World Academy of Science, Engineering and Technology
International Journal of Biotechnology and Bioengineering
Vol:12, No:06, 2018

Biofilm Text Classifiers Developed Using Natural Language Processing and Unsupervised Learning Approach

Authors : Kanika Gupta, Ashok Kumar

Abstract : Biofilms are dense, highly hydrated cell clusters that are irreversibly attached to a substratum, to an interface or to each other, and are embedded in a self-produced gelatinous matrix composed of extracellular polymeric substances. Research in biofilm field has become very significant, as biofilm has shown high mechanical resilience and resistance to antibiotic treatment and constituted as a significant problem in both healthcare and other industry related to microorganisms. The massive information both stated and hidden in the biofilm literature are growing exponentially therefore it is not possible for researchers and practitioners to automatically extract and relate information from different written resources. So, the current work proposes and discusses the use of text mining techniques for the extraction of information from biofilm literature corpora containing 34306 documents. It is very difficult and expensive to obtain annotated material for biomedical literature as the literature is unstructured i.e. free-text. Therefore, we considered unsupervised approach, where no annotated training is necessary and using this approach we developed a system that will classify the text on the basis of growth and development, drug effects, radiation effects, classification and physiology of biofilms. For this, a two-step structure was used where the first step is to extract keywords from the biofilm literature using a metathesaurus and standard natural language processing tools like Rapid Miner_v5.3 and the second step is to discover relations between the genes extracted from the whole set of biofilm literature using pubmed.mineR_v1.0.11. We used unsupervised approach, which is the machine learning task of inferring a function to describe hidden structure from 'unlabeled' data, in the above-extracted datasets to develop classifiers using WinPython-64 bit_v3.5.4.0Qt5 and R studio_v0.99.467 packages which will automatically classify the text by using the mentioned sets. The developed classifiers were tested on a large data set of biofilm literature which showed that the unsupervised approach proposed is promising as well as suited for a semi-automatic labeling of the extracted relations. The entire information was stored in the relational database which was hosted locally on the server. The generated biofilm vocabulary and genes relations will be significant for researchers dealing with biofilm research, making their search easy and efficient as the keywords and genes could be directly mapped with the documents used for database development.

Keywords : biofilms literature, classifiers development, text mining, unsupervised learning approach, unstructured data, relational database
Conference Title : ICBB 2018 : International Conference on Bioinformatics and Bioengineering
Conference Location : Toronto, Canada
Conference Dates : June 21-22, 2018