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 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 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 o Anaconda, os pacotes de linguagens do Tesseract e o Poopler.

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

Passo 2 – Instale o Anaconda. Após a instalação, reinicie o terminal para que ele atualize a PATH recém alterada. Outra opção é baixar o instalador pelo site.

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

```
export PATH="/usr/local/anaconda3/bin:$PATH"
```

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. 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.

```
brew install poppler
```

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

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
```