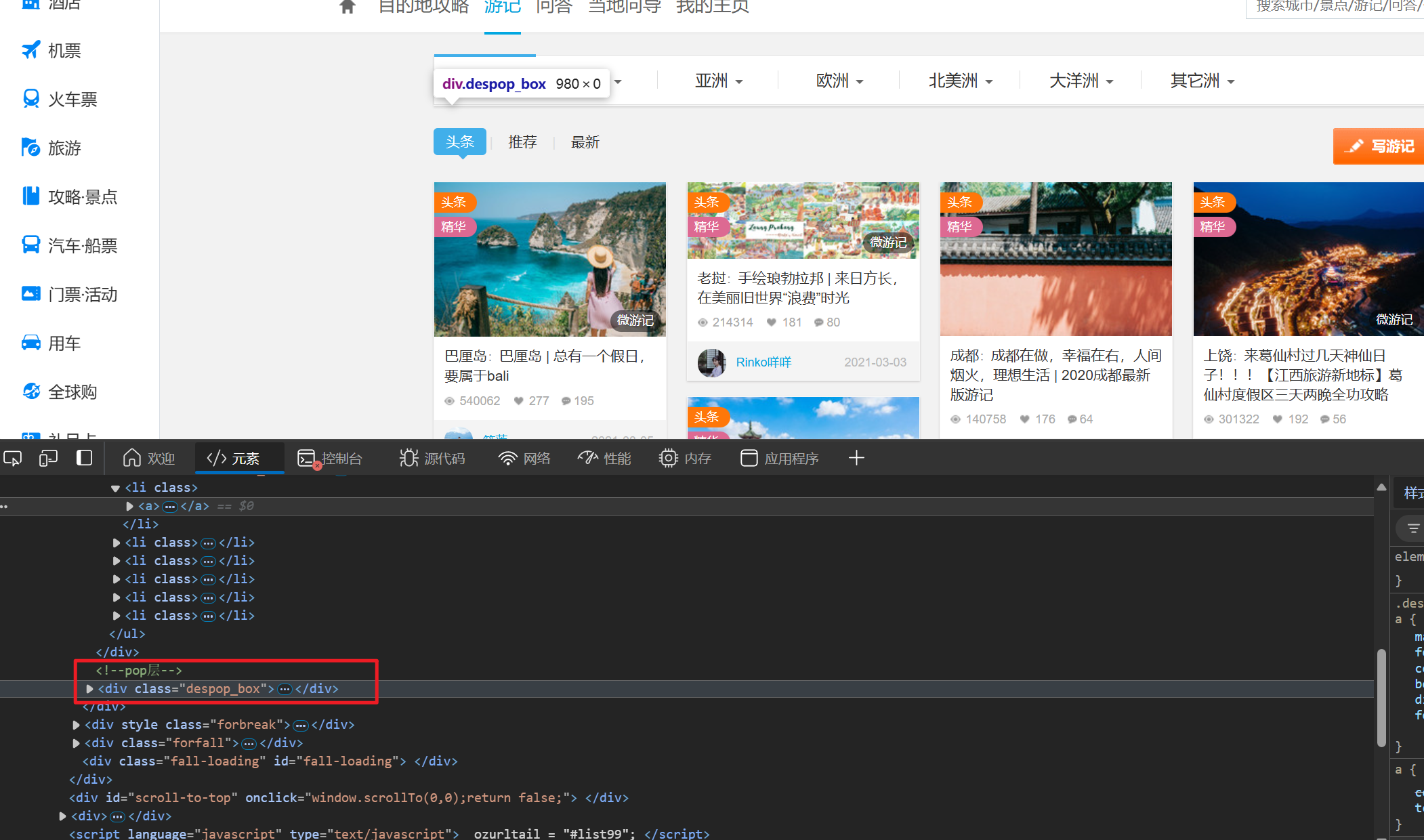
