目录

**爬虫框架2**

总体流程2

一、queue2

二、crawler\_controller2

三、list\_page\_handler3

1、流程3

2、接收的消息4

3、发送的消息4

四、detail\_page\_handler5

1、流程5

2、接收的消息6

3、发送的消息6

五、downloader8

1、流程8

2、判断url是否需要下载9

3、接收的消息10

4、发送的消息11

六、file\_handler12

1、流程12

2、接收的消息13

3、发送的消息13

总体流程

low queue

low queue

low queue

low queue

high queue

high queue

high queue

high queue

crawler\_controller

(start\_urls)

high queue

low queue

high queue

high queue

file\_handler

downloader

detail\_page\_handler

list\_page\_handler

1. queue:
2. 优先级队列借助redis的lpush(queue,message)和brpop([high\_queue,low\_queue])实现。

二、crawler\_controller:

1、list\_page\_url：各应用市场中各分类的首页url地址。

2、detail\_page\_url一般通过构造所得。例：<http://www.XXX.com/app/id=(0~1000000)>

三、list\_page\_handler:

list\_page\_handler

多

程

进

是

否

是

否

是

Receive list\_page\_url

是否在24h

内爬取过

否

Phantomjs打开网页

获取并发送页面中的详情页url

是否存在“加载更多”按钮

点击“加载更多”

市场(source)

是否已覆盖

是否存在“下一页”按钮

否

Redis(db=10)记录该url并设置过期时间

获取并发送下一页的页面url

是

等待接收消息

1、流程

2、接收消息：

-必需参数：

、list\_page\_url：【string】需要处理的列表页链接。

爬虫流程中为crawler\_controller给定或由上一页列表页解析得到。

、source：【string】页面源。

此时的source为shouji\_baidu\_com格式，方便导入市场的配置文件。

-有默认值非必需参数：

、message\_type：【string】该消息的类型（需做何种处理）。

默认为”normal”。”normal”类型消息不作处理。爬虫流程中应为”crawl”。

、level：【string】该消息级别。

默认为”low”。该项决定了向下一部件发送消息时送送到低优先级队列还是高优先级队列。

-无默认值非必需参数：

无

3、发送消息：

-必须项：

、detail\_page\_url：【string】下载链接。

由页面中解析得到。

、source：【string】页面来源。

此处格式为shouji\_baidu\_com。

-非必须项：

、level：【string】该消息级别。

由接收消息指定或默认。

、message\_type：【string】消息类型。

由接收消息指定或默认。

四、detail\_page\_handler：

detail\_page\_handler

多

程

进

否

是

否

是

否

是

Receive detail\_page\_url

是否在24h

内爬取过

Phantomjs打开网页

获取页面中的app信息

是否可以找到

下载链接

市场(source)

是否已覆盖

Redis(db=11)记录该url并设置过期时间

发送下载链接及相关信息

等待接收消息

1、流程

2、接收消息：

-必需参数：

、detail\_page\_url：【string】需要处理的详情页链接。

爬虫流程中为list\_page\_handler获取到的下载链接或crawler\_controller生成的链接。

、source：【string】页面源。

此时的source为shouji\_baidu\_com格式，方便导入市场的配置文件。

-有默认值非必需参数：

、message\_type：【string】该消息的类型（需做何种处理）。

默认为”normal”。”normal”类型消息不作处理。爬虫流程中应为”crawl”。

、level：【string】该消息级别。

默认为”low”。该项决定了向下一部件发送消息时送送到低优先级队列还是高优先级队列。

-无默认值非必需参数：

无

3、发送消息：

-必须项：

、download\_url：【string】下载链接。

由页面中解析得到。

-非必须项：

、source：【string】app来源。

由接收消息的source确定，此时格式应为shouji.baidu.com。

、detail\_url：【string】app所在的详情页地址。

由接收消息指定。

、download\_redirect\_urls：【list[string]】app的下载地址。

由download\_url处理得到。

、app\_name：【string】app名字。

由页面中解析得到。

、category：【string】app所属分类。

由页面中解析得到。

、download\_count：【string】app下载次数。

由页面中解析得到。

、upload\_time：【string】app上传时间。

由页面中解析得到。

、rating：【string】app评级。

由页面中解析得到。

、level：【string】该消息级别。

由接收消息指定或默认。

、message\_type：【string】消息类型。

由接收消息指定或默认。

、download\_folder：【string】下载存放目录。

、download\_file\_name：【string】下载文件名。

、download\_file\_type: 【string】下载文件类型。

、target\_folder: 【string】通知file\_handler处理的目标文件夹。

、method：【string】网络连接动作（GET/POST等 ）

由市场配置文件决定或默认。

、request\_headers：【string】使用requests库进行连接时的请求头。

由市场配置文件决定或默认。

、download\_cookie\_producer：【{‘url’:’’,’method’:’GET/POST’,’data’:{}}】生成session的参数。

由市场配置文件决定或默认。

、data：【dict】若method为post时要发送的数据。

由市场配置文件决定或默认。

五、downloader:

1、流程

构造访问session

（cookie等信息）

Get response header

downloader

多

程

线

是

否

Receive download\_url

判断是否

需要下载

等待接收消息

Get response content

将content写入文件

将重定向后url的md5存到redis(db12)

