### 爬虫答辩需求文档

1. 小组成员：曾易、杨洁、刘赛赛
2. 项目名称：研究北京各区域二手房的房价对比
3. 项目背景：
   1. 为想在北京买二手房的朋友们，提供各区域房子的参考价，便于根据自己的经济水平，定位到适合自己的买房区域
   2. 为想在北京卖房的朋友们，提供各区域卖房的参考价，便 根据自己的房子所在区域合理定价
4. 分工
   1. 爬取数据 —— 曾易（暂定）
      1. 方法：多任务异步协程（暂定）
   2. 分析数据 —— 杨洁（暂定）
      1. 方法：数据分析（暂定）
   3. 展示结论 —— 刘赛赛（暂定）
      1. 方法：制作ppt展示（暂定）
5. 功能模块的介绍
   1. requests模块
      1. 一个基于网络请求的模块，作用就是用来模拟浏览器发起请求
      2. requests是一个不支持异步的模块
   2. asyncio模块
      1. 用于多任务异步协程
   3. aiohttp模块
      1. 用于多任务异步协程
      2. 支持异步操作的网络请求的模块
   4. lxml模块
      1. 使用lxml中的etree的xpath方法进行解析
6. 业务逻辑
   1. 数据爬取分析
      1. 简单示例：爬取链家北京昌平地区的二手房数据
         1. 爬取数据：二手房标题、户型信息、总价、单价、封面图片
         2. url=[https://bj.lianjia.com/ershoufang/changping/pg%d/](https://bj.lianjia.com/ershoufang/changping/pg2/)
         3. 使用url模板进行页码拼接得到对应的new\_url,进行数据爬取
      2. 判断爬取的数据是否是动态加载的
         1. 经过搜索判定不是动态加载的数据
      3. 数据爬取：使用requests爬取数据
         1. 使用requests.get获取text数据，即返回的response数据
      4. 数据解析：使用lxml中的etree的xpath方法进行解析
         1. 实例化一个etree类型的对象，且将页面源码数据加载到该对象中
         2. 需要调用该对象的xpath方法结合着不同形式的xpath表达式进行标签定位和数据提取
         3. 获取到二手房标题、户型信息、总价、单价、封面图片的url
      5. 通过封面图片的url获取图片数据
         1. 使用requests.get获取对应content数据
         2. 将图片数据保存到指定目录下
      6. 将二手房标题、户型信息、总价、单价、封面图片的本地路径保存到指定文件中
      7. 注意事项：
         1. 爬取图片的url时，使用属性src可能爬取不到对应数据，根据抓包工具发现属性data-origina和属性src的值是相同的，通过验证，可以通过使用属性data-origina爬取到图片url
   2. 数据分析
      1. 通过爬取到的二手房单价信息，进行计算得到北京各区域的平均单价
      2. 通过计算得到的数据进行比较，绘制出区域对比的柱状图（暂定）