两个任务的想法

201805301351 软工 3班 孙昊

假期里根据学长们推荐的书和自己找的网上的资料，对python语言有了一些了解，也不知道讲啥，就讲一下我学习的过程和做任务的过程吧

最起初边理解边抄了一个自定义的log,为了自己更好的理解log的作用和用法（深入的东西并没有理解多好）。之前学安卓和java的时候也都用到过，相当于是复习了一下（就是怎么用。。。。在构造上并没有下功夫去了解。。惭愧。。）。

首先是网络爬虫，刚开始一点儿头绪没有，我买的书和从网上看的python的教学资料并没有写相关的东西。我去csdn上搜python完成网络爬虫，然后根据学长的提示可能用到的库有beautifulsoup，我看了下用这个库写的还挺多的，我就去搜beautifulsoup这个库，并搜到了一份资料，讲解beautifulsoup这个库的实际应用。

我用pip下载了beautisoup这个库，开始学习了reque库，了解到了urlopen的方法，只需要一个url就行，beautifulsoup库可以解析request返回的网页的数据，网页的前端我在大一刚开学的时候了解过，看那个HTML稍微没那么陌生，学到了那个find，findall方法，只需要根据HTML里的一些标签就可以找到相应的内容，比如soup.find\_all('a',attrs={'class':'qg-item qg-ing'})就可以找到a中的class='qg-item qg-ing'的项中的内容，还可以不提取内容，而提取里面的属性值，真的是非常方便。我还了解到真正的爬虫爬的时候，还可能用到DFS或者BFS，但是我爬天猫的打折商品的时候不太知道怎么加上DFS或者BFS，我是直接从商品的总览的那一栏就能找到所有的商品的url然后先找好所有的url，再直接挨个去搜url，从返回的界面信息里根据标签去找我所需要的内容，十分的方便。然后我再学了一下pymysql，跟数据库mysql的连接。在做java课设的时候安装过mysql，所以我就直接学了pymysql，重温了一下mysql的语句，然后实现了往里面写数据。写好数据以后，又得想怎么把它数据可视化，我想的就是把大分类的销售数量显示出来，根据要求中可能用到的库，和书中的资料，了解了一下matplotlib.pyplot和numpy。也就是了解了一下，能够稍微完成要求，画图的方法和用法实在是太多了，炫酷的非常炫酷，一时无法掌握，numpy也是比较复杂，两个我都是稍微了解了其中为数不多的几个参数，勉强实现了要求。

再就是第二个任务，这个任务就是完成了一个后想尝试一下另一个。隐约感觉应该是用到之前了解过的re，正则表达式，去稍微了解了一下正则表达式，也算是掌握了一些，但其中非常复杂的用法和式子，以及一些很理论性的东西还是不太明白。

