Projekt Data Analytics

Teilnehmer: Michael Clemens Heinrich, Thomas Voss

# Projekttitel: Wortanalyse der gemeinsamen Ministerialblätter 1950-2023

# Datensatz:

Auf der Website fragdenstaat.de sind die 2713 gemeinsamen Ministerialblätter von 1950-2023 sowohl im .pdf- als auch im .json-Format hinterlegt. Die pdf-Dateien sind eingescannte Dokumente, bei denen ein durch OCR entstandener digitaler Text hinterlegt ist. Dieser Text ist ebenfalls in den .json-Dateien enthalten. Im pdf-Format ist der Datensatz ca. 3,5 GB groß. Der reine Text im json-Format ist ca. 400 MB groß. Der Datensatz wird mithilfe der API von fragdenstaat.de heruntergeladen.

# Datenbereinigung:

Mithilfe von Spacy und weiteren Tools sollen aus den Dokumenten die inhaltlich relevanten Wörter identifiziert und passend zusammengefasst werden.

# Datenverarbeitung:

Für die einzelnen Wörter sollen grundlegende jahresbasierte und kumulierte Statistiken erhoben werden. Dazu gehören die absolute wie die relative Häufigkeit eines Wortes nach Jahr und kumulierte Häufigkeiten und die Standardabweichung der relativen Häufigkeiten. Daraus sollen Korrelationen zwischen verschiedenen Wörtern errechnet werden.

# Datenvisualisierung:

Mittels Dash soll in Python ein Dashboard mit den relevanten Statistiken erstellt werden.