**Python大作业个人文档**

楼思琦 5140379052

**代码部分：**抓取天猫商品信息

* 第三方库：bs4；urllib；json；requests；
* 实现效果：

输入是用户输入的商品名称，输出时一个包含商品list的json文件，每个商品中包含商品链接、商品名称、价格、评价、图片、销量六个信息。

* 实现方法：

1. 确定商品搜索网址。尝试搜索不同关键词，比较所得网址之间的异同，发现前面部分的形式是基本相同的，为“https://list.tmall.com/search\_product.htm?q=”+商品+“&click\_id=”+商品，后缀有较大差异，但是删去后缀后不影响搜索功能，所以拼出了搜索页面的URL。
2. 通过浏览器的“审查网页元素”功能，观察天猫网页的html元素信息。发现每个商品就是一个class=”product “的div，所以可以利用bs4的find和get[]功能抓取具体信息。
3. 所有信息抓取完后，利用json可以直接dump()输出dict的信息，把item类改成dict，其中的key分别是source,price,rate,name,pic,sale，其中pic是保存在本地的商品图片名称。

* 遇到的问题与解决方法：

1.拼接url时遇到中文的编码不对，导致url错误，无法抓取信息。原来是中文编码默认用的coding形式是cp936，而天猫的url用的编码用的是utf-8，所以加上encoding=utf-8即可。

2.刚开始尝试的是进入每一个商品的链接抓取具体信息，但是总是报Nontype can't find 的错误，后来查看网页源代码和审查网页元素进行对比后发现，许多在元素中出现的信息在源代码中找不到，说明并不是通过html直接呈现的，而是动态的，和后台有数据传送。然后回到主搜索页面进行对比，大部分信息都是静态的，所以选择直接抓取朱搜索页面的商品信息。但是虽然可以从主搜索页面找到评分页的链接跳转过去，但是评分的具体信息始终是动态的，所以只能选择采用默认值。

3.当销量较大时天猫会采用“1.3万笔”而不是“13000笔”的显示方式，只截取数字会获得错误信息。所以直接返回整个text。（尝试解析中文，判断“万”字后转换，但转换成数字会出现乱码，所以直接返回全部信息）

* 代码架构：

1.goods是3个conv的list

2.encodeName函数负责拼搜索页面的url

3. processor 负责抓取，其中调用了5个帮助函数get\_name, get\_picture, get\_rate, get\_price, get\_source分别抓取5块信息

4. save\_img\_to\_file负责把图片存到本地

5. Provider负责输出json

**文档部分：**小组文档的整理和撰写

**收获和感想：**

通过这次大作业，对用python写爬虫程序有了一定的了解，可以较熟练的运用reques和beautifulsoup等第三方库，对静态动态的网页也有所区分。而且由于是团队合作，在共同工作中学习和收获了很多，比如代码的规范性和对新工具如何开始学习运用等等。