

题目：

从上述网址中爬取《动物农场》所有章节的网址，再通过一个多线程爬虫将每一章的内容爬取下来。在本地创建一个“动物农场”文件夹，并将小说中的每一章分别保存到这个文件夹中。

代码：

```
import re
import requests
import os
from multiprocessing.dummy import Pool

#新建文件夹函数
def mkdir(path):
    folder = os.path.exists(path)
    if not folder:
        os.makedirs(path)
        print("Ok")
    else:
        print("the file is exit")

#爬取内容存入文件夹里
def query(url):
    url1 = "https://www.kanunu8.com/book3/6879/"
    chapter = requests.get(url1+url).content.decode('GBK')
    title1 = re.search('(.*)_动物庄园',chapter).group()
    Title = re.sub(" ", "", title1)
    content = re.findall('<[pbr /]{1,5}>\r\n\u3000\u3000(.*)<',chapter)
    filename = file+'\\'+Title+'.txt'
    print(filename)
    with open(filename, 'w', encoding='utf-8') as f:
        for i in range(0,len(content)):
            f.write(content[i])
            f.write("\n")

#创建文件夹
file = '动物庄园'
```

```
mkdir(file)
```

```
url = "https://www.kanunu8.com/book3/6879/"
```

```
html = requests.get(url).content.decode('GBK')
```

```
list1 = re.findall('<td width="25%"><a href="(.*?)">',html)
```

```
list2 = re.findall('<td><a href="(.*?)">',html)
```

```
url_list = list1+list2
```

```
print(url_list)
```

#多线程

```
pool = Pool(4)
```

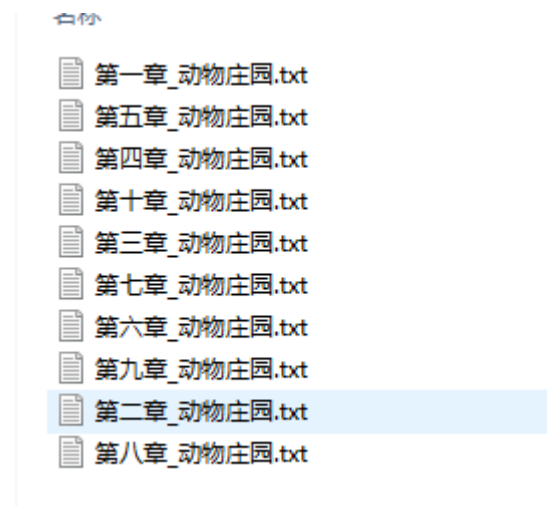
```
pool.map(query,url_list)
```

```
print("爬虫完成")
```

截图：



```
Ok
['131779.html', '131780.html', '131781.html', '131782.html', '131783.html', '131784.html', '131785.html', '131786.html', '131787.html', '131788.html']
动物庄园\第一章_动物庄园.txt
动物庄园\第三章_动物庄园.txt
动物庄园\第二章_动物庄园.txt
动物庄园\第四章_动物庄园.txt
动物庄园\第五章_动物庄园.txt
动物庄园\第九章_动物庄园.txt
动物庄园\第七章_动物庄园.txt
动物庄园\第八章_动物庄园.txt
动物庄园\第十章_动物庄园.txt
动物庄园\第六章_动物庄园.txt
爬虫完成
*** Repl Closed ***
```



```
📁
📄 第一章_动物庄园.txt
📄 第五章_动物庄园.txt
📄 第四章_动物庄园.txt
📄 第十章_动物庄园.txt
📄 第三章_动物庄园.txt
📄 第七章_动物庄园.txt
📄 第六章_动物庄园.txt
📄 第九章_动物庄园.txt
📄 第二章_动物庄园.txt
📄 第八章_动物庄园.txt
```