Clustering with Probabilistic Topic Models on Arabic Texts: A Comparative Study of LDA and K-Means

1 Introduction

由于阿拉伯语的特殊性，针对该语言文档的聚类所面临的困难也是特殊的。本文比较了阿拉伯语的特殊性对 LDA 和 K-means 的性能影响。

2 Document Clustering

LDA and Clustering

根据 [16]，**使用主题模型进行文档聚类有两种方法**。第一种方法使用主题模型对文档进行降维表示（单词表示 $\rightarrow$ 主题表示），然后在新的表示形式上应用标准聚类算法（比如 K-means）；而另一种方法直接使用主题模型，其思想是，在完成参数 ϕ 和 θ 的估计后，每个主题 z 就变成了一个新的聚类，被分配给这个聚类的文档在所有分配给主题 z 的文档中拥有最大概率。

本文使用的是第二种方法，该方法允许我们测量 LDA 相比于传统聚类算法（如 K-means）的性能。