OCR tesseract ocrmypdf

written by js | 2021-10-24

Tema

OCRiranje programom tesseract (slike) ili ocrmypdf (pdf).

tesseract OCR-ira u tekstualne datoteke.

ocrmypdf kreira PDF datoteke u kojima je moguće (bolje) pretraživati tekst (ako je već pretraživa po tekstu, ocrmypdf je neće OCR-irati jer nema potrebe. No, ako misliš da može bolje napraviti, upotrijebi "–force-ocr"). Koristi tesseract u podlozi.

Postupak – tl;dr

- `apt install tesseract tesseract-ocr-hrv ocrmypdf`
- *(instaliraj i dodatne jezike ako trebaš)*
- `ocrmypdf --force-ocr --sidecar "rjecnik" "ulazna_datoteka.pdf" "izlazna_datoteka.pdf"`
- `for f in *.jpg; do tesseract -l hrv --oem 3 "$f" "$f.txt"; done`
- `for f in *.pdf; do ocrmypdf -l hrv --oem 3 "$f" "ocr-$f"; done`
- `ocrmypdf -l hrv --sidecar rjecnik.txt --tesseract-oem 3 --title "Neki naslov dojde ovdje" --jpeg-quality 99 --png-quality 99 --force-ocr --image-dpi 300 --verbose --deskew --clean --oversample 300 --optimize 1 --rotate-pages --output-type pdfa ulazni_dokument.pdf izlazni_dokument.pdf`

Razno

Funkcionira i pod WSL

Slike

…

Video

…

---

ocr xocrx tesseract xtesseract ocrmypdf xocrmypdf xscan sken skeniranje