**간단한 이미지 크롤링**

1) **필요한 패키지**

from bs4 import BeautifulSoup

: 여러분이 좋아하시는 BeautifulSoup. 파이썬 크롤링할 때 꼭 사용하죠.

from urllib import request, parse

: 여러모로 필요한 함수들이 가득! 코드를 보며 차근차근 살펴봅시다.

from PIL import Image

: Python Image Library라고 합니다. 이미지(.jpg 확장자) 등을 다룰 때 사용합니다. Anaconda3을다운 받았으면 바로 사용이 가능하지만, 없다면 따로 받으시면 됩니다.

import numpy as np

: 이미지도 array 아니겠습니까! 꼭 필요하지요.

사용하는 패키지에 정답은 없기 때문에 편한 것으로 사용하시면 됩니다.

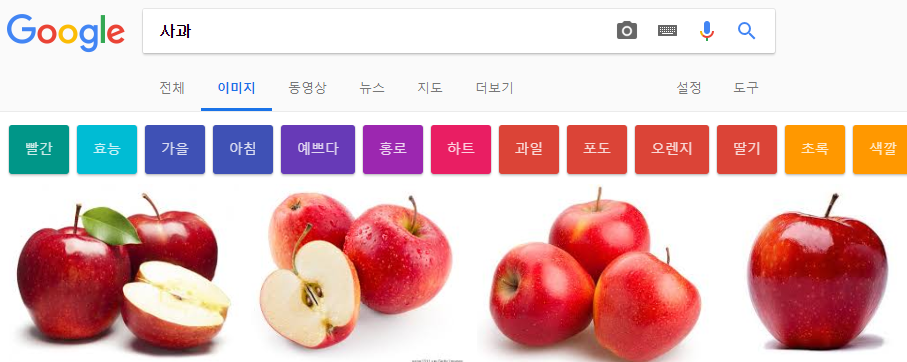
**2) 본격 크롤링1: 여기까지는 똑같다!**

예시로서 구글 이미지 검색을 통해 나오는 이미지를 크롤링해보고자 합니다. <Figure 1>과 같은 상황에서 크롤링하는 것이죠. 이미지 크롤링이라 할 지라도 파싱하는 과정까지는 다른 것과 동일합니다.

크롤링하고자 하는 사이트의 URL은 다음과 같습니다.

https://www.google.co.kr/search?q=%EC%82%AC%EA%B3%BC&newwindow=1&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiW1K-8qarcAhXJaN4KHaEtBi0Q\_AUICigB&biw=1920&bih=974

빨간색 글씨 부분에 있는 ‘q=’ 다음에 문자들은 ‘사과’라는 단어를 URL 인코딩 한 것입니다. 브라우저로 사이트를 켜서 보면 한글로 보이지만 실제로는 인코딩이 되어있습니다. 저 부분을 바꿔주면 검색어가 바뀌니 나중에 크롤링하실 때 참고하시면 되겠습니다.



Figure

word = parse.quote('사과') *# parse.quote(str): URL 인코딩 됨*

url='https://www.google.co.kr/search?q='+word+'&newwindow=1&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiW1K-8qarcAhXJaN4KHaEtBi0Q\_AUICigB&biw=1920&bih=974'

이번에 제가 겪은 일입니다만 크롤링하다 처음으로 ‘HTTPError: Forbidden’가 발생했습니다.

hdr = {'User-Agent': 'Mozillar/5.0', 'referer': 'https://www.google.co.kr'}  
req = request.Request(url, headers=hdr)

다음과 같은 코드로 해결할 수 있다고 합니다. 구글에서 접근한 것처럼 보이게 하는 것이라는데 블로그에서 본 것이라 확실하지는 않습니다. 아무튼 다음 코드와 같이 파싱까지 하면 다른 크롤링과 똑같은 과정이 끝납니다.

html = BeautifulSoup(request.urlopen(req), 'lxml')

