A Chinese Word Clustering Method Using Latent Dirichlet Allocation and K-means

1 Introduction

本文提出了一种基于 LDA 和 K-means 的中文单词聚类方法。K-means 作为一种无监督聚类算法，由于初始聚类中心是随机选择的，所以得到的聚类结果不唯一。为了解决词问题，我们通过 LDA 算法从每个句子中的名词中抽取主题，然后选择每个主题下概率最高的名词作为 K-means 的初始聚类中心，最后使用 K-means 对文本中的所有单词进行聚类。