# 如何对房产领域词典中的词进行分类

已有数据：

* 已经分类的文本，假设一共分为k个类别
* 需要进行分类的房产领域词汇，假设一共包含m个词（包括初步房产词典中从搜狗输入法得到的词语和LDA算法学到的新词）

得到结果：

房产词汇1:类别1中的权重，类别2中的权重，…，类别k中的权重

房产词汇2:类别1中的权重，类别2中的权重，…，类别k中的权重

…

房产词汇m:类别1中的权重，类别2中的权重，…，类别k中的权重

训练过程：

1. 对于每个已经分类的文本textj，假设类标为labelj，首先对text文本进行分词，找出其中和房产相关的词汇（直接和房产领域词典中的词进行匹配）words(0-n)，更新wordi在类别labelj下的权重值。
2. 对所有的已分类文本进行分析后，对房产词汇中关于类别的权重值进行归一化等处理。

用途：

可以给房产领域词典进行分类，而且词典中的每个词对应的不是简单的一个类标，而是各个类标下的一个分布。

应用：

根据用户输入的词汇判断用户的兴趣爱好进行推送。

局限和难点：

1. 词的分类非常依赖于文本分类的分类标准和准确度。
2. 如何根据文本来计算更新每个词汇的权重。