Web Crawler Practice

reCAPTCHA

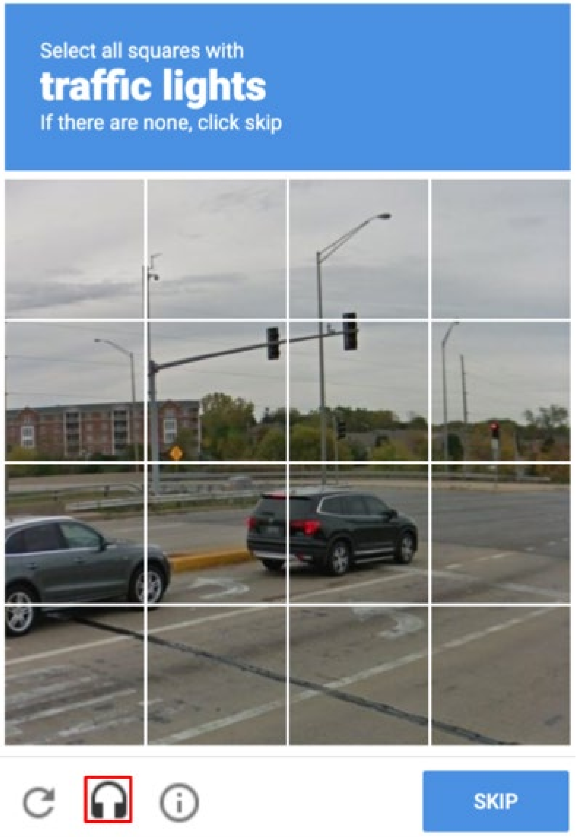
Dr. Chun-Hsiang Chan

Department of Geography,
National Taiwan Normal University

Outline

- reCAPTCHA
- Apply reCAPTCHA API
- Pytesseract

Have you ever seen this?

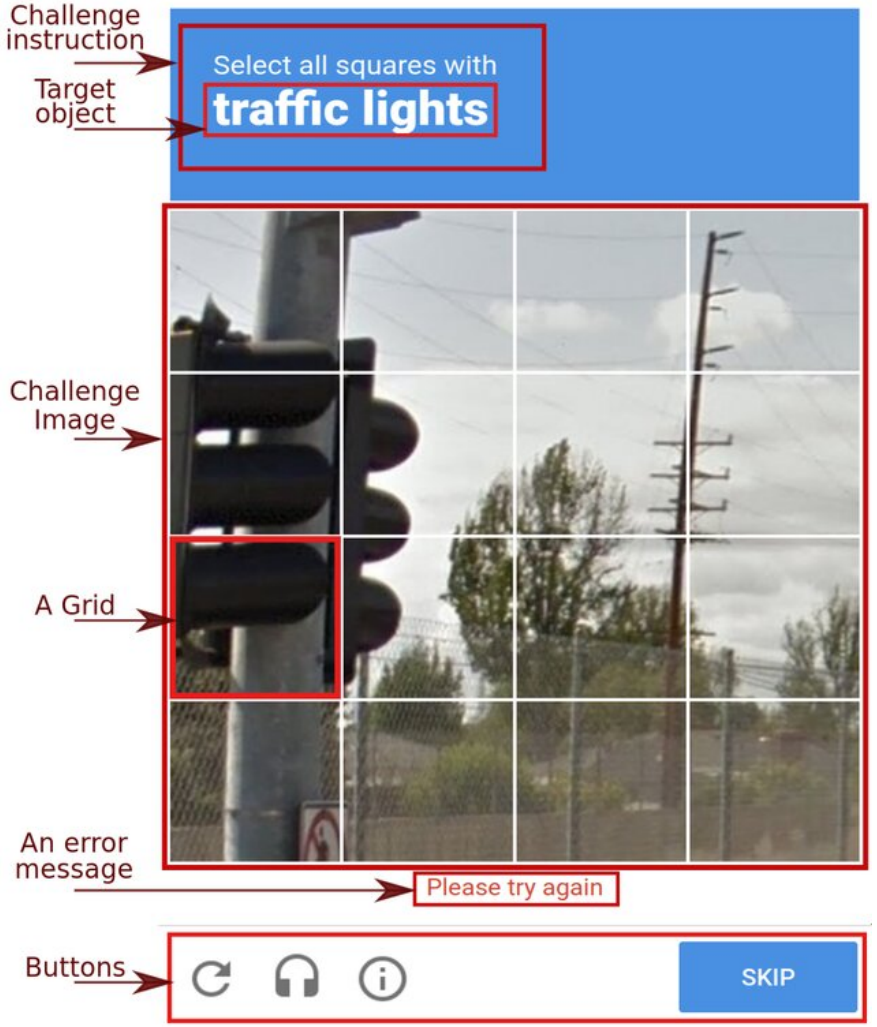


Select all squares with

traffic lights

If there are none, click skip

or



Select all squares with

traffic lights

Please try again

Source: An Object Detection based Solver for Google's Image reCAPTCHA v2

reCAPTCHA

reCAPTCHA protects your website from fraud and abuse without creating friction.

reCAPTCHA uses an advanced risk analysis engine and adaptive challenges to keep malicious software from engaging in abusive activities on your website. Meanwhile, legitimate users will be able to login, make purchases, view pages, or create accounts and fake users will be blocked.

Source: https://www.google.com/recaptcha/about/

The reCAPTCHA Advantage

- **Proven:** reCAPTCHA has been at the forefront of bot mitigation for over a decade and actively protects data for our network of five million sites.
- **Frictionless:** A seamless fraud detection service that stops bots and other automated attacks while approving valid users.
- **Adaptive:** reCAPTCHA's risk-based bot algorithms apply continuous machine learning that factors in every customer and bot interaction to overcome the binary heuristic logic of traditional challenge-based bot detection technologies.

