数据集处理Readme文档

一、文档目录

- txtTocsv.py #u.data数据转换为csv格式，转换完毕后的文档为data.csv，转换完毕后，打开，将第一行修改为这样，压缩包中data.csv文件为我已预处理好的文件，可以直接使用



* changecsv.py # 代码文档，所有内容在里面修改
* data\_1w.csv 已筛选出1w量级数据集
* data\_5k.csv 已筛选出5k量级数据集
* data\_1k.csv 已筛选出1k量级数据集

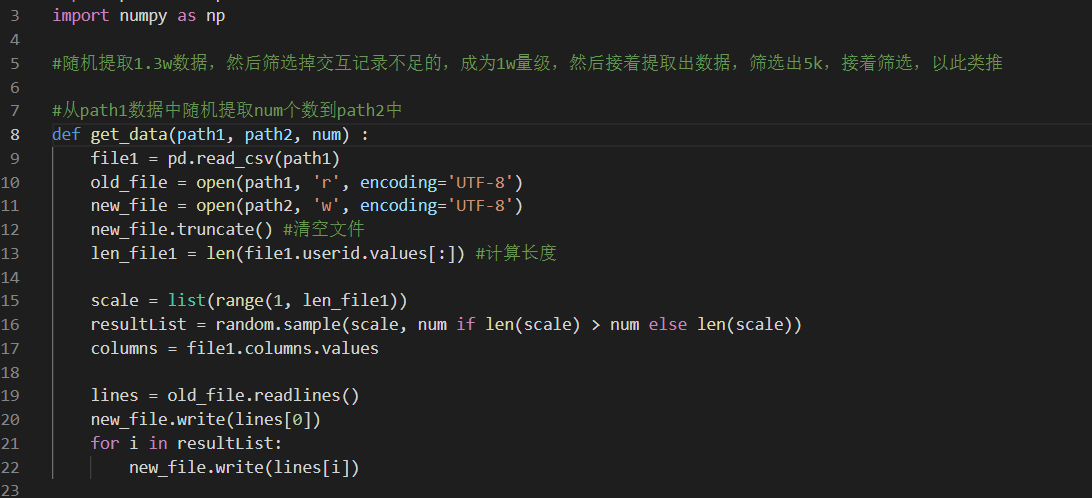
除了data.csv，其余数据文件不用创建，由代码自动生成

所有数据集均为随机生成，重新执行代码，得到的数据集会变化，所以提前保存好已收集到的数据集

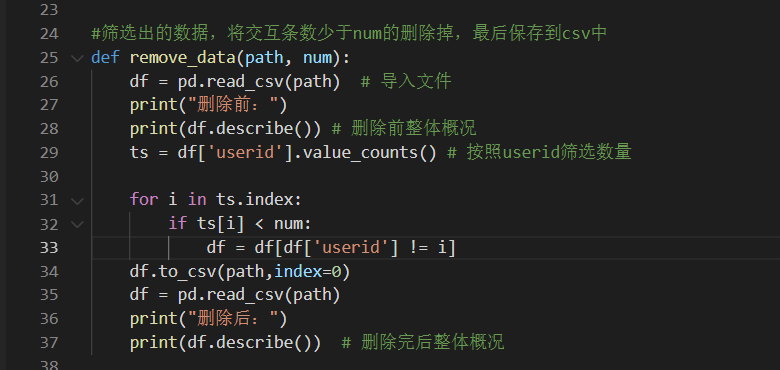
二、 代码解释

1. 导入库

Pip install pandas

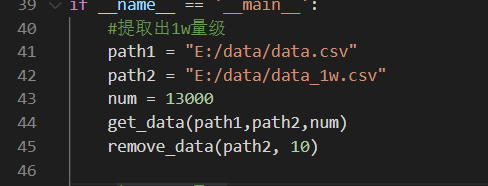


此函数目的是大数据集中随机提取部分数据到小数据集中

（3）

此函数目的是从提取到的数据集中删除掉不符合要求的数据，比如1w量级中删除掉少于10条交互数据的用户

三、具体使用以及实例展示



在main中修改文档目录，修改num为量级，这里选为13000，因为筛选数据后续需要剔除，所以处理完的文档大致是1w量级

处理结果：





5k， 1k量级类推