说明文档

320180941581 饶诗浩

我的数据预处理思路如下：

1. 用爬虫将数据下载到本地
2. 用Pandas解析本地json文件，先用列表保存所有文件的文件名
3. 第一步筛查：答题时间太短或太长的筛除，保留符合条件的文件名到另一个文件名列表
4. 被保留的文件名依据传感器种类分为3类
5. 根据每类数据的特点进行第二轮处理，如果将手机放在桌子上答题，则该文件被删除。删除条件为：取中间三分之一数据的方差，方差小于某个值，判断为答题方式不合理。
6. 处理异常值将超过上四分位的值1.5倍距离和低于下四分位1.5倍距离的值都视为异常值，按照中位数替换。
7. 保留预处理结果：用三个字典保存三种传感器数据，key为文件名，value为处理异常值后的数据对象。

补充：鉴于数据处理过程是边看图边处理，因此用了Jupyter，代码也是ipynb格式