Source: https://www.google.com/recaptcha/about/

Use Cases

Scarping
Content pilfering for ad revenue diversion or competitive use

Fraudulent Transactions
Purchase of goods or gift cards with stolen credit cards

Account Takeovers (ATO)
Credential stuffing to validate stolen accounts

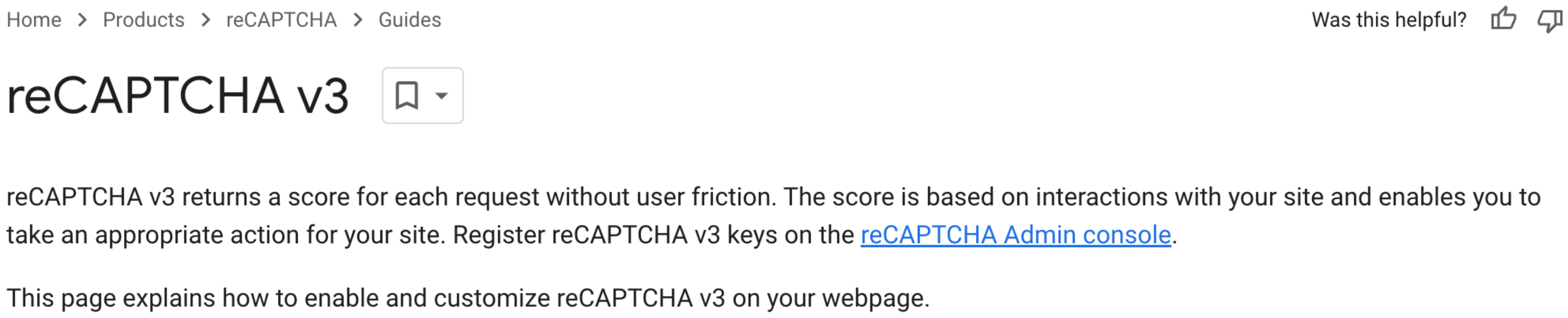
Synthetic Accounts
Creation of new accounts for promotion value or future misuse

