Instalação e Configuração do Tesseract OCR no Windows

MBA em Ciência de Dados – USP

Técnicas Avançadas de Captura e Tratamento de Dados

Autores: Jadson Oliveira, Damares Resende

Data: 15/05/2021

Tesseract OCR

Tesseract é um software de código aberto para o reconhecimento óptico de caracteres, originalmente desenvolvido pela Hewlett-Packard, desde 2006 é mantido pela Google, e atualmente hospedado no Github.

Este breve tutorial tem o intuito de demonstrar como realizar a instalação e configuração do Tesseract *OCR* dentro do ambiente Windows, para ser utilizado com a linguagem de programação Python por meio da biblioteca *pytesseract*.

Instalando o Tesseract OCR no Windows

Existem diversas formas para a instalação do Tesseract dentro do ambiente Windows, através da compilação do código fonte disponibilizado no Github ou por meio de instaladores desenvolvidos especialmente para facilitar tais configurações. Este tutorial tem foco na segunda opção, pela simplicidade e rapidez do processo. Serão apresentados duas formas de instalação, uma utilizando o ambiente conda (*que é a mais indicada*) e outra utilizando um instalador criado pela Universidade Mannhein.

Instalação utilizando o ambiente Conda:

Esta é a opção mais simples para quem utiliza o ambiente Conda para gerenciar pacotes, instalações e dependências. Siga os seguintes passos para instalação e configuração do Tesseract.

Passo 1 – Realize a instalação do motor de OCR Tesseract por meio do seguinte comando:

```
conda install -c conda-forge tesseract
```

Passo 2 – Após a instalação do tesseract por meio do Conda, será necessário a adição de dicionários para idiomas específicos a serem utilizadas pelo OCR. Para isso é indicado o uso dos modelos pré-treinados contidos no link oficial: https://github.com/tesseract-ocr/tessdata_best. Esses são os modelos, que atualmente, obtém a melhor acurácia no

processo de reconhecimento de caracteres. Utilize o link acima e faça o download dos dicionários que você deseja utilizar.

Passo 3 – Realize a adição dos modelos pré-treinados, isso pode ser realizado por meio de duas maneiras principais:

3.1 - Adicionando os arquivos de dicionário no diretório padrão do tesseract (normalmente segue o caminho: "/Library/bin/tessdata/por.traineddata");

3.2 - Atualizando a variável de ambiente que guarda a informação de onde o diretório de dicionários se encontra atualmente, para isso basta adicionar a seguinte linha de código no jupyter notebook, ou executar o comando similar no prompt do Windows:

```
%env TESSDATA_PREFIX=<caminho-diretorio-de-dicionarios-pre-treinados>
```

Instalação utilizando executável da Universidade Mannhein

Se já realizou a instalação do Tesseract por meio do Conda, ***ignore esta etapa****!*

Caso prefira realizar a instalação do motor de OCR por meio do executável disponibilizado pela Universidade de Mannhein, siga os seguintes passos:

Passo 1 – Faça o download da versão mais recente e estável do instalador mantido pela Universidade Mannhein (https://github.com/UB-Mannheim/tesseract/wiki). Atualmente, existem duas opções do instalador (32 e 64 bits), verifique qual a melhor delas para as suas configurações locais.

Passo 2 – Execute o instalador como administrador do sistema operacional (clique com o botão direito do mouse e escolha a opção "executar como administrador").

Passo 3 – Verifique os termos de uso, e siga as opções padrões de instalação do software. Marque também a opção para fazer download dos modelos pré-treinados (dicionários) para a língua portuguesa e outras linguagens de sua preferência.

Passo 4 – Atualize a variável de ambiente PATH com o caminho do executável do tesseract, de preferência no terminal que utilizará para inicializar o jupyter notebook, ou jupyter lab, ou mesmo no terminal inicializado do jupyter notebook. Para isso, basta executar a seguinte linha de comando, substituindo a parte r'caminho-para-tesseract.exe' pelo caminho do executável na sua máquina local (normalmente esse caminho é algo como: r'C:\\Program Files\\Tesseract-OCR\\tesseract.exe'):

```
pytesseract.pytesseract.tesseract_cmd = r'caminho-para-tesseract.exe'
```

Instalando a biblioteca pytesseract

Após a instalação do software Tesseract OCR, será necessário instalar o pacote pytesseract, que é um *wrapper* construído com o intuito de facilitar o uso do Tesseract por meio da linguagem de programação Python.

Para instalar a biblioteca pytesseract utilizando o gerenciador de dependências Conda, utilize o seguinte comando:

```
conda install -c conda-forge pytesseract
```

Para instalar a biblioteca pytesseract utilizando o gerenciador de dependências PIP, utilize o seguinte comando:

```
pip install pytesseract
```

Instalando bibliotecas complementares

Para a instalação de outras bibliotecas úteis e que são complementares à avaliação que está sendo proposta, basta executar os seguintes comandos:

Para quem utiliza o gerenciador de dependências **Conda**:

```
conda install -c conda-forge pdf2image
```

```
conda install -c conda-forge poppler
```

Para quem utiliza o gerenciador de dependências **PIP**:

```
pip install pdf2image
```

```
pip install poppler
```