Jordan University of Science and Technology

Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization

Authors: Rehab Duwairi, Mohammad Al-Refai, Natheer Khasawneh

Abstract: This paper compares and contrasts two feature selection techniques when applied to Arabic corpus; in particular; stemming and light stemming were employed. With stemming, words are reduced to their stems. With light stemming, words are reduced to their light stems. Stemming is aggressive in the sense that it reduces words to their 3-letters roots. This affects the semantics as several words with different meanings might have the same root. Light stemming, by comparison, removes frequently used prefixes and suffixes in Arabic words. Light stemming doesn't produce the root and therefore doesn't affect the semantics of words; it maps several words, which have the same meaning to a common syntactical form. The effectiveness of above two feature selection techniques was assessed in a text categorization exercise for Arabic corpus. This corpus consists of 15000 documents that fall into three categories. The K-nearest neighbors (KNN) classifier was used in this work. Several experiments were carried out using two different representations of the same corpus; the first version uses stem-vectors; and the second uses light-stem-vectors as representatives of documents. These two representations were assessed in terms of size, time and accuracy. The light stem representation was superior in terms of classifier accuracy when compared with stemming.