用户画像分类（性别）

基本情况：

1）特征筛选：选择有男女区分度的特征，选择的标准是该特征男女区分度（该特征在男性中占比和在女性中所占比之差）大于某阈值1（比如0.5），并且该特征在所有人群中占比不低于某阈值2（该特征不会过于少导致用户的特征向量过于稀疏，比如取0.001平均在1000人里该特征至少出现一次）

2）数据：需要分类用户近三个月的所有数据

有标注的样本：目前通过qq号或者淘宝号爬出来的，基本上男女比例3：1的样子，准确度未知。

3）模型：逻辑回归

4）训练阶段：

将有标注样本对应的mix_uid的行为数据下载到本地使用Python进行特征筛选，生成训练矩阵等，使用Sklearn下的逻辑回归模型来训练。

5）分类阶段：

用Hive Streaming读取训练好的模型参数，结合排序后用户的行为数据进行分类。

见用户画像分类里的streaming_nosklearn.py

过程与结果：

之前邮件发出过一版过程与结果：

1. 先根据解析出的 qq号/性别 列表匹配出 mix_uid/性别 列表,加上一个序列号，保存到一个用户/index/性别的字典中，存在一个文件里

样例

```
Y054AOYY78WZWTWDY5QYB62Z9US4RR3W    0    Male
55O6XT2ZYWL7YWYQ5C84X3VF29PXAPND    1    Male
US485633ATU91PWZZ019A7TA55EQQ408    2    Female
X1SU532WBZRVB15SQA0S22XNA303ZZUA    3    Female
XY8R0CV2WZVWM1T75QS1MPUT529Q13V5    4    Male
BU3R1116UUB9ZD1PT7X9YW28S2MB7B21    5    Female
13P6RRW7UWABBV59A40AAW8DA1000697    6    Male
```

2. 找到所有这些mix_uid的行为，如近90天，取 mix_uid,ActionType,ActionIndex这几列，联合ActionType_ActionIndex为特征key的维度

```
'mix_uid/ 特征key / gender分类'
0012C25V6YXZ0235N224O18SNYZAFZ82    c_54    Male
001YW18225PWS4USDTP5Y7XRZS9083U6    b_505692    Male
003X90C0BP7P0R844ZET7U2D14A48Z12    b_507712    Female
0088V21YVUUXZ0A9UQ5YSZ2CP10OU475    c_20    Female
00QB15Z011RPX120WX5Z9Z0X3O3X0ZS5    b_515978    Male
00WVYQSYXOC2QXY5100YW9T6AAF042R5    c_36    Male
00WYUT3W706TQXR320XXPDX308R3AW0W    c_35    Female
0112C941Q349V3V2T51B2Z0211Z5VQ10    c_7502  Male
01134VB462B3T2Z8WY4VZW295039T2U7    b_64    Male
```

3. 分析下所有行为标签来筛选特征，目前选择的标准：

1) 该行为标签在所有人中出现的比率大于0.001（至少1000个人中有一个标签）

2) 该行为标签在男女的lift值之差大于一定阈值（现在选的0.3）

lift（行为标签a，男） = 该样本在男性UV中出现的总次数/男性UV的总数

lift (行为标签a，女) = 该样本在女性UV中出现的总次数/女性UV的总数

diff = abs(lift(male) - lift(female)) / min(lift(male/female)) > 0.3

=> 有效行为标签当特征

这样筛选出特征来，保存到一个字典中，存在一个txt文件中，用于生成特征矩阵/向量

样例

f_2:564:: 0
4_4031188 1
b_130 2
c_44 3
f_6:501813:: 4

分析筛选后的特征： 特征/特征名/lift值之差

偏男性特征：

c_606 葫芦侠3楼 2.519822321
c_7345 掌上穿越火线 2.142698501
c_10708 掌上TGP 2.037941884
c_1012 掌上道聚城 2.037941884
c_615 多米音乐 1.828428651
c_4182 CF掌上穿越火线 1.828428651
c_253 拉卡拉 1.514158801
c_1412 分期乐 1.451304831

偏女性特征：

c_266 驴妈妈 -1.485920936
c_79 豆果美食 -1.545583039
c_1031 消灭星星 -1.620453128
c_4584 布丁动画 -1.651648999
c_183 楚楚街 -1.672862191
c_199 阿里巴巴 -1.692443599
c_9956 快读全本小说 -1.784231449
c_14264 视吧 -1.784231449
c_120 聚美优品 -1.871541843
c_9050 P1－最时尚的照片分享社区 -1.916813899
c_512 课程格子 -2.181978799
c_26921010 -2.181978799
c_713 宝宝钓鱼 -2.181978799
c_153 荔枝fm -2.252689439
c_514 有道云笔记 -2.409262998
c_6656 互动吧-发活动、找活动就上互动吧 -2.636547198
c_1499 InstaSize -3.917603598

