OCR ?

OCR stands for Optical Character Recognition. It is a widespread technology to recognise text inside images. OCR technology is used to convert virtually any kind of images containing written text into machine-readable text data.

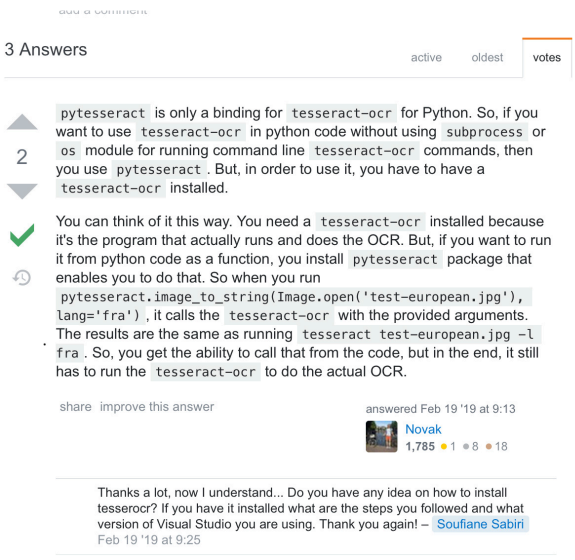
OCR python library
- pytesseract
- tesserocr
- baidu API

~~pytesseract~~ :
① 使用简单
② 进行中文识别要下载语言包
③ 支持 python 3.
④ 效果不理想 (?)

tesserocr :
① 使用人数多
② 比 pytesseract 快

选择 tesserocr 作为使用库

~~baidu~~ API :
① 数据多会导致时间慢
② 准确率高
③ 有一定额度免费



3 Answers

active oldest votes

2

`pytesseract` is only a binding for `tesseract-ocr` for Python. So, if you want to use `tesseract-ocr` in python code without using `subprocess` or `os` module for running command line `tesseract-ocr` commands, then you use `pytesseract`. But, in order to use it, you have to have a `tesseract-ocr` installed.

You can think of it this way. You need a `tesseract-ocr` installed because it's the program that actually runs and does the OCR. But, if you want to run it from python code as a function, you install `pytesseract` package that enables you to do that. So when you run `pytesseract.image_to_string(Image.open('test-european.jpg'), lang='fra')`, it calls the `tesseract-ocr` with the provided arguments. The results are the same as running `tesseract test-european.jpg -l fra`. So, you get the ability to call that from the code, but in the end, it still has to run the `tesseract-ocr` to do the actual OCR.

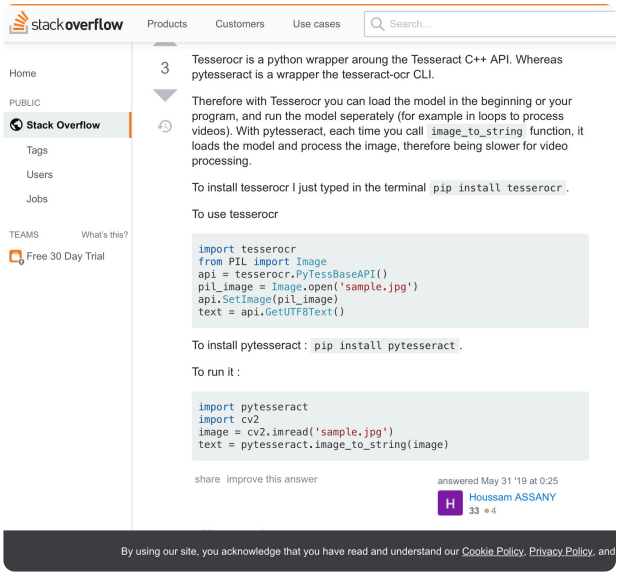
share improve this answer

answered Feb 19 '19 at 9:13

Novak

1,785 ● 1 ● 8 ● 18

Thanks a lot, now I understand... Do you have any idea on how to install tesserocr? If you have it installed what are the steps you followed and what version of Visual Studio you are using. Thank you again! – Soufiane Sabiri Feb 19 '19 at 9:25



stackoverflow Products Customers Use cases Search...

