AN EFFECTIVE APPROACH FOR WEB DOCUMENTS CLASSIFICATION USING FP-GROWTH AND NAÏVE-BAYES TECHNIQUES

RAJENDRA KUMAR ROUL & S.K.SAHAY

BITS, Pilani - K.K. Birla, Goa Campus, Zuarinagar, Goa - 403726, India

ABSTRACT

Exponential growth of the web increased the importance of web documents classification and data mining. To get the exact information, in the form of knowing what classes a web document belongs to, is expensive. Automatic classification of web documents is of great use to search engines which provides this information at a low cost. In this paper, we propose an approach for classifying the web documents using the frequent item word sets generated by the Frequent Pattern(FP) Growth technique. These set of associated words act as feature set. The final classification obtained after Naïve Bayes classifier used on the feature set. For the experimental work, we use Gensim package, as it is simple and robust. Results show that our approach can be effectively classifying the web documents.

KEYWORDS: Classification, FP-Growth, Gensim, Naïve Bayes, Vector Space Model