

# 网络爬虫——小说下载

## 使用python对小说进行下载：

(本次实战网站：<https://www.biqubao.com>笔趣阁)

这是一个盗版小说网站，只能提供在线浏览小说，不支持打包下载，本次实战就是要通过爬虫技术把一本《斗破苍穹》下载下来，仅供学习，支持正版。

## 对于目录的获取：

首先我们打开笔趣阁的网站，搜索到《斗破苍穹》的网页：

笔趣阁 > 玄幻小说 > 斗破苍穹最新章节列表



### 斗破苍穹

作者：天蚕土豆

状态：连载中, 加入书架, 直达底部

最后更新：2017-02-22 15:16:41

最新章节：第一章 五帝破空

这里是属于斗气的世界，没有花俏艳丽的魔法，有的，仅仅是繁衍到巅峰的斗气！  
新书等级制度：斗者，斗师，大斗师，斗灵，斗王，斗皇，斗宗，斗尊，斗圣，斗帝。

各位书友要是觉得《斗破苍穹》还不错的话请不要忘记向您QQ群和微博里的朋友推荐哦！

推荐阅读：极品仙帝在花都 不灭战神 绝世大神豪 唐砖 咫尺青风剑 三界红包群 农绣 宦臣为后 无限恐怖网 万兽自然

《斗破苍穹》正文		
第一章 陨落的天才	第二章 斗气大陆	第三章 客人
第四章 云岚宗	第五章 聚气散	第六章 炼药师
第七章 休！	第八章 神秘的老者	第九章 药老！
第十章 借钱	第十一章 坊市	第十二章 离他远点
第十三章 黑铁片	第十四章 吸掌	第十五章 修炼
第十六章 萧宁	第十七章 冲突	第十八章 玄阶高级斗技：八极崩
第十九章 残酷训练	第二十章 拍卖	第二十一章 二品炼药师谷尼
第二十二章 风卷决	第二十三章 争抢	第二十四章 一切待续

可以看到我们如果想要把整本小说下载下来，就要从这个网页着手，获取它的目录，从它的目录获取每一章的跳转链接，进而跳到那个链接去下载文本。所以首先我们分析这个网页的html，从中找到目录的跳转链接：

## 按F12：

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>...</head>
  <body>
    <div id="wrapper">
      <script>login();</script>
      <div class="ywtop">...</div>
      <div class="header">...</div>
      <div class="nav">...</div>
      <div class="box_con">...</div>
      <script>listindex();</script>
      <div class="box_con">
        <div id="list">
          <dl>
            <dt>《斗破苍穹》正文</dt>
            <dd>
              <a href="/book/13991/6262303.html">第一章 陨落的天才</a>
            </dd>
            <dd>
              <a href="/book/13991/6262304.html">第二章 斗气大陆</a>
            </dd>
            <dd>
              <a href="/book/13991/6262305.html">第三章 客人</a>
            </dd>
            <dd>...</dd>
            <dd>...</dd>
            <dd>...</dd>
            <dd>...</dd>
            <dd>...</dd>
            <dd>...</dd>
          </dl>
        </div>
      </div>
    </div>
  </body>
</html>
```

可以看到我们对应的目录在

的下分目录中，也就是说我们需要找到id是list的div目录，它的 $dl \rightarrow dt$ 目录中出现了小说的名字，而 $dl \rightarrow dd \rightarrow a$ 中存在我们需要的章节链接，因此我们需要对这个链接进行获取，之后把它加个总网站头就可以跳转到小说文本网页，我们接下来观察文本网页：

山崖之颠，萧炎斜躺在草地之上，嘴中叼中一根青草，微微嚼动，任由那淡淡的苦涩在嘴中弥漫开来...

举起有些白皙的手掌，挡在眼前，目光透过手指缝隙，遥望着天空上那轮巨大的银月。

“唉...”想起下午的测试，萧炎轻叹了一口气，懒懒的抽回手掌，双手枕着脑袋，眼神有些恍惚...

“十五年了呢...”低低的自喃声，忽然毫无边际的从少年嘴中轻吐了出来。

在萧炎的心中，有一个仅有他自己知道的秘密：他并不是这个世界的人，或者说，萧炎的灵魂，并不属于这个世界，他来自一个名叫地球的蔚蓝星球，至于为什么会来到这里，这种离奇经过，他也无法解释，不过在生活了一段时间之后，他还是后知后觉的明白了过来：他穿越了！

看到它的html:

```
▶ <div class="con_top">...</div>
▶ <div class="bookname">...</div>
▶ <div style="text-align: center;">...</div>
▲ <div id="content">
    月如银盘，漫天繁星。笔『趣阁』阁Ww』W. 0B i Q u G e . C N
    <br />
    <br />
    山崖之颠，萧炎斜躺在草地之上，嘴中叼中一根青草，微微嚼动，任由那淡淡的苦涩在嘴中
    弥漫开来...
    <br />
    <br />
    举起有些白皙的手掌，挡在眼前，目光透过手指缝隙，遥望着天空上那轮巨大的银月。
    <br />
    <br />
    “唉...”想起下午的测试，萧炎轻叹了一口气，懒懒的抽回手掌，双手枕着脑袋，眼神有些恍
    惚...
    <br />
    <br />
    “十五年了呢...”低低的自喃声，忽然毫无边际的从少年嘴中轻吐了出来。
    <br />
    <br />
    在萧炎的心中，有一个仅有他自己知道的秘密：他并不是这个世界的人，或者说，萧炎的灵魂，并不属于这个世界，他来自一个名叫地球的蔚蓝星球，至于为什么会来到这里，这种离
    奇经过，他也无法解释，不过在生活了一段时间之后，他还是后知后觉的明白了过来：他穿
    越了！
    ...
```

可以看到文本在

的目录中，也就是说，我们只需要找到id是content的目录就可以获取小说文本，剩下来就是对于一些格式的处理还有系统文件夹的基本处理了：

```

import requests
from bs4 import BeautifulSoup
import os

#小说列表网页以及主网页
target='https://www.biqubao.com/book/13991/'
server='https://www.biqubao.com'

#注意转化编码
req=requests.get(url=target)
req.encoding='gbk'
html=req.text
#找到list
div=BeautifulSoup(html,"html.parser")
list_tag=div.div(id='list')
#小说名
title=list_tag[0].dl.dt.string
#目标文件夹
save_path='F:/Python/novel/new'
dir_path=save_path+'/'+title
if not os.path.exists(dir_path):
    os.path.join(save_path,title)
    os.mkdir(dir_path)

for dd_tag in list_tag[0].dl.find_all('dd'):
    #章节名字
    chapter_name=dd_tag.string
    #章节网址
    chapter_url=server+dd_tag.a.get('href')
    c_req=requests.get(url=chapter_url)
    c_req.encoding='gbk'
    c_soup=BeautifulSoup(c_req.text,"html.parser")
    content_tag=c_soup.div.find(id='content')
    content_text=str(content_tag.text.replace('\xa0','\n'))

    with open (dir_path+'/'+chapter_name[:]+' .txt','w') as f:
        f.write(content_text)

```

虽然下载的速度有点慢，但是行得通，我们暂时就已经可以说入门了。