



# 数据获取作业

本次作业含两组题目，同学只需选取其中一组完成：

第一组：本组含一道题目（作业1），满分10分

第二组：本组含两道题目（作业2、作业3），满分分别为6分（作业2）、3分（作业3）

说明：

1. 不支持两组都做、合并计分。只要作业中出现作业1，则按选做第一组处理
2. 任意一道题目，若采用数据接口直接获取数据，则该题目满分减1分



# 第一组

作业1：爬取北京链家官网二手房数据

<https://bj.lianjia.com/ershoufang/>

具体爬取信息及存储要求：

- 要求爬取东城、西城、海淀、朝阳四个城区的二手房数据（每个城区爬取5页），提取楼盘名称、平米数、总价、单价
- 将信息保存在csv文件中（可以是一个，也可以是每城区一个）



## 第二组（作业2）


作业2：爬取豆瓣电影Top 250数据

<https://movie.douban.com/top250>

具体爬取信息及存储要求：

- 要求爬取全部250部电影的数据，提取电影名称、综述、评分、评价人数。  
各项如右图框中所示
- 将信息保存在csv文件中

1



肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]

导演: 弗兰克·德拉邦特 Frank Darabont 主演: 蒂姆·罗宾斯 Tim Robbins / ...  
1994 / 美国 / 犯罪 剧情

★★★★★ 9.7 2733274人评价

“希望让人自由。”



## 第二组（作业3）

作业3：爬取学堂在线的合作院校页面内容

<https://www.xuetangx.com/university/all>

具体爬取信息及存储要求：

- 要求爬取全部开课院校的学校名称和对应的课程数量
- 将信息保存在json文件中



# 作业提交方式及要求

以上作业以报告形式提交，报告中要求包含（1）：题目的核心代码；（2）存储的文件内容（每个文件截取50条数据即可）。

文件名为学号，文件格式为pdf，提交给助教同学。

作业提交截止时间：2022.11.27 24点。