Jordan University of Science and Technology

Arabic Text Categorization

Authors: R. M. Duwairi

Abstract: In this paper, we compare the performance of three classifiers for Arabic text categorization> In particular, the na?ve Bayes, k-nearest-neighbors (knn), and distance-based classifiers were used. Unclassified documents were preprocessed by removing punctuation marks and stopwords. Each document is then represented as a vector of words (or of words and their frequencies as in the case of the na?ve Bayes classifier). Stemming was used to reduce the dimensionality of feature vectors of documents. The accuracy of the classifier is compared using recall, precision, error rate and fallout. The results of the experimentations that were carried out on an in-house collected Arabic text show that the na?ve Bayes classifier outperforms he other two.