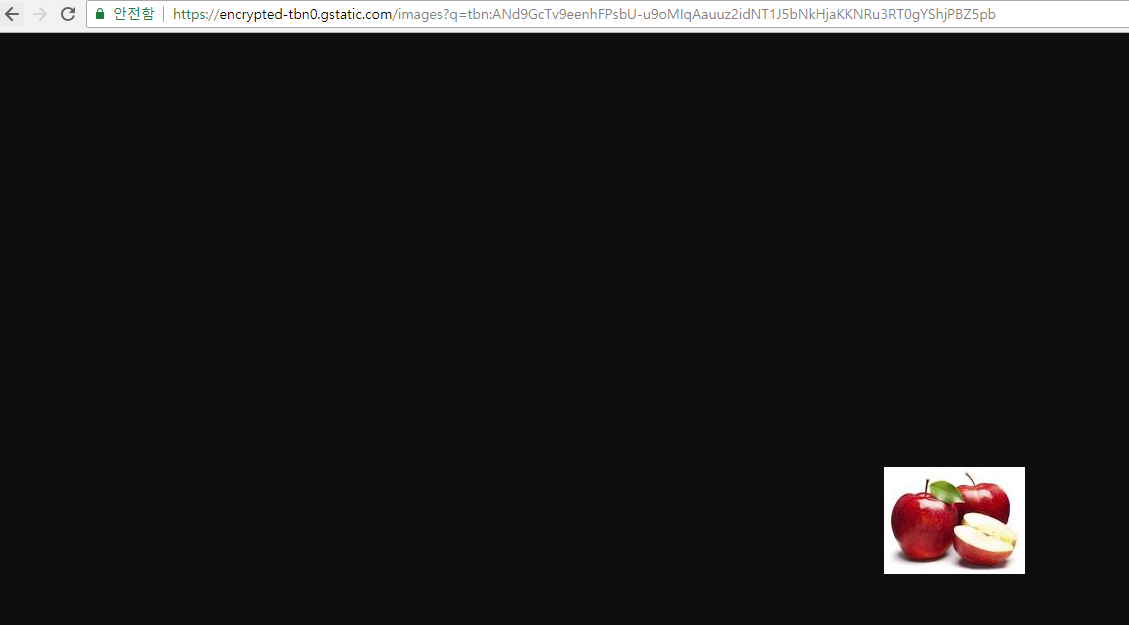
**3) 본격 크롤링2: 이미지는 어디에 있는가.**

HTML에서 이미지 태그는 ‘img’입니다. 일단 한번 살펴 보시죠.

img = html.find\_all('img')  
img[1]

Out: <img alt="사과에 대한 이미지 검색결과" height="107" src="https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTv9eenhFPsbU-u9oMIqAauuz2idNT1J5bNkHjaKKNRu3RT0gYShjPBZ5pb" width="141"/>

find\_all을 통해 ‘img’ 태그를 다 가지고 오면 총 21개에 태그가 가져와집니다. 첫번째 태그는 구글 로고 이미지기 때문에 빼고 생각하시면 됩니다. 아무튼 img[1] 즉, 첫번째 사과 사진 태그를 살펴보면 ‘src=’가 보이실 겁니다. 이 속성이 이미지 파일명입니다. URl 형태로 되어있으니 브라우저 창에 해당 주소를 입력해보면 다음과 같이 나타납니다.



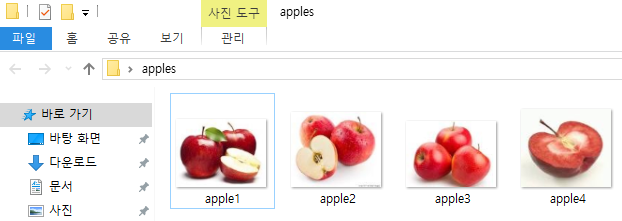
작게 잘라서 그렇지 사과 사진이 화면 가운데에 있을 겁니다. 원하는 사진이면 제대로 태그를 가져온 것입니다.

이제 할 일은 태그들에서 ‘src’ 속성을 가져와서 저장하는 것입니다.

file\_dir = 'C:/Users/user/Desktop/apples/'

count = 1  
for tag in img[1:]:  
 src = tag.get('src') *# .get(‘속성’): 각 태그에서 해당하는 속성을 가져옴  
 # request.urlretrieve(이미지 URL, filename=’저장할 이름’): 이미지 저장함*  
 request.urlretrieve(src, filename=file\_dir+'apple{}.jpg'.format(count))  
 count += 1

for 문으로 끝까지 다 저장해봤습니다. 저장할 이름에 숫자를 붙인 이유는 같은 이름으로 하면 같은 것으로 인식해 하나밖에 남지 않기 때문입니다. 잘 저장이 된 모습입니다.



**4) 이미지를 array로!**

이미지를 파이썬에서 사용하려면 array로 바꿔야하는데, 두가지 상황이 발생할 수 있습니다.

1. 내 컴퓨터에 저장되어있는 이미지를 가져 와 사용.

2. 이미지 URL을 바로 사용.

첫번째 상황은 다음과 같이 하면 간단합니다.

apple = Image.open(file\_dir+'apple1.jpg') *# Image.open(‘이미지 이름’)*img\_array = np.array(apple)  
img\_array.shape

Out: (107, 141, 3)

shape을 살펴봤을 때 3차원이고 마지막이 3이라면 잘 불러와 진 것입니다.

두번째 상황도 간단합니다.

apple = Image.open(request.urlopen(src))  
img\_array = np.array(apple)  
img\_array.shape

Out: (137, 137, 3)

위에 코드를 쭉 실행하셨다면 같은 결과가 나올 것입니다. urlopen을 해주는 것 이외에는 동일합니다.

코드는 따로 파이썬 파일로 첨부하였으니 한번 실행해보시길 바랍니다. 감사합니다!