False Posts
Posting of malicious links or misinformation propagation

Money Laundering
Bot generated ad click revenue on fraudulent websites

Apply reCAPTCHA API

- Apply here: https://developers.google.com/recaptcha/docs/v3

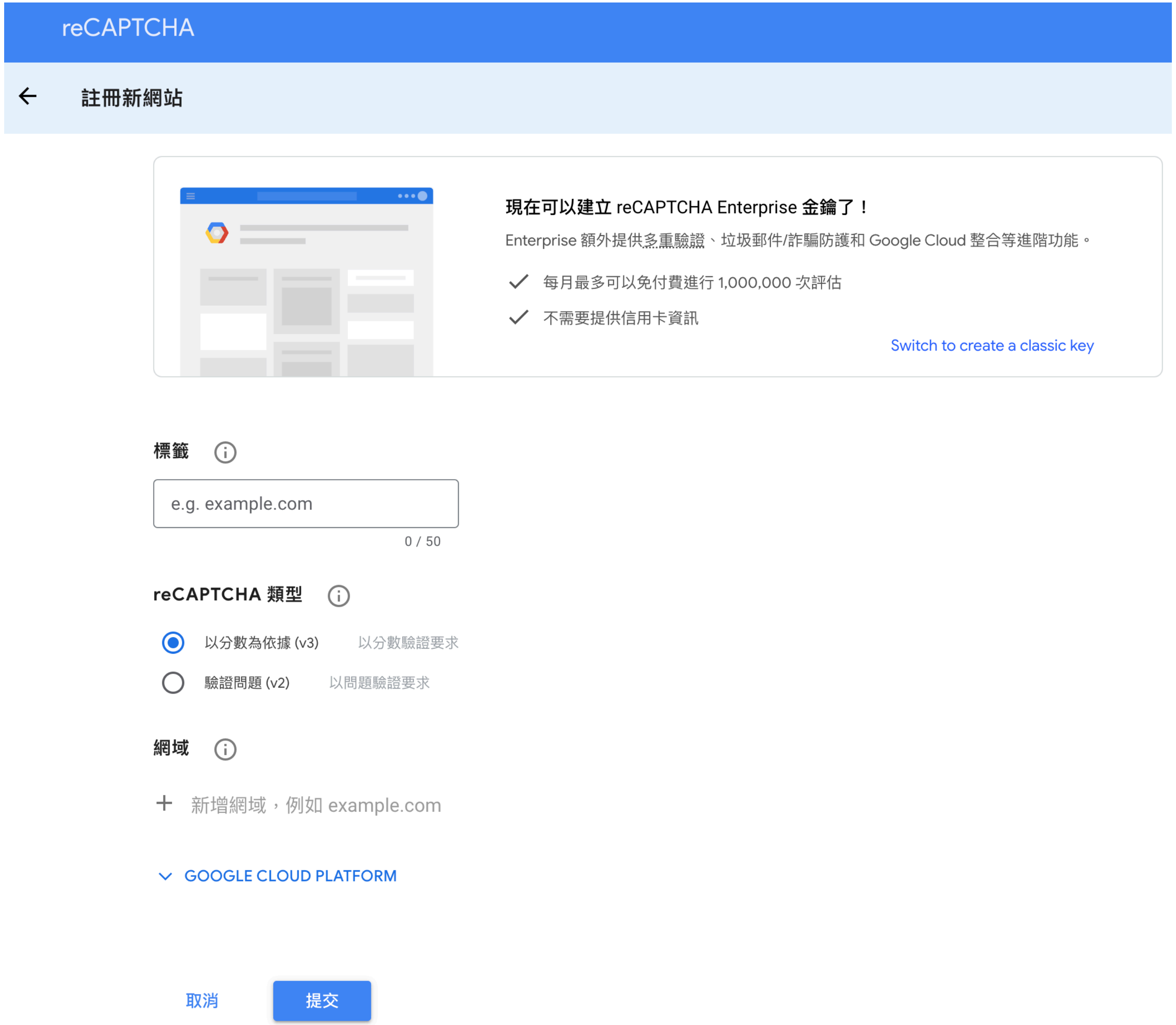


reCAPTCHA v3

reCAPTCHA v3 returns a score for each request without user friction. The score is based on interactions with your site and enables you to take an appropriate action for your site. Register reCAPTCHA v3 keys on the reCAPTCHA Admin console.

This page explains how to enable and customize reCAPTCHA v3 on your webpage.

