Collscience: A Text Mining Web Service Tool for Extracting Knowledge from Scientific Texts

Diogo de Jesus Soares Machado, Camilla Reginatto De Pierri, Roberto Tadeu Raittz

Federal University of Paraná

Abstract

Manual curation of scientific texts requires effort and, depending on the research, involves a large number of researchers, which may not be sufficient to deal with the number of articles published in a given area of knowledge. In addition, manual curation tends to be affected by disagreement over interpretation, meaning different readers may interpret a specific text in different ways. Because of this, applications like Text Mining have become common in scientific research. With the emergence of the Knowledge Discovery from Text (KDT) definition, the mining process has gained notoriety due to the possibility of discovering important concepts in a variety of subjects. There are many text mining tools that are effective at the moment, but with the use of large amounts of text processing becomes very slow. Therefore, to address the difficulty of high computational effort in word processing, we propose Collscience (Collective Conscience), a text mining web service tool that integrates the best of information retrieval concepts with agility. The algorithm consists of a text mining pipeline: 1) compiling the texts to a format based on the representation of proteins in the FASTA format; 2) process the compiled texts in FASTA format using the SWeeP method, a methodology for vectoring and projecting data in FASTA format; 3) cluster using an improved version of k-means, predicting cluster numbers using a strategy that combines hierarchical clustering and Fuzzy logic; 4) define TF-IDF term scores, considering each text cluster as a document unit; 5) Put the terms in order of importance, using TF-IDF scores as a basis; 6) select TF-IDF scores for the most important words, with a user-defined total; 7) Perform hierarchical grouping, with the TF-IDF scores of the selected terms as a parameter; 8) Generate a dendrogram with the result of clustering. We also developed a graphical interface for Collscience, in web application format, using the languages HTML, CSS, JavaScript and PHP. Collscience is an algorithm that has a text set as input and performs mining, returning a dendrogram that shows the word correlation. The method is currently in the implementation phase. As future prospects, we plan to add more output options to the tool as well as improve user interaction with the machine learning process.

Funding: CAPES, Fundação Araucária