# Word Way in Classificational MedLDA

## Abstract

本文探讨分类MedLDA模型如何进行后续的单词选择。为了实现这个目标，我们增强了原来的分类MedLDA模型，对每个单词赋予一个权重，因此每个文档的向量表示不再是所包含单词向量的简单加和，而是所包含的不同权重的单词向量的加和。在EM算法的M-Step我们使用梯度下降的方法优化这些权重。模型训练好之后，按照权重从大到小，我们选取不同比例的单词向量来对新文档进行分类。我们在真实的网页分类集上测试所提出的方法，结果总结为以下三点：1) 所有类别的平均分类效果在所有的单词向量被使用时最好。2) 很小一部分单词(在我们试验中小于2%)能够保留全部单词使用时几乎相同的效果。3) 这部分单词的比例随着样本中不同单词数量的增加而变小。这些结果表明在使用分类MedLDA模型进分类时一个更廉价的方式，仅保留一小部分权重高的单词向量，从而节约了绝大多数的内存开销。