Apply reCAPTCHA API



reCAPTCHA

← 註冊新網站

現在可以建立 reCAPTCHA Enterprise 金鑰了！

Enterprise 額外提供多重驗證、垃圾郵件/詐騙防護和 Google Cloud 整合等進階功能。

✓ 每月最多可以免付費進行 1,000,000 次評估

✓ 不需要提供信用卡資訊

Switch to create a classic key

標籤 ⓘ

e.g. example.com

0 / 50

reCAPTCHA 類型 ⓘ

● 以分數為依據 (v3)　以分數驗證要求

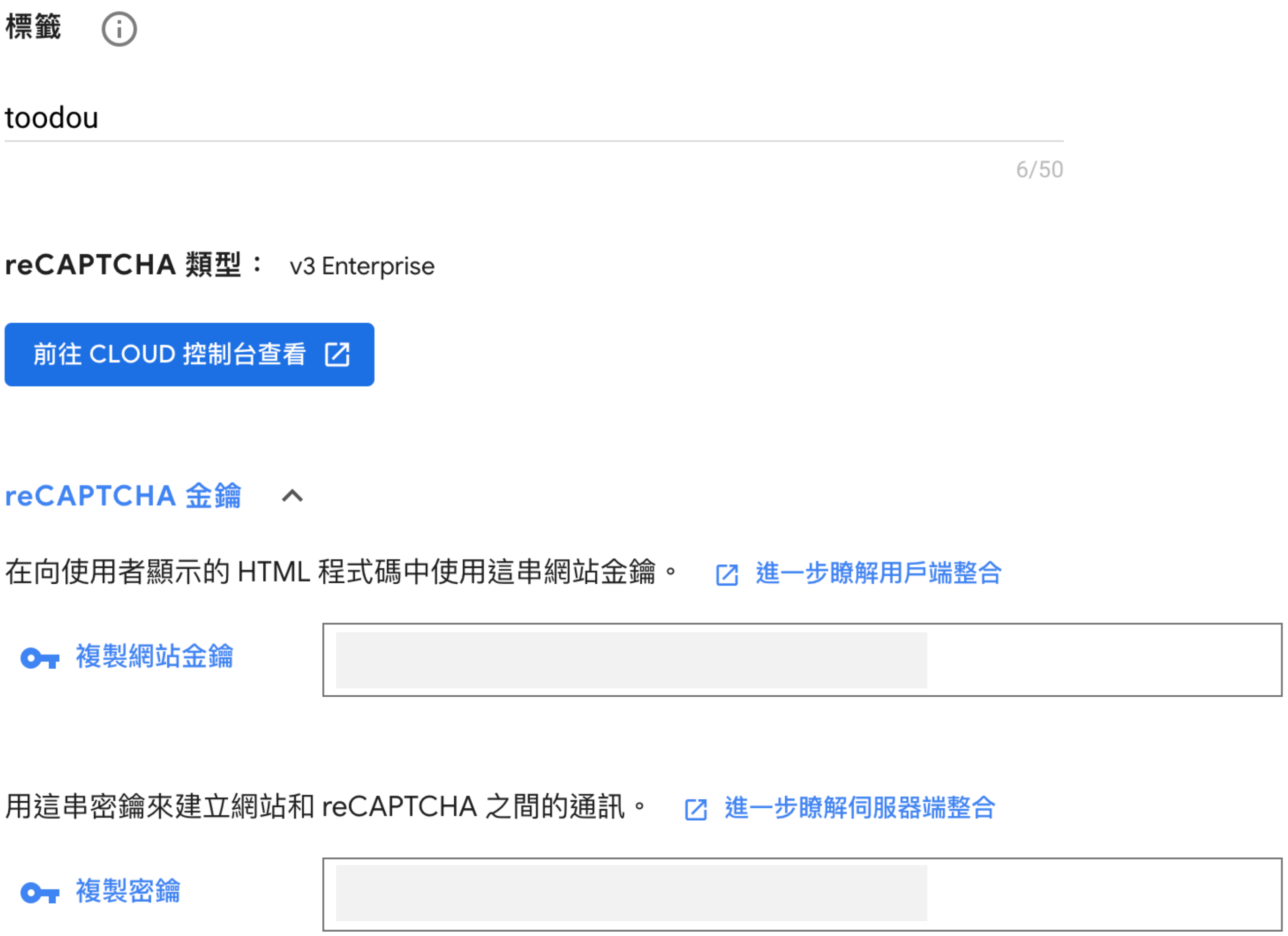
○ 驗證問題 (v2)　以問題驗證要求

網域 ⓘ

＋ 新增網域，例如 example.com

GOOGLE CLOUD PLATFORM

Apply reCAPTCHA API



標籤 ⓘ

toodou

6/50

reCAPTCHA 類型： v3 Enterprise

reCAPTCHA 金鑰

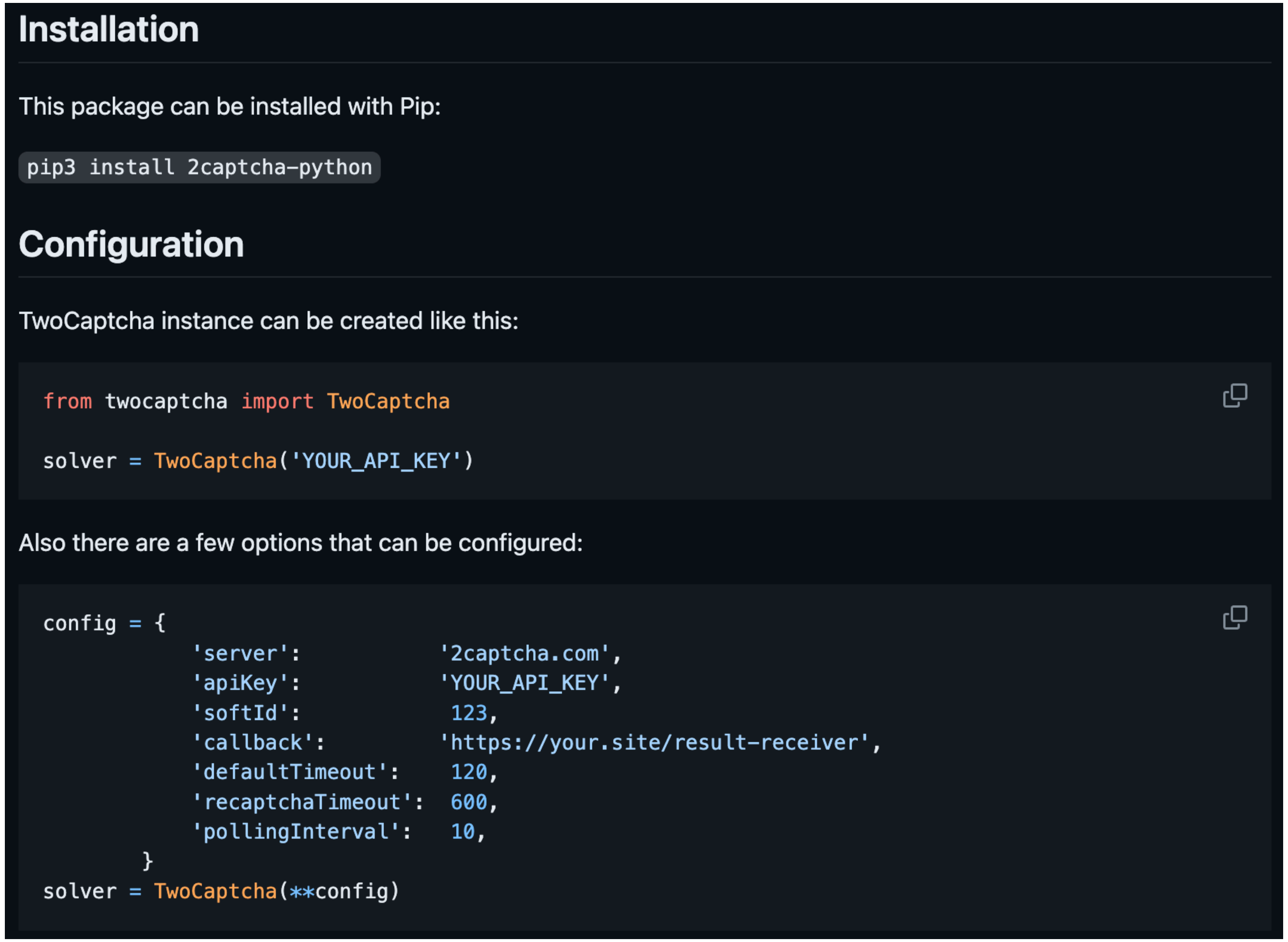
在向使用者顯示的 HTML 程式碼中使用這串網站金鑰。 進一步瞭解用戶端整合

用這串密鑰來建立網站和 reCAPTCHA 之間的通訊。 進一步瞭解伺服器端整合

reCAPTCHA Packages

Please refer to this GitHub:

https://github.com/2captcha/2captcha-python/tree/master?tab=readme-ov-file#recaptcha-v2



Installation

This package can be installed with Pip:

```
pip3 install 2captcha-python
```

Configuration

TwoCaptcha instance can be created like this:

