Tesserwrap Documentation

Release 0.1.1

Greg Jurman, et al

Jun 01, 2017

Contents

Tesserwrap is a ctypes/capi wrapper for Tesseract OCR.

class tesserwrap.**Tesseract**(*datadir=’‘, lang=’eng’*)

Tesseract OCR object.

Parameters

- **datadir** – Tesseract data-directory with Tesseract training data.
- **lang** – The language of the image(s) to be OCRed.

A simple example:

```
>>> from tesserwrap import Tesseract
>>> from PIL import Image

>>> img = Image.open("test.png")
>>> tr = Tesseract()
>>> tr.ocr_image(img)
'The quick brown fox jumps ove\n\n'
```

clear()

Clear the tesseract Image, and clean up any Tesseract run-data.

get_mean_confidence()

Returns the (average) confidence value between 0 and 100.

get_page_seg_mode()

Returns the page analysis mode from Tesseract

get_rectangle()

Get the bounding rectangle that tesseract is looking at inside of the image.

get_symbols()

Get a list containing all symbols in the OCR’d image. :returns: A list containing objects with the attributes:

> value: the string value of the symbol box: left, upper, right, and lower pixel coordinate confidence: confidence value between 0 and 100

get_text()

Get the text of the OCR’d image as a byte-string

get_textlines()

Get a list containing all lines in the OCR’d image. :returns: A list containing objects with the attributes:

> value: the string value of the line box: left, upper, right, and lower pixel coordinate confidence: confidence value between 0 and 100

get_utf8_text()

Get the text of the OCR’d image as a string.

This function is kept for backwards compatability with the 0.0 version of tesserwrap.

get_words()

Get a list containing all the words in the OCR’d image. :returns: A list containing objects with the attributes:

> value: the string value of the word box: left, upper, right, and lower pixel coordinate confidence: confidence value between 0 and 100

ocr_image(*image*)

OCR an image returning the UTF8 text data.

> **Parameters** **image** – image Image to be OCR’d by tesseract.

set_image(*image*)

Takes a PIL Image and loads it into Tesseract for further operations.

Note:: This function will automatically convert the image to Grayscale.

Parameters **image** – image Image to use in tesseract.

set_page_seg_mode(*mode=6*)

Set the page layout analysis mode.

Parameters **mode** – integer The page layout analysis mode. See PageSegMode class for options

set_rectangle(*left*, *top*, *width*, *height*)

Set the OCR detection bounding-box.

Parameters

- **left** – integer Pixels offset right from left of the image.
- **top** – integer Pixels offset down from the top of the image.
- **width** – integer Width of the bounding-box.
- **height** – integer Height of the bounding-box.

set_variable(*key*, *value*)

Set an internal Tesseract variable.

Parameters

- **key** – str Variable name to change.
- **value** – str New variable value.

CHAPTER 1

Indices and tables

- genindex
- modindex
- search

Python Module Index

Index