**Xpath**

之前说过，xpath是我最喜欢的，python 具有一些比较流行的解析库,例如 lxml , 使用的是 XPath 语法，是大众普遍认为的网页文本信息提取的爬虫利器之一。我也这么认为。

先来介绍一下xpath，其实他就是一个提取内容的东西，正则也可以提取内容，完成相同的工作，实现的功能也差不多，但是xpath明显比re更有优势，所以在网页分析上re也就退居二线，那究竟什么是xpath呢

XPath 是 XML路径语言（XML Path Language），支持 HTML，是一种用来确定XML文档中某部分位置的语言。XPath基于XML的树状结构，提供在数据结构树中查找节点的能力。Xpath 可以通过元素和属性进行导航，相比 正则表达式，它同样可以在 XML 文档中查询信息，甚至使用起来更加简单高效。

这是比较官方的话，我通常不喜欢这么介绍，通俗点讲就是我们获取网页内容，有些是我们需要的，有些是不需要的，我们要通过什么方法获取我们需要的内容呢？这个时候就需要xpath。

他有很多优点，让他脱颖而出：

1） 可在XML中查找信息   
2） 支持HTML的查找   
3） 通过元素和属性进行导航

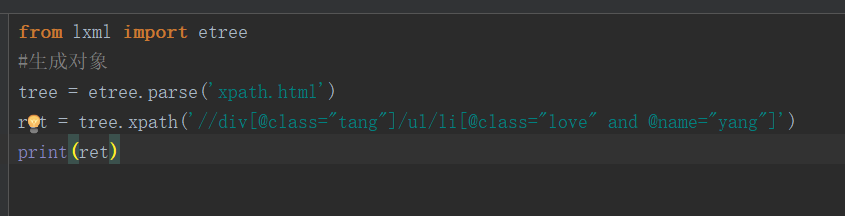
因为xpath属于lxml库模块，所以在使用xpath之前，我们要先安装库lxml，直接在cmd里面使用pip命令就可以

首先讲一下XPath的基本语法知识：   
四种标签的使用方法   
1) // 双斜杠 定位根节点，会对全文进行扫描，在文档中选取所有符合条件的内容，以列表的形式返回。   
2) / 单斜杠 寻找当前标签路径的下一层路径标签或者对当前路标签内容进行操作   
3) /text() 获取当前路径下的文本内容   
4) /@xxxx 提取当前路径下标签的属性值   
5) | 可选符 使用|可选取若干个路径 如//p | //div 即在当前路径下选取所有符合条件的p标签和div标签。   
6) . 点 用来选取当前节点   
7) .. 双点 选取当前节点的父节点

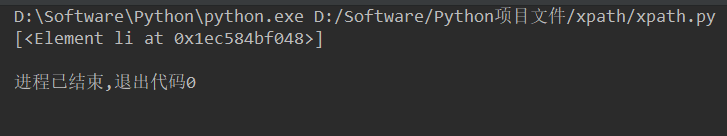
光看没用，我写了一个html我们来练习一下：



我们要来提取这里面的古诗内容，但是他在一个一个的节点里面，这时候我们就要用到xpath



其实代码不多，我们就是首先导入库，然后生成对象，我们需要提取的内容在哪里，我们就写哪里，然后用xpath获取，就是我们刚才说的用法，最后打印出来，我们来看看运行结果：

