语法：

|  |
| --- |
| python35 -m pip install wxPython -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.5.3PyGTK

PyGTK是基于GTK+的Python语言封装，它提供了各式的可视元素和功能，能够轻构创建具有图形用户界面的程序。PyGTK具有跨平台性，利用它编写的代码能够不加修改地稳定运行在各操作系统中，如Windows、MacOS、Linux等。

有关PyGTK库的更多介绍可访问 <http://www.pygtk.org>

安装语法：

|  |
| --- |
| python35 -m pip install pygtk -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

## 11.6机器学习方向

机器学习是人工智能领域的一个重要分支，Python语言也是机器学习和人工智能的重要基础语言。常用的三个高质量的机器学习框架：scikit-learn、 TensofFlow、Theano。

### 11.6.1scikit-learn

scikit-learn是一个简单且高效的数据挖掘和数据分析工具，它基于NumPy、SciP和matplotplib构建。scikit-learn项目最早由数据科学家David Cournapeau在2007年组织开发，它是Python语言中专门针对机器学习应用而发展起来的一款开源框架。scikit-learn的基本功能主要包括6个部分：分类、回归、聚类、数据降维、模型选择和数据预处理。scikit-learn也被称为sklearn。

有关scikit-learn库更多的介绍可访问 <http://scikit-learn.org/>

安装语法：

|  |
| --- |
| python35 -m pip install scikit-learn -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.6.2TensorFlow

TensorFlow是谷歌公司基于DistBelief进行研发的第二代人工智能学习系统，也是用来支撑著名AlphaGo系统后台框架，其命名来源于其自身的运行原理。Tensor（张量）指N维数组，Flow（流）指基于数据流图的计算，TensorFlow描述张量从流图的一端流动到另一端的计算过程。

TensorFlow的应用十分广泛，从语音识别或图像识别到机器翻译或自主跟踪等，既可以运行在数万台服务器的数据中心，也可以运行在智能手机或嵌入式设备中。

有关TensorFlow库的的更多介绍可访问 <http://www.tensorflow.org/>

### 11.6.3Theano

Theano为执行深度学习中大规模神经网络算法的运算而设计，擅长处理多维数组。Theano开发始于2007年，可以理解它是一个运算数学表达式的编译器，并可以高效运行在GPU或CPU上。Theano是一个偏向底层开发的库，更像一个研究平台而单纯的深度学习库。

有关Theano库的更多介绍可访问 <http://deeplearning.net/software/theano/>

安装语法：

|  |
| --- |
| python35 -m pip install theano -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

## 11.7Web开发方向

Web开发是Python语言流行的一个重要方向，主要用于服务器后端开发。根据Python Web开发框架的复杂程序。常用的Python第三方生态度有:Django、 Pyramid、Flask。

### 11.7.1Django

Django是Python生态中最流行的开源Web应用框架。Django采用模型（Model）、模板（Template）和视图（Views）的编写模式，称为MTV模式。其中，模型即数据存取层，用于处理与数据相关的所有事务，包括如何存取、如何验证有效性、包含哪些行为以及数据之间的关系等；模板即表现层，用于处理与表现相关的功能，通过模板方式定义页面风格；视图即业务逻辑层，用于存取模型及调取恰当模板的相关逻辑，是模型与模板的桥梁。

Django中提供了开发网站经常用到的模块，Django的开发理念是DRY（Don’t Repeat Yourself），用于鼓励快速开发、进而减少程序员建立一个高性能Web应用所花费的时间和精力，形成一种一站式解决方案。

有关Django库的更多介绍可访问 <https://www.djangoproject.com/>

安装语法：

|  |
| --- |
| python35 -m pip install django -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.7.2Pyramid

Pyramid是一个通用、开源的Python Web应用程序开发框架。它主要的目的是让Python开发者更简单地创建Web应用。相比Django，Pyramid是一个相对小巧、快速、灵活的开源PythonWeb框架。

