Jordan University of Science and Technology

Machine Learning for Arabic Text Categorization

Authors: R. M. Duwairi

Abstract: In this article, we propose a distance-based classifier for categorizing Arabic text. Each category is represented as a vector of words in an m-dimensional space, and documents are classified on the basis of their closeness to feature vectors of categories. The classifier, in its learning phase, scans the set of training documents to extract features of categories that capture inherent category specific properties; in its testing phase the classifier uses previously determined category-specific features to categorize unclassified documents. Stemming was used to reduce the dimensionality of feature vectors of documents. The accuracy of the classifier was tested by carrying out several categorization tasks on an in-house collected Arabic corpus. The results show that the proposed classifier is very accurate and robust.