Ilkogretim Online - Elementary Education Online, 2019; Vol 18 (Issue 1): pp. 636-644
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2019.527672

A Case Study For Protein Localization-Mining Pubmed Data Using Loctext

Zahrah Ayub Department of Computer Science and Engineering, Islamic University of Science and Technology, Jammu and Kashmir, India.

Asif Ali Banka Department of Computer Science and Engineering, Islamic University of Science and Technology, Jammu and Kashmir, India.

Muneer Ahmad Department of Computer Science Technical University of Munich, Germany.

Roohie Naaz Department of Computer Science and Engineering, NIT Srinagar, Jammu and Kashmir, India.

Correspondence: *Zahrah Ayub

Abstract
Data sharing and web 2.0 have revolutionized the whole new idea and context of thinking about data. Generation and consumption of data has become integral part of human life. Data generated is in various formats and most widely accepted is text data. The biomedical literature is exploding and it is complexity increases to keep track of publications in relevant areas of interest. In this work we attempt to implement LocText on PubMed data to illustrate relationships between protein and sub-cellular locations. The aim of experimentation is to match the UniPortKB accession numbers, GO Cellular Component identifiers to NCBI taxonomy identifiers available in PubMed data. The experiments are performed on a commodity cluster that comprise of 16 machines installed with Ubuntu 16.04 Operating System, Java 8, Hadoop 2.7, Spark2.1, Elasticsearch, Nalaf, LocText and Kibana.

Keywords LocText, Protein Localization, PubMed, Text Mining.

I. INTRODUCTION

Data sharing and web 2.0 have revolutionized the whole new idea and context of thinking about data. The interconnected devices have practically resulted in data intensive computing where data is now essential part of human life [1]. Generation and consumption of data has become integral part of human life. Raw data available to users is diverse and complex in nature, consisting primarily of un- structured and unsupervised data. Pre-processing of data to extract ordered representation of data for downstream consumption is challenging. Data generated is in various