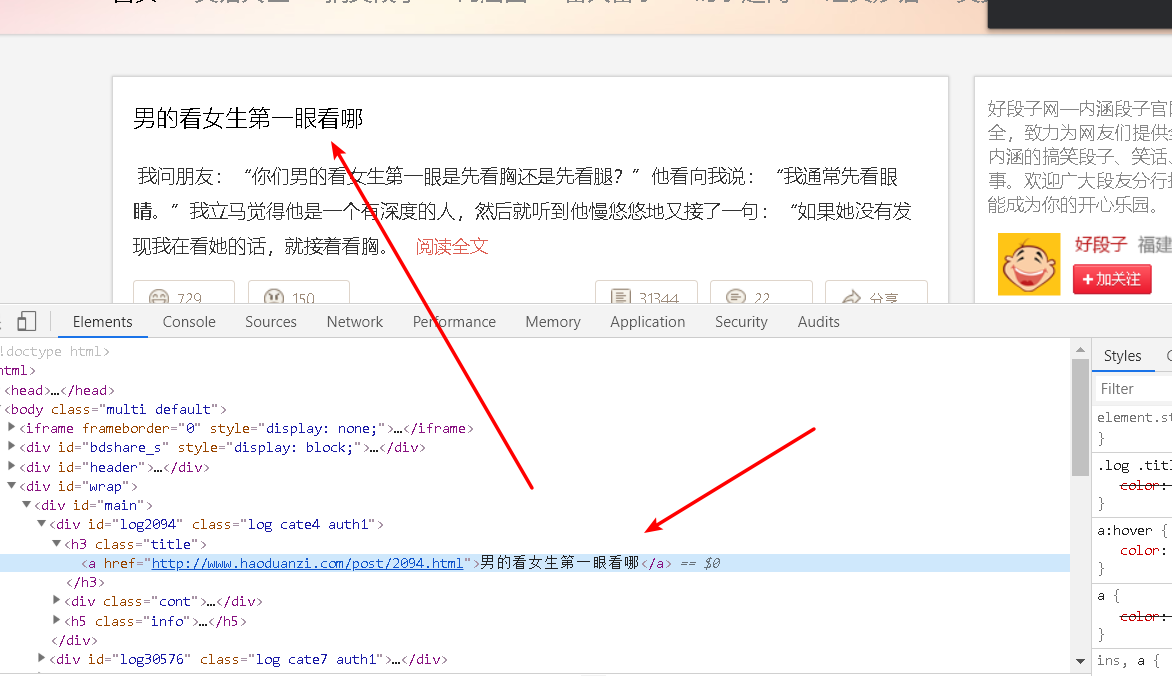


成功获得地址

我们写一个实例练练手，我们爬取的是好段子网站

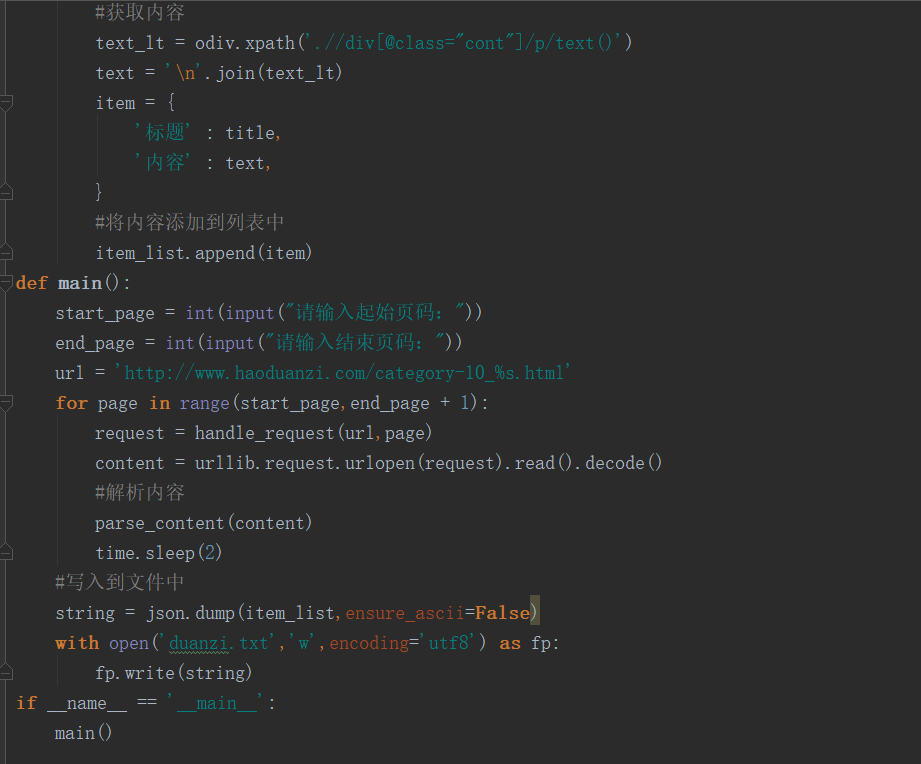
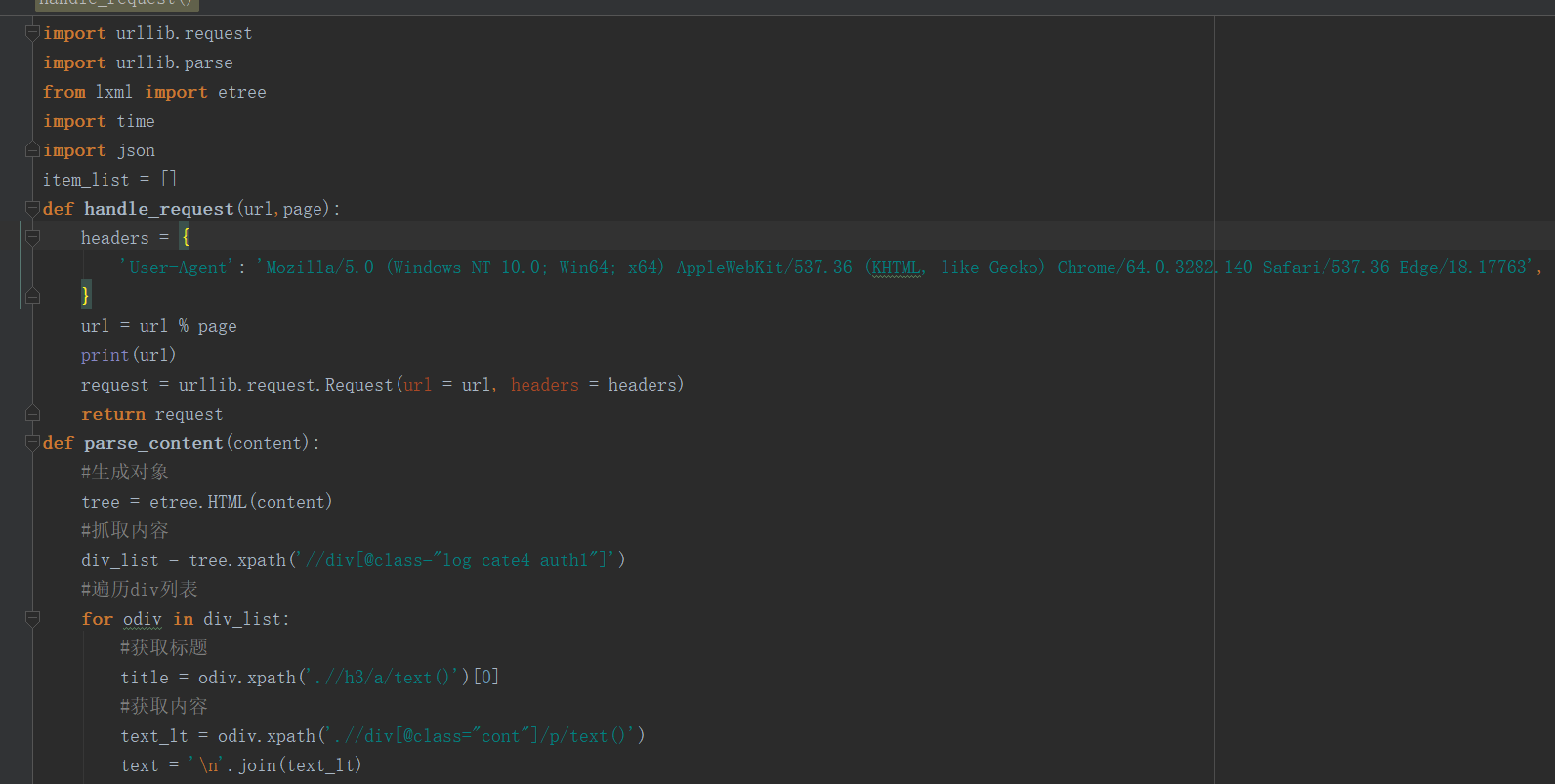


