## 实用编程期末项目报告

**姓名：刘云龙** **学号：20420182201980** **院系：化学化工学院**

**联系方式（QQ或电话）：13906040102**

|  |
| --- |
| 1. 项目标题   Python淘宝爬虫项目 |
| 1. 项目简介和功能描述（可附图说明，如需用到网络资源，请注明网址）   用python实现对任意淘宝商品的价格销量条件筛选，得到符合条件商品的链接并显示出来。输入有：1.要搜索的商品名称2.给出你能接受的价格区间（min，max）3.给出你要求最低的销量  然后进行淘宝登陆操作，扫取二维码  用爬虫爬取后，输出所有符合条件的商品销量，价格，以及链接。  用到的网络资源：  <https://blog.csdn.net/u010141562/article/details/67633211?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-2.nonecase&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-2.nonecase>  这是解决下滑滑动条问题（代码中down\_to\_bottom函数）时用到js脚本参考的csdn上提供的代码 |
| 1. 所用到的主要python模块，若用到其他软件，请注明所用软件的名称和版本号   selenium库  re模块  time模块 |
| 四．主要实现方法（包括：遇到的难点和解决方法，主要功能的实现方法等，也可以写些心得体会以及花费的总小时数等等。尽可能用文字和图形介绍，不要粘贴大量代码）  1.用chrome打开淘宝，遇到第一个难点，如何在输入框中输入你想要的商品。  查阅百度得知可以用sendkeys方法填写，然后用click方法点击搜索解决问题。  2.遇到登陆问题，尝试用账号密码登陆，但需要验证码。解决方法人为操作，用手机扫描二维码登陆。  3.接下来就是采集数据，翻页，如此循环。但遇到新问题，循环次数如何获取。通过正则在网页源码中/提取到了“总页数”中的数字，解决问题。  4.点击click下一页出现问题，翻页到一定页数开始需要填入验证码。搜索得知可以通过修改url后面的currentpage（淘宝url上s参数）以实现翻页，不需验证码，经验证，不论搜索什么商品，s变化都是44，第一页是0，第二页是44，依次递增。通过修改url实现翻页问题。  5.爬取数据发现每一页都只取得一个商品的数据。猜想是代码有bug或者是网速太慢没有加载出来，于是用time的sleep睡了10s仍然没有改观，确定是代码错误。发现xpath路径在这并不能直接copy而是需要人为写，不然复制后div[1]中的数字用find\_elements时永远只采集第一个数据，没有办法遍历整页所有商品数据。用xpath改写为采取所有class属性满足是商品的class属性（class=item J\_MouserOnverReq），解决遍历问题。  6.不设条件发现爬取数据只有每页近10个商品可获取信息，搜索得知必须下拉滚动条使之加载完成方能获取数据，所以不加载只能获取上面加载出来的数据，只有十几个。  参考csdn用js脚本实现了滑动条下拉，解决问题。  7.由此可以获取所有淘宝搜索商品的信息，下面进行信息筛选。对提取的销量（x人付款）和价格用正则提取出来，与输入的PriceMin,PriceMax，SalesVolumeMin进行比较，符合条件写入字典供输出，并把信息存在一个列表中。 |
| 1. 所开发程序的使用方法 2. 输入你要搜的商品，价格区间以及对销量的最低要求   2.手动登陆  3.查看爬取结果：符合条件的商品价格销量和对应url |
| 1. 意见和建议   我觉得老师的课做的很好，没有什么问题。上这门课最大的收货就是有了把编程应用到生活中的意识，进一步坚定了我转专业的决定。 |

**签名：刘云龙**

**日期：2020.5.27**

**说明：**

1. 此报告必须填写，未完成报告者没有分数。
2. 提交时需将报告，程序文件和相关资源（如音乐，图片等）压缩成一个zip文件（文件名: xxxx.zip，xxxx为提交人的学号），并将该文件作为附件发送到我的电子邮箱：[chongzhao2014@163.com](mailto:chongzhao2014@163.com)，邮件标题为《期末报告》，内容为提交人的姓名和学号。**请严格按要求发送邮件，否则统计时有可能会被遗漏。**
3. **第16周报告时需提交此报告的纸质版并签名**以备存档。电子档和程序须在本学期结束（**6月20日）**前发给我。