18. Development of text mining tools for information retrieval and extraction from patents

Tiago Alves, Hugo Costa and Miguel Rocha

Biomedical literature is composed of a large and ever increasing number of publications, written in natural language. Patents are a relevant fraction of these publications, considered important sources of information due to all the curated information available in the documents, from the granting process. Although being real technological libraries, their unstructured data turns the search of information within these documents a challenging task. Biomedical text mining is a scientific field that explores this task, creating methodologies to search and structure the information in the biomedical literature.

Information retrieval (IR) is one of the biomedical text mining tasks, in which the relevant information is obtained from an extensive collection of documents using several text retrieval methodologies. Getting all the information available on a patent document requires the download of all the bibliographic information as well as the respective patent PDF document, that is then converted into a machine-readable text by technologies as Optical Character Recognition (OCR).

To achieve that goal, a "patent pipeline" was developed and integrated into @note2, an open-source computational framework for biomedical text mining. The patent pipeline can be disintegrated into four different tasks that were used to build different new processes on @note2: the patent search and the retrieval of patent metadata (relative to bibliographic data) were introduced as a new IR Search tool and the retrieval of patent PDF files as well as the subsequent extraction of all the information from them were introduced as a new IR Crawling and a new PDF to text conversion processes, respectively.

A set of patents from the BioCreative V CHEMDNER task was used to test the developed pipeline, evaluating the framework performance and the real capacity to retrieve the requested patents and extract their unstructured information to machine readable text. The results were promising, and showed that is possible bringing the published patent information to the scientific community, allowing the posterior implementation of other biomedical text mining processes over these documents, automating the search of structured information on patents and taking advantage of all the great informative capacity applied on them.

Universidade do Minho
Escola de Engenharia

2017

Bioinformatics Open Days (BOD)

Book Conference

22-02-2017 to 24-02-2017