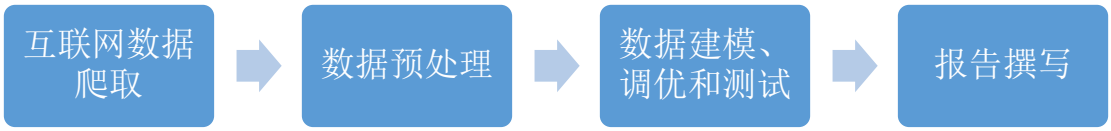
数据挖掘课程大作业要求

1. 项目任务综述

- 目标：完成一个全生命周期数据挖掘项目，覆盖以下步骤流程。



  - ✓ 互联网数据爬取：从主流互联网（如去哪儿网）爬取数据（如酒店评论），具体分组任务见下一节；
  - ✓ 数据预处理：如爬取内容中包含自然语言（如酒店评论），则需做分词数据转换为结构化数据，还需要进行缺失值填充和数据清洗等操作；
  - ✓ 数据建模、调优和测试：建立预测模型（如预测一条评论为好评或差评），调优超参数，并给出模型的性能；
  - ✓ 报告撰写：将主要流程撰写成文档，并上传代码和爬取的数据集。
- 截止日期：2021 年 6 月 19 日。
- 实现方式：不限，推荐使用 Python 语言。
- 提交方式：将文档、代码和数据集打包后上传 canvas 平台。

2. 分组和具体任务

- 分组：不超过 4 位同学组成一组，确保所有组员任务分配合理。
- 每一小组选择一个互联网网站作为分析目标，建议的网站如下（不限于）。

网站	任务
安居客	1. 爬取上海地区二手房价格（https://shanghai.anjuke.com） 2. 自建变量，如房间数、面积、朝向、有无电梯、与市中心距离、与地铁站距离、标题中文关键词等等 3. 建立房价预测模型
二手车	1. https://www.renrenche.com/ 2. 建立二手车变量，如品牌、型号、里程、车况等 3. 建立二手车价格预测模型
大众点评网	1. 爬取上海地区麦当劳餐厅概况和评论（http://www.dianping.com/shop/2394082） 2. 自建变量，如菜系、人均、与市中心距离、推荐词中文关键词、评论中文关键词等等 3. 建立好评差评分类模型（4 分及以上为好评，2 分及以下为差评）
前程无忧	1. 爬取上海地区职位薪资

	（ https://search.51job.com/list/020000,000000,0000,00,9,99,%2520,2,1.html） 2. 自建变量，如职位名关键词、公司名关键词、工作地点等等 3. 建立薪资预测模型（预测薪资下限或上限）
京东	1. 爬取某款商品的评论（https://item.jd.com/6946627.html） 2. 自建变量，如购买型号、购买时间、评论中文关键词等等 3. 建立好评差评分类模型（4 分及以上为好评，3 分及以下为差评）
豆瓣电影	1. 爬取豆瓣电影 Top 250 榜单（https://movie.douban.com/top250）中的电影短评（https://movie.douban.com/subject/1292052/comments） 2. 自建变量，如评论关键词、影片信息等等 3. 建立好评差评分类模型（4 分及以上为好评，2 分及以下为差评）

3. 评分标准

项目	占比
爬取数据量：至少 2 万条，条数越多、内容越全面得分越高	30%
特征提取：数据结构化后，所创建的特征是否有意义，是否利用了中文文本信息等	30%
模型性能：对于分类问题，AUC 越高越好；对于回归问题，R^2 越高越好	30%
项目报告：逻辑清晰易懂得分越高	10%