6.4 作业

2023 年 9 月 18 日

作业

1. 网页爬虫（5 分）

使用爬虫软件爬取 JD 上某个商品的名称、价格、评价数、店铺等信息，样本为 1000 条，保存为 CSV 文件格式，然后和本 ipynb 文件一起上传提交。

[ ]:

2. 数据预处理（1 分 +1 分 +1 分 +2 分 =5 分，附加题 2 分）

对上述爬取的数据进行预处理：1. 检查是否有重复样本 (行)，如果有，去重 2. 检查是否有缺失值，如果有，填补缺失值 3. 检查是否有异常值，如果有，修正异常值或者删除样本（行）4. 对于数据集进行哑变量处理，即根据商品信息，将品牌、类别等信息从商品名称中提取出来，使用 0 和 1 表示这些特征。5. 附加题：将采集的评价数从字符串改成数值。

[ ]: