Enterprise Platform and Integration Concepts
Research Group of Prof. Dr. Hasso Plattner
Bachelor Project 2015/2016

Text Mining for Biomedical Applications

Motivation

The current data deluge demands fast and real-time processing of large datasets to support various applications, also for textual data, such as scientific publications. Natural language processing (NLP) is the field of automatically processing textual documents and includes a variety of tasks such as tokenization (delimitation of words), part-of-speech tagging (assignment of syntactic categories to words), chunking (delimitation of phrases) and syntactic parsing (construction of syntactic tree for a sentence). Further, NLP also involves semantic-related tasks such as named-entity recognition (delimitation of predefined entity types, e.g., person and organization names), relation extraction (identification of pre-defined relations from text) and semantic role labeling (determining pre-defined semantic arguments). Processing and semantically annotating large textual collection is a time-consuming and tiresome task, which requires integration of various tools. In-memory database (IMDB) technology comes as an alternative given its built-in text analysis and machine learning components and its ability to process large document collections in real time.

Data curation is one application of NLP and consists on the development of a text mining pipeline for automatic extraction of predefined data from textual documents. For instance, biological databases need to extract precise data from scientific literature according to a existing template. Thus, given a predefined template and corresponding list of terminologies, a text mining application can automatically fill in the slots with the required information. A text mining pipeline usually includes three main components: (a) triage of relevant documents; (b) named-entity recognition, and (c) slot filling or relationship extraction. These tasks usually rely on machine learning methods, provided a corpus of annotated documents, i.e., supervised or semi-supervised learning, a set of previous curated data, i.e., distant supervision, or even when no data at all is available, i.e., unsupervised learning.

Project goals

- Build a text mining pipeline for integration of triage, named-entity recognition and relationship extraction tasks
- Develop a Web application to interact with the text mining pipeline
- Apply the system to large collections of documents, such as PubMed, a database of biomedical publications
- Evaluate the system on curation of data for external partners, e.g., cancer research

External partners

The project will be executed in cooperation with SAP SE and potentially further external partners. We expect knowledge exchange and visits of partners.

Skills

Participants should have knowledge of SQL, of at least one programming language (preferably C++, Python or Java) and of Web development, as well as interest in database technologies, machine learning and natural language processing.

Group structure and project start

The team will consist of 6-8 students and the project will start in October 2015.

Contact

Your are welcome to contact or visit us in room V0.01 (Villa), HPI Campus II:

Dr. Mariana Neves (mariana.neves@hpi.de)
Milena Krause (milena.kraus@hpi.de)
Dr. Matthias Uflacker (matthias.uflacker@hpi.de)
Prof. Dr. Hasso Plattner (office-epic@hpi.uni-potsdam.de)