EVALUATING THE EFFECT OF STEMMING IN CLUSTERING OF ARABIC DOCUMENTS

Omaia M. Al-Omari
Department of Computer Science
University of Hail
KINGDOM OF SAUDI ARABIA
omaiaomari@yahoo.com

ABSTRACT

In text mining, the concept of clustering is common and important to retrieve and categorize documents. Clustering techniques divers and many of them are applied on different languages but not on Arabic. K-means algorithm is a widely used clustering technique that seeks to minimize the average squared distance between points in the same cluster. This paper aimed to implement and evaluate the K-means algorithm on clustering of Arabic documents and estimate the effect of stemming on such clustering algorithm. The experimented work showed that the accuracy of clustering Arabic documents using the K-means algorithm varies from low to very good. The best achieved result was 69% of successful documents without stemming. Furthermore, the effect of stemming resulted in decreasing the accuracy of retrieving documents because the stemming is an abstract of a word which leads to miss-discriminating of documents. The best result scored with stemming was 55% of successful documents, when applying the same thresholds.

Keywords: clustering Arabic documents, Stemming, K-means.