与Django一样，Pyramid仍然面向较大规模的Web应用，但它更关注灵活性，开发者可以灵活选择所使用的数据库、模板风格、URL结构等内容。

有关Pyramid库的更多介绍可访问 <https://trypyramid.com/>

安装语法：

|  |
| --- |
| python35 -m pip install pyramid -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.7.3Flask

Flask是轻量级Web应用框架，相比Django和Pyramid，它也被称为微框架。使用Flask开发Web应用十分方便，甚至几行代码即可建立一个小型网站。Flask核心十分简单，并不直接包含诸如数据库访问等的抽象访问层，而是通过扩展模块形式来支持。

有关Flask库的更多介绍可访问 <http://flask.pocoo.org/>

安装语法：

|  |
| --- |
| python35 -m pip install flask -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

## 11.8游戏开发方向

游戏开发是一个有趣的方向，在游戏逻辑和功能实现层面，Python已经成为重要的支撑性语言。有Python中有三个常用的生态库：Pygame、Panda3D、cocos2d。

### 11.8.1Pygame

Pygame是在SDL库基础上进行封装的、面向游戏开发入门的Python第三方库，除了制作游戏外，还用于制作多媒体应用程序。其，中SDL（Simple DirectMedia Layer）是开源、跨平台的多媒体开发库，通过OpenGL和Direct3D底层函数提供对音频、键盘、鼠标和图形硬件的简洁访问。

Pygame是一个游戏开发框架，提供了大量与游戏相关的底层逻辑和功能支持，非常适合作为入门库理解并实践游戏开发。

有关Pygame库的更多介绍可访问 <http://www.pygame.org>

安装语法：

|  |
| --- |
| python35 -m pip install pygame -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.8.2Panda3D

Panda3D是一个开源、跨平台的3D渲染和游戏开发库，简单说，它是一个3D游戏引擎，由迪士尼和卡耐基.梅隆大学娱乐技术中心共同开发。Panda3D支持Python和C++两种语言，但对Python支持更全面。

Panda3D支持很多当代先进游戏引擎所支持的特性，如法线贴图、光泽贴图、HDR、卡通渲染和线框渲染等。

有关Panda3D库的更多介绍可访问 <http://www.panda3d.org>

安装语法：

|  |
| --- |
| python35 -m pip install panda3d -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.8.3cocos2d

Cocos2d是一个构建2D游戏和图形界面交互式应用的框架，它包括C++、JavaScript、Swift、Python等多个版本。Cocos2d基于OpenGL进行图形渲染，能够利用GPU进行加速。Cocos2d引擎采用树形结构来管理游戏对象，一个游戏划分为不同场景，一个场景又分为不同层，每个层处理并响应用户事件。

有关cocos2d库的更多介绍可访问 <http://python.cocos2d.org/>

安装语法：

|  |
| --- |
| python35 -m pip install cocos2d -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

## 11.9更多第三方库

Python语言有十几万个第三方库，覆盖信息技术几乎所有领域。即使在每个方向，也会有大量的专业人员开发多个第三方库来给出具体设计。除了本章所提到的方向外，还有以下几个第三方库也是比较常用的。

### 11.9.1PIL

PIL库是Python语言在图像处理方面的重要第三方库，支持图像存储、处理和显示，它能够处理几乎所有的图片格式，可以完成对图像的缩放、剪截、叠加以及向图像添加线条、图像和文字等操作。使用Python语言处理图像相关的程序，首选PIL库。

PIL库主要可以完成图像归档和图像处理两方面功能需求。

图像归档：对图像进行批处理、生成图像预览、图像格式转换等。

图像处理：图像基本处理、像素处理、颜色处理等。

有关PIL库的更多介绍可访问 <http://pillow.readthedocs.io>

安装语法：

|  |
| --- |
| python35 -m pip install pillow -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.9.2SymPy

SymPy是一个支持符号计算的Python第三方库，它是一个全功能的计算机代数系统。SymPy代码简洁、易于理解、支持符号计算、高精度计算、模式匹配、绘图、解方程、微积分、组合数学、离散数学、几何学、概论与统计、物理学领域的计算和应用。

