题一：基于 Hadoop Map-Reduce 的 word count 程序

- 环境描述：本题目需要运行在 Hadoop 2.4.x 环境下，必须采用 Map-Reduce 的编程模型。
- 题目描述：在英文环境下，给定一篇文章，统计每个单词出现的次数，单词不区分大小写，标点符号不进行统计。
- 程序设计约束：程序需要两个输入参数，第一个为输入文件的路径，第二个为输出文件的路径，输出文件的格式为：

```
Word1:110
Word2:20
```

  每个单词的统计数据在一行中，单词和该单词在给定文章中出现的次数用西文冒号分隔(:)。

- 评判标准：
  1. 程序准确度。
  2. 程序运行速度。
- 提交材料：

  本题目需要提交如下的材料：
  1. 程序代码。
  2. 程序 jar 包。
  3. 报告。报告需要涵盖程序设计思路，实现方案，测试结果等。

题二：基于 Hadoop Map-Reduce 的推荐系统。

- 环境描述：本题目需要运行在 Hadoop 2.4.x 环境下，必须采用 Map-Reduce 的编程模型。
- 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。
- 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）。如果各参赛队伍受制于硬件环境，可以在该数据集中截取一部分进行处理，具体截取长度由各参赛队伍自行确定。
- 程序设计约束：程序需要两个输入参数，第一个为输入文件的路径，即指向 ratings.dat 文件的路径。第二个为输出文件的路径，输出文件的格式为：

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

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。

- 评判标准：

对所有数据做十折交叉验证，测试以下指标，并取平均值进行评判。

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。

- 提交材料：

  本题目需要提交如下的材料：

  1. 程序代码。
  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。

题三：K-频繁项集挖掘并行化算法

- 环境描述：本题目需要运行在 Apache Spark 1.0.1 环境下，使用 Java 或者 Scala 进行编程开发。
- 题目描述：在规定的 Chess 标准数据集上，规定 K = 8，支持度 support = 85%，进行 1-频繁项集到 K-频繁项集的挖掘。
- 数据集：本题目将采用 Chess 标准数据集 apriori_data，具体下载地址见大赛网站 http://cloud.seu.edu.cn。
- 程序设计约束：程序需要两个输入参数，第一个为数据集路径，第二个为输出文件夹路径。1-频繁项集到 K-频繁项集的结果放在 K 个文件中，文件名分别为 result-1,result-2,...,result-8(K=8)，每个文件的格式为：

```
a,b,c:0.85
a,b,d:0.90
```

  项集和支持度用西文冒号(:)分割，项集中如果有多个元素则用西文逗号分割(,)。
- 评判标准：
  1. 程序准确度。
  2. 程序运行速度。
- 提交材料：

  本题目需要提交如下的材料：
  1. 程序代码。
  2. 程序 jar 包。
  3. 报告。报告需要涵盖程序设计思路，实现方案，测试结果等。

题四：莎士比亚文集词频统计并行化算法

- 环境描述：本题目需要运行在 Apache Spark 1.0.1 环境下，使用 Java 或者 Scala 进行编程开发。
- 题目描述：在给定的莎士比亚文集上（多个文件），根据规定的停词表，统计出现频度最高的 100 个单词。
- 数据集：shakespear 文集，具体下载地址见大赛网站 http://cloud.seu.edu.cn。
- 停词表：stopword.txt，具体下载地址见大赛网站 http://cloud.seu.edu.cn。
- 程序设计约束：程序需要三个输入参数，第一个为数据集路径（即 shakespear 文件夹的路径，文件夹中的文件名为固定文件名），第二个为停词表路径，第三个为输出文件路径。输出文件的格式为：

  ```
  Word1
  Word2
  ```

  每个单词独立一行。
- 评判标准：
  1. 程序准确度。
  2. 程序运行速度。
- 提交材料：

  本题目需要提交如下的材料：
  1. 程序代码。
  2. 程序 jar 包。
  3. 报告。报告需要涵盖程序设计思路，实现方案，测试结果等。