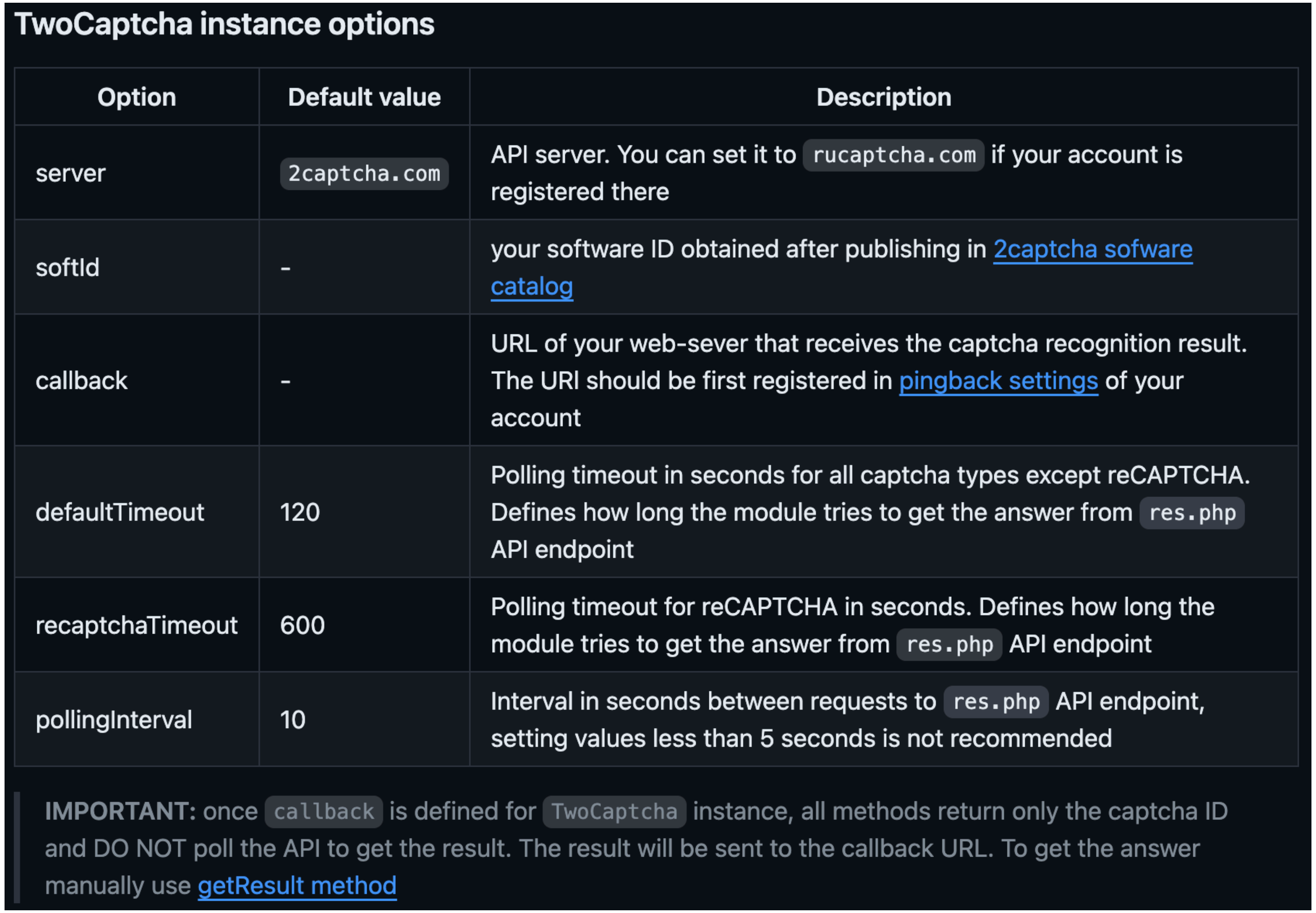
```
from twocaptcha import TwoCaptcha

solver = TwoCaptcha('YOUR_API_KEY')
```

Also there are a few options that can be configured:

```
config = {
            'server':           '2captcha.com',
            'apiKey':           'YOUR_API_KEY',
            'softId':            123,
            'callback':         'https://your.site/result-receiver',
            'defaultTimeout':    120,
            'recaptchaTimeout':  600,
            'pollingInterval':   10,
        }
solver = TwoCaptcha(**config)
```

