How to Extract Text from Images with Python

Learn to extract text from images in 3 lines of codes

Costas Andreou
Apr 11 · 2 min read ★

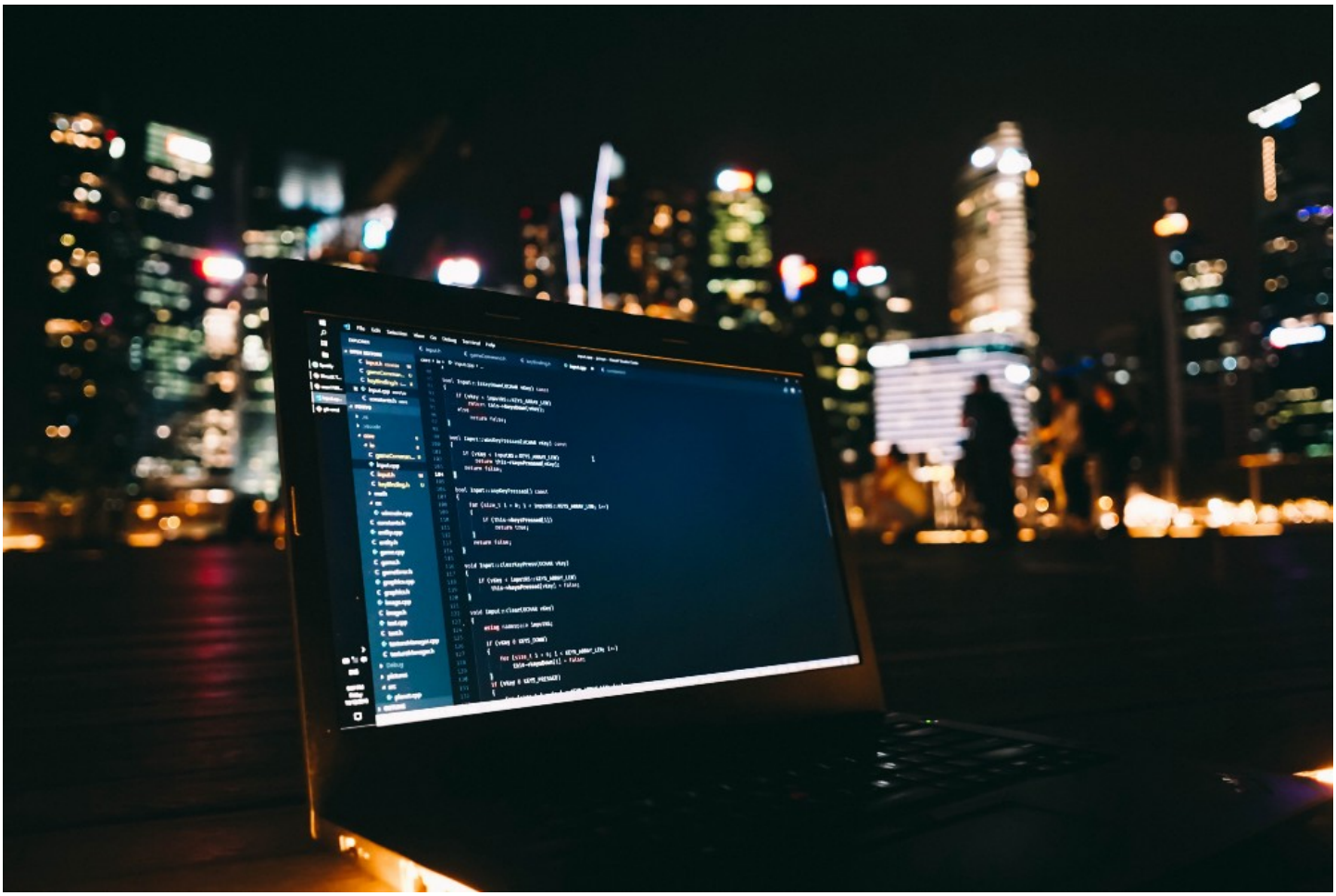

Photo by Sean Lim on Unsplash

In this short article, I am going to show you how you can use the power of Python to extract text from images. The applications of this technique are endless. Some examples include:

- Data mining for Machine Learning (ML) projects

- Taking pictures of receipts and reading the content for processing

. . .

The Python Library

To address this problem, we are going to be using a library known as Python Tesseract. From the library's website:

> *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. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. Additionally, if used as a script, Python-tesseract will print the recognized text instead of writing it to a file.*

. . .

Setting Things Up

When it comes to setting up Python libraries to use, it's usually a one-step process. With PyTesseract, however, we will need to do two things:

1. Install the Python Library
2. Install the Tesseract application

Firstly, to install the Python Library, simply open your command line window and type:

```
pip install pytesseract
```

Then, head to this website, download and install the Tesseract OCR executable. At the time of this writing, I am using the 64-bit Alpha Build v5.0.0, compiled on 2020–03–28.

We will need to know where we install this, as we will need to let your python script know.

Once you've followed the above, you're ready to get started.

. . .

The Python Code

As promised, with 3 lines of code, you will be able to read the text out of a picture:

```
import pytesseract

pytesseract.pytesseract.tesseract_cmd = r'C:\Program
Files\Tesseract-OCR\tesseract'

print(pytesseract.image_to_string(r'D:\examplepdf2image.png'))
```

. . .

If you liked the above article, you might also like:

Learn How to Quickly Create UIs in Python

Finally a library you can pick up in under 10 minutes

towardsdatascience.com

Estimating a Software Deadline Is Really Hard — Let's Talk About Why

The 5 laws you need to know for planning

medium.com

Data Science Machine Learning Startup Programming Ocr