Semantischer Fingerabdruck für juristische Textdokumente

Studiengang: BSc in Informatik | Vertiefung: Data Engineering
Betreuer: Prof. Dr. Jürgen Vogel
Experte: Andreas Dürsteler

Automatisierte Extraktion von Informationen aus Texten des Schweizer Bundesgerichts um erweitertes Suchen auf Urteilen zu ermöglichen

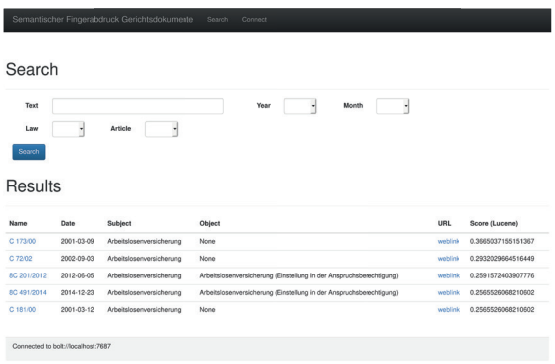
Ausgangslage

Gerichtsurteile von diversen Schweizer Gerichten sind öffentlich zugänglich und werden bei der Bearbeitung von Rechtsfällen verwendet. Diese Urteile sind nur mit einer Volltextsuche oder nach Titel durchsuchbar. Dies macht die Suche nach spezifischen Urteilen oder Urteilen mit ähnlichem Inhalt schwer und zeitaufwändig. Für diese Arbeit wurde ein Datensatz an Dokumenten (Urteilen) zur Verfügung gestellt. Dieser Datensatz besteht aus rund 117'000 Dokumenten des Schweizer Bundesgerichts aus den Jahren 2000 bis und mit 2018.

Ziel der Arbeit

Im Rahmen dieser Arbeit sollen grösstenteils automatisiert Urteile mit Informationen (z.B. Thema, betroffenes Recht) ergänzt werden. Die verwendete "juristische Sprache" verleiht den Dokumenten eine vergleichbare Struktur im Aufbau und in der Wortwahl was eine automatisierte Kennzeichnung ermöglichen soll. Die Dokumente enthalten von sich aus keine Metainformationen - die gewünschten Informationen müssen also aus dem Text extrahiert werden.

Um die Resultate zu präsentieren soll auf den resultierenden Daten ein kleines Interface aufgebaut werden. Dies beinhaltet eine einfache Suche sowie die Darstellung von Dokumenten inklusive Links zu inhaltlich ähnlichen Dokumenten.



Suchoberfläche

Ergebnisse

Im Laufe der Arbeit wurden drei Verfahren für die Extraktion von Informationen angewendet:

- Dokumente in Klassen unterteilen basierend auf ähnlichen Inhalten unter Verwendung von k-means
- Gemäss definierten Mustern referenzierte Rechtsartikel und andere Urteile extrahieren
- Passende Schlüsselwörter (Keywords) pro Dokument mithilfe von annotierten Dokumenten und Linearer Regression finden

Die Verfahren liefen aus Gründen der Einfachheit sowie Performance über eine gewählte Subdomäne von rund 2500 Dokumenten, es bestehen jedoch Möglichkeiten zur Skalierung auf den ganzen Datensatz. Die optimale Anzahl k-means Klassen konnte nicht eindeutig bestimmt werden und der R2 Score von den gefundenen Keywords liegt bei den Test Daten bei tiefen 2%. Nichtsdestotrotz konnten nützliche Informationen extrahiert werden, auch wenn die Verfahren noch Verbesserungspotenzial haben.

Nebst den extrahierten Informationen hat die Arbeit ausserdem einen Einblick in mögliche Techniken, Ideen und zu extrahierende Informationen gegeben. Ein Teil der 2500 Dokumente wurde zudem im Rahmen der Arbeit manuell mit Informationen ergänzt (Gold Standard).

Ausblick

In weiteren Schritten könnten die Verfahren sicher verbessert und verfeinert werden. Im Datensatz gibt es zudem weitere interessante Informationen zum Extrahieren. Es wäre auch denkbar die Verfahren auf Dokumente von anderen Gerichten (z.B. Erstinstanzen) durchzuführen und die Resultate zu vergleichen.

Jan Ackermann