Topic Search - Suchmaschine für ein spezifisches Thema

Studiengang: BSc in Informatik | Vertiefung: Data Engineering
Betreuer: Prof. Dr. Erik Graf
Experte: Pierre-Yves Voirol (Abacus Research AG)

Das World Wide Web besteht aus einem Übermass an Daten. Um aktiv nach passenden Informationen in dieser Datenmenge zu suchen, braucht es Suchmaschinen. Im Rahmen dieser Bachelorthesis wird eine semantische Suchmaschine realisiert, die auf eine bestimmte Domäne beschränkt ist. Zur Umsetzung wird neben regelbasierten Ansätzen auch auf Tools aus dem Bereich von Deep Learning zurückgegriffen.

Einleitung

Bei Topic Search handelt es sich um eine Applikation, welche im Web verfügbare, themenspezifische Dokumente ermittelt und über ein User Interface passende Resultate zu einem Suchbegriff liefert. Dies ermöglicht es, die Suchergebnisse auf Inhalte aus einzelnen Domänen, beispielsweise dem Thema „erneuerbare Energien", zu fokussieren. Durch die inhaltliche Eingrenzung wird das Suchumfeld verkleinert, was präzisere Resultate liefert. Zusätzlich zur Keyword-Suche ist es möglich, die Resultate nach deren Inhalt zu filtern. Durch die Filterung können die Resultate, basierend auf den im Text vorkommenden Personen, Orten, Kantonen oder auch Organisationen, eingeschränkt werden.

Ziele

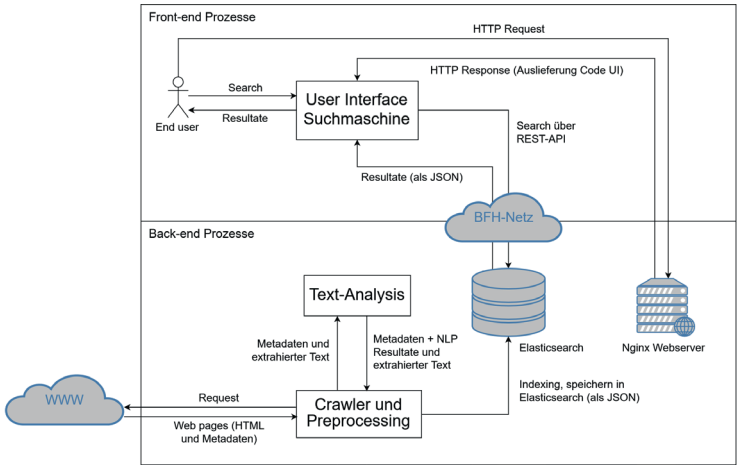
Das Ziel dieser Bachelor Thesis war es, einen Prototyp für ein lauffähiges System zu entwickeln, welcher die verschiedenen Prozesse durchläuft, die für eine domänenspezifische Suche relevant sind. Unter anderem ging es darum, mittels Crawling passende Ressourcen im Internet zu finden. Mithilfe dieser Daten und passenden Bibliotheken zur Textverarbeitung wurden weiterführende Textanalysen durchgeführt. Im Anschluss sollten die aufbereiteten Daten an einem geeigneten Ort gespeichert und schlussendlich dem Benutzer in passender Form dargestellt werden, um das Sucherlebnis zu optimieren.

Die Anwendung

Im Umfang dieser Arbeit wurde das System so umgesetzt, dass alle spezifischen Teilschritte auf einem Server laufen. Der Vorgang ist automatisiert, so dass während dem Crawling bereits das Preprocessing und die Weiterverarbeitung durchlaufen werden. Anschliessend werden die gesammelten Daten als Dokumente in Elasticsearch, einer quelloffenen Suchmaschine, indexiert. Über das UI kann der Benutzer entsprechend die Resultate abrufen. Das UI wird komplett clientseitig gerendert und über einen Webserver ausgeliefert.

Nathalie Bandi

Tamara Burri



Prozess-Architektur