# 大数据可视化技术实验1-2内容

实验1：报告上交电子版包含：

1个.ipynb文件（包含数据和原始图形），文件名为:学号+姓名.xlsx

1个word文件，文件名为:学号+姓名.docx，内容包实验截图、必要公式、简要观点说明、叙述，公式和图形插入后需要按照学位论文格式，附上公式编号、图编号、标题和必要说明（例如，图1 XXXX随时间变化曲线等，你认为必要的说明内容）

Word文件报告要有实验报告封皮、姓名学号等信息，无字数多少要求和限制。

题目：基于用户Pearson相似性的协同过滤推荐算法实现与验证

使用MovieLens数据集，实现如下目标：

1. 读入评分数据集数据

2. 评分数据集切块分10组

3. 由输入list确定邻居数量[1,5,10,15,20,25,30,35,40,45,50,55,60]

4. 所有用户两两配对

5. 对用户，物品，评分进行索引方便后续计算（可选用矩阵或字典）

6. 实现获取某用户评分过物品的函数

7. 实现获取user item，的评分的函数

8. 实现计算皮尔逊相似性函数

9. 实现函数按照配对计算用户间两两相似性，返回结果

10. 实现函数对预测目标（user,item）,选择user的排序邻居集合，要求user邻居对item评过分数

11. 实现预测公式函数

12. 实现计算所有用户平均评分函数，返回用户平均分字典

13. 实验计算误差(MAE)函数

14. 实验计算误差(RMSE)函数

对于步骤2中的10个数据data，每个作为测试集，其他9块作为训练集，进行如下循环

1. 对于每个测试集Q,计算用户间相似性（使用函数9），
2. 对于每个预测集T中的预测目标（user，item）中的user，选择所有邻居中对item评过分（使用函数7），要求评分必须在测试集Q中，选则符合条件邻居中相似性高的前60个（由3条件给定的需要最大数量确定），如果剩余邻居数量不足60，全取为预测目标（user，item）的邻居
3. 对预测集T中，每个预测目标user，对3条件给定的邻居数量进行循环，依次取[1,5,10,15,20,25,30,35,40,45,50,55,60]邻居信息，调用函数11进行预测，将预测结果保存，包括预测目标，需求邻居数量，预测结果。
4. 对于每个预测集T中的预测目标（user，item），读取C步骤中，的预测结果（包含不同邻居数量下一组预测），读取真实评分，计算预测集T中不同邻居数量下的预测误差MAE

计算10次实验不同邻居数量下MAE的均值

输出算法在不同邻居数量下的MAE曲线，横坐标邻居数量，纵坐标MAE值（采用Matplotlib和seaborn包均可，保存成pdf文件）

讨论使MAE最小的邻居取值

实验3

在你已经实现Pearson协同过滤的基础上，

1）实现Cosine协同过滤，画出不同邻居数量下和Pearson协同过滤的对照曲线

2） 仅使用用户在训练集里的平均评分作为对未知评分的预测，标记为AverageRating, 画出AverageRating, Pearson，Cosine三种方法在不同邻居数量下的对照曲线

实验4

利用Python从键盘读取一个整数，作为用户编号，基于预测的评分和真实的评分，在屏幕上输出下列数据：

用户编号：

所有这个用户的预测结果（邻居数量最优的情况下），以预测评分排序(降序)

1. 物品名(电影名) 整体样本平均评分 评分人数 预测评分 真实评分

2. 物品名(电影名) 整体样本平均评分 评分人数 预测评分 真实评分

3. 物品名(电影名) 整体样本平均评分 评分人数 预测评分 真实评分

4. 物品名(电影名) 整体样本平均评分 评分人数 预测评分 真实评分

5. 物品名(电影名) 整体样本平均评分 评分人数 预测评分 真实评分

6. 物品名(电影名) 整体样本平均评分 评分人数 预测评分 真实评分