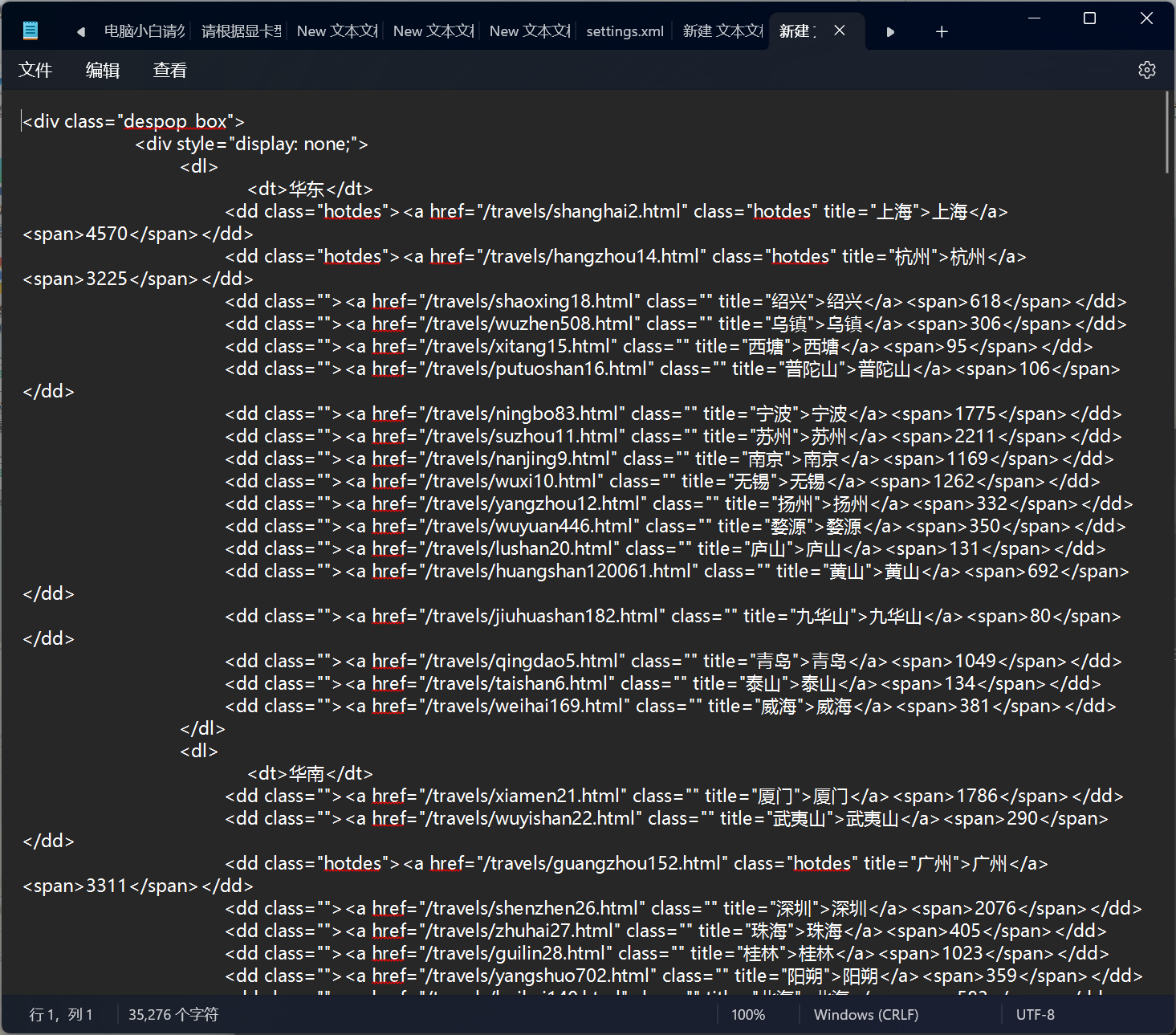
