Jordan University of Science and Technology

Classifying Arabic Text Using KNN Classifier

Authors: Amer Al-Badarneh, Emad Al-Shawakfa, Khaleel Al-Rabab'ah, Safwan Shatnawi, Basel Bani-Ismail

Abstract: With the tremendous amount of electronic documents available, there is a great need to classify documents automatically. Classification is the task of assigning objects (images, text documents, etc.) to one of several predefined categories. The selection of important terms is vital to classifier performance, feature set reduction techniques such as stop word removal, stemming and term threshold were used in this paper. Three term-selection techniques are used on a corpus of 1000 documents that fall in five categories. A comparison study is performed to find the effect of using full-word, stem, and the root term indexing methods. K-nearest ? neighbors classifiers used in this study. The averages of all folds for Recall, Precision, Fallout, and Error-Rate were calculated. The results of the experiments carried out on the dataset show the importance of using k-fold testing since it presents the variations of averages of recall, precision, fallout, and error rate for each category over the 10-fold.