Package Index > **pytesseract** > **0.2.0**

pytesseract 0.2.0

Python-tesseract is a python wrapper for Google's Tesseract-OCR

Download
pytesseract-0.2.0.tar.gz

Python Tesseract
================

Python-tesseract is an optical character recognition (OCR) tool for python.
That is, it will recognize and "read" the text embedded in images.

Python-tesseract is a wrapper for `Google's Tesseract-OCR Engine <https: github.com="" tesseract-ocr="" tesseract="">`_. It is also useful as a
stand-alone invocation script to tesseract, as it can read all image types
supported by the Python Imaging Library, including jpeg, png, gif, bmp, tiff,
and others, whereas tesseract-ocr by default only supports tiff and bmp.
Additionally, if used as a script, Python-tesseract will print the recognized
text instead of writing it to a file.

USAGE

Quickstart

.. code-block:: python

```
try:
import Image
except ImportError:
from PIL import Image
import pytesseract

pytesseract.pytesseract.tesseract_cmd = '<full_path_to_your_tesseract_executable>'
# Include the above line, if you don't have tesseract executable in your PATH
# Example tesseract_cmd: 'C:\\Program Files (x86)\\Tesseract-OCR\\tesseract'

# Simple image to string
print(pytesseract.image_to_string(Image.open('test.png')))

# French text image to string
print(pytesseract.image_to_string(Image.open('test-european.jpg'), lang='fra'))

# Get bounding box estimates
print(pytesseract.image_to_boxes(Image.open('test.png')))

# Get verbose data including boxes, confidences, line and page numbers
print(pytesseract.image_to_data(Image.open('test.png')))
```

Support for OpenCV image/NumPy array objects

.. code-block:: python

```
import cv2

img = cv2.imread('/**path_to_image**/digits.png')
print(pytesseract.image_to_string(img))
# OR explicit beforehand converting
print(pytesseract.image_to_string(Image.fromarray(img))
```

Add the following config, if you have tessdata error like: "Error opening data file..."

.. code-block:: python

```
tessdata_dir_config = '--tessdata-dir "<replace_with_your_tessdata_dir_path>"'
# Example config: '--tessdata-dir "C:\\Program Files (x86)\\Tesseract-OCR\\tessdata"'
# It's important to add double quotes around the dir path.

pytesseract.image_to_string(image, lang='chi_sim', config=tessdata_dir_config)
```

Functions

* **image_to_string** Returns the result of a Tesseract OCR run on the image to string

* **image_to_boxes** Returns result containing recognized characters and their box boundaries

* **image_to_data** Returns result containing box boundaries, confidences, and other information. Requires Tesseract 3.05+. For more information, please check the `Tesseract TSV documentation <https: github.com="" tesseract-ocr="" tesseract="" wiki="" command-line-usage#tsv-output-currently-available-in-305-dev-in-master-branch-on-github="">`_

Parameters

``image_to_data(image, lang=None, config=", nice=0, output_type=Output.STRING)``

* **image** Object, PIL Image/NumPy array of the image to be processed by Tesseract

* **lang** String, Tesseract language code string

* **config** String, Any additional configurations as a string, ex: ``config="-psm 6"``

* **nice** Integer, modifies the processor priority for the Tesseract run. Not supported on Windows. Nice adjusts the niceness of unix-like processes.

* **output_type** Class attribute, specifies the type of the output, defaults to ``string``. For the full list of all supported types, please check the definition of `pytesseract.Output`_ class.

.. _pytesseract.Output: src/pytesseract.py

INSTALLATION

Prerequisites:

- Python-tesseract requires python 2.5+ or python 3.x
- You will need the Python Imaging Library (PIL) (or the Pillow fork).
Under Debian/Ubuntu, this is the package **python-imaging** or **python3-imaging**.
- Install `Google Tesseract OCR <https: github.com="" tesseract-ocr="" tesseract="">`_
(additional info how to install the engine on Linux, Mac OSX and Windows).
You must be able to invoke the tesseract command as *tesseract*. If this
isn't the case, for example because tesseract isn't in your PATH, you will
have to change the "tesseract_cmd" variable ``pytesseract.pytesseract.tesseract_cmd``.
Under Debian/Ubuntu you can use the package **tesseract-ocr**.
For Mac OS users. please install homebrew package **tesseract**.

| Installing via pip:
Check the `pytesseract package page <https: pypi.python.org="" pypi="" pytesseract="">`_ for more information.

.. code-block:: bash

```
$ (env)> pip install pytesseract
```

| Or if you have git installed:

.. code-block:: bash

```
$ (env)> pip install -U git+https://github.com/madmaze/pytesseract.git
```

| Installing from source:

.. code-block:: bash

```
$> git clone https://github.com/madmaze/pytesseract.git
$ (env)> cd pytesseract && pip install -U .
```

LICENSE

Python-tesseract is released under the GPL v3.

CONTRIBUTERS

- Originally written by `Samuel Hoffstaetter <https: github.com="" h="">`_
- `Juarez Bochi <https: github.com="" jbochi="">`_
- `Matthias Lee <https: github.com="" madmaze="">`_
- `Lars Kistner <https: github.com="" sr4l="">`_

File	Type	Py Version	Uploaded on	Size
pytesseract-0.2.0.tar.gz (**md5**)	Source		2018-01-31	151KB

Author: Matthias Lee
Home Page: https://github.com/madmaze/python-tesseract
Keywords: python-tesseract OCR Python
License: GPLv3
Categories

- **Programming Language :: Python**
- **Programming Language :: Python :: 2**
- **Programming Language :: Python :: 3**

Package Index Owner: madmaze
DOAP record: pytesseract-0.2.0.xml