有关SymPy库的更多介绍可访问： <http://www.sympy.org/>

安装语法：

|  |
| --- |
| python35 -m pip install sympy -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.9.3NLTK

NLTK是一个非常重要的自然语言处理Python第三方库，它支持多种语言，尤其对中文支持良好。NLTK可以进行语料处理、文本统计、内容理解、情感分析等我种应用，具备非常可靠的应用价值。

有关NLTK的更多介绍可访问 <http://www.nltk.org/>

安装语法：

|  |
| --- |
| python35 -m pip install nltk -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.9.4WeRoBot

WeRoBot是一个微信公众号开发框架，也称为微信机器人框架。WeRoBot可以解析微信服务器发来的消息，并将消息转换成Message或者Event类型。其中，表示用户发来的消息，如文本消息、图片消息：Event则表示用户触发的事件，如关注事件、扫描二维确事件。在消息解析、转换完成后。WeRoBot会将消息转交给Handler进行处理，并将Handler的返回值给微信服务器，进而实现完整的微信机器人功能。

有关WeRoBot库的更多介绍可访问 <http://werobot.readthedocs.io>

安装语法：

|  |
| --- |
| python35 -m pip install werobot -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

### 11.9.5MyQR

MyQR是一个能够产生基本二维码、艺术二维码和动态效果二维码的Python第三方库。有关MyQR库的更多介绍可访问 <https://github.com/sylnsfar/qrcode>

安装命令：

|  |
| --- |
| python35 -m pip install myqr -i http://pypi.douban.com/simple --trusted-host pypi.douban.com |

## 11.10课后习题

### 11.10.1选择题

|  |  |  |
| --- | --- | --- |
| 1) | 以下选项中用于网络爬编程的是（ ） | |
|  | A. | pandas |
|  | B. | PyQt5 |
|  | C. | scrapy |
|  | D. | Django |

|  |  |  |
| --- | --- | --- |
| 2) | 以下选项中用于科学计算/数据分析是（ ） | |
|  | A. | PyGTK |
|  | B. | numpy |
|  | C. | PyQt5 |
|  | D. | Panda3D |

|  |  |  |
| --- | --- | --- |
| 3) | 以下选项中用于文本处理的是（ ） | |
|  | A. | pandas |
|  | B. | Flask |
|  | C. | request |
|  | D. | beautifulsoup4 |

|  |  |  |
| --- | --- | --- |
| 4) | 以下选项中用于数据可视化的是（ ） | |
|  | A. | numpy |
|  | B. | PyQt5 |
|  | C. | scipy |
|  | D. | matplotlib |

|  |  |  |
| --- | --- | --- |
| 5) | 以下选项中用于用户图形界面的是（ ） | |
|  | A. | scrapy |
|  | B. | PyQt5 |
|  | C. | Django |
|  | D. | Cocos2d |

|  |  |  |
| --- | --- | --- |
| 6) | 以下选项中用于游戏开发的是（ ） | |
|  | A. | NLTK |
|  | B. | PIL |
|  | C. | PyQt5 |
|  | D. | Pygame |

|  |  |  |
| --- | --- | --- |
| 7) | 以下选项中用于机器学习的是（ ） | |
|  | A. | TensorFlow |
|  | B. | numpy |
|  | C. | SymPy |
|  | D. | openpyxl |

|  |  |  |
| --- | --- | --- |
| 8) | 以下选项中用于Web开发的是（ ） | |
|  | A. | Django |
|  | B. | scrapy |
|  | C. | requests |
|  | D. | TensorFlow |

|  |  |  |
| --- | --- | --- |
| 9) | 以下选项中用于处理图像的是（ ） | |
|  | A. | beautifulsoup4 |
|  | B. | PyQt5 |
|  | C. | matplotlib |
|  | D. | PIL |

|  |  |  |
| --- | --- | --- |
| 10) | 以下选项中用于开发网络爬虫的是（ ） | |
|  | A. | numpy |
|  | B. | scrapy |
|  | C. | wxPython |
|  | D. | Flask |