# 1、Spark

1. 代码



1. 本题数据流程的简短描述

·读入数据

·对可能存在共同好友的陌生人（a和b为好友、a和c为好友=>b和c是可能存在共同好友的陌生人）添加一条记录，贡献为1。

·对已经存在的好友添加一条记录，贡献为 -inf

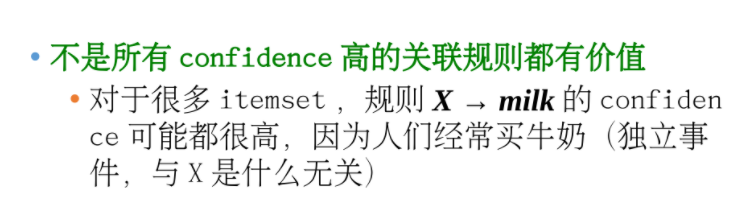
·对数据进行reduce 过滤掉贡献为负数的记录，剩下的就是存在共同好友的陌生人。

·对数据进行groupByKey分组，并按照出现次数降序、id升序进行排序，取前10个。

·根据题目条件输出。

# 关联规则

1. confidence 的一个缺点是它忽略了 P r(B)。请解释为什么这是一个缺点？以及为什么 lift 和conviction 没有这个缺点？



= 描述的是表示A与B的关联度，当lift较高时A与B出现的概率越不互相独立。而 也与类似。

1. lift指标是对称的。

因为，所以不是对称指标。

因为，所以是对称指标。

因为，所以不是对称指标。

1. 满足的有confidence，完美蕴含时就是1；
2. ('DAI93865 -> FRO40251', 1.0)

('GRO85051->FRO40251',0.999176276771005)

('GRO38636->FRO40251',0.9906542056074766)

('ELE12951->FRO40251',0.9905660377358491)

('DAI88079 -> FRO40251', 0.9867256637168141)

1. ('(DAI23334, ELE92920) -> DAI62779', 1.0)

('(DAI31081, GRO85051)-> FRO40251', 1.0)

('(DAI55911, GRO85051) -> FRO40251', 1.0)

('(DAI62779, DAI88079) -> FRO40251', 1.0)

('(DAI75645, GRO85051) -> FRO40251', 1.0)

# 局部敏感哈希

|  |  |  |
| --- | --- | --- |
| 行号 | S1 | S2 |
| 1 | 0 | 0 |
| 2 | 1 | 1 |
| 3 | 0 | 1 |
| 4 | 1 | 0 |

S1 和 S2 的 Jaccard 相似度 = 1/3

从第一行开始循环：h(S1) = 2, h(S2) = 2

从第二行开始循环：h(S1) = 1, h(S2) = 1

从第三行开始循环：h(S1) = 2, h(S2) = 1

从第四行开始循环：h(S1) = 1, h(S2) = 3

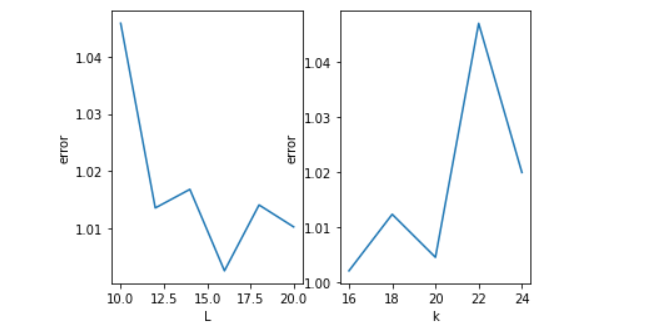
随机循环重排列为 S1 和 S2 产生相同 min-hash 值的概率 = 1/2

# LSH 用于近似近邻搜索

考虑逆命题“返回的点不是一个(c, )-ANN”的概率”

lsh 平均搜索用时： 1.8248265027999877

线性搜索 平均搜索用时： 8.27710964679718



线性搜索line-7464line-7551line-8196line-12444line-21780line-22509line-25289line-25549line-28251line-28351

lshlsh-7464lsh-7551lsh-12444lsh-18531lsh-21780lsh-22509lsh-25289lsh-25549lsh-28351lsh-37765

1. 代码



本人承诺本次作业的解答独立完成。