# 《大数据挖掘》课程实验3

## 实验题目：音乐推荐算法优化

## 数据集介绍

* user\_artist\_data.txt：2420万条用户播放艺术家歌曲次数

格式：(用户ID 艺术家ID 播放次数)

* artist\_data.txt: 160+万个艺术家的ID和名字

格式：（艺术家ID 艺术家名字）

* artist\_alias.txt: 拼写错误的艺术家ID到该艺术家规范ID的映射（19万条记录）

格式：（拼写等错误的艺术家ID 艺术家正确ID）

## 实验内容要求：(本实验满分10分)

### 实验代码（5分）

1. 预处理：
   * 找出artist\_data.txt和artist\_alias.txt:文件中数据不规范的类型，分别统计不同类型的记录数量：（0.5分）

例如，缺少艺术家ID：100条

* + 根据artist\_alias.txt的映射关系，将user\_artist\_data.txt进行更新。（0.5分）

1. 统计播放量最高的30名艺术家ID及其被播放总次数（0.5分）
   * 按照播放次数降序输出。
   * 显示格式：序号，艺术家名字，总播放次数
2. 基于讲义中UV分解方法编码实现算法，画出每次迭代的误差平方和曲线。（1.5分）
3. 采用梯度下降法提高上述UV分解方法效率，从程序收敛速度和误差曲线下降趋势对比讲义中的方法（2分）

### 实验报告（5分）

## 评分标准：

1. 1人1组独立完成。
2. 实验代码完整实现上述每个步骤要求，验收时可以正常运行，能够解释代码的细节原理，不与其他人雷同，代码实际运行结果与实验报告相符。
3. 实验报告记录每个步骤的详细实现过程，实验结果展示清晰合理，文字格式规范。
4. 最终提交可运行的代码和实验报告。
5. 实验报告文件名规范：**姓名-学号-实验3**.docx/pdf
6. 代码压缩文件命名格式：**姓名-学号-实验3代码**.zip