大数据分析第二次作业

（最晚提交时间 2025.05.09,18:00 之前提交到微助教上）

1. 设计 MapReduce 算法来实现下列功能。其中输入是整数构成的大文件，输出是（20 分）。

(a)所有整数的平均值

(b) 整数集合，但每个整数只出现一次。

2. 假定有 100 个项，编号是 1 到 100。同时有 100 个购物篮，编号也是 1 到 100。当且仅当 b 能被 i 整除时，项 i 放入购物篮 b 中。例如项 1 放入所有的购物篮，项 2 只放入 50 个偶数编号的购物篮中，其他类推。所有能整除 12 的数所对应的项组成第 12 号购物篮中的项集{1,2,3,4,6,12)，那么请回答（20 分）

(a)如果支持度阈值是 5，哪些项是频繁的?

(b) 如果支持度阈值是 5，哪些项对时频繁的?

(c) 所有购物篮中的项的数目之和是多少?

(d) {5,7}-->2 这条关联规则的可信度是多少?

3. 假定有 100 个项，编号从 1 到 100，同时有 100 个购物篮，编号也是从 1 到 100。当且仅当 i 能被 b 整除时，项 i 放入购物篮 b 中。例如第 12 号购物篮中的项是{12,24,36,48,60,72,84,96}。使用 A-Priori 算法应用到该数据集，找出所有频繁项集，其中支持度阈值为 5（20 分）。

4. 下面给出了 12 个购物篮组成的集合。每个购物篮都是由项 1 到项 6 的三个

项组成：

{1,2,3}

{2,3,4}

{3,4,5}

{4,5,6}

{1,3,5}

{2,4,6}

(1,3,4}

{2,4,5}

{3,5,6}

{1,2,4}

{2,3,5}

{3,4,6}

假定支持度阈值为 4。在 PCY 算法的第一轮扫描中，我们使用一个具有 11 个桶的哈希表，集合{i,j}会哈希到桶 i*j mod 11，那么请回答（40 分）

(a) 不管采用什么方式，计算每个项及每个项对的支持度

(b) 哪个项对会哈希到哪个桶中?

(c) 哪些桶是频繁的?

(d) 在 PCY 算法的第二次扫描中，哪些项对会被计数?