**JsonPath**

今天学习的是jsonpath，首先要明白什么是json，他就是一种数据交换格式，它既让人们很容易的进行阅读和编写，同时也方便了机器进行解析和生成，适用于进行数据交互的场景，比如网站前台与后台之间的数据交互，淘宝就是这样的，后面我们会爬取淘宝的内容

Python从2.7开始就自带了json模块，我是3.7的，直接import json就可以使用了，我们经常会用到一个网站，就是json在线解析网站：<https://www.json.cn/#>

经过先前的学习，我知道了，json模块提供了四个功能，dumps，dump，loads，load，看起来很像，但用法和用途不一样，他们都是用于字符串和python数据类型键进行转换，我们一个个看，然后用代码理解。

1、json.loads()：

从json到python的转化

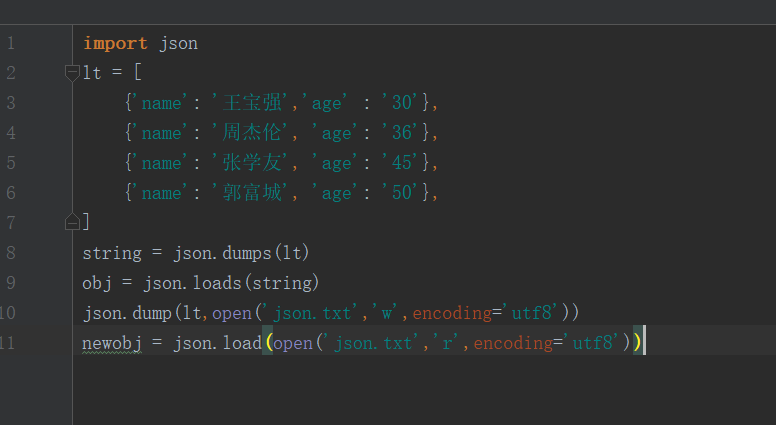
2、json.dumps()：

从python到json的转化

