Webbasiertes Natural Language Interface für Datenbanken

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

Wie simpel wäre es, wenn man ohne technisches Wissen mit Daten interagieren könnte? In dieser Arbeit wurde versucht, eine Schnittstelle von natürlicher Sprache zu Datenbanken zu entwickeln. Dies ist besonders nützlich für Personen, welche technische Abfragesprachen nicht beherrschen. Die resultierende Webapplikation kann Single-Table Abfragen mit einfachen WHERE-Klauseln, Aggregationen und Gruppierungen verarbeiten.

Unsere Vision

Wir entwickeln ein Interface der natürlichen Sprache zu einer Datenbank. Die Schnittstelle nimmt eine Fragestellung in textueller, englischer Form entgegen und wandelt diese zu einer syntaktisch korrekten Abfrage in SQL um. Dazu wird ein maschinell erlerntes Modell verwendet. Das resultierende Query wird anschliessend gegen die Zieldatenbank ausgeführt und die zurückgegebenen Daten tabellarisch dargestellt.

Herausforderung

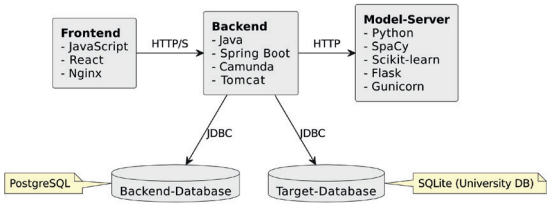
Eine der Herausforderungen dieser Bachelor-Thesis ist die Vielfältigkeit der natürlichen Sprache. Sie ist nicht formal, es gibt keine abgeschlossene Menge an kombinierbaren Ausdrücken und die allermeisten Wörter sind mehrdeutig. Auf der anderen Seite gibt es viele Variationen von passenden SQL-Ausdrücken, aber alle müssen immer syntaktisch korrekt sein. Das maschinell erlernte Model kann eine so grosse Variation an Kombinationen niemals alleine verstehen, es ist zu wenig mächtig. Es braucht deshalb eine Kombination mit statisch hinterlegten Regeln, welche aber die Generalisierbarkeit der Applikation für jede relationale Datenbank weiterhin gewährleisten. Zudem müssen möglichst viele Informationen über die Schematik der Daten extrahiert und dargestellt werden können. Diese werden einerseits benötigt, um einen Datensatz für das Training des Modells zu generieren. Andererseits sollen sie auch für die Nutzerinteraktion verwendet werden, sodass die Eingabe der Abfrage ohne Vorwissen gemacht werden kann. Zugleich müssen die Möglichkeiten und Grenzen der Applikation den Benutzer:innen jederzeit bekannt sein. Bei all dem muss die Applikation jederzeit performant bleiben.

Resultat

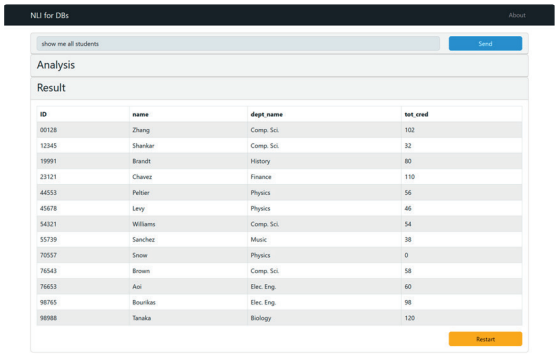
Nach State-of-the-Art Architekturvorgaben wurde eine Webapplikation mit drei Komponenten entwickelt. Die Webapplikation führt Benutzer:innen durch den Prozess, von der Eingabe der Abfrage bis zum Resultat. Ausserdem verlangt es laufend Rückmeldungen und misst so die Performance des maschinell erlernten Modells. Ein Backend steuert den Benutzerprozess. Und der intelligente Teil der Applikation kümmert sich um die Sprachverarbeitung. Der Model-Server verwendet SpaCy für POS-Tagging und erkennt regelbasiert Tabellennamen und Attribute. Danach ordnet ein eigener Classifier der Anfrage das passende SQL-Statement zu. Dieser wurde auf Basis des invertierten Datenbankindex der Zieldatenbank trainiert. Die Applikation kann Single-Table Abfragen mit einfachen WHERE-Klauseln, Aggregationen und Gruppierungen verarbeiten.

Timon Borter
timon.borter@gmx.ch

Mario Schläppi
mario.schlaeppi@gmail.com



Die webbasierte Datenbankschnittstelle.



Technologieübersicht.