BA/MA Theses: Sprachtechnologie für das Transparenzportal

Das Transparenzportal Hamburg (http://transparenz.hamburg.de/) ist das im Hamburgischen Transparenzgesetz (HmbTG) geforderte Informationsregister, über das alle per Gesetz veröffentlichungspflichtigen Informationen anonym recherchiert werden können. Es ist der zentrale Zugang zu aktuellen Daten und Informationen der Hamburger Verwaltung und stellt, um die leichte Auffindbarkeit der gesuchten Inhalte zu gewährleisten, eine Suche über den Volltext aller Datensätze bereit.
Im Transparenzportal stehen ca. 70'000 Dokumente seit 2012 zur Verfügung, es kommen kontinuierlich neue Dokumente hinzu, über 10'000 pro Jahr. Die Dokumente kommen aus heterogenen Quellen (viele Protokolle u.ä., aber auch Baupläne, Excel-Tabellen usw.) und sind bereits digitalisiert und OCRed, d.h. falls es Textdokumente sind, steht der Text zur Verfügung. Die Dokumente sind anonymisiert, d.h. Privatpersonen sind geschwärzt. Die Dokumente sind bereits in einem Apache SolR Index vorindiziert, welcher ohne Einschränkungen verfügbar gemacht werden kann, z.B. als Dump oder als Vollindexkopie.
In diesem Umfeld ergeben sich zwei Themen, welche bei erfolgreicher Umsetzung direkt in das Portal einfließen können. Beide Themen sind je nach Definition des Umfangs als BA oder MA bearbeitbar. Viele der sprachtechnologischen Einzelkomponenten sind in Teilen schon vorhanden, müssen jedoch auf den Anwendungsfall angepasst werden.
Wir sind für ähnlich gelagerte Themenvorschläge offen.

1. Semantische Suche im Transparenzportal
In den Dokumenten werden ggf. Begrifflichkeiten verwendet, welche den Laiensuchern nicht bekannt sind. Abhilfe schaffen könnte eine automatische Erweiterung der Suche: z.B. können im Fall von 0 Treffern ähnliche Begriffe angezeigt werden, welche tatsächlich im Index vorkommen. Hierzu werden Wortähnlichkeiten aus einem großem Hintergrundkorpus verwendet, z.B. siehe: http://ltmaggie.informatik.uni-hamburg.de/jobimviz/
Zum Beispiel ergibt "Kloake" im Transparenzportal 0 Treffer. Die drei ähnlichsten Wörter "Schlamm", "Brühe", und "Abwasser" sind jedoch im Transparenzportal vorhanden.
Dies kann man auch auf Anfragen ausweiten, welche nicht 0 sondern nur wenige Treffer haben, ferner können wir die Ähnlichkeiten auch aus den Transparenzportaldokumenten selbst ziehen.
Mögliche Erweiterungen:

- Mehrwortbegriffe mit bekannten Methoden identifizieren und zur Einschränkung der Suche einsetzen, falls es zu viele Treffer gibt, z.B. bei "Anstalt" anbieten: "Sozialtherapeutische Anstalt", "Anstalt Hamburger Stadtentwässerung" ...
- Kompositazerlegung einsetzen, z.B. sollte "Siel" auch "Sielabgabengesetzes" finden, jedoch nicht "Fantasieland".

2. Organisationsfinder
Wir könnten in den Dokumenten Organisationen (z.B. Verwaltungseinheiten, Firmen welche mit Ausschreibungen betraut sind usw.) mit Namenserkennung automatisch auszeichnen und zu Suchanfragen diese als Suchfacetten anbieten: Während das Portal http://suche.transparenz.hamburg.de/ auf der rechten Seite Facetten nach Metadaten (Kategorie, Dateiformat etc.) bietet, könnten wir hier Firmen oder Organisationseinheiten hinzufügen. Dies wäre ähnlich zum vorangegangenen new/s/leak-Projekt, siehe http://www.newsleak.io/ .
Mögliche Erweiterungen:

- Relation zwischen Suchbegriff und den Organisationen erkennen, z.B. Suchbegriff "Luftqualität" liefert "HaLM betreibt Messstationen"
- Möglichkeiten zur manuellen Korrektur der Erkennung für die kontinuierliche Verbesserung der Namenserkennungsgenauigkeit

Bei Fragen und Interesse bitte Chris Biemann (LT, biemann@informatik.uni-hamburg.de) und Lothar Hotz (HITeC, hotz@informatik.uni-hamburg.de) kontaktieren. Stand: 9/2017