爬虫部分

本次爬取的商品目标是天猫网站上的商品，搜索关键词为“手机”，使用软件是Spyder含python3的版本。实验环境的配置导入模块有re、requests，requests第三方模块需要自行安装（在cmd指令输入‘pip install requests’进行安装）。

首先进入天猫，在搜索栏输入“手机”后，可以直接显示商品信息不用登录验证，在浏览器打开F12开发人员工具刷新页面可以直接获取当前页面的URL还可以看到从服务器发出的请求是用的GET模式，在python代码中新建一个变量’index\_url’来储存URL，用导入模块的requests.get(index\_url)来模仿浏览器向服务器请求获取数据，爬取到当前页面的数据信息后，用正则表达式re.findall()把其他标签和无关数据过滤掉并且对商品名称和价格数据进行提取。目前就做到这一步，但是如何将提取出来的数据进行持久化还没弄通透，还没能力将数据存储到本地的excel表格和txt文档中。还有就是将提取到的数据通过接口连接进入HTML网页和算法当中整合的功能，由于个人能力目前还没能做到。后面在修改代码的时候加了一个header模仿浏览器防止反爬。





源代码：

Import re

Import requests

header = {}

index\_url = ‘’

response = requests.get(index\_url, headers = header )

html = response.text

name = re.findall(r’title=”(.\*?)”data-p’, html)

price = re.findall(r’</b>(.\*?)</em>’, html)

print(name)

print(price)