Python 데이터 분석과 이미지 처리

나동빈

중급 Captcha Hacking 3 – 데이터 정제

데이터 정제하기: 색상별로 이미지 추출 (utils.py)

```
BLUE = 0
GREEN = 1
RED = 2

# 특정한 색상의 모든 단어가 포함된 이미지를 추출합니다.
def get_chars(image, color):
    other_1 = (color + 1) % 3
    other_2 = (color + 2) % 3

    c = image[:, :, other_1] == 255
    image[c] = [0, 0, 0]
    c = image[:, :, other_2] == 255
    image[c] = [0, 0, 0]
    c = image[:, :, color] < 170
    image[c] = [0, 0, 0]
    c = image[:, :, color] != 0
    image[c] = [255, 255, 255]

    return image
```

중급 Captcha Hacking 3 – 데이터 정제

데이터 정제하기: 테스트 (test.py)

```
import cv2
import utils

image = cv2.imread('image_5.png', cv2.IMREAD_COLOR)
blue = utils.get_chars(image.copy(), utils.BLUE)
green = utils.get_chars(image.copy(), utils.GREEN)
red = utils.get_chars(image.copy(), utils.RED)
cv2.imshow('Image Gray', blue)
cv2.waitKey(0)
cv2.imshow('Image Gray', green)
cv2.waitKey(0)
cv2.imshow('Image Gray', red)
cv2.waitKey(0)
```

중급 Captcha Hacking 3 – 데이터 정제

트레이닝 데이터 만들기: 전체 이미지에서 왼쪽부터 단어별로 추출(utils.py)

```
import cv2

# 전체 이미지에서 왼쪽부터 단어별로 이미지를 추출합니다.
def extract_chars(image):
    chars = []
    colors = [BLUE, GREEN, RED]
    for color in colors:
        image_from_one_color = get_chars(image.copy(), color)
        image_gray = cv2.cvtColor(image_from_one_color, cv2.COLOR_BGR2GRAY)
        ret, thresh = cv2.threshold(image_gray, 127, 255, 0)
        # RETR_EXTERNAL 옵션으로 숫자의 외각을 기준으로 분리
        contours, _ = cv2.findContours(thresh, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)

        for contour in contours:
            # 추출된 이미지 크기가 50 이상인 경우만 실제 문자 데이터인 것으로 파악
            area = cv2.contourArea(contour)
            if area > 50:
                x, y, width, height = cv2.boundingRect(contour)
                roi = image_gray[y:y + height, x:x + width]
                chars.append((x, roi))

    chars = sorted(chars, key=lambda char: char[0])
    return chars
```

중급 Captcha Hacking 3 – 데이터 정제

트레이닝 데이터 만들기: 이미지를 (20 x 20) 크기로 통일 (utils.py)

```
import numpy as np

# 특정한 이미지를 (20 x 20) 크기로 Scaling 합니다.
def resize20(image):
    resized = cv2.resize(image, (20, 20))
    return resized.reshape(-1, 400).astype(np.float32)
```

중급 Captcha Hacking 3 – 데이터 정제

트레이닝 데이터 만들기: 데이터 만들기 (make_train_data.py)

```
import os
import cv2
import utils

# training_data 폴더 생성 및 그 내부에 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 폴더 생성
image = cv2.imread("image_5.png")
chars = utils.extract_chars(image)

for char in chars:
    cv2.imshow('Image', char[1])
    input = cv2.waitKey(0)
    resized = cv2.resize(char[1], (20, 20))

    if input >= 48 and input <= 57:
        name = str(input - 48)
        file_count = len(next(os.walk('./training_data/' + name + '/'))[2])
        cv2.imwrite('./training_data/' + str(input - 48) + '/' +
                    str(file_count + 1) + '.png', resized)
    elif input == ord('a') or input == ord('b') or input == ord('c'):
        name = str(input - ord('a') + 10)
        file_count = len(next(os.walk('./training_data/' + name + '/'))[2])
        cv2.imwrite('./training_data/' + name + '/' +
                    str(file_count + 1) + '.png', resized)
```