OCR .PDF datoteke

written by js | 2022-11-15

- `ocrmypdf -O 3 --tesseract-oem=3 --language=hrv --force-ocr --output-type pdf --sidecar OCR.txt inputFile.pdf temp.pdf`

Po završetku će datoteka “OCR.txt” imati tekst iz datoteke “inputFile.pdf”. Datoteka “temp.pdf” je upravo to.

Naravno, kvaliteta i kvantitela izvučenog teksta ovisi o kvaliteti originalnog dokumenta, kvaliteti skeniranja, …

Šanse za bolji OCR možeš malo povećati auto-levelom ulazne datoteke: https://jednostavno.somware.org/1582/pdf-kontrast-i-to/

https://ocrmypdf.readthedocs.io/en/latest/index.html

---

xpdf xocr ocr optical character recognition ocriranje xocriranje