Abstract

Text categorization, TC, is a process of labeling natural language texts with one or several categories from a predefined set. TC is a supervised learning where the set of categories and examples of documents belonging to those categories is given. The task of automatic TC is assigned an electronic document to several categories, based on a training set of labeled documents. The research objectives are, to formulate a K-Nearest Neighbor (KNN) algorithm for the automatic and suitable classification of any Holy Quran Tafseer segment; to identify relevant categories of Holy Quran Tafseer in the form of number classes; and to retrieve, Tafseer of verses of the Holy Quran in Malay language. Hence, this research aims to automatically categorize the Tafseer of verses of Holy Quran using the KNN algorithm as a technique to solve text categorization. This research has been designed to classify different verses in the Holy Quran. The first phase is to pre-process the Arabic text and then change the word in Arabic to Malay word. After that, categorize classes based on the cosine similarity between a test document and specific training documents. The majority of the same kind of nearest neighbors
decides the category of the test sample and calculates precision and recall for a collection of documents. The result shows the outperform of TC using the KNN algorithm is one of the best algorithm for categorization Tafseer of Holy Quran. Furthermore, this study contributes in building a classifier to Tafseer Al-Quran in Malay language.

References

17. Milan Sonka, Vaclav Hlavac, and Roger Boyle. Image processing, analysis, and
Categorization of ‘Holy Quran-Tafseer’ using K-Nearest Neighbor Algorithm


Index Terms

Computer Science

Algorithms

Keywords

Text categorization, K-Nearest Neighbor algorithm