reCAPTCHA Packages

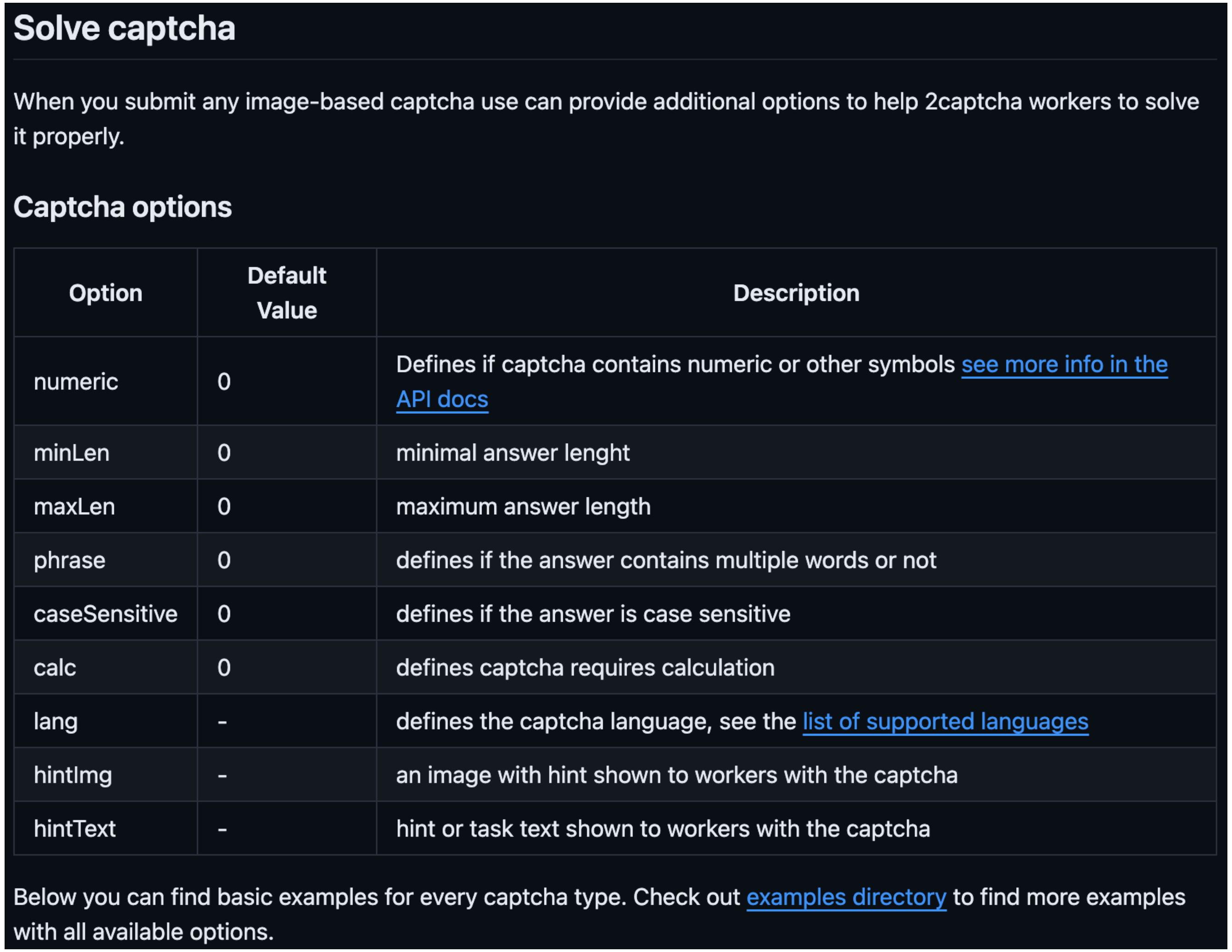


TwoCaptcha instance options

Option	Default value	Description
server	`2captcha.com`	API server. You can set it to `rucaptcha.com` if your account is registered there
softId	-	your software ID obtained after publishing in 2captcha sofware catalog
callback	-	URL of your web-sever that receives the captcha recognition result. The URl should be first registered in pingback settings of your account
defaultTimeout	120	Polling timeout in seconds for all captcha types except reCAPTCHA. Defines how long the module tries to get the answer from `res.php` API endpoint
recaptchaTimeout	600	Polling timeout for reCAPTCHA in seconds. Defines how long the module tries to get the answer from `res.php` API endpoint
pollingInterval	10	Interval in seconds between requests to `res.php` API endpoint, setting values less than 5 seconds is not recommended

> **IMPORTANT:** once `callback` is defined for `TwoCaptcha` instance, all methods return only the captcha ID and DO NOT poll the API to get the result. The result will be sent to the callback URL. To get the answer manually use getResult method

reCAPTCHA Packages



Solve captcha

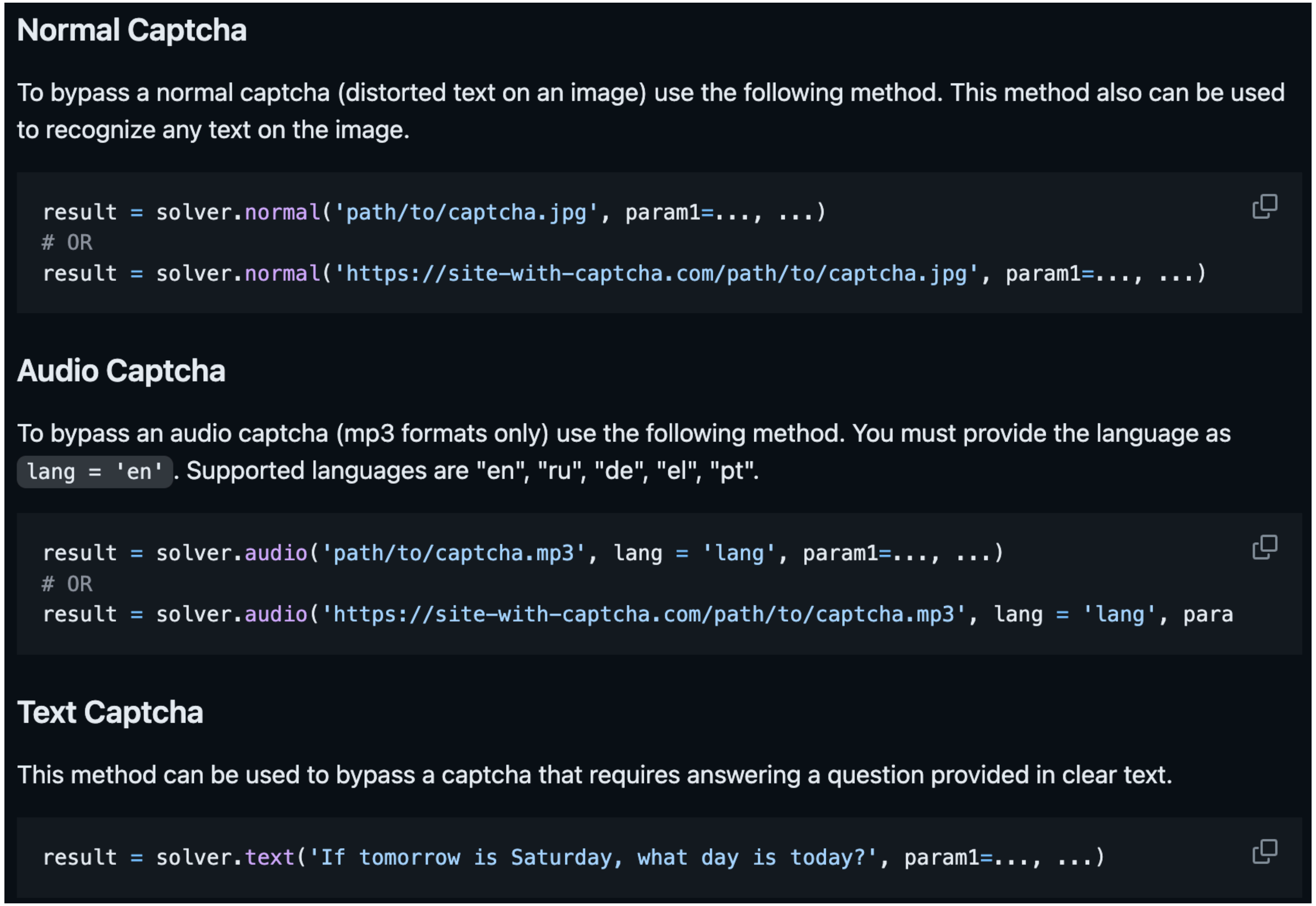
When you submit any image-based captcha use can provide additional options to help 2captcha workers to solve it properly.

