BioReader: a text mining tool for performing classification of biomedical literature

BioReader: a text mining tool for performing classification of biomedical literature

Scientific data and research results are being published at an unprecedented rate. Many database curators and researchers utilize data and information from the primary literature to populate databases, form hypotheses, or as the basis for analyses or validation of results. These efforts largely rely on manual literature surveys for collection of these data, and while querying the vast amounts of literature using keywords is enabled by repositories such as PubMed, filtering relevant articles from such query results can be a non-trivial and highly time consuming task. We here present a tool that enables users to perform classification of scientific literature by text mining-based classification of article abstracts. BioReader (Biomedical Research Article Distiller) is trained by uploading article corpora for two training categories - e.g. one positive and one negative for content of interest - as well as one corpus of abstracts to be classified and/or a search string to query PubMed for articles. The corpora are submitted as lists of PubMed IDs and the abstracts are automatically downloaded from PubMed, preprocessed, and the unclassified corpus is classified using the best performing classification algorithm out of ten implemented algorithms. BioReader supports data and information collection by implementing text mining-based classification of primary biomedical literature in a web interface, thus enabling curators and researchers to take advantage of the vast amounts of data and information in the published literature. BioReader outperforms existing tools with similar functionalities and expands the features used for mining literature in database curation efforts. The tool is freely available as a web service at http://www.cbs.dtu.dk/services/BioReader.