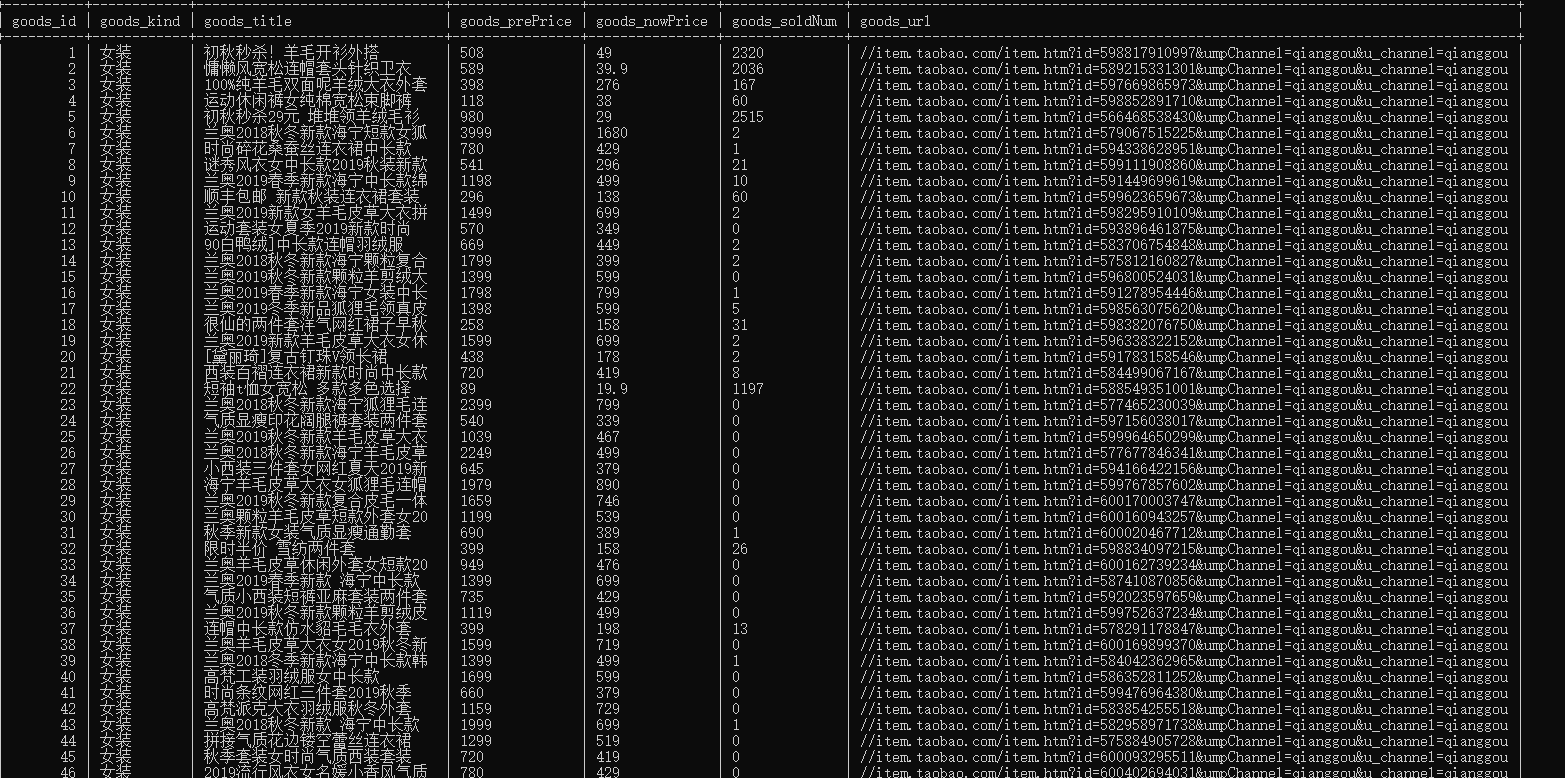
先有了一个完成任务的思路：我先从文本中读入数据，利用我学习到的re，去找出性别，年龄，就诊时间，症状和所得疾病。然后先去将这些数据整理到数据库中。然后对获得的数据离散化，所有的疾病离散为A,B,C,D……，所有的症状离散为1,2,3,4……。对于分类，我思考后。。。还是很难有一个时间复杂度很美好的算法。于是准备用一个比较暴力的方法，把相关因数定为2，有2个相同症状的病我就认为相关，然后为了方便实现这种分类，我打算反着存，先开一个字典，每个症状对应着一个列表，里面是包含该症状的疾病。然后两个for循环嵌套枚举所有2个症状的组合，再去判断里面有没有出现过两次的疾病名称，所有在两个症状中都出现的疾病就归结为一组了。

我感觉思路起码可以实现要求，但是在实际写的过程中，我万万没想到，给的数据竟然有乱入，时间突然变了格式我还可以忍，到后面偶尔出现个“p\_男职。业：”和“p\_女职员，”甚至到后面还有人年龄的地方对应着其他的东西，我就实在是改不动了。

