软件学院 2020 年数据分析/挖掘课程编程作业 1（20 分）

1. [交易数据] 本次作业利用交易数据集合，开展针对用户交易的聚类分析，有助于客户画像分析。

字段名称: uid, **sldatime**, pno, cno, cmrid, **vipno**, id, **pluno**, bcd, pluname, spec, pkunit, dptno, dptname, **bndno**, bndname, **qty**, **amt**, disamt, ismmx, mtype, mdocno, isdel

示例记录 1:16072913541329219, **2016-07-29 13:54:22**,13,8323,男[45 以上], **2900003115009**, 2, **22002240**, 200328600506004228, 红油桃（中）, , 千克, **22002**, 桃,,,0.422,5.06,0.0,0,,,0

示例记录 2:16060809581811553, 2016-06-08 09:58:40, 18,8334, 女[18-25], 2900001575201, 5, 34150006, 6926458841290, MSU 男童平脚裤 74129, 1*1, 盒, **34150**, 男童裤, 34224.0, 真想你, 1.0, 27.9, 0.0,0,,,0

字段说明:订单编号,**购买时间**,收银员编号,收银机编号,性别年龄,**会员编号**,商品单内编号,**商品编号**,条码,商品名称,包装规格,商品单位,商品类型编号,商品类型名称,品牌编号,品牌名称,购买数量,**金额**,是否打折,是否促销,促销类型,促销单号,是否更正

商品类别结构 **pluno 22002240**：商品类别结构可由商品编号构建，商品编号的前两位，前三位，前四位，前五位为商品逐渐细化的品类。例如"红油桃"的商品编号为 22002240，则其品类由粗到细为 22：蔬果课；220：水果；2200：实果类：22002：桃。

   a) 方法 1 (5 分)
      i. 对每个会员编号的交易记录进行分组，根据商品类别 **pluno** 的第 4 级品类结构，对购买金额 **amt** 进行求和汇总，该用户所有购买产生该用户在第 4 级品类结构的汇总特征。例如**商品编号** 22002240 产生了第 4 级结构 22002，需要对每个用户购买了品类为 22002 的商品金额进行求和汇总。
      ii. 通过 Jaccard 系数计算任意两个会员之间的相似度：若会员 **2900003115009** 购买了品类 22002 的 100 元商品、 品类 34150 的 120 元商品；而会员 **2900001575201** 购买了品类 22002 的 200 元商品、 品类 10113 的 180 元商品，则二者之间的 Jaccard 相似度=100/（120+200+180）=0.2，其中分子 100 为二者共同购买了相同(交集)品类商品的金额，分母 120、200、180 分别为二者购买品类 34150、品类 22002 和品类 10113 的并集商品金额。
      iii. 通过 K-means 算法对用户进行聚类，并利用 Silhouette Coefficient (SC)和 Compactness (CP)这两个指标寻解最优聚类结果。
   b) 方法 2 (5 分)
      i. 对每个会员编号的交易记录进行分组，针对商品类别 **pluno 生成** 4 级品类结构，进行购买金额 **amt** 的求和汇总，以此产生 4 个商品类别的汇总特征，例如**商品编号** 22002240 产生特征名称为 22、220、2200、22002 等 4 个品类，然后针对每个品类进行汇总金额值；
      ii. 通过 Jaccard 系数计算任意两个会员之间的相似度：首先对第一级品类结构，如上 a.ii 方法计算一个相似度 sim_1；依此类推对第二、三、四级品类结构计算相似度 sim_2、sim_3、sim_4；然后通过求平均值（sim_1+sim_1+sim_1+sim_1）/4 即为二者的最终平均值。
      iii. 通过 K-means 算法对用户进行聚类，并利用 Silhouette Coefficient (SC)和 Compactness (CP)这两个指标寻解最优聚类结果。
   c) 方法 3：按照 cluster.pdf 进行聚类设计，求解最优聚类结果。(5 分)
   d) 方法 4：比较、讨论上述算法的优缺点(5 分)。
   e) 方法 5：如有可能设计一个改进更优的聚类算法；(奖励分：5 分)
2. 数据集下载地址：

   链接：https://pan.baidu.com/s/18xjDDjcZYY6yqsbecDrkOw 提取码：ng6a
3. 提交方式：

提交日期：2020/04/28 日 23：59PM，提交内容发送至 tongjidam20@163.com，每个作业提交内容以学号+hw1.zip 作为命名方法；其中包括 5 个子目录，命名方式分别为 q1,q2,q3,q4 和 q5，每个子目录包括对应目的代码和 word 报告。其中报告包括 1）代码运行结果屏幕拷贝；2）讨论分析部分，例如绘制 cluster.pdf 的 Figure 3，分析距离分布情况，还可以作图比较不同聚类簇个数对应的 SC/CP 曲线图；3）性能比较图表，例如以 cluster.pdf 中的 table3 来比较上述 5 个题目的结果。