**通过fasttext训练客户的分类**

对于客户的消费习惯进行分类，是听鉴平台重要的功能之一，因为前期缺少客户数据，采用了人工标注的方式进行分类，随着数据的不断累积，可以通过机器对于标注的信息不断的进行学习，以达到机器自动标注的目的；本文中就是根据已有的分类数据进行一次机器学习的过程；本文中所涉及到的内容较多，如果想做好机器学习的训练工作，需要提前对于相关知识有一些了解，机器学习的引擎使用的是facebook推出的开源工具：fasttext, 分词使用的是国内优秀的中文分词工具：jieba；系统的运行环境：linux+python3.5以上；

本文中所有的代码，可访问: <https://github.com/xiaobo0903/tjcategory.git> 获得

**一、数据预处理**

1、原始训练素材格式要求：

1. 原始训练素材中不能有时间戳内容,如[2020-10-12 08:01:02.234] …。
2. 每个训练素材的后缀必须是.txt，文件名不建议使用中文；

2、建立数据处理目录tjdata(如果存在，不需要重新建立)  
3、按不同的训练类型，分别建立子目录，并把训练素材拷入到相应的子目录中；

# mkdir tjdata

如: 不定型子目录 mkdir bdx

因为需要训练出6个用户分类，所以需要建立6个子目录：

# cd tjdata

# mkdir bdx cdx jjx lzx xgx ylx

bdx：不定型

cdx : 冲动型

jjx：经济型

lzx：理智型

xgx：习惯型

ylx：疑虑型

建立完成后，把训练素材拷入到相应的目录中；（训练素材的命名规则要求，参考第一条中描述的内容）

3、把automatictxt.sh文件拷入到tjdata目录中，并确认其有执行权限；

执行命令：

# sh automatictxt.sh

执行完成后会新建6个文件，bdx.txt cdx.txt jjx.txt lzx.txt xgx.txt ylx.txt

4、把新生成的6个文件进行分词处理；

分词引擎使用的是jieba分词，其安装方式网上可查找；

分词的Python程序是：tjs.py

# git clone <https://github.com/xiaobo0903/tjcategory.git>

该仓库中已有tjs.py程序；

执行的方法：

# python tjs.py <文件目录:tjdata>

目录后不要带/, 否则文件会生成在原文件的目录下；

举例说明：python tjs.py tjdata, 执行后会生成一个新目录tjdata\_cut，分词处理后的数据会保存在tjdata\_cut目录中。

5、合并训练素材

进入分词后的文本目录中(例:tjdata\_cut)，把执行程序mergetxt.sh拷入该目录中，在目录中执行mergetxt.sh，会把所有的分词后的内容打标签后生成一个all.dat文件；

# sh mergetxt.sh

**二、训练新模型**

如想提前了解 fasttext的知识，可叁考：

<https://blog.csdn.net/feilong_csdn/article/details/88655927?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromBaidu-1.control&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommendFromBaidu-1.control>

1、fasttext安装方法：

# git clone <https://github.com/facebookresearch/fastText.git>

# cd fastText

# make

2、训练模型

# ./fasttext supervised -input all.dat -output qicheGZLX -lr 0.5 -epoch 25 -wordNgrams 5 -bucket 200000 -dim 50 -loss hs

-input: 输入的就是训练数据，就是通过上面步骤生成的all.dat文件；

-output: 是生成的模型名称，本汽车分类模型的名称是: qicheGZLX

训练完成后，会生成qicheGZLX.bin和qicheGZLX.vec两个文件， 就是训练出来的模型；