

## 第十章

# python第三方库概览

# 本章大纲

1. 第三方库的获取和安装
2. PyInstaller库(必选)
3. 第三方库: jieba库(必选)、wordcloud库 (可选)

## 1.第三方库获取和安装

Python第三方库依照安装方式灵活性和难易程度有三个方法：**pip工具安装、自定义安装和文件安装。**

### **pip工具安装：**

最常用且最高效的Python第三方库安装方式是采用pip工具安装。pip是Python官方提供并维护的在线第三方库安装工具。

**pip install <拟安装库名>/pip3 install <拟安装库名>**

## 自定义安装：

自定义安装指按照第三方库提供的步骤和方式安装。第三方库都有主页用于维护库的代码和文档。以科学计算用的numpy为例，开发者维护的官方主页是：

<http://www.numpy.org/>

浏览该网页找到下载链接，如下：

<http://www.scipy.org/scipylib/download.html>

进而根据指示步骤安装。

## 文件安装：

为了解决这类第三方库安装问题，美国加州大学尔湾分校提供了一个页面，帮助Python用户获得Windows可直接安装的第三方库文件，  
链接地址如下：

<http://www.lfd.uci.edu/~gohlke/pythonlibs/>

这里以scipy为例说明，首先在上述页面中找到scipy库对应的内容。选择其中的.whl文件下载，这里选择适用于Python 3.5版本解释器和32位系统的对应文件：scipy-0.17.1-cp35-cp35m-win32.whl，下载该文件到D:\pycodes目录。然后进入该目录使用pip安装。

## pip工具使用：

### 1、执行pip -h将列出pip常用的子命令

```
C:\Users\11397>pip -h

Usage:
  pip <command> [options]

Commands:
  install           Install packages.
  download          Download packages.
  uninstall         Uninstall packages.
```

2、pip支持安装（install）、下载（download）、卸载（uninstall）、列表（list）、查看（list）、查找（search）等一系列安装和维护子命令。

## 2. pyinstaller库概述

PyInstaller是一个十分有用的Python第三方库，它能够在Windows、Linux、Mac OS X等操作系统下将Python源文件打包，变成直接可运行的可执行文件。通过对源文件打包，Python程序可以在没有安装Python的环境中运行，也可以作为一个独立文件方便传递和管理。

安装方式如下：

```
C:\Users\11397>pip install pyinstaller_
```

1、使用PyInstaller库对Python源文件打包十分简单，使用方法如下：

`:>PyInstaller <Python源程序文件名>`

执行完毕后，源文件所在目录将生成dist和build两个文件夹。最终的打包程序在dist内部与源文件同名的目录中。

2、可以通过-F参数对Python源文件生成一个独立的可执行文件，如下：

`:>PyInstaller -F <Python源程序文件名>`

```
C:\Users\11397>cd Desktop  
C:\Users\11397\Desktop>pyinstaller -F 000.py
```



PyInstaller有一些常用参数

参数	功能
-h, --help	查看帮助
--clean	清理打包过程中的临时文件
-D, --onedir	默认值，生成dist目录
-F, --onefile	在dist文件夹中只生成独立的打包文件
-i <图标文件名.ico >	指定打包程序使用的图标（icon）文件

### 3. jieba库的概述

- 1、由于中文文本中的单词不是通过空格或者标点符号分割，中文及类似语言存在一个重要的“分词”问题。
  - 2、jieba（“结巴”）是Python中一个重要的第三方中文分词函数库。
- 安装方式：

```
C:\Users\11397>pip install jieba_
```

3、jieba库的分词原理是利用一个中文词库，将待分词的内容与分词词库进行比对，通过图结构和动态规划方法找到最大概率的词组。除了分词jieba还提供增加自定义中文单词的功能。

4、jieba库支持三种分词模式：**精确模式**，将句子最精确地切开，适合文本分析；**全模式**，把句子中所有可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；**搜索引擎模式**，在精确模式基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

## 4. jieba库与中文分词

1、jieba.lcut(s)是最常用的中文分词函数，用于精准模式，即将字符串分割成等量的中文词组，返回结果是列表类型。

```
>>> import jieba
>>> ls = jieba.lcut("全国计算机等级考试python科目")
>>> print(ls)
['全国', '计算机', '等级', '考试', 'python', '科目']
```

2、jieba.lcut(s, cut\_all = True)用于全模式，即将字符串的所有分词可能均列出来，返回结果是列表类型，冗余性最大。

```
>>> import jieba
>>> ls = jieba.lcut("全国计算机等级考试python科目", cut_all = True)
>>> print(ls)
['全国', '国计', '计算', '计算机', '算机', '等级', '考试', 'python', '科目']
```

3、jieba.lcut\_for\_search(s)返回搜索引擎模式，该模式首先执行精确模式，然后再对其中长词进一步切分获得最终结果。

```
>>> import jieba
>>> ls = jieba.lcut_for_search("全国计算机等级考试Python科目")
>>> print(ls)
['全国', '计算', '算机', '计算机', '等级', '考试', 'Python', '科目']
```

搜索引擎模式更倾向于寻找短词语，这种方式具有一定冗余度，但冗余度相比全模式较少。

4、jieba.add\_word()函数，顾名思义，用来向jieba词库增加新的单词。

```
>>> import jieba
>>> jieba.add_word('奥里给')
>>> import jieba
>>> jieba.add_word("奥里给")
>>> ls = jieba.lcut("作为一个年轻人，我们要努力，加油，奥里给！")
>>> print(ls)
['作为', '一个', '年轻人', '，', '，', '我们', '，', '要', '，', '努力', '，', '，', '加油',
'，', '，', '奥里给', '，', '！']
```

## 5. wordcloud库概述

1、wordcloud库是专门用于根据文本生成词云的Python第三方库，十分常用且有趣。

2、装wordcloud库在Windows的cmd命令行使用如下命令：

```
C:\Users\11397>pip install wordcloud
```

3、wordcloud库的使用十分简单，以一个字符串为例。其中，产生词云只需要一行语句，在第三行，并可以将词云保存为图片。

```
from wordcloud import WordCloud  
txt = 'i like python i am learning python'  
wd = WordCloud().generate(txt)  
wd.to_file('test.png')
```



在生成词云时，wordcloud默认会以空格或标点为分隔符对目标文本进行分词处理。对于中文文本，分词处理需要由用户来完成。一般步骤是先将文本分词处理，然后以空格拼接，再调用wordcloud库函数。

```
import jieba
from wordcloud import WordCloud
txt = '程序设计语言是计算机能够识别和理解用户操作意图的一种交互体系，\
他按照特定规则组织计算机指令，是计算机能够自动进行各种运算处理'
ls = jieba.lcut(txt)
txtwords = ' '.join(ls)
wd = WordCloud(font_path='C:\\Windows\\Font\\simkai.ttf',
               width=800,height=600).generate(txtwords)
wd.to_file('word.png')
```

## 6. WordCloud对象创建的常用参数

参数	功能
font_path	指定字体文件的完整路径，默认None
width	生成图片宽度，默认400像素
height	生成图片高度，默认200像素
mask	词云形状，默认None，即，方形图
min_font_size	词云中最小的字体字号，默认4号
font_step	字号步进间隔，默认1
max_font_size	词云中最大的字体字号，默认None，根据高度自动调节
max_words	词云图中最大词数，默认200
stopwords	被排除词列表，排除词不在词云中显示
background_color	图片背景颜色，默认黑色

## 7. WordCloud类常用方法

方法	功能
generate(text)	由text文本生成词云
to_file(filename)	将词云图保存为名为filename的文件