JData 算法大赛 FAQ

参赛篇

1、 Q：非大陆用户如何通过实名认证并参加京东 JData 算法大赛？

A： 目前 DF 平台实名认证功能仅针对大陆用户，非大陆用户报名参赛并被要求实名认证时，请提供 DF User ID、以及本人有效证件（护照、学校 ID 卡、驾驶证、台胞证中任意一种）的照片或扫描件，发送至 contact@datafountain.cn，我们将为您进行人工验证。

赛题篇

1、 Q：赛题数据量级

A：本次赛题数据共 6 个 csv 文件。

文件名	记录条数（不含表头）
JData_Action_201602.csv	11,485,424
JData_Action_201603.csv	25,916,378
JData_Action_201604.csv	13,199,934
JData_Comment.csv	558,552
JData_Product.csv	24,187
JData_User.csv	105,321

文件 MD5 校验：

6a645e8db9f8a532777ffc7894be8d7d JData_Action_201602.csv

d08a7bdb40bd9841bcf412d567785c34 JData_Action_201603.csv

9a035deaf8b08d51a9f6aa242fb1a261 JData_Action_201604.csv

06b6ae7dd5bdf7f9a72714e5458d5399 JData_Comment.csv

de71f22ac871a52daa13d1c5011ee655 JData_Product.csv

9fe9dbc1dfef366bbc6f8b989da26e28 JData_User.csv

2、 Q：赛题数据问题件的编码格式

A：提供的数据文件编码格式为 GBK。

3、 Q：评测数据中一个用户是否存在购买多个商品？

A：不存在，一个用户（user_id）只购买一个所提供的候选商品集合中的商品（sku_id）；用户购买候选商品集合之外的商品无需提交。

4、 Q：关于商品集合 P 和 S

A：（1）商品集合 P 为候选商品集合（JData_Product.csv），即参赛者预测的结果中的 sku 需要在集合 P 中；

（2）商品全集 S=商品表 Product 中 sku+行为表 Action 中 sku+评价表中 sku；

5、 Q：评论数据是当天的还是累计的

A：评论数据为截止到当日 23:59:59 的累计数据

6、 Q：关于 model_id 的解释

A：model_id 表明用户在页面上点击了哪一个位置，只有当 type=6 时 model_id 才可能有值；

数据中可能存在一些空值，有可能是异常值，也有可能是在页面上点击了一个空白的位置产生的数据，请参赛者自行理解并处理。

7、 Q：提交数据文件有什么要求？

A：（1）提交数据文件需要去重处理，即：提交结果文件中用户 user_id 唯一，不存在重复；

（2）预测用户下单的商品 sku_id，必须在商品子集 P 中；

（3）提交数据文件格式为 csv，编码为 utf-8，第一行为表头，即：第一列为“user_id”，第二列为“sku_id”，提交数据文件截图如下：

	A	B	C
1	user_id	sku_id	
2	1000	147796	
3	10000	166707	
4	100002	9702	
5	100006	154636	
6	100008	169819	
7	100016	109083	
8	100031	86842	
9	100036	36371	
10	100037	154636	
11	100039	145946	
12			
13			
14			
15			

8、 Q：关于日期和时间格式

A：日期格式统一为"yyyy-mm-dd"，时间格式统一为"yyyy-mm-dd hh:mi:ss"

9、 Q：关于用户表中存在 3 条用户注册时间为空，9 条用户注册时间晚于 2016 年 4 月 15 日情况

A：此情况为真实生产环境中的异常数据，请选手自行理解并处理。

10、Q：关于 Product 表中 cate 值全为 8 是否有问题情况

A：数据没有问题，请选手自行理解。

11、Q：关于评测算法的详细说明以及示例

A：（1） A 榜阶段，$F1_1$ 正确率=预测 A 榜 user_id 正确数量/提交数量，召回率=预测 A 榜 user_id 正确数量/A 榜数据总量；$F1_2$ 正确率=预测 A 榜 user_id+sku_id 正确数量/提交数量，召回率=

