Abstract

With huge amounts of biomedical data being generated day by day extracting statistical information about the chemicals mentioned in such huge databases manually is tedious and time consuming. Our system is mainly designed for naive users, which aims to automate data collection and knowledge extraction from chemical literature in a user friendly and efficient way on the hadoop platform. The system downloads the abstracts related to the disease of interest from Pubmed database. The text of the abstracts is then extensively parsed for chemicals such as protein/gene names and chemical compound names and classified into different classes. This analysis would prove to be helpful in various biomedical and pharmaceutical industries. The extraction of important information will be done using the Ling Pipe API wherein a training dataset is given to this Ling Pipe which classifies the extracted bioentities in the respective classes. The system being deployed on hadoop platform provides a scalable and distributed system which processes huge number of abstracts in a short time and with high efficiency. The system also provides a user friendly user interface for easy use of the hadoop system for non
Hadoop based Text Mining System for Identification of Chemicals Associated with Disease of Interest

technical users.

References

2. Su Yan, W. Scott Spangler, and Ying Chen - Chemical Name Extraction Based on Automatic Training Data Generation and Rich Feature Set
3. Spiros Papadimitriou, Jimeng Sun - DisCo: Distributed Co-clustering with MapReduce A Case Study Towards Petabyte-Scale End-to-End Mining.
16. alias-i.com/lingpipe/ Ling pipe
17. M. Karthikeyan and Renu Vyas, Practical Chemoinformatics.
18. Anabel Usi, Joaquim Cruz, Jorge Comas, Francesc Solsona and Rui Alves CheNER: a
tool for the identification of chemical entities and their classes in biomedical literature, 2015.

**Index Terms**

Computer Science  
Information Sciences

**Keywords**

Text mining, chemicals, hadoop, LingPipe, bioentities, data extraction, data classification.