测试案例

# 数据准备

本实验涉及三个数据集，即用户、资讯、用户对教育资讯的评价分数，所有数据集数据通过随机模拟的方式生成。

## 用户数据集

用户数据集用于模拟系统中的用户信息，是系统中其他功能的数据基础，正常情况下，该数据集来源于系统初始化数据和用户提交注册，此处为了验证说明问题，采用按照一定生成规则尽可能模拟真实情况下的用户注册数据。用户信息包含以下字段：

* **用户ID：**根据序列顺序生成5位的数字编码，不足位数的编号左补0，例如00001、00002等，以此类推。
* **用户等级：**随机生成1-20级，当用户参与讨论的问题越多、答案被采纳的次数越多，等级越高。
* **姓名：**根据百家姓列表及常用男性人名汉字列表和女性人名常用汉字列表，随机组合成2位或者3位的人名。
* **性别：**男性和女性随机生成性别，并与姓名相匹配，例如“赵红”这个名字一般是女性，“李强”这个名字一般是男性。
* **地址：**通过城市列表、道路列表等元数据，加上随机生成的门牌号，组合成用户的地址信息。
* **电话：**通过手机号前3位固定列表，加上8位随机数字，模拟用户的联系电话。
* **邮件：**通过随机模拟邮箱前缀的字符，加上固定邮箱后缀列表，模拟用户的电子邮件地址，例如：food\_desk@163.com。

按以上字段及规则随机生成100条用户信息。

## 用户标签数据集

## 教育资讯数据集

教育资讯是用于收集用户喜好，采集用户对于某类新闻的敏感度和关注度的载体数据。正常情况下，系统通过每天在互联网上采集具有针对性的、公开的教育新闻包含以下字段：

1）资讯ID，根据序列顺序生成5位的数字编码，不足位数的编号左补0。如00001、

00002等，以此类推。

2）资讯标题，通过抓取“教育新闻网”（网址：http://jijiao.jyb.cn/xw/qy/）的新闻数据，将标题字段按规则抓取。

3）资讯内容，通过抓取的新闻数据，将内容字段按规则抓取下来。

4）发布时间，通过抓取的新闻数据，生成相应新闻的发布时间。

按以上字段规则，抓取生成100条新闻数据。

## 用户评价数据集

该数据集是用户对教育资讯的评价，包含以下字段：

1）用户ID，关联用户表的用户ID。

2）新闻ID，关联教育资讯表的资讯ID。

3）分数，随机生成1-10之间的整数。

按以上字段规则，随机生成100条用户评价数据。

# 建立用户画像

## 调查问卷

## 收集用户喜好

## 结论

# 采集新闻数据

## 数据来源

## 爬虫设计

## 抓取策略

## 结论

# 推荐相似用户

## **测试目的**

通过分析家长和老师对教育资讯的评价，动态为每一位用户推荐与之相似的用户。

## 测试方法

指定一个用户ID，从用户评价的数据集中，选取部分用户的评价数据作为分析样本，为简单起见，选择3个用户共17条数据。然后利用皮尔逊相似度算法分别计算这些用户与指定用户的相似度，取值范围为-1~1之间的小数，具有两位精度，值越大表示相似度越高，值越小表示相似度越低。

## 输入数据

指定00001用户，并根据以下裁剪数据计算出与00001用户最为相似的用户

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 用户ID | 新闻ID | 评分 |
| 1 | 00001 | 00001 | 2.5 |
| 2 | 00001 | 00002 | 3.5 |
| 3 | 00001 | 00003 | 3.0 |
| 4 | 00001 | 00004 | 3.5 |
| 5 | 00001 | 00005 | 2.5 |
| 6 | 00001 | 00006 | 3.0 |
| 7 | 00002 | 00001 | 3.0 |
| 8 | 00002 | 00002 | 4.0 |
| 9 | 00002 | 00004 | 5.0 |
| 10 | 00002 | 00005 | 3.5 |
| 11 | 00002 | 00006 | 3.0 |
| 12 | 00003 | 00001 | 3.0 |
| 13 | 00003 | 00002 | 3.5 |
| 14 | 00003 | 00003 | 1.5 |
| 15 | 00003 | 00004 | 5.0 |
| 16 | 00003 | 00005 | 3.5 |
| 17 | 00003 | 00006 | 3.0 |

## 计算方法

通过皮尔逊相似度的算法实现，可以计算出任一两个用户的相似度，皮尔逊公式如下（来源于维基百科）：

C:\Users\Administrator\Desktop\2017-02-04_151555.jpg

其中，E是数学期望，cov表示协方差，δX和δY是标准差。

## 输出数据

通过计算得出，00002用户与00001的相似度为：0.7470，00003用户与00001的相似度为0.3961，所以针对00001用户，推荐与之更加相似的00002用户。

## 结论

# 为用户推荐教育资讯

## 测试目的

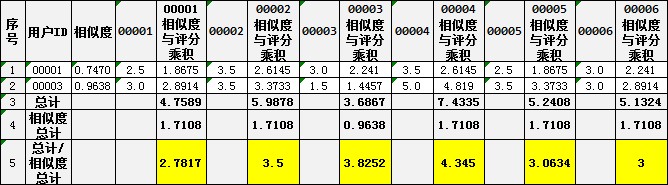
根据与指定用户相似用户的关联数据集，计算出用户可能会感兴趣的教育资讯。

## 测试方法

## 输入数据



## 输出数据



## 结论

上图中最后一行黄色标出的数字，即为给用户预测的教育资讯评分，从大到小排列，即可计算出用户可能最感兴趣的教育资讯。

# 全文搜索

## 搜索数据源定义

## 关键词拆分

## 搜索结果展示

# 总结

根据以上实验结果，得出以下结论：

1. 根据用户的基础数据，并对数据进行分析，可以正确得出用户的基本“画像”，即用户标签。根据这些标签，可以正确对用户进行分组和归类。
2. 根据每个家长用户对教育咨询的评价，通过基于用户的协同过滤推荐算法，可以正确计算用户的相似度，可以正确的给每一个用户推荐与之相似的用户列表，并按照相似度从高到低排序。
3. 根据每个家长用户之间的相似度，以及对教育资讯的评价，通过加权的算法预测出该用户对没阅读过的资讯的评价，可以正确得出为该用户推荐的资讯列表，结果按照预测分数从高到低排序。

根据以上测试结果，验证了推荐算法在家长圈的应用中起到了关键的推荐作用。对于每一个家长用户，系统都可以向他推荐与之相似度最高的用户，这就可以使有相同话题的家长在共同的圈子里相互分享、讨论解决家庭教育问题；同时，对于每一个家长用户，系统都可以向他推荐他未阅读过的文章，从而积累该用户对于该类文章的评价，积累更多该类用户的相似数据，使系统更加智能。