预测 A 榜 user_id+sku_id 正确数量/A 榜数据总量；

（2） B 榜阶段，$F1_1$ 正确率=预测 B 榜 user_id 正确数量/提交数量，召回率=预测 B 榜 user_id 正确数量/B 榜数据总量；$F1_2$ 正确率=预测 B 榜 user_id+sku_id 正确数量/提交数量，召回率=预测 B 榜 user_id+sku_id 正确数量/B 榜数据总量；

示例：假设评测数据为

	A	B	C
	user_id	sku_id	A/B
	1	1	A
	10	2	A
	11	1	A
	100	2	A
	101	3	A
	1000	1	B
	1001	2	B
	10000	1	B
	10001	2	B
	100000	3	B

参赛者小 a 提交的结果为

user_id	sku_id
1	1
10	1
1000	1
1001	1
10000	1
10001	1

那么参赛者小 a 的 A 榜阶段得分计算如下：

$F1_1$ ：正确率=2/6，召回率=2/5，

$F1_1$ =6*(2/5)*(2/6)/(5*(2/5)+(2/6))

$F1_2$ ：正确率=1/6，召回率=1/5，$F1_2$ =5*(1/5)*(1/6)/(2*(1/5)+3*(1/6))

总得分为： 0.4*$F1_1$ +0.6*$F1_2$ =0.24825

参赛者小 a 在 B 榜阶段得分计算如下：

$F1_1$ ：正确率=4/6，召回率=4/5，

$F1_1$ =6*(4/5)*(4/6)/(5*(4/5)+(4/6))

$F1_2$ ：正确率=2/6，召回率=2/5，$F1_2$ =5*(2/5)*(2/6)/(2*(2/5)+3*(2/6))

总得分为： 0.4*$F1_1$ +0.6*$F1_2$ =0.49651

12、Q：关于浏览、点击有什么区别

A：（1）浏览，打开商详页至页面加载完成；

（2）点击，在商详页中点击某个元素或位置。

13、Q：关于提交数据后无法计算得分给出的异常提示

A：选手提交结果文件后，评测系统会首先对所提交的结果文件作必要的合法性检查，通过后才能计算并给出得分。对于没有通过合法性检查的结果文件，系统会给出如下的错误类型提示：

（1）大小错误，是指文件大小超出限制，本次比赛的限制为<2MB；

（2）格式错误，是指结果中的字段数不对、字段分隔符、换行符不对，请按照比赛说明中要求的格式编写结果文件，换行符推荐用\n；

（3）编码错误，是指结果为非文本格式或包含非法字符、字符集错误等，推荐使用 UTF-8 无 BOM 编码的文本格式；

（4）逻辑错误，是指结果文件的内容在逻辑上不符合比赛要求，本次比赛的提交结果中不能包含重复的 user_id，每一行上只能包含一个 sku_id。

14、Q：行为表里为什么出现了一些重复数据？

A：行为数据，均为生产环境中的真实数据，请选手自行理解并处理。

15、Q：关于 Action 表中 user_id 说明

A :Action 表中 user_id 为浮点型 ,转为整型后与 User 表中 user_id 一一对应 ,请参赛选手自行转换。提交结果时，请提交整型数据。

16、Q：关于旧数据使用说明

A：旧数据是指 2017-04-04 12:00:00 之前下载的数据，该数据可以用来分析，但严禁用来训练模型进行比赛，如资料审核过程中，一旦发现，将取消排名资格。

奖励篇

1、 Q：获奖参赛队伍每个成员是否都能拿到 Special Offer？

A：奖项中所涉及的 Special Offer 均是真 Special Offer，只要获奖参赛团队中成员通过专家评委的考核，就能拿到 Special Offer，参赛团队成员通过几人，京东就敢发出几个 Special Offer；获得冠亚季军的每个团队至少发出一个 Special Offer。