Ask Company

크롤링 차근차근 시작하기

이미지 파일 다운로드와 이미지 처리

여러분의 파이썬/장고 페이스메이커가 되겠습니다.

주된 크롤링 대상

HTML 문서 + JSON

이미지

PDF, EXCEL 등 여러 정적인 파일들

이미지

1. 이미지를 다운받기
2. 고화질의 이미지를 받더라도, 경우에 따라 작은 용량으로 줄일 필요가 있음. → 썸네일 처리
3. 여러 개의 파일로 쪼개진 경우 → 하나로 합치기

   ex) 네이버 웹툰

4. 이미지를 다른 포맷으로 변환하기 (jpg, png 등)

파이썬 이미지 라이브러리

~~PIL : Python Image Library~~ (레거시)

Pillow : PIL Fork

PILKit : PIL 유틸리티 컬렉션

Wand : ImageMagick 파이썬 바인딩

Pillow

설치: pip install pillow

PIL 프로젝트의 대체 프로젝트 : PIL과 호환

활용 예 : **이미지 썸네일 생성하기**, 다수 이미지 합성하기, 다른 이미지 포맷으로 변환하기, 회전하기 등

장고에서는 models.ImageField 필드를 쓸 때, Pillow 설치 필수

Python Imaging Library Handbook

http://effbot.org/imagingbook/pil-index.htm

웹에서 자주 쓰이는 이미지 포맷

jpg : 주로 사진을 저장할 때

이미지 품질 옵션 : 0 (저) ~ 100 (고)

대개 60~80 선에서 타협

gif : 움직이는 이미지. 저품질

png : 투명지원되는 이미지 포맷

Case 1) 이미지 다운받기

```
import os
import requests

# image_url = 'https://ee5817f8e2e9a2e34042-3365e7f0719651e5b8d0979bce83c558.ssl.cf5.rackcdn.com/python.png'
image_url = 'https://bit.ly/1KTaQws'

res = requests.get(image_url) # 사이트에 따라 headers 추가 지정
image_data = res.content      # bytes 타입

filename = os.path.basename(image_url)  # URL에서 파일명 획득

with open(filename, 'wb') as f:
    f.write(image_data)
```

Case 2) 이미지 품질 낮추기 / 포맷 변경

이미지 모드 : RGB, RGBA, CMYK

```
LIGHT_YELLOW = (255, 255, 224)   # RGB color

with Image.open('python3.png') as im:
    im.save('python3.jpg', quality=80)   # quality 옵션은 jpg에서만 유효
    im.save('python3_another.png')

    with Image.new('RGBA', im.size, LIGHT_YELLOW) as canvas:
        # alpha채널을 살리며, canvas 베이스에 im를 합성
        canvas_im = Image.alpha_composite(canvas, im)
        canvas_im.save('python3_bg_white.jpg')
```

https://pillow.readthedocs.io/en/3.1.x/reference/Image.html#PIL.Image.Image.save

참고: 이미지 포맷별 최대 지원 크기

jpg

$2^{16}-1$ (65,535) 픽셀

png

$2^{31}-1$ (2,147,483,647) 픽셀 (signed)

http://www.libpng.org/pub/png/spec/iso/index-object.html#11IHDR

Case 3) 이미지 가로/세로 크기 줄이기 (1)

resize(size, resample=0)

리사이징된 "Image 복사본" 생성

원본의 가로/세로 비율 무시, 지정 크기로 강제 리사이징

thumbnail(size, resample=3)

원본 "Image 객체"를 변경

원본의 가로/세로 비율 유지하면서, 지정 크기로 리사이징

이미지는 크기를 줄이거나 늘리거나, 약간의 변경도 모두 손실

Case 3) 이미지 가로/세로 크기 줄이기 (2) - 썸네일

```
from PIL import Image

# image thumbnail
with Image.open('python3.png') as im:
    print('current size : {}'.format(im.size))
    size = (300, 300)
    im.thumbnail(size) # 원본 사이즈 유지하며, 원본 변경
    thumb_im.save('python3_thumb.png')  # png format
```

Case 4) 이미지 이어 붙이기

```
from PIL import Image

WHITE = (255, 255, 255)

with Image.open('img1.jpg') as im1:
    with Image.open('img2.jpg') as im2:
        # 이미지 2개를 세로로 이어서 붙일려고 합니다.
        width = max(im1.width, im2.width)
        height = sum(im1.height, im2.height)
        size = (width, height)

        with Image.new('RGB', size, WHITE) as canvas:
            canvas.paste(im1, box=(0, 0))              # left/top 지정
            canvas.paste(im2, box=(0, im1.height))  # left/top 지정
            canvas.save('canvas.jpg')
```

https://pillow.readthedocs.io/en/latest/reference/Image.html#PIL.Image.Image.paste

연습문제

네이버웹툰 이미지 로컬에 다운받기

```
import os
import requests
from bs4 import BeautifulSoup

ep_url = 'http://comic.naver.com/webtoon/detail.nhn?titleId=20853&no=1164&weekday=tue'
html = requests.get(ep_url).text
soup = BeautifulSoup(html, 'html.parser')

for tag in soup.select('.wt_viewer img'):
    img_url = tag['src']
    img_name = os.path.basename(img_url)
    headers = {'Referer': ep_url}
    img_data = requests.get(img_url, headers=headers).content

    with open(img_name, 'wb') as f:
        f.write(img_data)
```

경고: 다운받은 이미지는 절대 유포하시면 안 됩니다.
→ 저작권 위반

미션

- 한 네이버 웹툰 에피소드 내 이미지를 모두 다운받아, 하나의 이미지로 만들어보세요.
- 미션 풀이는 다음 에피소드에서 ~ :D

인생은 짧습니다.
파이썬/장고를 쓰세요.

여러분의 파이썬/장고 페이스메이커가 되겠습니다.

- Ask Company