Captcha options

Option	Default Value	Description
numeric	0	Defines if captcha contains numeric or other symbols see more info in the API docs
minLen	0	minimal answer lenght
maxLen	0	maximum answer length
phrase	0	defines if the answer contains multiple words or not
caseSensitive	0	defines if the answer is case sensitive
calc	0	defines captcha requires calculation
lang	-	defines the captcha language, see the list of supported languages
hintImg	-	an image with hint shown to workers with the captcha
hintText	-	hint or task text shown to workers with the captcha

Below you can find basic examples for every captcha type. Check out examples directory to find more examples with all available options.

reCAPTCHA Packages



Normal Captcha

To bypass a normal captcha (distorted text on an image) use the following method. This method also can be used to recognize any text on the image.

```
result = solver.normal('path/to/captcha.jpg', param1=..., ...)
# OR
result = solver.normal('https://site-with-captcha.com/path/to/captcha.jpg', param1=..., ...)
```

Audio Captcha

To bypass an audio captcha (mp3 formats only) use the following method. You must provide the language as `lang = 'en'`. Supported languages are "en", "ru", "de", "el", "pt".

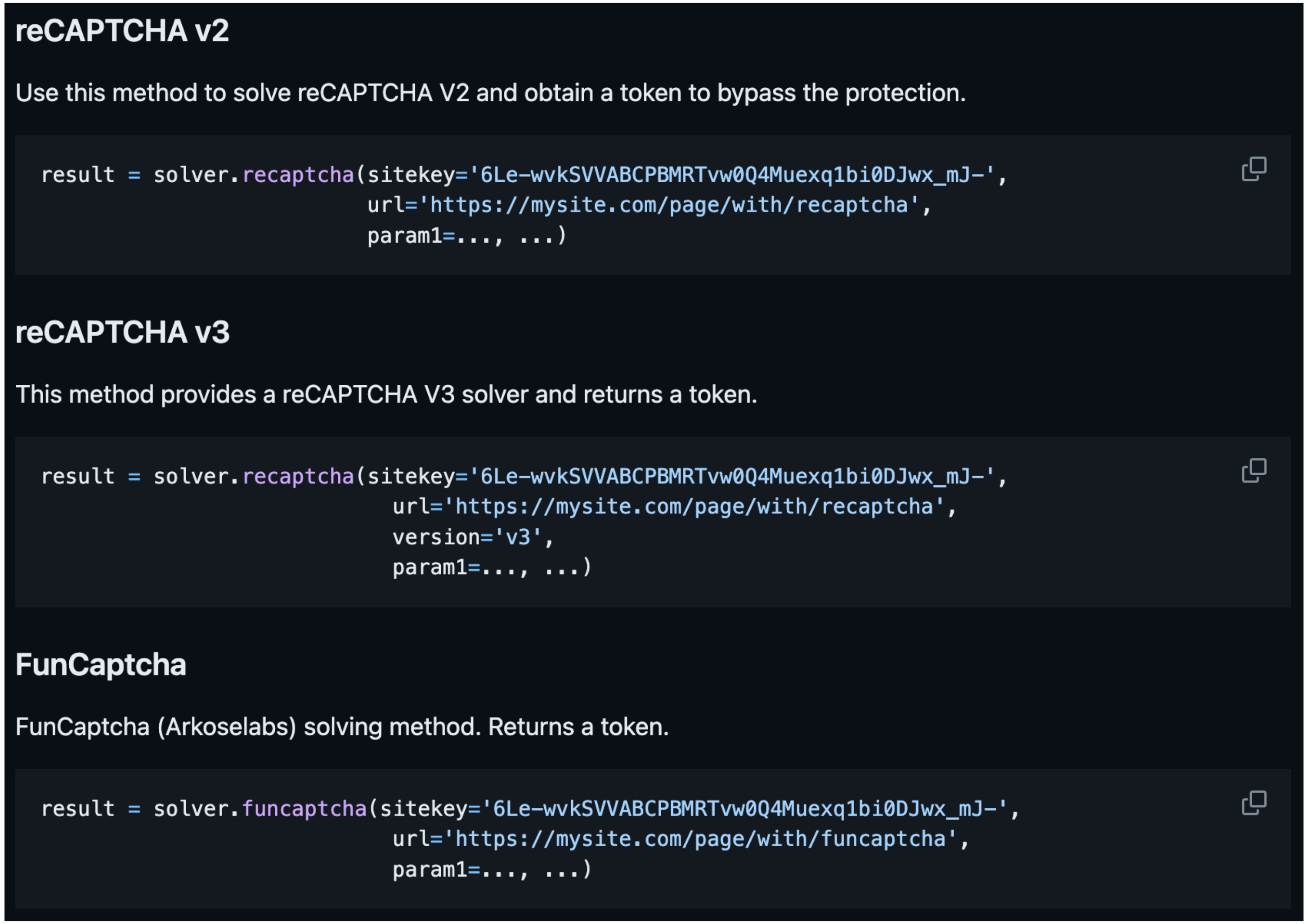
```
result = solver.audio('path/to/captcha.mp3', lang = 'lang', param1=..., ...)
# OR
result = solver.audio('https://site-with-captcha.com/path/to/captcha.mp3', lang = 'lang', para
```

