Create a simple OCR reader with TessaractOCR

Create a command-line tool to do OCR and extract text from a given scanned document image.

Input

Scanned document images (include English only). Your command needs to accept the following file formats.

- PNG
- JPEG
- PDF

Output

Extracted texts as a text file.

Interface

```
python your_code.py --input=./test.pdf --output=output.text --verbose
```

- –input : input file
- –output : output text file
- –verbose : verbose mode (output detailed logs)

Requirements

- Before doing the OCR process with TessaractOCR, you should do pre-processing to improve OCR accuracy.
- After doing the OCR process with TessaractOCR, you should do post processing to do text correction to remove OCR mis-recognition.

What we want to check

- **Clarity**: You can write clear code that any devs could read and understand in one go
- **Simplicity** : You can write gimmick-free and straightforward code with no ambiguities
- **Defensiveness**: You can cover edge cases and treat user inputs with care

Regurations

- Use Click as an command line interface builder
- Use Poetry to install required thirdparty packages

- Use yapf and isort to format python codes
- Use logging package to do output. Never use `print` for log output.
- Use `.gitignore` to exclude unnesesarry files.

```
from logging import getLogger
logger = getLogger(__name__)
```

- Upload your code to GitHub and send the URL of the repository to us.