，我们需要获取标题和内容，并且实现爬取多少页自定的内容，我们先来看标题和内容都在网站的哪里

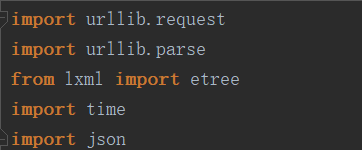
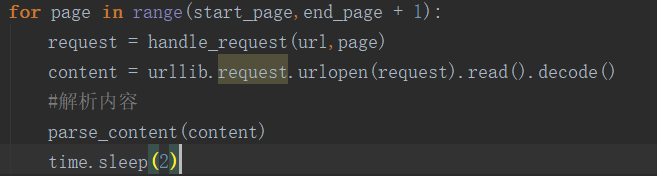
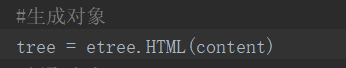
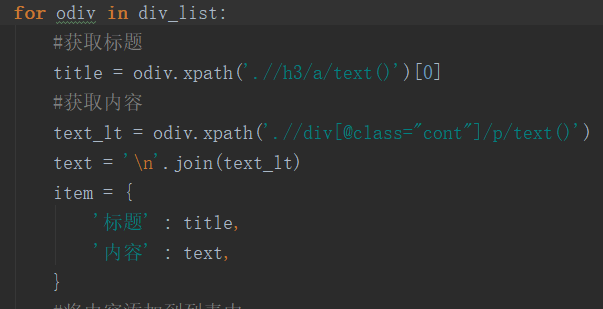
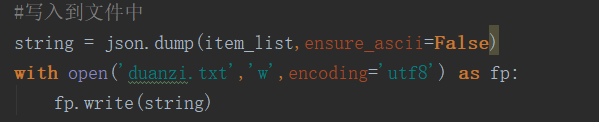




