Abstract

Text categorization is the process of automatically assigning labels or categories to new or previously unseen text documents. The text documents may be unstructured or semi structured in nature. In our work, we have used concepts of natural language processing for text categorization. That is, a lexical approach for text categorization. We have developed an algorithm which automatically classifies articles into their categories. The algorithm identifies tokens and assigns them weights in the abstracts of journal articles. We have implemented our approach using K Nearest Neighbor (KNN) classifier as it is the most widely used classifier in research. The proposed algorithm Lexical KNN (LKNN) has been evaluated on two datasets. One is set of journal articles of computer science discipline and the other is a collection of medical documents (Ohsumed collection). The experimental results show that our proposed algorithm Lexical KNN (LKNN) performs better than the other existing classifiers.


29. www.lextek.com/manuals/onix/stopwords2.html


Index Terms

Computer Science

Information Sciences

Keywords

Text categorization, K Nearest Neighbor (KNN), Lexical Analysis, Tokens.