World Academy of Science, Engineering and Technology
International Journal of Health and Medical Engineering
Vol:11, No:07, 2017

Incorporating Information Gain in Regular Expressions Based Classifiers

Authors : Rosa L. Figueroa, Christopher A. Flores, Qing Zeng-Treitler

Abstract : A regular expression consists of sequence characters which allow describing a text path. Usually, in clinical research, regular expressions are manually created by programmers together with domain experts. Lately, there have been several efforts to investigate how to generate them automatically. This article presents a text classification algorithm based on regexes. The algorithm named REX was designed, and then, implemented as a simplified method to create regexes to classify Spanish text automatically. In order to classify ambiguous cases, such as, when multiple labels are assigned to a testing example, REX includes an information gain method Two sets of data were used to evaluate the algorithm's effectiveness in clinical text classification tasks. The results indicate that the regular expression based classifier proposed in this work performs statically better regarding accuracy and F-measure than Support Vector Machine and Naïve Bayes for both datasets.

Keywords : information gain, regular expressions, smith-waterman algorithm, text classification
Conference Title : ICMIH 2017 : International Conference on Medical Informatics and Healthcare
Conference Location : Rome, Italy
Conference Dates : July 17-18, 2017