Instalação e Configuração do Tesseract OCR no MacOS/Linux

MBA em Ciência de Dados – USP

Técnicas Avançadas de Captura e Tratamento de Dados

Autores: Damares Resende, Jadson Oliveira

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 Unix, que inclui os sistemas operacionais MacOS e Linux, para ser utilizado com a linguagem de programação Python por meio da biblioteca *pytesseract*.

Instalando o Tesseract OCR no MacOS

Existem diversas formas para a instalação do Tesseract. Nesse tutorial iremos mostrar três formas: através do ambiente Conda, do ambiente Colab, e baixando diretamente do repositório. Outras formas estão disponíveis no GitHub do projeto Tesseract.

Importante: recomendamos fortemente que ambientes virtuais sejam utilizados. Sua criação depende do tipo de ambiente usado. Conda, por exemplo, cria o ambiente com "**conda create --name nomeescolhido**", e o ativa com "s**ource activate nomeescolhido**". Já o virtualenv cria com "**virtualenv nomeescolhido**" e ativa com "**source nomeescolhido/bin/activate**". Em ambos casos, o ambiente virtual deve sempre ser ativado antes de executar o projeto. Esse é um mecanismo Python usado para deixar os projetos independentes e evitar problemas de conflitos de dependências de pacotes.

Instalação utilizando o ambiente Conda:

Essa é a forma padrão usada no curso. Para isso é requisito instalar o Conda. Caso esse requisito já esteja satisfeito, pule para o passo 3. Os passos completos estão abaixo.

Passo 1 - Instale o Homebrew. Ele é um programa para gerenciamento de pacotes análogo ao “apt-get” de alguns ambientes Linux. Com ele poderemos instalar os pacotes de linguagens do Tesseract e o Poppler.

```
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
```

Passo 2 – Baixe o instalador pelo site e siga as instruções. Se preferir, instale o Anaconda ou Miniconda pelo brew:

Anaconda:

```
brew install --cask anaconda
```

Adicione a path "`/usr/local/anaconda3/bin`" em "`/etc/paths`"

```
/usr/local/anaconda3/bin/conda init zsh
```

Reinicie o terminal

Miniconda:

```
brew install --cask miniconda
```

Adicione a path "`/opt/miniconda3/condabin/conda`" em "`/etc/paths`"

```
/opt/miniconda3/condabin/conda init zsh
```

Reinicie o terminal

Passo 3 – Instale o Tesseract.

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

Passo 4 – Instale o suporte à língua portuguesa. Com o comando abaixo todas as línguas são instaladas.

```
brew install tesseract-lang
```

Passo 5 – Instale o Pytesseract, o wrapper para o Tesseract em Python.

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

Passo 6 – Instale também as dependências para leitura de PDFs. Elas não são pré-requisitos do Tesseract, mas são utilizadas em algumas atividades deste módulo.

```
brew install poppler
```

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

Instalação utilizando o ambiente Colab:

Outra forma simples de instalar o Tesseract para quem usa o Colab. Siga os passos abaixo.

Passo 1 – Instale o Tesseract e suas dependências

```
apt install tesseract-ocr
```

```
apt install tesseract-ocr-por
```

```
apt install libtesseract-dev
```

```
apt install poppler-utils
```

Passo 2 – Instale em seu ambiente Python as dependências de projeto

```
pip install pytesseract
```

```
pip install pdf2image
```

Instalação utilizando o projeto do repositório:

Uma terceira forma de instalar o Tesseract é baixando-o diretamente do repositório. As instruções abaixo são baseadas no tutorial do GitHub do projeto usando o **brew**. Siga os passos abaixo.

Passo 1 - Instale o Homebrew. Ele é um programa para gerenciamento de pacotes análogo ao "apt-get" de alguns ambientes Linux. Com ele poderemos instalar os pacotes de linguagens do Tesseract e o Poppler.

```
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
```

Passo 2 – Instale as dependências do Tesseract

```
brew install automake autoconf libtool
```

```
brew install pkgconfig
```

```
brew install icu4c
```

```
brew install leptonica
```

Passo 3 – Baixe o projeto do repositório e configure-o. Ele já vem com a linguagem em português disponível.

```
git clone https://github.com/tesseract-ocr/tesseract/
```

```
cd tesseract
```

```
./autogen.sh
```

Passo 4 - Instale o Pytesseract, o wrapper para o Tesseract em Python.

```
pip install pytesseract
```

Passo 5 – Instale também as dependências para leitura de PDFs. Elas não são pré-requisitos do Tesseract, mas são utilizadas em algumas atividades deste módulo.

```
brew install poppler
```

```
pip install pdf2image
```

Troubleshooting (resolvendo problemas)

Nesse módulo estamos usando o Tesseract 4.1.1 e o Pytesseract 0.3.7. A versão Python não deve influenciar, mas recomendamos que seja maior ou igual à 3.7. Siga os passos abaixo para verificar sua versão e atualizá-la caso não seja compatível.

Ambiente Conda:

Passo 1 – Verifique a versão atual do Tesseract

```
conda list tesseract
```

Passo 2 – Atualize os pacotes

```
conda update conda
```

```
conda update tesseract
```

```
conda update pytesseract
```

Passo 3 – Confirme as novas versões

```
conda list tesseract
```

Ambiente Virtualenv (Pip):

Passo 1 – Verifique a versão atual do pip e Tesseract

```
pip3 --version
```

```
pip freeze | grep tesseract
```

Passo 2 – Atualize o pip e os pacotes

```
python3 -m pip install --upgrade pip
```

```
pip install --upgrade tesseract
```

```
pip install --upgrade pytesseract
```

Passo 3 – Confirme as novas versões

```
pip freeze | grep tesseract
```

Importante: até o presente momento, a versão mais atual do Tesseract no repositório pip é a 0.1.3, que é bem antiga. Se você o instalou com o pip, atualizar o pip não atualizará o Tesseract. A solução é desinstalá-lo e baixar o projeto diretamente do repositório, como descrito anteriormente. Desinstale o pacote com "**pip uninstall tesseract**" e atualize o brew com "**brew update**", depois instale o Tesseract com os dados do repositório.

Instalando o Tesseract OCR no Linux

Existem diversas formas para a instalação do Tesseract. Nesse tutorial iremos mostrar a forma mais simples: através do ambiente Conda. Outras formas estão disponíveis no GitHub do projeto Tesseract.

Instalação utilizando o ambiente Conda:

Essa é a forma mais simples de instalar o Tesseract. Para isso é requisito instalar o Conda. Caso esses requisitos já estejam satisfeitos, pule para o passo 2. Os passos completos estão abaixo.

Passo 1 – Instale o Anaconda pelo instalador do site.

Passo 3 – Instale o Tesseract.

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

Passo 4 – Instale o suporte à língua portuguesa.

```
sudo apt-get install tesseract-ocr-por
```

Passo 5 – Instale o pytesseract. Ele é um wrapper para o Tesseract em Python.

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

Passo 6 – Instale também as dependências para leitura de PDFs. Elas não são parte do Tesseract, mas são utilizadas em algumas atividades deste módulo.

```
sudo apt-get install -y poppler-utils
```

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