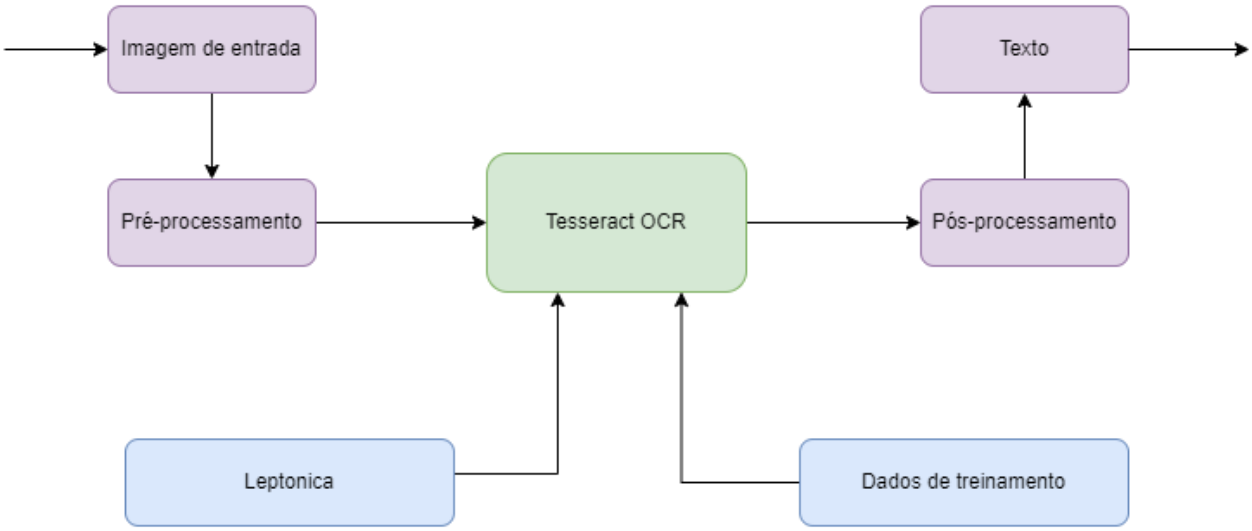
Visão Computacional: reconhecimento de texto com OCR e OpenCV

cursos.alura.com.br/course/visao-computacional-reconhecimento-texto-ocr-opencv/task/113541

O Tesseract OCR é um mecanismo de reconhecimento de texto de código aberto, que está sob a licença Apache 2.0, e podemos encontrar seu repositório no GitHub. O uso do Tesseract OCR pode ser feito de diversas maneiras para retirar textos de imagens ou extrair textos diretamente de PDFs, por exemplo. Por suportar uma ampla variedade de idiomas e ser compatível com muitas linguagens de programação, é interessantíssimo entender como ele trabalha por trás dos panos.



Seu funcionamento ocorre em partes e podemos observar na imagem que os processos em roxo dependem da imagem que passamos de entrada, então temos:

1. imagem de entrada;
2. pré-processamento da imagem;
3. passagem pelo Tesseract OCR;
4. pós processamento da imagem e retorno em texto, como já vimos em aula.

Em azul, na parte inferior da imagem, podemos ver duas entradas no Tesseract OCR que não são feitas por nós, a Leptonica e os dados de treinamento. Para entender melhor o que significa cada entrada e saída, vamos começar com as entradas em azul que **não** são feitas pelo usuário.

Leptonica é uma biblioteca de código aberto contendo software amplamente útil para aplicativos de processamento e análise de imagens. Ela é usada no Tesseract OCR e no OpenCV e pode ser usada em diversas aplicações.Seu repositório no GitHub trás diversas opções de aplicações e implementações, além de várias informações extras sobre a biblioteca.

Os **dados de treinamento**, que também estão na entrada do Tesseract OCR de cor azul na imagem, são de todos os idiomas que podemos utilizar e já são pré-treinados, ou seja, cada idioma pode ser importado para o projeto de forma rápida. Esses dados de treinamento estão disponíveis na pasta “tessdata” do repositório do GitHub do próprio Tesseract OCR.

Voltando aos passos que estão em roxo na imagem, que são gerados a partir da imagem de entrada, temos:

- **Imagem de entrada:** gerada a partir da aquisição da imagem por um scanner/câmera que lê o documento e transforma em dados binários. O software de OCR analisa a imagem digitalizada e classifica as áreas claras como plano de fundo e as áreas escuras como texto.
- **Pré-processamento:** limpeza da imagem, remoção de erros, troca de cores, a fim de prepará-la para a leitura.
- **Reconhecimento de texto com Tesseract OCR:** Os dois principais tipos de algoritmo de OCR ou processos de software que um software de OCR usa para reconhecimento de texto são chamados de **reconhecimento de padrões** e **detecção de recursos.**
    - **Reconhecimento de padrões** - Os programas OCR são alimentados com exemplos de texto em várias fontes e formatos que são usados para comparar e reconhecer caracteres no documento digitalizado.
    - **Detecção de recursos** - Os programas de OCR aplicam regras relacionadas aos recursos de uma letra ou número específico para reconhecer caracteres no documento digitalizado. Os recursos podem incluir o número de linhas angulares, linhas cruzadas ou curvas em um caractere para comparação. Por exemplo, a letra maiúscula “A” pode ser armazenada como duas linhas diagonais que se encontram com uma linha horizontal no meio.
- **Pós-processamento:** Após a análise, o sistema converte os dados de texto extraídos em um arquivo informatizado. Alguns sistemas de OCR podem criar arquivos PDF anotados que incluem versões anteriores e posteriores do documento digitalizado.
- **Texto:** Resultado final da extração.

Como o Tesseract OCR pode ser utilizado em diversas linguagens de programação, no Python podemos utilizá-lo com o pytesseract, biblioteca própria que foi feita para essa finalidade.

Discutir no Fórum Próxima Atividade