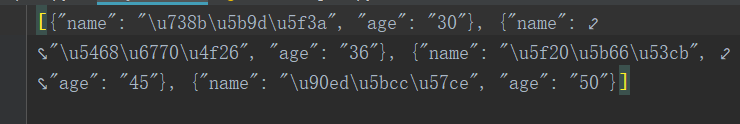
3、json.dump()：

将python内置类型序列化为json对象然后写入文件

4、json.load()：  
 读取文件中json格式的字符串元素转化成python类型

直接贴上联系代码，包括了所有类型：  
 

下面是运行结果：



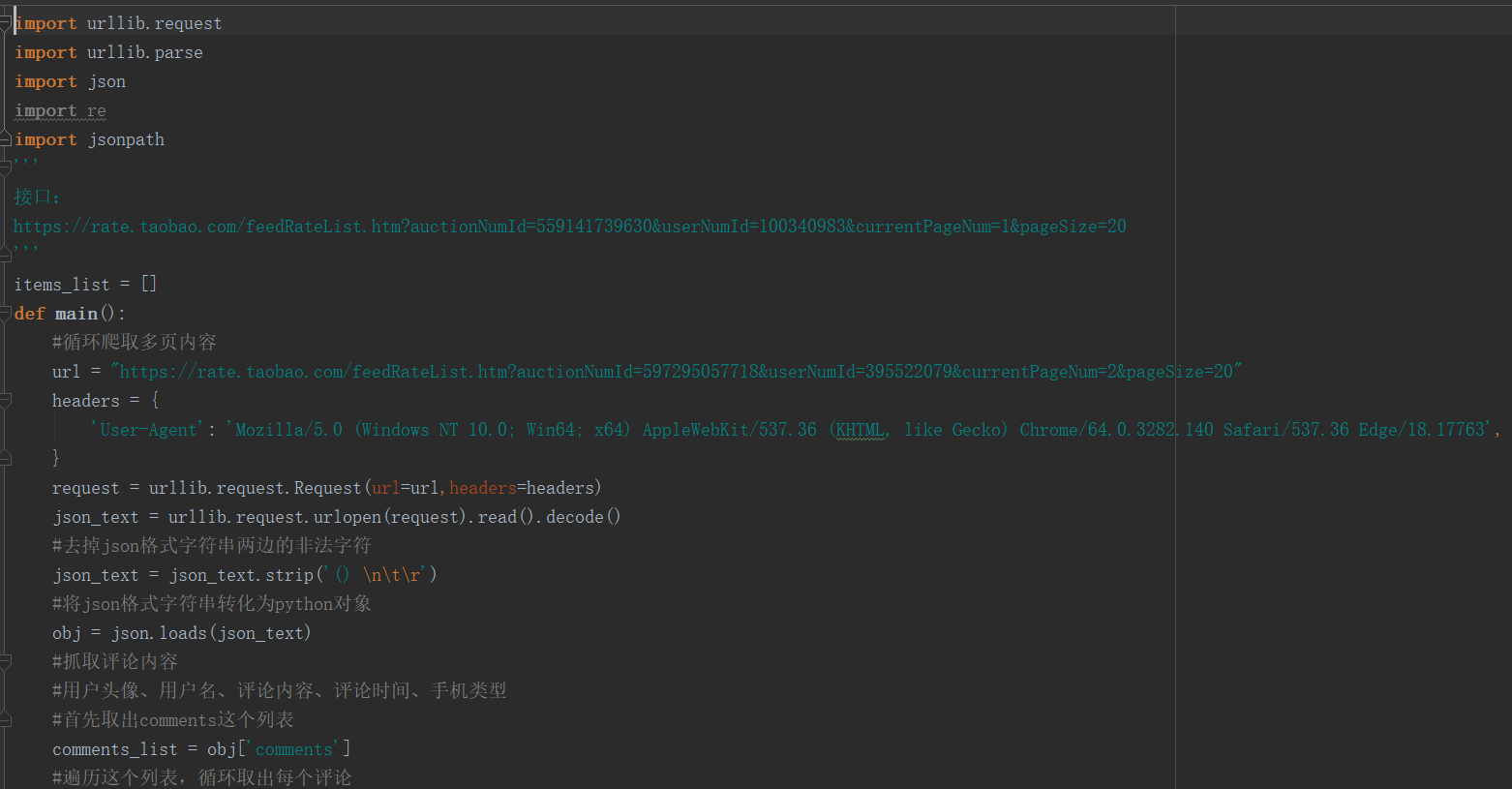
这个练习就实现了各个类型的转化，很方便

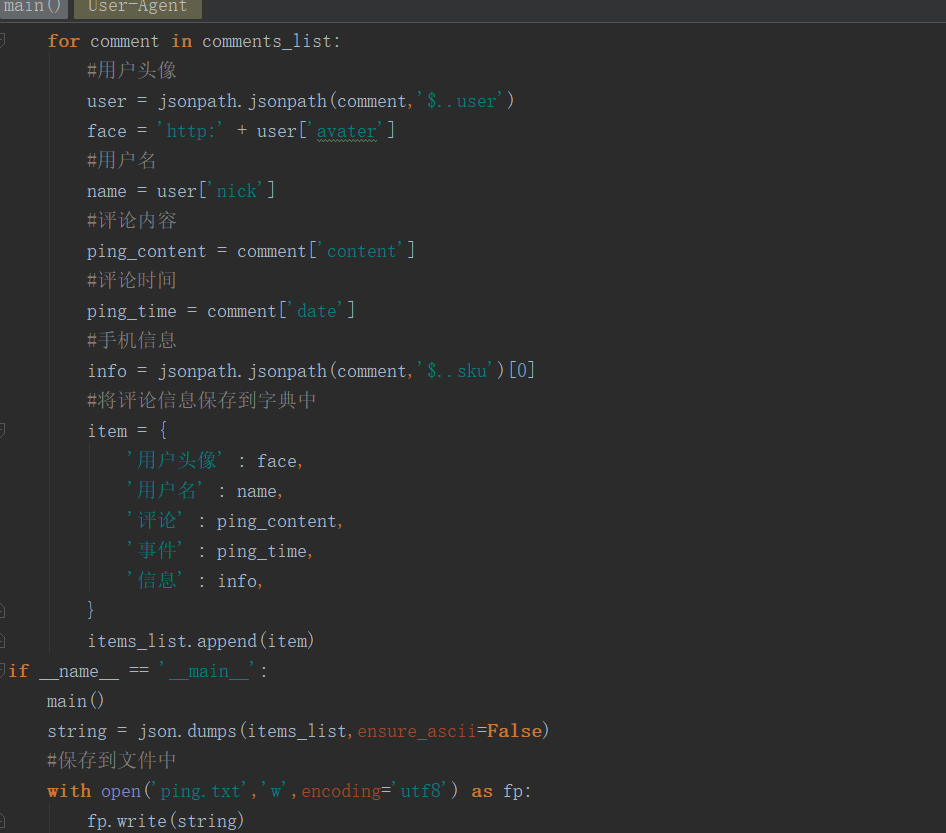
然后我们开始看jsonpath，jsonpath就是一种信息抽取类库，就是从json文档中抽取指定的工具，我们可以这么用，当我们在网上爬取json格式的数据的时候，太多了太杂了，我们就可以用jsonpath来提取我们想要的数据，其实后面我们会学到xpath，xpath是和xml紧紧的联系在一起的，可以这么说，jsonpath对于json来说，就相当于xpath对于xml，为了对比，我在网上找了一张他俩的对比图如下：  


