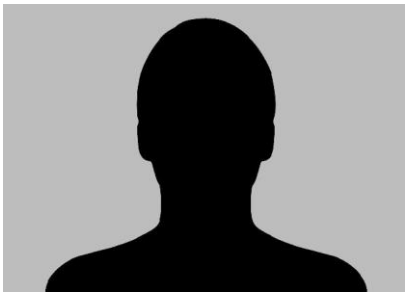
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

InIT Institut für angewandte Informationstechnologie

KenSpace: Explorative und komplexe Suchen auf unstrukturierten Dokumenten



Diplomierende
Pascal Severin Andermatt
Stefan Brunner

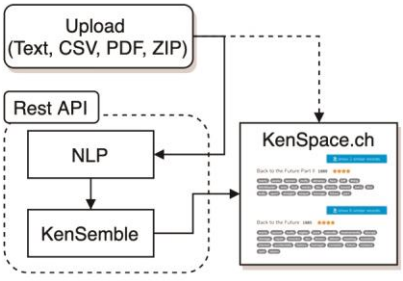
Dozent
Andreas Weiler

In der heutigen Welt werden immer mehr Daten produziert, die öffentlich zugänglich sind. Um die Daten zu finden, müssen komplexe Suchmechanismen, wie eine unscharfe Suche, eingesetzt werden. Zusätzlich werden die Resultate meist in einer fast endlos langen Liste dargestellt, welche zu einem Datenüberfluss für den Anwender führt. Trotz der ständig wachsenden Datenmenge ändern sich die Suchmechanismen nur minimal. Eine grosse Herausforderung bei der Begrenzung des Datenüberflusses besteht darin, die Daten zu kategorisieren und sie einer kleinen Menge statischer Facetten zuzuordnen.

Um dieses Problem zu lösen, werden wir in unserer Arbeit eine explorative Suche auf unstrukturierten Dokumenten umsetzen und auswerten. Das Ziel dieser Bachelorarbeit ist es, herauszufinden, wie dynamische Facetten generiert werden können und ob sie dem Benutzer beim Finden und Explorieren seiner Daten helfen. Das resultierende Produkt wird als Prototyp entwickelt und im World Wide Web (WWW) zur Verfügung gestellt. Zudem wird eine Representational State Transfer (REST) API für Dritte freigegeben. Um das Ziel der Generierung der dynamischen Facetten zu erreichen, evaluieren wir die zwei Natural Language Processing (NLP) Bibliotheksklassen: Natural Language Toolkit (NLTK) und spaCy. Die generierten Vorschläge werden mit dem KenSemble-Verfahren erstellt. Dieses Verfahren kombiniert das K-Means Clustering mit dem Latent Dirichlet Allocation (LDA). Zusätzlich werden auch weitere Methodiken des Unsupervised Learning wie das hierarchische Clustering analysiert.

Explorative Suche von KenSpace mit Hilfe von Facetten

Um die ganze Applikation zu evaluieren, wird eine User Study mit 23 Personen durchgeführt und es werden Case Studies mit Datensätzen von AirBnB und weiteren kritisch analysiert. Resultierend stellt sich heraus, dass das KenSemble-Verfahren sich für die explorative Suche eignet und die Benutzer die Applikation gegenüber anderen bevorzugen. Das Verfahren erreicht einen *F-Score* von 0,6 mit Hilfe von Filmdaten als Dokument-Basis. Basierend auf den Ergebnissen ist ein Potential zur Verwendung der dynamischen Facetten im Alltag möglich. Jedoch müsste die Applikation noch einige Optimierungen durchlaufen, um sie produktiv für jegliche eigene Daten zu verwenden.



Ablauf von KenSpace