OCR w Pythonie – moduły, przykłady i instalacja Tesseract

1. Popularne moduły OCR w Pythonie

• pytesseract – Najpopularniejsza biblioteka OCR w Pythonie. Interfejs do Tesseract OCR (Google).

• easyocr – OCR oparty na deep learning. Nie wymaga zewnętrznej instalacji Tesseract.

• ocrmypdf – Dodaje warstwę OCR do skanowanych plików PDF.

• kraken – Zaawansowany OCR oparty na RNN. Dobrze radzi sobie z rękopisami.

2. Porównanie: pytesseract vs easyocr

Poniższy kod analizuje trzy obrazy z polskim tekstem i porównuje wyniki OCR z użyciem pytesseract i easyocr.

```
import pytesseract
from PIL import Image
import easyocr

images = ["polski_tekst1.png", "polski_tekst2.png",
"polski_tekst3.png"]

# (dla Windows ustaw ścieżkę do tesseract.exe)
# pytesseract.pytesseract.tesseract_cmd = r"C:\Program Files\Tesseract-
OCR\tesseract.exe"
reader = easyocr.Reader(['pl'], gpu=False)

for img_path in images:
    print("=" * 60)
    print(f"Obraz: {img_path}")
    image = Image.open(img_path)

    print("\nWynik pytesseract:")
    print(pytesseract.image_to_string(image, lang="pol").strip())

    print("\nWynik easyocr:")
    for box in reader.readtext(img_path):
        print(box[1])
```

3. Instalacja Tesseract OCR

Windows

1. Pobierz instalator: https://github.com/UB-Mannheim/tesseract/wiki

2. Zainstaluj Tesseract OCR (np. tesseract-ocr-w64-setup-v5.3.3.exe).

3. Wybierz język polski w sekcji „Additional language data”.

4. Dodaj ścieżkę do katalogu instalacyjnego (np. C:\Program Files\Tesseract-OCR) do zmiennej PATH.

5. Sprawdź instalację: uruchom w cmd `tesseract --version`.

Linux (Ubuntu / Debian)

sudo apt update
sudo apt install tesseract-ocr tesseract-ocr-pol

macOS (Homebrew)

brew install tesseract
brew install tesseract-lang

Sprawdzenie języków

Użyj komendy: tesseract --list-langs
Powinien pojawić się język "pol".