Abstract

This paper describes the impact of dataset characteristics on the results of Arabic document classification algorithms using TF-IDF representations. The experiments compared different stemmers, different categories and different training set sizes, and found that different dataset characteristics produced widely differing results, in one case attaining a remarkable 99% recall (accuracy). The use of a standard dataset would eliminate this variability and enable researchers to gain comparable knowledge from the published results.

References

On the Impact of Dataset Characteristics on Arabic Document Classification

- Ababneh, J., Almomani, O., Hadi, W., El-Omari, N. and Al-Ibrahim, A. 2014. Vector
Space Models to Classify Arabic Text. In International Journal of Computer Trends and Technology (IJCTT), vol. 7, no. 4.

**Index Terms**

Computer Science  
Artificial Intelligence

**Keywords**

Dataset; TF-IDF representation; Arabic Stemmers; Arabic document classification