我们已经找到这些内容，那我们就用xpath获取，我们直接上代码：

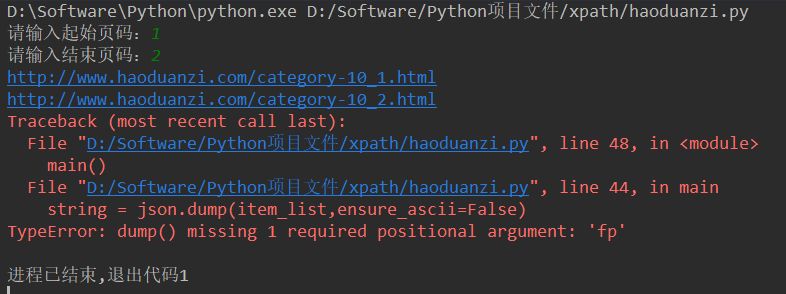


分析代码：

* 首先我们导入我们需要的库
* 
* 根据以往的经验我们先定义一个字典，用来存放我们所爬取的数据
* 我们先来定义主函数，这里是我们爬虫的主程序，我们需要实现用户手动输入页码，所以先input然后输入url地址
* 我们就用一个for循环，来遍历我们的内容，在这个for循环里面，我们和往常一样，发送请求，获取响应内容
* 
* 然后我们就需要解析内容，这个时候我们新建一个类，用来单独解析内容
* 对了，刚才忘说了，在这个for循环里面发送请求内容的时候，我们还需要新建一个类，这个类单独用来发送请求
* 
* 把我们用户输入的页码和url当作参数传进去
* 先来到发送请求这个类
* 在这里面还是老规矩，请求头，经过我们分析url后，我们需要拼接url，才能达到我们爬取的页数
* 所以在这里我们拼接一下url
* 
* 然后发送请求，最后把我们的请求返回到主函数
* 在主函数接收后，我们获取到了响应内容，这个时候我们再来解析内容
* 我们来到解析内容的函数
* 和刚才一样，我们需要生成对象，这里有一点，刚才我们获取本地内容，用的是etree.parse，在获取网页内容的时候，我们需要用到，etree.HTML
* 
* 然后我们就开始抓取内容，先获取大头部
* 然后我们遍历这个大头部
* 
* 在这里我们获取标题和内容，经过我们分析后，我们写出xpath，获取到了内容，然后我们定义一个字典，用来存储我们的内容，最后将内容添加到我们一开始定义的列表里
* 然后这个列表已经有我们的数据了，其实我们也就爬取到了，最后我们写入到文件中，保存到本地
* 

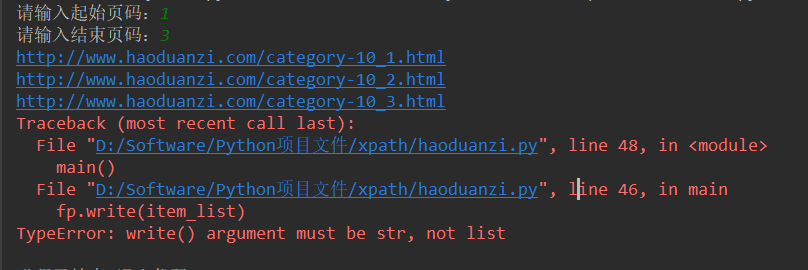
我们来看看运行结果：

哦，我们遇到一个错误：

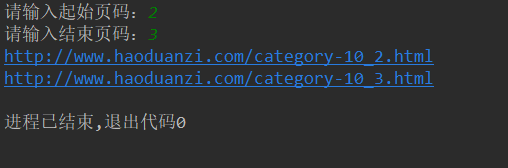


他说我们把数据转换成json的时候报错了，我们使用的是dump，原来我们应该使用dumps，我们再来试试

又遇到问题了==



很简单的问题，python写入的内容是要字符串类型的修改，再来

 代码成功运行，没有报错，但是，是一个空列表，我们刚才明明已经获取到了内容啊，后来我又看了看网站，404 not found。。。。。。真尴尬，网站崩了，但是xpath刚刚学习开始，具体用法我们已经了解了，在以后的爬虫里，我每一个基本都用的xpath，所以以后希望越用越精通吧，其实我觉得所有的到现在的学习都是在为Scrapy做铺垫，哈哈，我觉得他最好用了，好了，先到这里！