MicroPyramid Follow
Python, Django, Android and IOS, reactjs, react-native, salesforce, AWS development services company
Oct 30, 2017

Extract text with OCR for all image types in python using pytesseract

What is OCR?

Optical Character Recognition(OCR) is the process of electronically extracting text from images or any documents like PDF and reusing it in a variety of ways such as full text searches.

In this blog, we will see, how to use 'Python-tesseract', an OCR tool for python.

pytesseract:

It will recognize and read the text present in images. It can read all image types—png, jpeg, gif, tiff, bmp etc. It's widely used to process everything from scanned documents.

Installation:

```
$ sudo pip install pytesseract
```

Requirements:

* Requires python 2.5 or later versions.
* And requires Python Imaging Library(PIL).

Usage:

From the shell:

```
$ ./pytesseract.py test.png
```

Above command prints the recognized text from image ‘test.png’.

```
$ ./pytesseract.py -l eng test-english.jpg
```

Above command recognizes english text.

In Python Script:

```
import Image
from tesseract import image_to_string


print image_to_string(Image.open('test.png'))
print image_to_string(Image.open('test-english.jpg'),
lang='eng')
```

The article was originally published at MicroPyramid blog