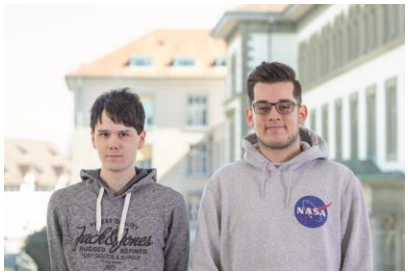
Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Texploration: Automatische Analyse von grossen Textsammlungen



Diplomierende
Dominik Steiner
Gëzim Zeneli

Dozent
Mark Cieliebak

Texploration ist eine Webplattform für die Analyse von grossen Textsammlungen, welche im letzten Jahr im Rahmen einer Projekt- und Bachelorarbeit entwickelt wurde. Die Applikation automatisiert verschiedene Aufgaben im Bereich der Datenanalyse und Klassifikation.
Diese Arbeit verfolgte zwei Ziele: Zum einen sollten verschiedene Bereiche des Quellcodes verbessert und überarbeitet werden und zum anderen sollten bestehende Algorithmen optimiert und neue hinzugefügt werden.
Der Fokus lag zu Beginn auf dem Engineering-Teil. Es wurde ein Refactoring im Backend durchgeführt, welches zu einer starken Verbesserung der Codequalität führte. Ausserdem wurden durch den Einsatz eines Flask Server (Python Webserver) die Bereiche Node.js und Python stärker gekapselt. Im Anschluss wurde noch ein Konzept für den Einsatz von Microservices in der Texploration entwickelt.
Im zweiten Teil hingegen, wurde der Fokus hauptsächlich auf die Optimierung der bestehenden Algorithmen gelegt. Zusätzlich wurde die neue Komponente Language Detection eingeführt. Mithilfe der Language Detection kann der Benutzer sehen, ob sein Korpus mehrsprachig ist. Des Weiteren wurde die Komponente Topic Modeling überarbeitet und optimiert. Neu funktioniert das Topic Modeling nicht nur für englische Texte, sondern auch für andere Sprachen. Ausserdem wurde für Korpora, die grösser als 1‘000‘000 Dokumente sind, die Laufzeit optimiert, indem ein Random Sampling betrieben wird. Zusätzlich wurden noch verschiedene Varianten aufgezeigt, wie das Topic Modeling in einem nächsten Schritt verbessert werden könnte.
In der Zukunft sollte als nächstes das Microservices-Konzept umgesetzt werden. Dies würde vielversprechende Erweiterungsmöglichkeiten garantieren. Einerseits würde es die Architektur vereinfachen und somit auch den Aufwand für das Einarbeiten eines frischen Entwicklers reduzieren. Andererseits können danach noch schneller und einfacher Module erweitert oder erstellt werden.

Ausschnitt einer Word Cloud