Comparison of Different Distance Measure Methods in Text Document Clustering

Yin Min Tun
Yinmintun2013@gmail.com
University of Computer Studies, Mandalay, UCSM

Clustering text document is an unsupervised learning method to find common groups. The clustering of text documents are the special issue in text mining for unlabelled train documents. Fortunately, there are many proposed features and methods to resolve this problem. The framework of text document classification consists of: input text document, pre-processing, feature extraction and clustering. The common classification methods are: self-organization map, k-means and mixture of Gaussians. The correlation of resulted clusters is based on selecting a distance measure method. The main focus of this paper is to present different exiting distance measure methods along with k-means clustering for text document clustering. The experiment performed k-means clustering on the Newsgroups dataset and measure clustering entropy to evaluate the different distance measure methods.