Home

PUBLIC

Stack Overflow

Tags

Users

Jobs

TEAMS What's this?

Free 30 Day Trial

3

Tesserocr is a python wrapper aroung the Tesseract C++ API. Whereas pytesseract is a wrapper the tesseract-ocr CLI.

Therefore with Tesserocr you can load the model in the beginning or your program, and run the model seperately (for example in loops to process videos). With pytesseract, each time you call `image_to_string` function, it loads the model and process the image, therefore being slower for video processing.

To install tesserocr I just typed in the terminal `pip install tesserocr`.

To use tesserocr

```
import tesserocr
from PIL import Image
api = tesserocr.PyTessBaseAPI()
pil_image = Image.open('sample.jpg')
api.SetImage(pil_image)
text = api.GetUTF8Text()
```

To install pytesseract : `pip install pytesseract`.

To run it :

```
import pytesseract
import cv2
image = cv2.imread('sample.jpg')
text = pytesseract.image_to_string(image)
```

share improve this answer

answered May 31 '19 at 0:25

H Houssam ASSANY

33 ● 4

By using our site, you acknowledge that you have read and understand our Cookie Policy, Privacy Policy, and

pytesseract 和 tesserocr 的对比.

Part I:

web service → django framework. 搭建

① 安装 Django

pip3 install django

安装完毕之后可以使用

pip3 -m django --version

检查版本 (3.0.4)

② 使用 pyCharm 创建 Django 项目

③ Django 项目目录结构

- __init__.py: 空文件. 声明所在目录为一个 python 包
- setting.py: 配置信息.
- urls.py: 声明请求url的映射关系.
- wsgi.py: 程序和 web 服务器的通信协议
- manage.py: 命令行工具, 用来和 Django 项目交互

settings.py 存放默认配置

④ 测试 server

python3 manage.py runserver

打开 127.0.0.1:8000

⑤ App

An app is a web application that does something - e.g. a weblog system, a database of public records.

在项目目录中执行

python3 manage.py startapp ocr_letters

⑥ App 的目录结构

- admin：对应应用后台管理配制文件
- apps：应用配置文件
- models：数据模块
- tests：编写测试脚本
- views：视图层，直接和浏览器进行交互

在 settings.py 中的 INSTALLED_APPS 里注册新建的 ocr_letters app.

⑦ 设计 APP：

需求是用户上传一张 png/jpg 图片

如何在 Django 中创建一个用户可以上传图片的视图？

(1) 在 models.py 中创建数据模型.

(2) 在 settings.py 中添加上传信息

(3) 数据库迁移

```
python3 manage.py makemigrations
python3 manage.py migrate
```

(4) 在 urls.py 添加 url 配置

(5) 创建 forms.py 添加表单信息

(6) 创建视图函数

(7) 创建 HTML 模板

遇到的 bug：

① No Reverse Match

Django cannot find a matching url pattern for the url you've provided in any of your installed app's urls.

