----------------------------------------------------------------用户指南-----------------------------------------------------------

参数名（args） 说明（instructions）

folder\_name \* 该参数表示：自定义存储html文件的根目录名称，即文件夹名称。

url \* 该参数表示：url前半部分，即，页码数字之前的部分。如：https://coinmarketcap.com/2/ 的url前半部分为数字2之前的部分。

asyn 默认值为False。该参数表示目标网站是否为异步加载。若为异步加载，则 asyn = True。若不为异步加载，则 asyn = False。

start\_page 默认值为空值。该参数表示：想要爬取的起始页码（这里的页码不是经过数学运算的页码，而是你在页面中看到的），即，第几页。

end\_page 默认值为空值。该参数表示：想要爬取的终止页码（这里的页码不是经过数学运算的页码，而是你在页面中看到的），即，第几页。

normal\_page\_number 默认值为True。该参数表示：url中的页码是否存在数学运算。

若存在数学运算，请将值修改为False

如：百度贴吧url中的页码：第一页1或0，第二页50，第三页100。

https://tieba.baidu.com/f?kw=pubg&ie=utf-8&pn=50

若url页码是经过数学运算的，需修改参数：normal\_page\_number = False，

my\_requests.auto\_requests(folder\_name = XXX, url = XXX, normal\_page\_number = False)

url\_numbers 默认值为空值。该参数表示：以列表或生成器形式表示的经数学运算后的url中的页码。

若url页码需要数学运算，请将数学运算后的页码以列表或生成器形式传入模块。示例如下：

百度贴吧url：https://tieba.baidu.com/f?kw=pubg&ie=utf-8&pn=50 第一页为1或0，第二页为50，第三页为100。

列表形式：

x = [(i-1) \* 50 for i in range(start\_page, end\_page + 1)]

my\_requests.auto\_requests(folder\_name = XXX, url = XXX, normal\_page\_number = False, url\_numbers = x)

生成器形式：

x = ((i-1) \* 50 for i in range(start\_page, end\_page + 1)) 注意：生成器形式最外层是圆括号，列表形式最外层是方括号。

my\_requests.auto\_requests(folder\_name = XXX, url = XXX, normal\_page\_number = False, url\_numbers = x)

url\_rear 默认值为' / '。该参数表示：url后半部分，即，页码数字之后的部分。默认为"/"，若不为“/”，则可根据需求设定。

如人民网网址： it.people.com.cn/index2.html

其中：index后的数字‘2’表示页码，页码后的部分为：‘.html’。故，这里，需设定：url\_rear = '.html'。

完整代码：

url = 'it.people.com.cn/index2'

rear = '.html'

my\_requests.auto\_requests(folder\_name = XXX, url = url, url\_rear = rear)

nap 默认值 = 0。该参数表示单页休眠时间（单位：秒）： 即，若需下载多页，每页之间需间隔几秒后再进行下一页的下载。

若需添加每页下载后休眠30秒：

my\_requests.auto\_requests(folder\_name = XXX, url = XXX, nap = 30)

sleep 默认值 = 0 。该参数表示批次休眠时间（单位：秒）：若需多次下载某些页面，该参数表示，每完整下载完一个任务后需要间隔多久再进行下一次下载。

示例如下：

若想每5分钟下载一次1至5页，则sleep = 300, 即5\*60=300

my\_requests.auto\_requests(folder\_name = XXX, url = XXX, sleep = 300)

注意：用户若想循环多次下载网页必须将sleep修改为非零值，此程序禁止用户连续无休止爬取网页。

ini\_count 默认值 = 0。该参数为子目录文件夹（sub\_folder）的编号数字。

此模块下载目录结构如下：

folder\_name:

sub\_folder\_name\_1:

file\_1

file\_2

sub\_folder\_name\_2:

file\_1

file\_2

single\_folder 默认值为True。该参数表示：用户是否将所有下载文件保存在同一根目录下（即，不创建任何以时间戳命名的子目录）。

单一文件目录模式下，程序会自动检查即将下载的文件是否已经存在，若存在，则不会重复下载。若用户想对同一页面下载多次，程序将以时间戳区分不同批次的下载。

若用户想通过以不同时间戳命名的子目录保存文件，请将此参数修改为False。

file\_type 默认保存格式为html。该参数表示：下载页面保存的文件格式。若用户想保存为其他格式，请以字符串形式提交文件格式名。

my\_requests.auto\_requests('coinmktcap',url,file\_type = '.json')

备注：带星号（\*）参数为必填参数，其余参数为选填参数。

教程：

该模块所有参数一览：

auto\_requests(folder\_name, url, asyn=False, start\_page = '', end\_page = '', normal\_page\_number = True, url\_numbers='', url\_rear = '/',

nap = 0, sleep = 0, ini\_count = 0, single\_folder = True, file\_type = ‘.html’)

单页下载示例：（该页面下载一次后程序结束）

from my\_module import my\_requests

url = 'https://coinmarketcap.com/'

my\_requests.auto\_requests('coinmktcap',url)

单页多次下载示例：（程序将每隔30秒下载一次url对应的网页）

from my\_module import my\_requests

#全部存放在一个同一根目录下方法：

url = 'https://coinmarketcap.com/'

my\_requests.auto\_requests('coinmktcap',url,sleep = 30)

#创建以时间戳命名的子目录，并将下载文件分别放置于子目录中的方法：

my\_requests.auto\_requests('coinmktcap',url,sleep = 30,single\_folder = False)

常规页面多页多次下载示例：（程序将每隔30秒下载一次1-3页，每一页之间并无间歇）

from my\_module import my\_requests

url = 'https://coinmarketcap.com/'

my\_requests.auto\_requests('coinmktcap', url, start\_page = 1, end\_page = 3, sleep = 30,single\_folder = False)

非常规url页码编号网页的多页多次下载示例：

from my\_module import my\_requests

url = 'https://tieba.baidu.com/f?kw=pubg&ie=utf-8&pn='

start\_page = 1

end\_page =3

x = ((i-1) \* 50 for i in range(start\_page, end\_page + 1))

#上面使用的是生成器方式。若用户想以列表形式传入url中的页码：x = [i \* 50 for i in range(start\_page, end\_page + 1)]

my\_requests.auto\_requests('tieba', url, start\_page = 1, end\_page = 3, normal\_page\_number = False, url\_numbers = x, sleep = 10,single\_folder = False)

异步加载网站的下载示例：（用户无需手动在浏览器中查找json文件，只需将asyn参数设置为True即可。）

from my\_module import my\_requests

url = 'https://tieba.baidu.com/f?kw=pubg&ie=utf-8&pn='

start\_page = 1

end\_page =3

x = ((i-1) \* 50 for i in range(start\_page, end\_page + 1))

my\_requests.auto\_requests('tieba', url, asyn = True, start\_page = 1, end\_page = 3, normal\_page\_number = False, url\_numbers = x, sleep = 10,single\_folder = False)