反思一下失败的原因，要么是提取信息的方向错了，re不通（我感觉可能性不是很大），要么就是我re学的太浅了，处理这些东西还是不太熟练（比较可能），还有就是我处理的时候逻辑写的不太好（也比较可能），因为我在写之前没有想过会有信息乱了会怎么办，现在想起来，或许可以直接跳过信息不太符合的那一行不要了。应该可行。

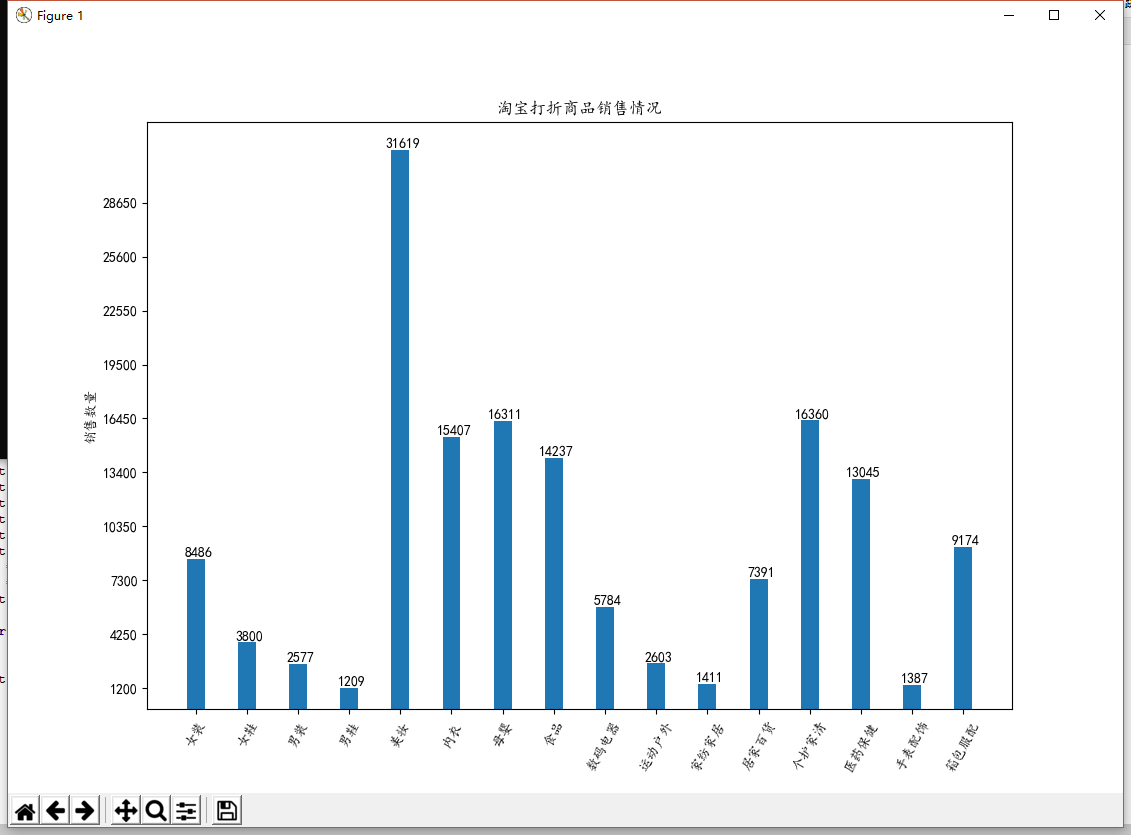
哇真的是，写这个报告的时候想出来的办法，之前怎么就突然卡壳了呢。刚刚改了改，成了，起码把数据都存入数据库中了。但惭愧的说。。写这个报告的时候，我拖了比较久，。。。。所以，后面我想的那些分类。。就没大有时间写了。。。。

附一：从<https://qiang.taobao.com/>下载打折促销的商品信息





附二：根据这些信息做的图：



附三：第二个任务存到数据库的数据：

