Gupta and Kumar (2017) Can J Biotech, Volume 1, Special Issue, Page 24
https://doi.org/10.24870/cjb.2017-a12

Canadian Journal of Biotechnology

ISSN 2560-8304

Poster Presentation OPEN ACCESS

Category: Bioinformatics

Text-Mining Applications for Creation of Biofilm Literature Database

Kanika Gupta and Ashok Kumar*

Centre for Systems Biology and Bioinformatics, South Campus, Sector-25, Panjab University, Chandigarh 160025, INDIA

*Corresponding author: ashokbiotech@gmail.com, ashokkumar@pu.ac.in

Abstract

The massive information hidden in the biomedical field, in the form of publications is growing exponentially therefore it is not possible for researchers and practitioners to keep themselves updated with all the developments in any specific field. Manual effort to transform unstructured text into structured is a laborious process. Automatic techniques for relation extraction provide a solution to the problem. Text Mining is one such technique defined as ‘‘the discovery by computer of new, previously unknown information, by automatically extracting and relating information from different written resources, to reveal otherwise ‘hidden’ meanings’’. The current study is focused on biofilm literature where 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. Biofilm research has become a very important field, due to their high mechanical resilience and resistance to antibiotic treatment, they constitute a significant problem in both industry and health care.

So in the present research published corpora of 34306 documents for biofilm was collected from PubMed database along with non-indexed resources like books, conferences, newspaper articles, etc. and these were divided into five categories i.e. classification, growth and development, physiology, drug effects and radiation effects. These five categories were further individually divided into three parts i.e. Journal Title, Abstract Title, and Abstract Text to make indexing highly specific. Text-processing was done using the software Rapid Miner_v5.3, which tokenizes the entire text into words and provides the frequency of each word within the document. The obtained words were normalized using Remove Stop and Stem Word command of Rapid Miner_v5.3 which removes the stopping and stemming words. The obtained words were stored in MS-Excel 2007 and were sorted in decreasing order of frequency using Sort & Filter command of MS-Excel 2007. The words are visualization through networks obtained by Cytoscape_v2.7.0. Now the words obtained were highly specific for biofilms, generating a controlled biofilm vocabulary and this vocabulary could be used for indexing articles for biofilm (similar to MeSH database which indexes articles for PubMed). The obtained keywords information was stored in the relational database which is locally hosted using the WAMP_v2.4 (Windows, Apache, MySQL, PHP) server. The available biofilm vocabulary will be significant for researchers studying biofilm literature, making their search easy and efficient.

Citation: Gupta, K. and Kumar, A. Text-Mining Applications for Creation of Biofilm Literature Database [Abstract]. In: Abstracts of the NGBT conference; Oct 02-04, 2017; Bhubaneswar, Odisha, India: Can J biotech, Volume 1, Special Issue, Page 24.
https://doi.org/10.24870/cjb.2017-a12