Text Captcha

This method can be used to bypass a captcha that requires answering a question provided in clear text.

```
result = solver.text('If tomorrow is Saturday, what day is today?', param1=..., ...)
```

reCAPTCHA Packages



reCAPTCHA v2

Use this method to solve reCAPTCHA V2 and obtain a token to bypass the protection.

```
result = solver.recaptcha(sitekey='6Le-wvkSVVABCPBMRTvw0Q4Muexq1bi0DJwx_mJ-',
                          url='https://mysite.com/page/with/recaptcha',
                          param1=..., ...)
```

reCAPTCHA v3

This method provides a reCAPTCHA V3 solver and returns a token.

```
result = solver.recaptcha(sitekey='6Le-wvkSVVABCPBMRTvw0Q4Muexq1bi0DJwx_mJ-',
                            url='https://mysite.com/page/with/recaptcha',
                            version='v3',
                            param1=..., ...)
```

FunCaptcha

FunCaptcha (Arkoselabs) solving method. Returns a token.

```
result = solver.funcaptcha(sitekey='6Le-wvkSVVABCPBMRTvw0Q4Muexq1bi0DJwx_mJ-',
                            url='https://mysite.com/page/with/funcaptcha',
                            param1=..., ...)
```

A Simple One – Pytesseract

- Tesseract is an open-source text recognition engine developed by Google, capable of converting text from images into editable text.
- Pytesseract is often used to implement text recognition functionality in Python applications, allowing developers to easily extract text from images for further processing or analysis.

Install Packages

```
>>> brew install imagemagick
>>> brew install tesseract --all-languages
>>> pip install pytesseract
```

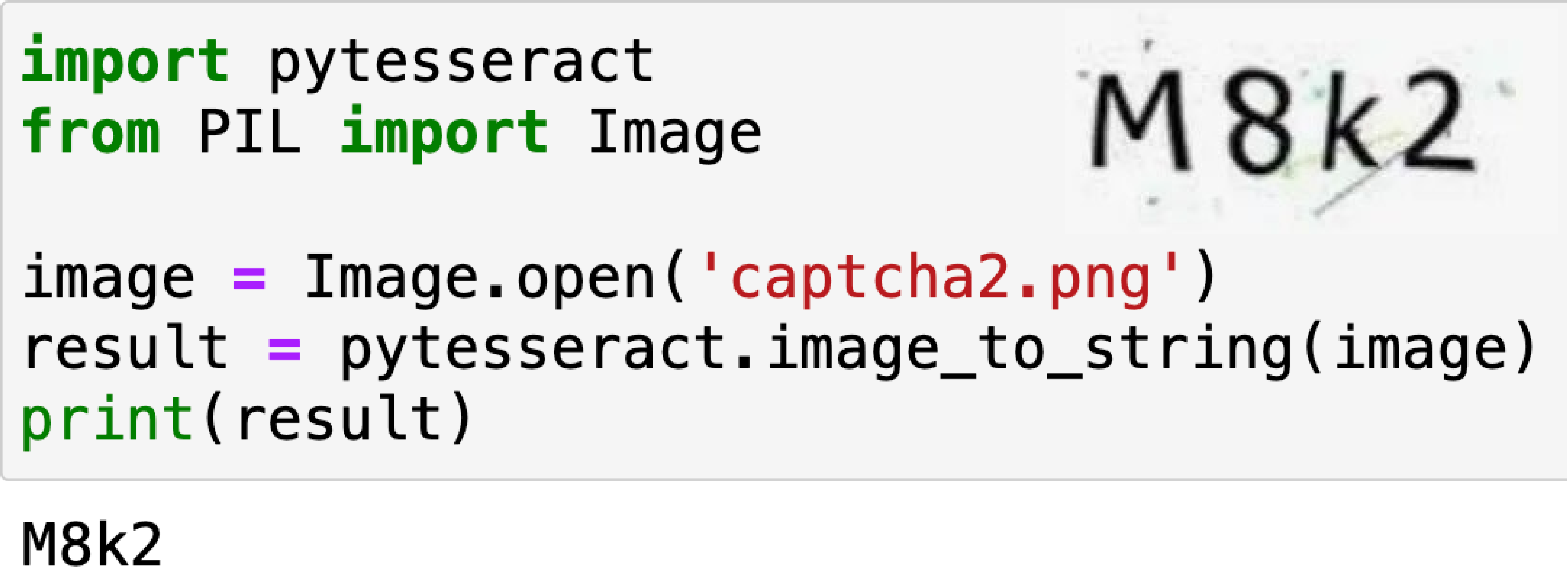
```python
import pytesseract
from PIL import Image

image = Image.open('captcha2.png')
result = pytesseract.image_to_string(image)
print(result)
```

V 9 A m
j F 5 4 s x

Source: https://developer.huawei.com/consumer/cn/forum/topic/0201129060817890054

Example



```python
import pytesseract
from PIL import Image

image = Image.open('captcha2.png')
result = pytesseract.image_to_string(image)
print(result)
```

```
M8k2
```

The End

Thank you for your attention!

Email: chchan@ntnu.edu.tw
Web: toodou.github.io

SCAN ME