

Bachelor-/Masterarbeit

“Entwicklung einer Suchmaschine für Datensätze” zu vergeben!

Um was geht es?

Die Anzahl an Datensätzen ist in den letzten Jahren dramatisch gestiegen. WissenschaftlicherInnen und EntwicklerInnen stehen daher häufig vor der Herausforderung, passende Datensätze im Web zu finden. Das Ziel der Abschlussarbeit ist es daher, Metadaten über Datensätze zu crawlen, semantisch zu repräsentieren und durchsuchbar zu machen (vgl. [1]). Im Unterschied zum Google Dataset Search [1,2] sollen die Daten öffentlich verfügbar sein, semantisch modelliert sein (sofern möglich mit Links zu anderen Dingen im Web) und letztendlich mittels keywords o.ä. durchsuchbar sein.

Das Vorgehen:

1. Wir extrahieren die Metadaten über Datensätze aus öffentlichen APIs und ggf. großen Datensätzen wie Common Crawl.
2. Die extrahierten Fakten modellieren wir in RDF und erstellen damit einen RDF knowledge graph. Dieser ist über das Web erreichbar und jeder kann mittels keywords nach Datensätzen suchen.

Neben diesen Kernelementen kann das entwickelte System erweitert werden. Beispielsweise kann ein User-Interface gebaut werden, um die Datensätze zu durchsuchen. Zudem können Daten in dem erstellten Wissensgraphen zu anderen Wissensgraphen in der Linked Open Data cloud mittels graph-basierter und String-basierter Algorithmen verlinkt werden (um z.B. Datensätze mit Publikationen zu verbinden).

Die Herausforderung besteht in der Extraktion von Fakten aus Webseiten, der Indizierung der Daten und der Entwicklung einer passenden Suche.

[1] <https://toolbox.google.com/datasetsearch>

[2] <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/77547c8d2a7fba472e76c774028cf2b3c0afdb8a.pdf>

Welche Voraussetzungen sind notwendig?

- Erfahrung im Umgang mit großen Datenmengen vorteilhaft
- Erfahrung im Text Mining/Natural Language Processing von Vorteil, aber nicht notwendig
- Mindestens mittlere Kenntnisse in Python oder einer ähnlichen Programmier-/Skriptsprache.

Im Falle guter Resultate können die Ergebnisse der Arbeit gerne in Zusammenarbeit mit den Betreuern wissenschaftlich veröffentlicht werden.

Kontakt:

Dr. Michael Färber | Anna Nguyen
michael.farber@kit.edu | anna.nguyen@kit.edu