我们可以看出来，json结构清晰，可读性高，复杂度低，非常容易匹配，优点还是不少的，但是我个人更倾向于使用xpath。

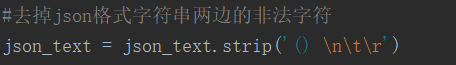
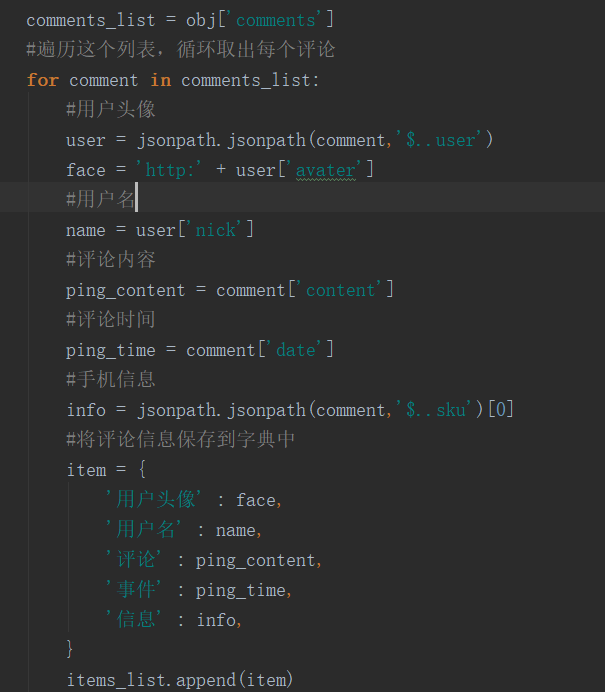
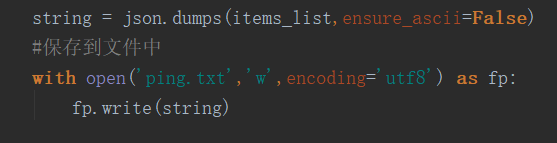
废话不多说，上实例，今天我们利用jsonpath爬取淘宝评论，先上代码，然后分析过程。

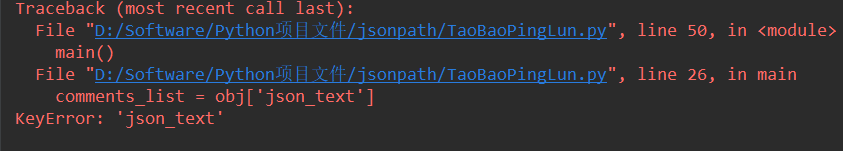
代码如下：





分析代码：

* 老规矩我们先导入库，既然使用json就导入json和jsonpath，还有一直在用的request和parse
* 然后我们来分析整个网页，本来我以为直接把url放上去就ok了，但是，通过分析发现，我们翻页的时候，上面的url地址不会变，这样我就想是不是ajax请求，那就抓包，经过抓包果然是ajax，我们获取到了接口：<https://rate.taobao.com/feedRateList.htm?auctionNumId=559141739630&userNumId=100340983&currentPageNum=1&pageSize=20>，我们可以看出来currentPageNum就是页码，PageSize就是页码里面有多少数据
* 那么好了老规矩，先定义一个主函数，在这里是我们爬虫主程序，估计这个程序不会太大，一条龙
* 我们先把url，也就是刚才的接口放上来，然后还是一样的请求头，构建请求，发送请求，这个时候不出意外我们已经获取到了内容
* 然后就是用jsonpath获取我们想要的数据了
* ，通过print内容我们发现有非法字符，那么我们就去掉
* 然后就是将json格式字符串转化为python对象
* 接下来就是抓取评论内容了，对了，在主函数前面要定义一个字典，用来存放我们的数据，我们首先要遍历这个列表，分别取出用户头像、用户名、评论内容、评论时间、手机类型
* 
* 然后将数据添加到字典里
* 最后就是保存到本地文件，在这里要注意，我们要用json格式保存
* 然后，大功告成！

总结：这个东西也是我第一次接触，类似bs4，也遇到不少问题，但是这个真的蛮好用的，有一次我列表里面没有数据我直接运行，结果报错了，典型的keyerror，所以以后一定要注意，另外还有注意的就是结束时候如果想以json格式保存数据，那么在dumps里面要传入两个参数，一个是我们的数据，另一个是固定的，ensure\_ascii=False，因为这里的中文中的ascii字符码，不是真正的中文，在dumps序列化时对中文默认使用的ascii编码，想要输出真正的中文需要ensure\_ascii=False。