数据仓库与数据挖掘 作业 2

1. 作业概述

该项目要求同学在掌握数据挖掘常用算法的基础上，对实际的数据集进行有意义的探索，每组从附录提供的三个数据集中选择一个，结合课程讲授内容，提出数据挖掘问题，并进行挖掘和分析。作业评分主要依据作业**完成程度**、**问题的价值与新颖性**、**技术合理性与难度**、**结果正确性**等。

2. 作业要求

1) 每组选择至少 2 种不同类别的知识（如分类、聚类、关联等），提出至少 4 个问题。本次作业不限定所用方法、模型、编程语言、工具软件等，可根据需要选择使用，也可根据需要使用额外的数据集，但需要在报告中特别说明，并提供参考文献等必要引用。
2) 本次作业需提交内容包括：
   a) 实验报告，不超过 3 页（含 3 页），必须使用提供的模板，至少包含以下内容：(1) **详细描述所提出的问题**、**阐述问题的合理性或价值**、**详细描述解决问题的思路和具体方法**、**挖掘结果与分析**。(2) **组员信息**和**组员分数调整**（可选，具体办法与作业 1 相同）。
   b) 源代码
   c) README 文件：包括代码运行环境和运行步骤
3) 鼓励组内成员之间充分讨论和合作，但是禁止各组之间的合作，若发现各组之间作业有抄袭现象，则一律记为 0 分。

3. 作业提交

1) 提交文件命名为：姓名 1_学号 1_姓名 2_学号 2_姓名 3_学号 3.zip，其中压缩包内的文件目录结构为：

```
src
    xx.py
    tt.cpp
    ……
doc
    xxx.pdf（必须提交模板生成的 pdf 文档）
README
```

2) 请于 **2018 年 1 月 2 日 23:59 (含)** 之前提交到网络学堂，不接受其它任何形式的提交；若逾期 3 日内提交，则扣除 30%的分数；若逾期 7 日内提交，则扣除 50%的分数；超过 7 日的，则本次作业记为 0 分。届时助教会在网络学堂上开补交窗口。

数据集

1. KKBox 音乐

该数据集是音乐软件 KKBox 的用户在一段时间内的听歌记录，由 6 个表组成，实际作业中只需要考虑以下 4 个表即可：

1) train.csv: 用户听歌记录，用 target 字段表示用户是否听了这首歌。
2) song.csv: 歌曲的详细信息，如类别、作曲者等。
3) member.csv: 用户的个人信息。
4) song_extra_info.csv: 歌曲的额外信息，包括歌名等。

每个表中详细的字段信息需参阅网址：https://www.kaggle.com/c/kkbox-music-recommendation-challenge/data

该数据集上的常见挖掘问题示例：

1) 根据用户听歌记录、歌曲信息和用户个人信息，挖掘哪些用户是相似的，从而可以向用户推荐兴趣一致的其他用户
2) 根据用户听歌记录、歌曲信息和用户个人信息，挖掘哪些歌曲是相似的，从而可以自动生成不同主题的歌单，也可以向用户推荐其感兴趣的歌曲
3) ……

2. Tmall

该数据集是天猫部分用户的行为日志，实际作业中可以只考虑以下 3 个表：

1) user_info.csv: 用户个人信息，包括年龄和性别等。
2) user_log.csv: 用户何时对何产品进行了何种操作，操作包括点击、加入购物车、购买或标记为喜欢。
3) train.csv: 用户是否会在某一商家重复购买，用 label 字段表示。

每个表中详细的字段信息需参阅网址：https://tianchi.aliyun.com/datalab/dataSet.htm?id=5

该数据集上的常见挖掘问题示例：

1) 哪些商品会经常被一起购买？
2) 同一商家的用户可以分为哪几类，这几类有什么显著特征，从而可以为商家促销提供重要依据。
3) 预测用户是否会在某个商家重复购买。
4) ……

3. Global Firepower 军事力量

该数据集是 2017 年世界上部分国家和地区的军事力量排名数据，只有一个表 GlobalFirePower_multiindex.csv，其中军事力量被分为 Man Power, Air Power, Army Strengths, Naval Power, Financial Resources, Logistical Resources, Natural Resources 和 Geography 这 8

大类，每个大类下又有细分的类别，详细的字段解释可以参照网址：https://www.kaggle.com/blitzr/gfp2017/。另外，也可以参考其他数据集，例如经济数据等，联合军事数据一起分析，可能会得到更有趣的结论。

该数据集上的常见挖掘问题示例：

1) 哪些国家的军事力量部署是相似的？可以进一步结合其他数据分析原因。
2) 军事力量的哪些方面具有很强的相关性，也可以进一部分分析其中的因果关系。
3) 根据某个国家一部分军事力量的值预测其他军事力量的值。
4) ……