* 样本格式:

类别名称<separator>text

例如：新闻资讯\001新浪网为全球用户24小时提供全面及时的中文资讯，内容覆盖国内外突发新闻事件、体坛赛事、娱乐时尚、产业资讯、实用信息等，设有新闻、体育、娱乐、财经、科技、房产、汽车等30多个内容频道，同时开设博客、视频、论坛等自由互动交流空间。

* 生成标准样本及词典文件、类别标记文件

一层分类使用cn.clickwise.clickad.sample.Sampler 来完成此功能，使用方法是：

Usage:<field\_num> <sample\_field\_index> <label\_field\_index> <separator> <gendict> <genlabeldict> <gensample>

field\_num : 输入的字段个数

sample\_field\_index: 样本体所在的字段，从0开始，即0表示第一个字段

label\_field\_index: 标记所在的字段，从0开始，即0表示第一个字段

separator:字段间的分隔符，001 表示 字符001，blank 表示\s+ 即连续空格 ,tab 表示

gendict:生成的词典路径,从1开始

genlabeldict:生成的标记索引路径，从1开始

gensample:生成的样本路径

脚本使用示例：

cat train\_text.txt | java –cp user\_click.jar cn.clickwise.clickad.sample.Sampler 2 1 0 001 gendict.txt genlabeldict.txt gensample.txt

注：生成的词典单词和索引之间，类别索引文件类别和索引之间，样本类别索引和单词之间均用separator分隔开

生成的样本格式为：

labelIndex<separator>[<wordIndex:count> ]\*

labelIndex:类别的索引

wordIndex:某个单词的索引

count：该文档中该单词出现的次数

不同的单词之间用空格分隔开

* 生成训练模型
* 生成medlda可执行文件的命令是：

g++ \*.cpp svmlight/\*.cpp svm\_multiclass/\*.cpp -o medlda

* 将样本(例如上步的gensample.txt)分成训练样本和测试样本：

echo data/mmtrainSample.txt data/mtrainSample.txt data/mtestSample.txt | php simRandom.php

mmtrainSample.txt为所有的样本(例如上步的gensample.txt)

mtrainSample.txt 生成的训练样本

mtestSample.txt 生成的测试样本

* 训练模型：

./medlda est 120 4 10 16 64 temp random

est 表示这是训练样本的选项

120 主题设置的数目

4 类别的个数

10 inference的循环次数

16 初始的C值

64 DELTA\_ELL的值，即模型预测错误的惩罚

配置文件为setting.txt 文件内容示例：

var max iter 20 ##最大的循环次数

var convergence 1e-4 ##程序终止的阈值

em max iter 10 ##em算法的最大循环次数

em convergence 1e-4 ##em 算法的终止阈值

model C 16.0 ##初始的C 值

init alpha 0.1 ##初始alpha值

svm\_alg\_type 2 ##svm 算法的类型，解决medlda的子问题

alpha 1 ## ESTIMATE\_ALPHA值

inner-cv false ##是否进行交叉检验

inner\_foldnum 4 ##交叉检验的次数

cv\_paramnum 10 ##

25.0

29.0

33.0

37.0

41.0

45.0

49.0

53.0

57.0

61.0

train\_file: data/mtrainSample.txt ##训练样本路径

test\_file: data/mtestSample.txt ##测试样本路径

* 测试模型：

./medlda inf 4 temp120\_\*

生成的结果文件位于temp120\_\* 的evl-performance.dat中，格式为

accuracy: 0.1032110092

perword likelihood1: -10.7591865769

perword likelihood2: -11.0659411910

1 1

2 2

2 2

2 1

2 2

2 2

2 2

13 13

第一列为预测类别，第二列为真实类别(实际分类时可以赋予任何值)