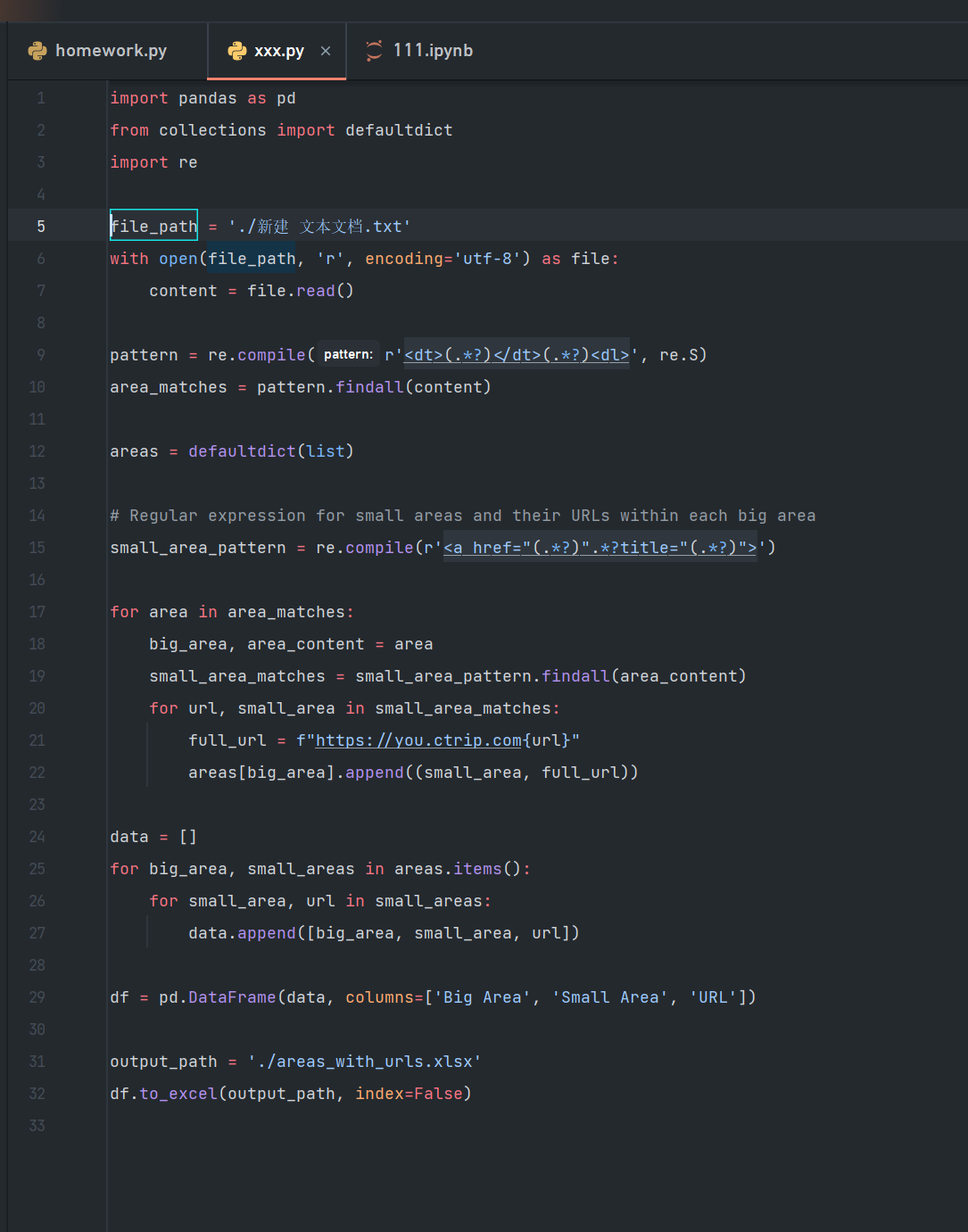
第四题解题过程