错误分析：网上的教程使用的App名字和自己的项目不一样导致

图片上传存储功能已实现

Part 2：tesserocr 实现文字识别

① MACOS 安装 tesserocr

brew install tesseract

brew install tesseract-lang

② Terminal 测试 tesserocr

tesseract 1.jpg result -l eng

缺陷：像素低的模糊图片无法识别

Module Not Found Error

运行Django之后出现的问题

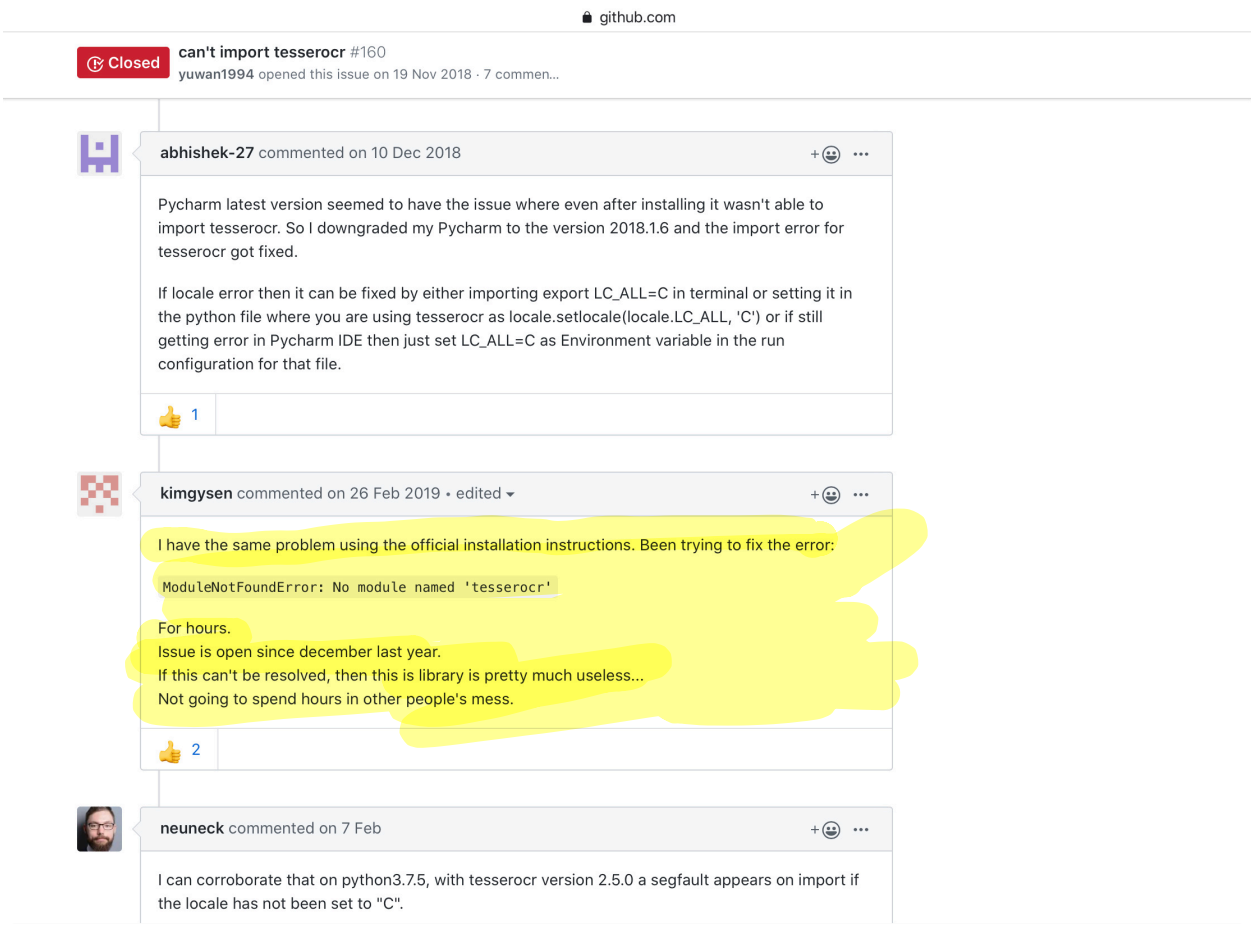
这个error debug了很久....

在pycharm证实这个库是被导入了的，但是不知道为什么库是找不到的....

在terminal运行也是有结果的....

上网查找：

发现有类似错误，如下图



解决方法：换用 pytesseract 库

sudo pip3 install pytesseract

sudo pip3 install Pillow

Pillow 库用来处理图片.

问题解决. pytesseract 的结果和在 terminal 中运行的结果一样

③ 获取结果后进行字符串处理，
移除空行，空格

④ 返回 Json 结果

Part 3：存储识别结果到 MySQL 数据库

时区不对

① 安装 pymysql

```
sudo pip3 install pymysql
```

② 建立 data model 已便存入数据库

test _init_.py

1364 Integrity Error.

Autofield.

pymysql → 数据库连接

brew install mysql-client

mysqlclient

pip install mysql-client

Django

create database