1. 实验目的

给定评课社区网站 ，需要设计一个网站遍历策略，爬取至少200个课程的详细信息，记录于 json 格式的文件中，需要提取的信息入下图红框所示：



1. 实验思路

## 整体结构

crawler.py为爬虫源代码

pksq.json为存放课程信息的文件

## module

1. requests：代理发送请求，获得响应文
2. BeautifulSoup：解析页面源代码，用以提取想要的html对象，如url、text等
3. json：将字典对象转换为json格式、
4. time：计算程序运行时间

## function

def crawl(url):

def get\_url(url):

if \_\_name\_\_ == '\_\_main\_\_':

1. crawl：爬取课程主页的文本
2. get\_url：爬取主网页的所有课程主页链接并调用crawl函数爬取课程信息

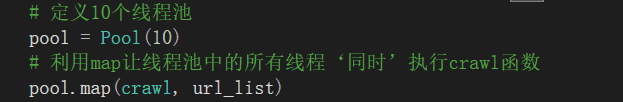
并返回本次爬取的课程信息组成的列表

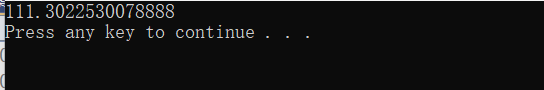
1. main：主函数，用于翻页直至课程数大于200

## 具体思路

主函数从page=1开始

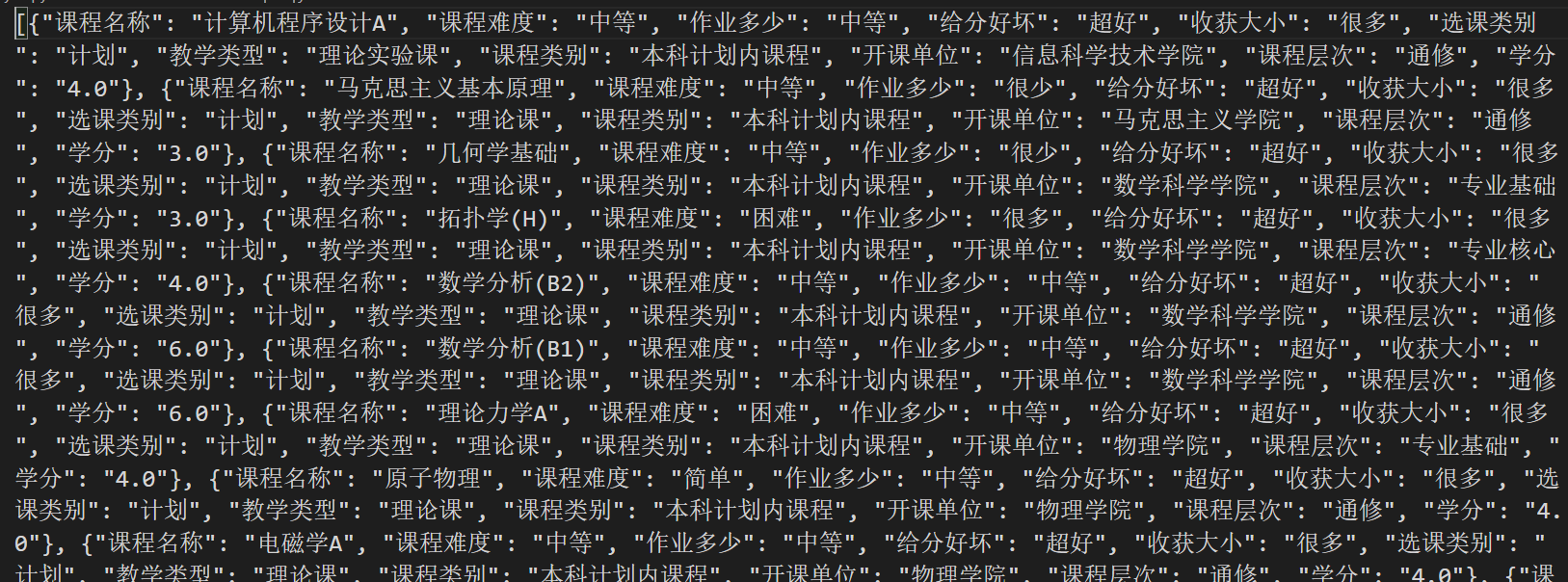
1. 首先，调用函数get\_url获取该page的所有课程页面的url链接并返回爬取的课程信息组成的列表，并将其与主函数的总结果列表lec\_list相加，同时在开头和结尾调用time库计时并输出
2. 然后，在函数get\_url中调用函数crawl返回课程信息组成的对象，将其加入列表
3. 计算总结果列表lec\_list的长度，如果大于200，转为json格式并保存到目标文件，程序结束，否则page+=1，转(1)
4. 缺点与改进

原本为单线程，速度过慢，改为多线程爬虫（在函数get\_url中实现多线程调用crawl函数），设定线程数为10

所用时间(单位s)：

是单线程时间的左右，比大的原因推测为python线程转换需要一定时间，总体上时间效率改进显著，但是有时会遇到网站反爬措施限制，此时需减少线程数或者设置爬一个网页一次后停顿时间

1. 实验结果

详见pksq.json文件，以下为部分截图

1. 参考资料
2. <http://t.csdn.cn/x1jYI>
3. <http://t.csdn.cn/g0q1f>
4. <http://t.csdn.cn/kGZEv>