* 样本格式:

类别名称<separator>text

例如：文化娱乐|视频\001爱奇艺-中国领先的视频门户，高清影视剧，网络视频在线观看 高清视频,正版视频,热门电影,经典大片,热播电视剧,综艺,动漫,纪录片,音乐,MV,在线观看,免费观看,网络视频,视频搜索,在线视频,点播 爱奇艺（iQIYI.COM）,网络视频播放平台；是国内首家专注于提供免费、高清网络视频服务的大型视频网站。爱奇艺影视内容丰富多元，涵盖电影、电视剧、综艺、纪录片、动画片等热门剧目；视频播放清晰流畅，操作界面简单友好，真正为用户带来“悦享品质”的观映体验

* 生成标准样本及词典文件、类别标记文件

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

Usage:<field\_num> <sample\_field\_index> <label\_field\_index> <level\_num> <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

level\_num: 标记的层数

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.SampleMulLevel 2 1 0 2 001 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。不同层次的类别名称和类别索引之间用”|”分隔开。

生成的样本格式为：

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

mullabelIndex:多层类别的索引,如1|2

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
* 生成可执行文件svm\_multiclass\_learn和svm\_multiclass\_classify的命令是：

make

将样本(例如上步的gensample.txt)分成训练样本和测试样本：

echo example4/nsim\_train.txt example4/sim\_train.txt example4/sim\_test.txt |php simRandom.php

nsim\_train.txt 为所有的样本(例如上步的gensample.txt,要稍微修改一下分隔符)

sim\_train.txt 生成的训练样本

sim\_test.txt 生成的测试样本

注：nsim\_train.txt、sim\_train.txt、sim\_test.txt的格式均为：

<fristLabelIndex> <secondLabelIndex> [<wordIndex:count> ]\*

firstLabelIndex:第一层类别的索引

secondLabelIndex:第二层类别的索引

wordIndex：单词的索引

count:单词在该文档里出现的次数

不同的单词之间用空格分开，第一二层类别索引之间，类别索引、单词之间也用空格分开例如：

1 1 1:2 2:2 3:2 4:2 5:1 6:1 7:1 8:1 9:2 10:2

表示第一层类别索引为1

第二层类别索引也为1

该文档的单词为：1:2 2:2 3:2 4:2 5:1 6:1 7:1 8:1 9:2 10:2

还需要一个所有类别索引的文件：例如example4/lll.txt，在svm\_struct\_api.c中的read\_all\_labels()方法中读取。

格式为：第一层类别索引\_第二层类别索引，例如

1\_1

2\_2

1\_3

3\_4

4\_5

5\_6

6\_7

7\_8

此文件可以从SampleMulLevel生成的类别索引文件稍微做些修改生成。即去掉类别名称部分，并稍微修改分隔符，由”|”变为”\_”

* 训练模型：

./svm\_multiclass\_learn -c 5000 example4/sim\_train.txt example4/model >> whole\_log.txt

5000为初始的C值，

example4/sim\_train.txt训练样本，

example4/model生成的训练模型

* 测试模型：

./svm\_multiclass\_classify example4/sim\_test.txt example4/model example4/predictions

example4/sim\_test.txt为测试样本

example4/model 生成的模型

example4/predictions 生成的预测结果