UNIVERSITÄT BIELEFELD
Faculty of Technology

Bachelor's Thesis

Towards knowledge extraction: data acquisition and analysis of scientific literature

Semantic Computing Group
Moritz Blum
mblum@techfak.uni-bielefeld.de

Natural language texts contain a lot of relevant information that does not exist in structured form, such as in a database or in a knowledge graph. Thus, collecting documents about a particular domain of interest is a relevant task. Existing databases of publications are usually very large and cover many different fields of research.
The aim of this bachelor's thesis is to systematically (and legally) retrieve publications from existing databases (e.g., Google Scholar, Elsevier, or Web of Science), to preprocess them (e.g., parse PDF files, clean textual artifacts), and to analyze them. For the analysis, natural language processing will be applied to derive key properties of the derived dataset (e.g., regarding named entities). Moreover, the main findings shall be presented through visualizations.
This leads to the following research questions:

- What is a systematic approach for querying scientific publication databases using literature references?
- What text preprocessing techniques are required for applying natural language processing to scientific publications?
- What are the properties of the resulting dataset?

This thesis is part of a collaborative interdisciplinary research project that involves researchers from the field of material science. The dataset collected will be used to create prototypes for extracting knowledge from it. The findings of this thesis will serve as the foundation for upcoming experiments.

Related literature and tools

- Tshitoyan, V., Dagdelen, J., Weston, L. et al. Unsupervised word embeddings capture latent knowledge from materials science literature
- Dan Jurafsky and James H. Martin: Speech and Language Processing
- `http://chemdataextractor.org`
- `www.spacy.io`

The Semantic Computing Group researches and develops methods that enable machines to acquire, process, and understand data from natural language and knowledge graphs. Using methods from *natural language processing* and *machine learning*, we are aiming at machines that are capable of knowledge acquisition by reading various kinds of data. In particular, the group focuses on methods for information extraction, knowledge graph completion, semantic parsing, ontology learning, sentiment analysis, entity linking, as well as question answering.

More information is available at:
https://uni-bielefeld.de/fakultaeten/technische-fakultaet/arbeitsgruppen/semantic-computing

Interested? @mail to `mblum@techfak.uni-bielefeld.de`