ETH zürich

Datensammlung

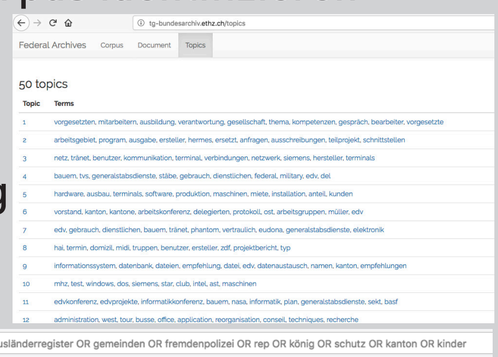
Die Datengrundlage des vorgestellten Werkzeugs bilden zum einen retrodigitalisierte Quellenbestände wie Korrespondenzen, Berichte, Listen, Vorträge und Protokolle aus dem Bundesarchiv Bern und zum anderen wissenschaftlich-technische Publikationen und (verwaltungs-)interne Veröffentlichungen.

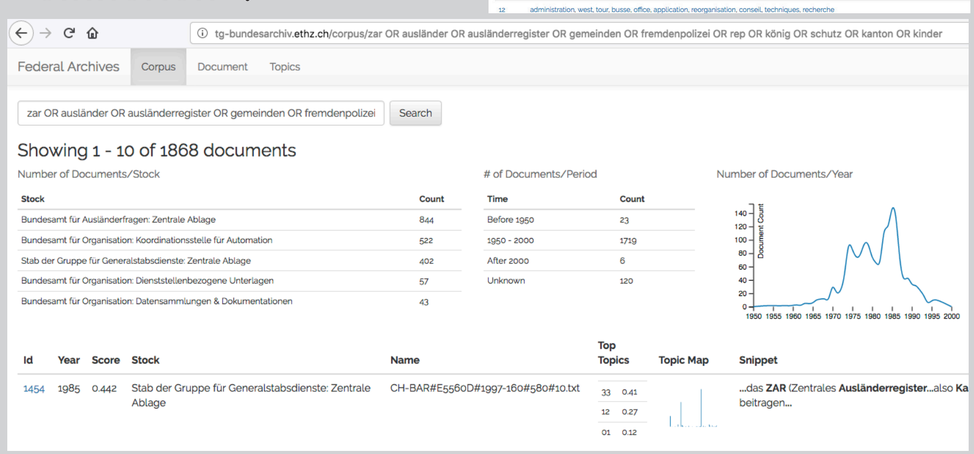
Datenaufbereitung

1. Initialisierung einer SQLite Datenbank zur permanenten Speicherung der Text- und Metadaten.
2. Extraktion der OCR-Textebenen mit Hilfe von PDFMiner und Pandas.
3. Datumsangaben werden anhand von regulären Ausdrücken extrahiert.
4. Die Dokumente werden zur besseren Wiederauffindbarkeit mit den Archivsignaturen aus dem Bundesarchiv versehen.
5. Die Sprache des Texts wird mit langdetect erkannt.
6. Der Text wird mit Hunspell einer automatischen Rechtschreibprüfung unterzogen. Darüber hinaus werden die Sätze mit einem Sprachmodul plausibilisiert, das mit 200.000 deutschsprachigen Wikipedia-Seiten trainiert worden ist.
7. Mit dem Stanford POS Tagger und NLTK werden lexikalisch-grammatischen Einheiten klassifiziert.

Analyse

- Wichtige Wörter im Korpus identifizieren (Tf-idf)
- Topic Modelling (LDA) →
- Dokumentenverteilung nach Organisations-einheiten ↓





Visualisierung

- Visualisierung der Suchergebnisse ↑
- Parametrisierbare Volltextsuche ↓
- Visualisierung der Topic-Wahrscheinlichkeit auf Dokumenten- und Korpusebene

Topic Modelling & explorative Suche

Im Rahmen des Projekt «Aushandlungszonen. Computer und Schweizerische Bundesverwaltung, 1960 - 1990» erarbeitet die Professur für Technikgeschichte der ETH Zürich in Zusammenarbeit mit den ETH Scientific Services Werkzeuge (Parametrisierbare Volltextsuche & Topic Modelling), um einen laufend wachsenden Quellenkorpus zu durchdringen.

D GESS