表 7‑25 下载器中间件代码

|  |
| --- |
| # -\*- coding: utf-8 -\*-  import random  import base64  # 随机的User-Agent  class RandomUserAgent(object):    user\_agent\_list = [  "Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_10\_1) AppleWebKit/537.36\  (KHTML, like Gecko) Chrome/41.0.2227.1 Safari/537.36",  "User-Agent:Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10\_6\_8; en-us)\  AppleWebKit/534.50 (KHTML, like Gecko) Version/5.1 Safari/534.50",  "Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_8\_2) AppleWebKit/537.13\  KHTML, like Gecko) Chrome/24.0.1290.1 Safari/537.13",  "Mozilla/5.0(compatible;MSIE9.0;WindowsNT6.1;Trident/5.0",  "Mozilla/5.0(Macintosh;IntelMacOSX10.6;rv:2.0.1)Gecko/20100101Firefox/4.0.1",  "Mozilla/5.0(WindowsNT6.1;rv:2.0.1)Gecko/20100101Firefox/4.0.1",  "Mozilla/5.0 (compatible; MSIE 10.0; Windows NT 6.2; ARM; Trident/6.0)"  ]  def process\_request(self, request, spider):  useragent = random.choice(user\_agent\_list)  request.headers.setdefault("User-Agent", useragent)    class RandomProxy(object):    PROXIES = [  {'ip\_port': '111.8.60.9:8123', 'user\_passwd': 'user1:pass1'},  {'ip\_port': '101.71.27.120:80', 'user\_passwd': 'user2:pass2'},  {'ip\_port': '122.96.59.104:80', 'user\_passwd': 'user3:pass3'},  {'ip\_port': '122.224.249.122:8088', 'user\_passwd': 'user4:pass4'}  ]  def process\_request(self, request, spider):  if proxy['user\_passwd'] is None:  # 没有代理账户验证的代理使用方式  request.meta['proxy'] = "http://" + proxy['ip\_port']  else:  # 对账户密码进行base64编码转换  base64\_userpasswd = base64.b64encode(proxy['user\_passwd'])  # 对应到代理服务器的信令格式里  request.headers['Proxy-Authorization'] = 'Basic ' + base64\_userpasswd  request.meta['proxy'] = "http://" + proxy['ip\_port'] |

表 7‑27 激活中间件代码（在settings脚本种追加）

|  |
| --- |
| DOWNLOADER\_MIDDLEWARES = {  'TipDMSpider.middlewares.DownloaderMiddleware': 310,  } |

表 7‑28 关闭内置下载器中间件RobotsTxtMiddleware

|  |
| --- |
| DOWNLOADER\_MIDDLEWARES = {  'TipDMSpider.middlewares.DownloaderMiddleware': 310,  'scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware':0,  } |

表 7‑30 停用RefererMiddleware中间件代码（在settings脚本种追加）

|  |
| --- |
| SPIDER\_MIDDLEWARES =  {  'scrapy.spidermiddlewares.referer.RefererMiddleware': 0,  } |