向file\_handler发送消息

该md5的文件

是否已存在

2、判断url是否需要下载

是

是

是

downloader\_url

response

是

是

否

否

否

是

是

否

否

是

否

否

否

状态码是否

为200

需要下载

获取重定向后的url

不需要下载

是否具有

content-type

Content-type

是否符合期望

该md5的文件

是否已存在

该md5的文件

是否已存在

是否具有

content-md5

Redis(db=12)中

是否有该url的md5

是否具有

last-modified

last-modified

是否已改变

3、接收消息：

-必需参数：

、download\_url：【string】需要处理的下载链接。

爬虫流程中为detail\_page\_handler获取到的下载链接。

-有默认值非必需参数：

、message\_type：【string】该消息的类型（需做何种处理）。

默认为”normal”。”normal”类型消息不作处理。爬虫流程中应为”crawl”。

、level：【string】该消息级别。

默认为”low”。该项决定了向下一部件发送消息时送送到低优先级队列还是高优先级队列。

、download\_folder：【string】下载存放目录。

默认为./download\_folder/tmp/（可在config.py中更改）。

、download\_file\_name：【string】下载文件名。

默认为“当前线程id\_下载时间戳”（可在downloader.py中更改）。

、download\_file\_type: 【string】下载文件类型。

默认为apk。（在config.py中定义了文件扩展名及content-type）

、target\_folder: 【string】通知file\_handler处理的目标文件夹。

默认为./download\_folder/（可在config.py中更改）。

、method：【string】网络连接动作（GET/POST等 ）

默认为GET。 （可在config.py中更改）

、request\_headers：【string】使用requests库进行连接时的请求头。

默认为：{‘User-Agent’:’AndroidDownloadManager/4.1.1(Linux;U;Android 4.1.1;Nexus S Build/JRO03E)’}（可在config.py中更改）。一般用来设置user-agent，也可用于设置cookie。不过cookie的问题一般交由session解决（见download\_cookie\_producer）。

-无默认值非必需参数：

、download\_cookie\_producer：【{‘url’:’’,’method’:’GET/POST’,’data’:{}}】生成session的参数。

有些下载链接有特殊下载要求，可以借由requests库先打开前置页面或发送数据以获得下载所需必要信息。

、data：【dict】若method为post时要发送的数据。

、source：【string】app来源。

在downloader中非必需，因为只是写记录文件时的可选信息。但由于source 在前面部件中为必需，所以爬虫流程传过来应该都有此项。

、detail\_url：【string】app所在的详情页地址。

同上

、category：【string】app所属分类。

、download\_count：【string】app下载次数。

、upload\_time：【string】app上传时间。

、rating：【string】app评级。

、app\_name：【string】app名字。

4、发送消息：

-必须项：

、file\_name：【string】下载的文件名。

由接收消息指定或默认文件名

-非必须项：

、source：【string】app来源。

由接收消息指定。

、detail\_url：【string】app所在的详情页地址。

由接收消息指定。

、download\_redirect\_urls：【list[string]】app的下载地址。

由download\_url处理得到。

、app\_name：【string】app名字。

由接收消息指定。

、category：【string】app所属分类。

由接收消息指定。

、download\_count：【string】app下载次数。

由接收消息指定。

、upload\_time：【string】app上传时间。

由接收消息指定。

、rating：【string】app评级。

由接收消息指定。

、level：【string】该消息级别。

由接收消息指定或默认。

、message\_type：【string】消息类型。

由接收消息指定或默认。

、target\_folder：【string】目标目录（即cell-xx所在目录）。

由接收消息指定或默认。

六、file\_handler

1、流程

detail\_page\_handler

多

程

进

否

是

是

否

app[‘new’]=’True’

Receive file name

写app信息文件与记录文件

是否可以找到

下载链接

是否通过

jarsigner校验

是否有相同md5

文件存在

等待接收消息

删除文件

计算文件md5

移动到对应目录并重命名

删除文件

2、接收消息：

-必需参数：

、file\_name：【string】需要处理的文件名字。

爬虫流程中为downloader下载下来的文件名字。（downloader中默认为“当前线程id\_下载时间戳”，app信息中的下载时间由此解析得到。）

-有默认值非必需参数：

、message\_type：【string】该消息的类型（需做何种处理）。

默认为”normal”。”normal”类型消息不作处理。爬虫流程中应为”crawl”。

、level：【string】该消息级别。

默认为”low”。该项决定了向下一部件发送消息时送送到低优先级队列还是高优先级队列。file\_handler不向下发送消息，所以可以无视该项，一般是由上面部件带下来的参数。

、target\_folder：【string】目标目录（即cell-xx所在目录）。

默认为./download\_folder/（可在config.py中更改）。

-无默认值非必需参数：

、source：【string】app来源。

在file\_handler中非必需，因为只是写记录文件时的可选信息。但由于source 在前面部件中为必需，所以爬虫流程传过来应该都有此项。

、detail\_url：【string】app所在的详情页地址。

同上。

、download\_redirect\_urls：【list[string]】app的下载地址。

同上；包括整个重定向前后的下载地址。

、app\_name：【string】app名字。

、category：【string】app所属分类。

、download\_count：【string】app下载次数。

、upload\_time：【string】app上传时间。

、rating：【string】app评级。

3、发送消息：

无