Lawsearch

Information Retrieval, Lab Project, SS 2019
University of Leipzig

Goal of the search engine

- Suche in deutschen Gesetzestexten
- Suche in amerikanischen Rechtsfällen

German corpus

- xml/html/(pdf)
- einzelne xml-Dokumente (Namen unique)
- 6419 Einträge; ca. 0,5 GB
- existierende Suchmethoden: Titelsuche/Volltextsuche
- englische Übersetzungen für einige Dokumente verfügbar (nur html/pdf)

German Corpus

1. Header
   - Abkürzung für den Titel
   - **Ausfertigungsdatum**
   - **langer Titel**
   - **kurzer Titel** (optional)
   - Fußnoten
   - Stand
2. Inhaltsübersicht (optional)
3. Abschnitt (min.1)
   - **Titel des Abschnitts**
   - Nummer des Abschnitts
4. Text (min.1)
   - Nummer des Paragraphs
   - **Titel des Paragraphs**
   - **Inhalt des Paragraphs (Text/Tabellen)**

English corpus

- jsonl
- pro Staat ein jsonl-Dokument
- insgesamt ca. 260000 Einträge; ca. 3,4 GB

English corpus

- id
- **Name des Falls**
- Namensabkürzung
- **Datum des Urteils**
- Aktennummer
- erste Seite
- letzte Seite
- Zitierungen
  - Typ
  - **Zitierung**
- Band
- **Reporter**
- Gericht
  - ID
  - **Name**
  - **Namensabkürzung**
  - url (meist null)
  - **slug**

English corpus

- Gerichtsbarkeit
  - ID
  - **slug**
  - Name
  - **vollständiger Name**
  - whitelisted (boolean)
- Falldaten
  - **Richter**
  - **Anwälte**
  - Meinungen
    - **Autor**
    - **Text**
    - Typ
  - **belteilgte Parteien**
  - **Hauptpunkt**
  - Status

Current status

- Java Server
- Frontend
- Apache Lucene