① 由于八爪鱼中无法选中具体地区，因此我通过原生网页中把这些地区的代码复制下来，通过八爪鱼循环访问这些链接来爬取数据。

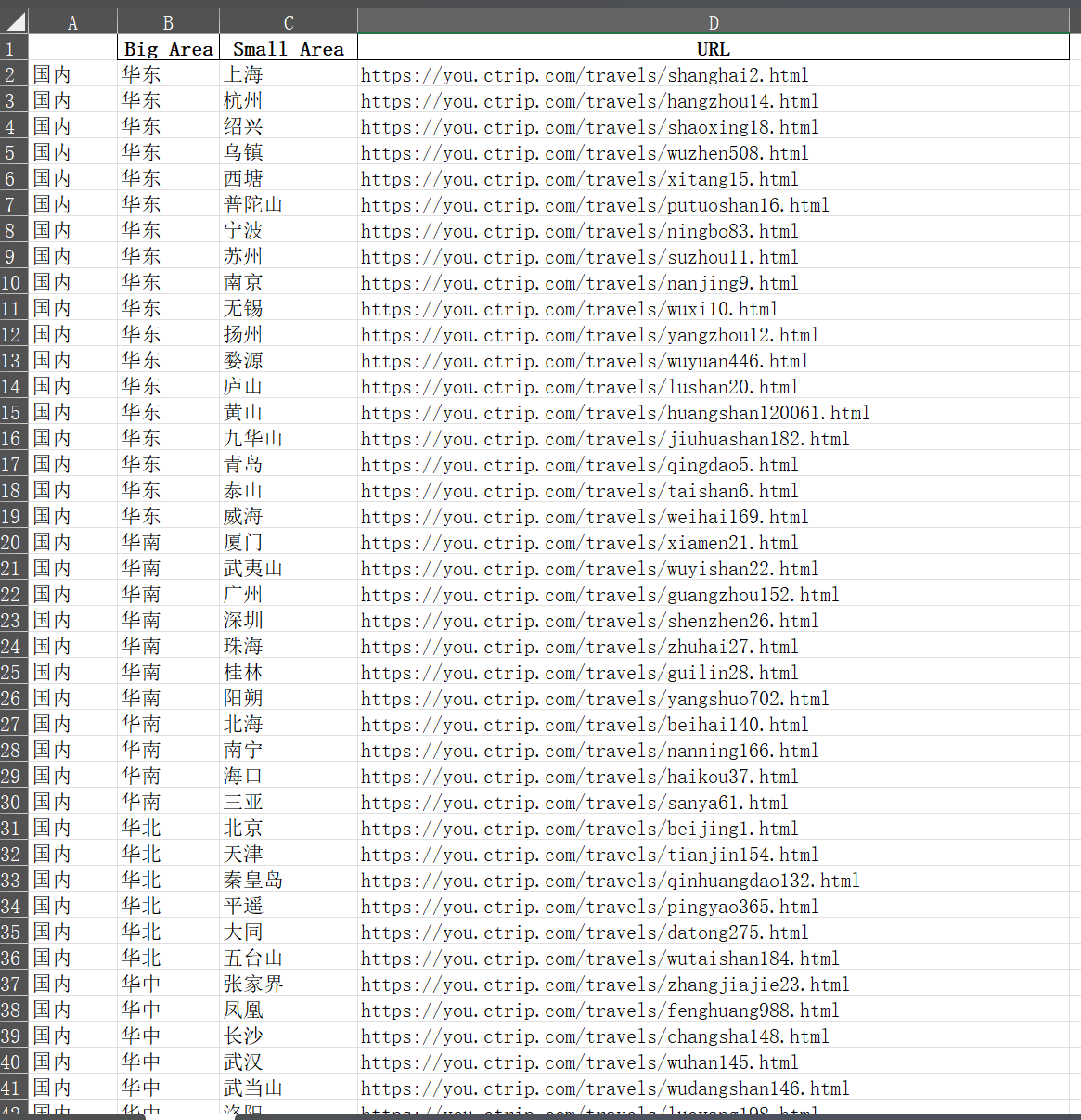




② 用python代码把对应div下的url和地区提取出来，得到结果如下



代码截图



结果图

一级分类（国内、亚洲等）在源码中和对应的二级分类地区不好对应起来，所以手动添加了第一列

③ 利用八爪鱼循环遍历这些地区的URL，访问各个地区，爬取数据

④ 一个游记和评论是一对多的关系，在表格中，要把所有评论显示出来（每个评论一行），会存在数据冗余（即下面效果）

游记1 评论1

游记1 评论2

游记1 评论3

…………

游记2 评论1

…………

因此手动改，把每一个评论都设为一个列，变成效果如下：

游记1 评论1 评论2 评论3 …… 评论10

游记2 评论1 评论2 评论3 …… 评论10

…………

⑤ 由于数据实在太多，因此我爬取了每一个二类地区的第一篇游记，共108篇



⑥ 导出csv的文件在pycharm中读取如下（直接打开会乱码）



⑦ 将其第一列与之前的地区excel表格内连接

把一类地区连接上去，得到结果如下

