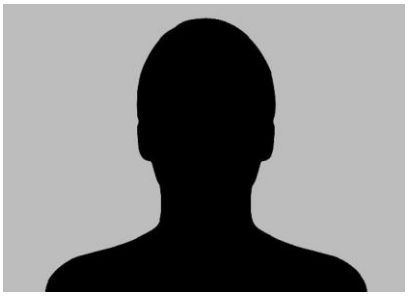
Zurich University
of Applied Sciences

School of
Engineering

InIT Institute of Applied
Information Technology

KenSpace: Exploratory and complex searches on unstructured documents



Diplomierende
Pascal Severin Andermatt
Stefan Brunner

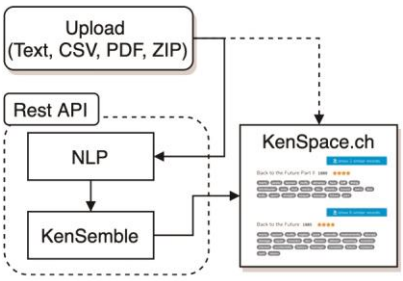
Dozent
Andreas Weiler

Nowadays, more and more data is being produced and made publicly available. To find the data, complex search mechanisms, such as a fuzzy search, must be used. In most cases, the results are presented in an almost endlessly long list, which leads to a data overload for the user. Despite the ever-increasing amount of data, the search mechanisms change only minimally. A significant problem to limit the data overflow is the categorization of the data and its assignment to a small set of static facets.

To solve this problem, we are going to implement and evaluate an explorative search on unstructured documents. The goal of this bachelor thesis is to find out how dynamic facets can be generated and if they can help the user to find and explore his data. The resulting product will be developed as a prototype and made available in the World Wide Web (WWW). Furthermore, a Representational State Transfer (REST) API is released for third parties. To achieve the goal of generating the dynamic facets, we evaluate the two Natural Language Processing (NLP) library classes: Natural Language Toolkit (NLTK) and spaCy. The generated suggestions are created with the KenSemble method. This procedure combines the K-Means clustering with the Latent Dirichlet Allocation (LDA). Additionally, other methods of Unsupervised Learning like hierarchical clustering are analyzed.

Exploratory search of KenSpace with the help of facets

To evaluate the whole application, a user study with 23 persons will be performed and case studies with data sets from AirBnB and others will be critically analyzed. The result is that the KenSemble method is suitable for explorative search and users prefer the application over others. The procedure achieves an *F-Score* of 0.6 using film data as a document basis. Based on the results, a potential for using the dynamic facets in everyday life is possible. However, the application would still have to go through some optimizations in order to use this productively for any own data.



Process of KenSpace