**代码说明**

**1 用户-电商行为数据**

脚本名称：getItem.sh

关联的表：tmp\_latend\_days

输出目录：/user/bre/guangming.li/latend\_2/item/（可配置）

用到的脚本：filterCateCn.py

使用范例：sh getItem.sh output\_dir（output\_dir表示自定义的输出目录）

说明：从冷集群的tmp\_latend\_days表中的数据中提取出gid,cate\_index,date信息，即得到用户-电商行为数据

**2 用户-媒体行为数据**

**2.1 媒体类别数**

脚本名称：hive操作，无脚本

关联的表：news\_from\_hbase

输出目录：/user/bre/guangming.li/test\_3\_cate/distinct\_catevec.data（冷集群）

用到的脚本：无

使用范例：无

说明：从news\_from\_hbase表中，选择nb\_cid, nb\_iid, np\_bfdcatveclist字段，分别为cid, iid和媒体类别名称（利用three\_news\_mapping.txt文件，可转换为类别代号）

**2.2 用户行为数**

脚本名称：hive操作，无脚本

关联的表：raw\_kafka\_input\_dt0

输出的表：latend\_media\_kafka

用到的脚本：无

使用范例：无

说明：从raw\_kafka\_input\_dt0表中，选择gid, customer, iid, l\_date字段

**2.3 用户-媒体行为数**

脚本名称：createTraining.sh

关联的表：latend\_media\_kafka, latend\_media\_index

输出表：latend\_media\_training

用到的脚本：无

使用范例：sh createTraining.sh output\_dir

说明：从Mini集群的latend\_media\_index表（cid, iid, index）和latend\_media\_kafka（gid, cid, iid, l\_daate）中的数据中提取出gid,cate\_index,date信息，即得到用户-媒体行为数据

**2.4 区分媒体和电商**

脚本名称：closemedia.sh

关联的表：latend\_media\_training

输出目录：/user/bre/guangming.li/latend/data/media/（可配置）

用到的脚本：closemedia.py

使用范例：sh closemedia.sh output\_dir

说明: 为区分媒体和电商的数据，以及后续安排媒体和电商的顺序位置，为Mini集群的latend\_media\_training表（gid, cate\_index, date）中的数据添加唯一标识，具体为在date字段后添加’#‘标记媒体数据

**3 潜在需求预测模型-数据提取**

脚本名称：getEleven.sh（调用getEleven.py文件）

关联目录：/user/bre/baoquan.zhang/latentIndex/reduce/output/part-00000

输出目录：/user/bre/guangming.li/latend/data/datasplit/source/（可配置）

用到的脚本：getEleven.py getEleven.py

使用范例：sh getEleven.sh 859 output\_dir filt 10 5

其中859是待处理的类别标号；output\_dir表示自定义的输出目录；filt表示是否进行数据过滤（主要针对行为数较少的情况）操作，如果该参数为“filt”则表示要进行数据预处理，其他任何字符串则表示不进行预处理；10表示对电商类别总行为数的限制；5表示对媒体类别总行为数的限制

说明：从电商和媒体拼接后的数据中，提取出关注月份的数据，输出数据格式为：gid 关注月份下一个月的行为数 关注月份所有电商与媒体的类别的行为数

**4 潜在需求预测模型-打标签**

脚本名称：formatData.sh

关联目录：/user/bre/guangming.li/latend/data/datasplit/source/part-00000

输出目录：/user/bre/guangming.li/latend/data/formatdata/part-00000（可配置）

用到的脚本：formatData.py formatData.sh

使用范例：sh formatData.sh 859 output\_dir 3 1.5 0.6

其中859是待处理的类别标号；output\_dir表示自定义的输出目录；3是对行为数的限制，低于此限制的，仅当下月行为数超过上涨阈值时才认为潜在需求，否则低于下降阈值时也认为是潜在需求；1.5是上涨阈值，表示下月行为数相对该月行为数超过1.5倍时，才认为是潜在需求；0.6是下降阈值，表示下月行为数相对该月低于60%时，认为是潜在需求

说明：根据提取出关注月份的数据，打潜在需求标签

**5 潜在需求预测模型-划分训练集和测试集（从此步以后是单机操作）**

脚本名称：splitTrainTest.py

关联目录：从第4步的输出目录down下来的数据

输出文件：res\_train\_xxx.data res\_test\_xxx.data

用到的脚本：splitTrainTest.py

使用范例：python splitTrainTest.py zero.data 0.7

其中zero.data表示标签为0的数据集合，输出文件格式为res\_train\_前缀加上文件名（不包括后缀）；0.7表示分别取70%作为训练集，30%为测试集

说明：从带标签的数据中，分别按照0和1划分训练集和测试集

**6 潜在需求预测模型-训练模型-baseline**

脚本名称：1\_preLatend\_baseline.py

关联文件：train.data test.data

输 出：baseline结果

用到的脚本：1\_preLatend\_baseline.py

使用范例：python 1\_preLatend\_baseline.py

说明：根据训练集训练模型，输出在测试集上的基线指标结果

**7 潜在需求预测模型-训练模型-lasso**

脚本名称：2\_preLatend\_lasso.py

关联文件：train.data test.data

输 出：lasso结果

用到的脚本：2\_preLatend\_lasso.py

使用范例：python 2\_preLatend\_lasso.py

说明：根据alpha值的不断变换，输出相应类别维度的相关系数

**8 潜在需求预测模型-训练模型-topN组合**

脚本名称：3\_preLatend\_lasso\_top\_N\_features.py

关联文件：train.data test.data

输 出：不同特征数的分类结果

用到的脚本：3\_preLatend\_lasso\_top\_N\_features.py

使用范例：python 3\_preLatend\_lasso\_top\_N\_features.py

说明：根据不同的类别数量，输出相应的分类准确率

**9 潜在需求预测模型-特征组合**

脚本名称：4\_select\_features.py

关联文件：train.data test.data

输 出：不同特征数的分类结果

用到的脚本：4\_select\_features.py

使用范例：python 4\_select\_features.py

说明：根据特征相关系数的强弱，从强到弱排序，然后依次输入模型进行训练，如果加入某特征后效果变好，则保留该特征，否则跳过