Improving Arabic Text Categorization using Normalization and Stemming Techniques

Abstract

Text Categorization is a technique for assigning documents based on their contents to one or more pre-defined categories. Achieving highest categorization accuracy remains one of the major challenges and it is also time consuming. We proposed approach to tackle these challenges. The proposed approach uses Frequency Ratio Accumulation Method (FRAM) as a classifier. Its features are represented using bag of word technique and an improved Term Frequency (TF) technique is used in features selection. The proposed approach is tested with known datasets. The experiments are done without both of normalization and stemming, with one of them, and with both of them. The obtained results of proposed approach are generally improved compared to existing techniques. The performance attributes of proposed Arabic Text Categorization approach were considered: Accuracy, Recall, Precision and F-measure (F1). The averages of the obtained results are 97.50%, 97.50%, 97.51%, and 97.49% respectively using normalization.

References
15. https://pythonhosted.org/Tashaphyne/Tashaphyne.normalize-module.html
18. https://pypi.python.org/pypi/Tashaphyne/
Improving Arabic Text Categorization using Normalization and Stemming Techniques


Index Terms
Computer Science Artificial Intelligence

Keywords

Arabic text categorization, Frequency ratio accumulation method (FRAM), Bag-Of-Word (BOW), Features selection, Term and document frequency.