ocror-detector

Ziel

Aus 20,000 PDF schnell diejenigen mit unbrauchbarer OCR filtern

Idee

- Einfache Basiswerte (# Worte, Verteilung Wortlänge, # Zeilen ...)
- In Relation setzen
- Filtern der PDF-Metadaten anhand der Messwerte

Probleme

- evtl. nicht englisch
- Rauschen durch Formeln, Tabellen, bibliografische Referenzen
- Silbentrennung
- Boilerplate (Wasserzeichen, Kopfzeilen usw)