爬取三国演义的文本内容（对bs4的练习

笔记本: reptile_draft
创建时间: 2021/7/3 15:45 **更新时间:** 2021/7/3 15:45
作者: 134exetj717
URL: about:blank

```
# -*- codeing = utf-8 -*-
# @Time : 2021/7/3 13:54
# @Author : 希杰
# @file : bs4解析基础.py
# @software : PyCharm


import lxml
import requests
from bs4 import BeautifulSoup
import os
#需求：爬取三国演义小说的所有章节标题和章节内容
if __name__ == '__main__':
    #对首页的页面数据进行爬取
    url = 'https://www.shicimingju.com/book/sanguoyanyi.html'
    headers = {
        'User-Agent': 'Mozilla / 5.0(Windows NT 10.0;  Win64;  x64; rv: 89.0)
Gecko / 20100101  Firefox / 89.0'
    }
    page_text = requests.get(url=url,headers=headers).content   #content 表示中文
字符串
    #在首页中解析出章节的标题和详情页的url
    #1.实例化BeautifulSoup对象，需要将页面源码数据加载到该对象中
    soup = BeautifulSoup(page_text,'lxml')
    #解析章节标题和详情页的url
    li_list = soup.select('.book-mulu > ul >li')
    fp = open('./sanguo.txt','w',encoding='utf-8')
    for li in li_list:
        title = li.a.string
        detail_url = 'https://www.shicimingju.com'+li.a['href']
        #对详情页发起请求、解析出章节内容
        detail_page_text = requests.get(url=detail_url,headers=headers).content
        #解析出详情页中相关的章节内容
        detail_soup = BeautifulSoup(detail_page_text,'lxml')
        #有些 class不能看作属性，我们不能写成class=''，不然会报错，应该用attrs，像下面
这样
        div_tag = detail_soup.find('div',attrs={'class':'chapter_content'})
        #解析到了章节的内容
        content = div_tag.text  #若为.text表示内容字符串，但如果是contents，表示对应
的列表
        if not os.path.exists('./sanguo.txt'):
            os.mkdir('./sanguo.txt')
        fp.write(title+':'+content+'\n')
        print(title,'爬取成功')
```