基于 Hadoop Map-Reduce 的推荐系统

- **环境描述：**本题目评测时将会运行在 Hadoop 2.4.x，JDK-1.7.0 环境下，必须采用 Map-Reduce 的编程模型，使用 Java 进行编程。

  评测使用的集群硬件配置如下表：

Nodes	14
Cores	336
Memory per node	64GB
CPU per node	Intel(R) Xeon(R) CPU E5-2697 v2 @ 2.70GHz × 4

- **题目描述：**本题目为基于公开数据集 MovieLens 数据集上的用户评价数据，计算用户对其未看过，并且可能会看的电影的评分。同时请各参赛队伍考虑数据稀疏性问题和恶意用户问题，使推荐系统在上述极端情况下具有较好的性能（考核的重要依据，具体请参见评判标准） 。
- **数据集：**本题目将采用推荐系统常用数据集 MovieLens 10M 数据集，（http://files.grouplens.org/datasets/movielens/ml-10m.zip）。请勿使用 tags 中的数据。本数据集包含来自 71567 个用户对 10681 部电影的 10000054 评分记录。关于数据集的详细说明请见：

  http://files.grouplens.org/datasets/movielens/ml-10m-README.html
- **程序设计约束：**程序需要三个输入参数，第一个为训练集的路径。第二个为测试集的路径，第三个为输出文件夹的路径，以上路径均为 HDFS 路径。输出文件格式为：

```
UserID1:MovieID1:5
UserID1:MovieID2:4
UserID2:MovieID1:4.5
```

  1. 用户对一部电影的打分在一行中，用户 ID，电影 ID，评分用西文冒号(:)分割。
  2. 程序的第一个以及第二个参数为文件在 HDFS 上的完整路径，如：

     hdfs://<name node uri>/input/testdata.dat
  3. 程序的第三个参数为文件夹路径，如：

hdfs://<name node uri>/ContestResult/PlayerXX/

在评测时评测程序会自动将文件夹下以 part-???格式命名的文件合并为单个文件进行评测。

- **程序评测方式：**

每个参赛组提交的程序 jar 包将会以如下格式的命令运行，进行最终的评测。（**请不要将数据集路径，jar 包路径等参数写死在程序中，否则提交的程序在组委会统一测试平台上可能无法运行！**）

```
${HADOOP_HOME/bin}./hadoop \
jar <path to your jar file>
cn.edu.seu.cloud.jn2.Main
hdfs://<path to training dataset> \
hdfs://<path to test dataset> \
hdfs://<output directory path>
```

- **评判标准：**

1. 均方根误差（$RMSE$）。设预测的用户评分集合表示为$\{p_1, p_2, \dots, p_N\}$，对应的实际用户评分集合为$\{q_1, q_2, \dots, q_N\}$，则均方根误差 $RMSE$ 定义为：

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(p_i - q_i)^2}{N}}$$

本项评判标准为所有测试集用户的平均 RMSE。

2. 平均绝对误差（MAE）。设预测的用户评分集合表示为$\{p_1, p_2, \dots, p_N\}$，对应的实际用户评分集合为$\{q_1, q_2, \dots, q_N\}$，则平均绝对误差 MAE 定义为：

$$MAE = \frac{\sum_{i=1}^{N}|p_i - q_i|}{N}$$

本项评判标准为所有测试集用户的平均 MAE。

3. 程序运行速度。
4. 数据稀疏性问题。

本题目的原始数据集的稀疏度在 1.3%左右，为了检验推荐算法在较为稀疏的数据上的表现，将通过随机删减数据的方法将稀疏度下降至 1.1%，0.9%，

并分别执行推荐算法，计算 RMSE 和 MAE。

5. 恶意用户问题。

   在推荐系统中，存在一系列的恶意用户，其行为主要体现在随机打分或为多部电影打相同的分。为了检验推荐算法是否可以减少或避免恶意用户的影响，将通过随机加入恶意用户的方法，比率为 5%，10%，并分别执行推荐算法，计算 RMSE 和 MAE。

6. 总分计算方法：

   本题目将分别在正常数据集情况下，稀疏化数据集情况下以及恶意用户数据集情况下考察各队伍的推荐算法得分，并进行加权求和计算总得分，三种情况的权重相同。单一情况下的推荐算法得分将通过计算各队伍的推荐算法 RMSE、MAE 以及运行时间并进行相应的排序与给分，按照 0.4、0.4、0.2 的权重进行相加求得推荐算法得分。

- **提交材料：**

本题目需要提交如下的材料：

1. 程序代码。要求提供包含完整目录结构的 src 代码包，并且提供编译方法说明。因为在必要时，会重新编译选手程序进行评测，请在提交代码时附带编译方法说明。
2. 程序 jar 包。
3. 报告。报告需要涵盖程序设计思路，实现方案，测试结果（包含数据分析过程，以及实验测试结果 RMSE，MAE 等）等。

- **友情提醒：**

1. 评测用的环境如前所述，请注意 Hadoop、JDK 版本。如果选手使用高版本的 JDK，请在编译程序时，编译为 JDK-1.7 兼容的格式。评测将使用 Hadoop 2.4.1 版本，更高级版本的 API 将不会被支持。
2. 评测时将会数据主办方划分的训练集和测试集。因此请选手在实现程序时，勿将数据集的信息（如数据集路径）硬编码在程序里，而是在运行时动态从输入文件中获取。
3. 评测将使用程序自动化地运行选手程序、判断结果，因此请务必遵守“程序设计约束”中要求。不符合格式要求的输出文件，可能导致评测失败，出现

这种情况时责任选手自负。

4. 请附带源代码的编译说明。在 jar 包无法运行的情况下，评测人员会尽量尝试从源代码重新编译运行。