数据集说明:

数据集 1　ques_info.txt

包含训练集中涉及到的所有问题列表，每一行代表一个问题的相关信息，每一行有 7 列，列之间采用 /tab 分隔符分割。

数据格式如下：

[问题 ID　问题创建时间　问题标题的单字编码序列　问题标题的切词编码序列　问题描述的单字编码序列　问题描述的词编码序列　问题绑定的话题 ID]

1.问题 ID，格式为 Qxxx。

2.问题创建时间，格式为 D3-H4。

3.问题标题的单字编码序列，格式为 SW1,SW2,SW3,...,SWn ，表示问题标题的单字编码序号。

4.问题标题的切词编码序列，格式为 W1,W2,W3,...,Wn ，表示问题标题的切词编码序号，如果问题标题切词后为空，则用 -1 进行占位。

5.问题描述的单字编码序列，格式为 SW1,SW2,SW3,...,SWn ，表示问题描述的单字编码序号，如果问题没有描述，则用 -1 进行占位。

6.问题描述的切词编码序列，格式为 W1,W2,W3,...,Wn ，表示问题描述的切词编码序号，如果问题没有描述或者描述切词后为空，则用 -1 进行占位。

7.问题绑定的话题 ID，格式为 T1,T2,T3,...,Tn ，表示问题绑定的话题 ID 的编码序号， 如果问题没有绑定的话题，则用 -1 进行占位。

数据集 2　user_info.txt

包含训练集中用户相关特征信息，每一行代表一个用户的相关信息， 每一行有 21 列，列之间采用 /tab 分隔符分割。

数据格式如下：

1.用户 ID，格式为 Mxxx。

2.性别。

3.创作关键词的编码序列，格式为 W1,W2,W3,...,Wn ，表示创作关键词的编码序号，如果创作关键词为空，则用 -1 进行占位。

4.创作数量等级。

5.创作热度等级。

6.注册类型。

7.注册平台。

8.访问频率，有五种取值［new | daily | weekly | monthly | unknow］，分别对应为［新用户 | 日活用户 | 周活用户 | 月活用户 | 未知］。

9.用户二分类特征 A，两种取值 0 或 1。

10.用户二分类特征 B，两种取值 0 或 1。

11.用户二分类特征 C，两种取值 0 或 1。

12.用户二分类特征 D， 两种取值 0 或 1。

13.用户二分类特征 E， 两种取值 0 或 1。

14.用户分类特征 A，格式为 MDxxx。

15.用户分类特征 B，格式为 BRxxx。

16.用户分类特征 C，格式为 PVxxx。

17.用户分类特征 D，格式为 CTxxx。

18.用户分类特征 E，格式为 PFxxx 。

19.用户的盐值分数。

20.用户关注的话题，格式为 T1,T2,T3,...,Tn ， 表示用户关注话题的序列编号（最多 100 个），如果关注话题为空，则用 -1 进行占位。

21.用户感兴趣的话题，格式为 T1:0.2,T2:0.5:T3,-0.3,...,Tn:0.42 ， 表示用户感兴趣的话题序列编号及喜好程度分数（最多 10 个），如果感兴趣话题为空，则用 -1 进行占位。

数据集 3 train.txt

包含用户最近 1 个月的邀请数据，每一行代表一个问题邀请的相关信息，每一行有 4 列，列之间采用 / tab 分隔符分割

数据格式如下：

[Qxxx Mxxx D3-H4 label]

1.邀请的问题 ID, 格式为 Qxxx。

2.被邀请用户 ID, 格式为 Mxxx。

3.邀请创建时间, 格式为 D3-H4。

4.邀请是否被回答, 如果值为 1 表示邀请被回答, 值为 0 表示邀请没有被回答。