中性特征（没有明显区分性，会被过滤掉的特征, abs(diff) < 0.3 ）：

c_12 腾讯新闻 -0.007716965
c_18 微信 -0.008219422
c_19 QQ -0.008318916
c_2 淘宝 -0.00910989
c_90 优酷视频 -0.010174679
c_223 安居客 -0.0124478
c_42 百度贴吧 -0.01537301
c_621 掌阅iReader -0.015525148

4. 这样根据两个字典和2里的行为数据流，初始化生成一个全0矩阵feature_mat，维度 UV数*特征数
   根据行为数据流，每读到一行，改写
   feature_mat[Dict_uv[mix_uid]][Dict_feature[feature_key]] = 1

   目前只把在所有特征中，有大于n条特征的uv当做有效uv进行训练和测试（现在取n > 10）
   就组织成了所有样本的特征矩阵

训练数据 X = 特征矩阵 （feature_mat）
y = X每行对应的mix_uid对应的性别（同样存在1中的字典里）

因为逻辑回归相当于 X*w = y 来估算出 w的模型参数，即下图的公式 x为X矩阵，y为y矩阵
theta为w矩阵，来用优化方法找到theta让误差E尽量小，估计出的w矩阵即为逻辑回归的参数。

$$x=\begin{bmatrix}x_1\\ \dots\\ x_m\end{bmatrix}=\begin{bmatrix}x_{11} & \dots & x_{1n}\\ \vdots & \ddots & \vdots\\ x_{m1} & \dots & x_{mn}\end{bmatrix}, y=\begin{bmatrix}y_1\\ \dots\\ y_m\end{bmatrix}, \theta=\begin{bmatrix}\theta_0\\ \dots\\ \theta_n\end{bmatrix}$$

$$A=x\bullet\theta=\begin{bmatrix}x_{10} & \dots & x_{1n}\\ \vdots & \ddots & \vdots\\ x_{m1} & \dots & x_{mn}\end{bmatrix}\bullet\begin{bmatrix}\theta_0\\ \dots\\ \theta_n\end{bmatrix}=\begin{bmatrix}\theta_0x_{10}+\theta_1x_{11}+\dots+\theta_0x_{1n}\\ \dots\\ \theta_0x_{m0}+\theta_1x_{m1}+\dots+\theta_0x_{mn}\end{bmatrix}$$

$$E=h_\theta(x)-\mathrm{y}=\begin{bmatrix}g(A_1)-y_1\\ \dots\\ g(A_m)-y_m\end{bmatrix}=\begin{bmatrix}e_1\\ \dots\\ e_m\end{bmatrix}=g(A)-y$$

5. 训练模型，引用Python机器学习包，sklearn的线性模型 逻辑回归
from sklearn.linear_model import LogisticRegression

classifier = LogisticRegression(class_weight='balanced')
用该模型来fit X，y矩阵就行，然后交叉验证准确度
metric = cross_val_score(classifier, X, y,cv=5,scoring='accuracy')

目前交叉验证结果：
阈值1，选取行为标签在所有人中出现的比率大于阈值0.001
阈值2，该行为标签在男女的lift值之差大于一定阈值（现在选的0.3） 有721维特征
阈值3， 目前只把在所有特征中，有大于n条特征的uv当做有效uv进行训练和测试（目前n>5）

交叉验证准确度为 69~70%的样子。
下面是10-Folds Cross Validation的结果
[0.69747899 0.69467787 0.68814192 0.69187675 0.68160598 0.70028011
  0.69719626 0.68878505 0.68381665 0.69597755]
在已经有的样本条件下我试了很多不同的参数和特征组合，差不多这样已经是最优的了。

如果换用其他模型比如SVM能够获得更高的交叉验证准确性，但是在分类阶段更复杂，可解释和分析性也不如逻辑回归模型强。

6. 对以后的样本做分类

使用Python的Hive Streaming，先读取训练好的逻辑回归模型参数，再读取用户的数据，可以分批读取来做分类（比如day_id > 20160701 and province = “jiangsu” 来对江苏省近三个月所有用户做性别分类）

如果做2分类score >= 0.5归于1类即Male，<0.5归于0类即Female。

或者不直接分类，保存score，以后想要女性人群比如200万，就取Top200万score小的mix_uid就可以。

每次测试一批新的样本，保存

mix_uid / score / gender_class到指定hive表里