```
# -*- coding:utf-8 -*-
import time,os,traceback,random
import requests,re
from bs4 import BeautifulSoup

Agent =['Firefox'
        ,'Chrome/10'
        ,'Mozilla/5.0'
        ]
def ProcName(name):
    pat = r'[<|>|/|\|||:|"|*|?]+'
    pat = re.compile(pat)
    return pat.sub('',name)
def GetHtmlText(url):#此处可以增加 proxies 代理服务器，只不过目前还没有
    try:
        r = requests.get(url,headers={'User-Agent':Agent[random.randint(0,Agent.__len__()-1)]},timeout = 20)
        r.raise_for_status()
        r.encoding = 'utf-8'
        return r.text
    except:
        traceback.print_exc()
```

```
        return None


def FindIndex(name):#返回目标小说目录页
    url =
"http://zhannei.baidu.com/cse/search?&s=2872930369481595 15&q="+str(name
)+"&click="+str(random.randint(1,3))+'&nsid='
    text = GetHtmlText(url)
    if text==None:
        print("文本为空，无法解析")
        exit()
    soup = BeautifulSoup(text.encode('utf-8'),'html.parser')
    list = soup.find_all(name = 'a',attrs = {"cpos" :"title","title":name})#一个 list
    url = []
    for i in list:
        url.append(i["href"])
    return url


def Write(storpath,tag):#传入存储路径和标签即可
    if None==tag:
        print('小说写入失败，原因是小说最后一层超链接无法获取')
    a = tag.a#标签的属性使用 tag['title']来获得，标签下的搜索使用 tag.children
来实现
```

```
        storpath += "/"+ProcName(a.string)+".txt"
        if os.path.exists(storpath) and os.path(storpath).ST_SIZE>0:
            return 1
        url = 'http://www.biquge.com/'+a['href']
        text = GetHtmlText(url)
        if text==None:
            print("最后一层文本获取失败!")
        soup = BeautifulSoup(text.encode('utf-8'),'html5lib')
        novel = soup.find_all("div",attrs={"id":"content"})
        text = novel[0].text
        #开始写入
        with open(storpath,'w',encoding='utf-8') as f:
            f.write(str(text))
            f.close()
        return 0

def Spider(url,path):#爬取并且存储

        text = GetHtmlText(url)
        if text == None:
            print("文本为空，无法解析")
            exit()
```

```
    num = 0
    soup = BeautifulSoup(text,'html5lib')
    list = soup.find_all(["dd","dt"])
    nowpath = ""
    flag = 0
    for i in list:
        if i.name == "dt":
            nowpath = path+"/"+ProcName(i.string)
            if os.path.exists(nowpath):
                pass
            else:
                os.mkdir(nowpath)
        else:
            flag = Write(nowpath,i)
            num+=1
            if num%10==0 and (not flag):
                time.sleep(random.randint(3,15))#爬虫每爬几个就休眠
        print("\r 当前进度: {:.2f}%".format(num * 100 / len(list)), end="")
    return ""


def main():
    name = ProcName(input("请输入要爬取的小说的名字:"))
```

```
    url = FindIndex(name)#爬取搜索结果，在其中查找目录页，并且返回
    if 0==len(url):
        print("查无此小说")
        exit()
    path = "E:/小说/"
    if os.path.exists(path):
        pass
    else:
        os.mkdir(path)
    path = "E:/小说/" +name
    if os.path.exists(path):
        pass
    else:
        os.mkdir(path)
    Spider(url[0],path)
main()
```