분석용 데이터
획득 전략

웹 이미지 데이터 확보 기법

학습 목표

학습 목표

- 이미지 데이터의 원본 URL 정보를 수집할 수 있다.
- 인증이 필요 없는 서버에서 이미지 정보와 인증이 필요한 서버에서 이미지 정보를 수집할 수 있다.

학습 내용

- 다양한 이미지 데이터 수집하기 - 이론
- 다양한 이미지 데이터 수집하기 - 실습

이미지 데이터 수집을 위해 알아야 하는 내용

1) 작업 순서

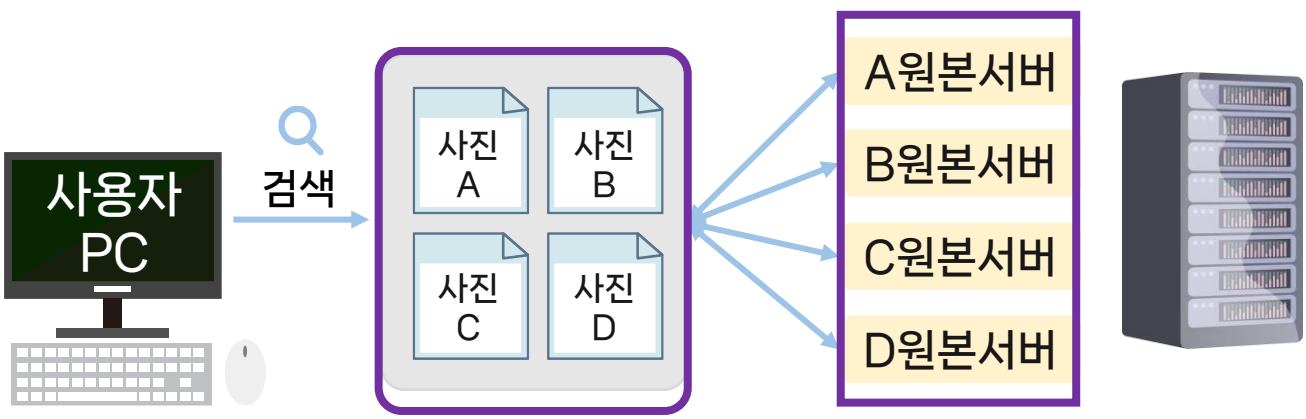
01 검색창에 키워드 입력 후 검색

02 이미지 클릭

03 이미지 URL 추출

04 이미지 다운로드 후 저장

2) 사진의 원본 URL 주소 추출 → 사진 다운로드



이미지 데이터 수집을 위해 알아야 하는 내용

3) 원본 이미지의 URL 주소 필요

이미지 URL 주소는 img 태그의 src 속성값

```
<img src='https://cdn.abc.com/aaa.jpg' alt="고양이" ..... >
```

태그 이름 : img

URL 주소 : https://cdn.abc.com/aaa.jpg

4) 주의사항

(1) 수집 대상 여부 확인

수집 대상 URL 맞음

```
https://search.pstatic.net/common/?src=http%..._8584
.jpg&type=ofullfill340_600_png
```

수집 대상 아님

```

```

주의사항

수집 대상에 맞는 URL을 구분하여 이미지 데이터 수집 코드를 작성

이미지 데이터 수집을 위해 알아야 하는 내용

(2) 이미지 URL에 한글이 들어가면 오류 발생

🔍 오류가 발생하는 경우

[https://search.pstatic.net/common/fill340_파이썬기초.png]

🔍 정상 수집되는 경우

[https://search.pstatic.net/common/fill340_python100.png]

주의사항

주소에 한글이 포함되어 있을 경우, 변환작업이 되도록 코드를 작성

(3) 웹 브라우저 정보

🔍 웹 브라우저 정보가 필요한 서버일 경우

➡ 이미지를 다운로드 받을 때 웹 브라우저 정보가 필요할 수 있음

🔍 웹 브라우저 정보가 필요한 정보를 주지 않을 경우

➡ HTTPError: HTTPErrror 403: Forbidden 에러발생

인증정보가 필요 없는 사이트 정보 수집하기

- 한 페이지에 보이는 이미지 개수

```
# 다양한 이미지 데이터 수집하기
# 필요한 모듈과 라이브러리를 로딩하고 검색어를 입력 받습니다
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.chrome.service import Service
import urllib.request
import urllib
import time
import sys
import math
import os
import random

#사용자에게 필요한 정보들을를 입력 받습니다.
print("=" *80)
print(" 이 크롤러는 이미지 정보를 수집합니다")
print("=" *80)

query_txt = input('1.크롤링할 이미지의 키워드는 무엇입니까?: ')
cnt = int(input('2.크롤링 할 갯수는 몇건입니까?: '))
real_cnt = math.ceil(cnt / 20)
f_dir=input('3.파일이 저장될 경로만 쓰세요(예: c:\\py_temp\\ ) : ')
if f_dir =='' :
        f_dir = "c:\\py_temp\\"
```

인증정보가 필요 없는 사이트 정보 수집하기

- 새 폴더 생성

```
28 #파일을 저장할 폴더를 생성합니다
29 n = time.localtime()
30 s = '%04d-%02d-%02d-%02d-%02d-%02d' % (n.tm_year, n.tm_mon, n.tm_mday, n.tm_
31
32 img_dir = f_dir+s+'-'+query_txt
33 os.makedirs(img_dir)
34 os.chdir(img_dir)
35
```

- 폴더로 이동

```
28 #파일을 저장할 폴더를 생성합니다
29 n = time.localtime()
30 s = '%04d-%02d-%02d-%02d-%02d-%02d' % (n.tm_year, n.tm_mon, n.tm_mday, n.tm_
31
32 img_dir = f_dir+s+'-'+query_txt
33 os.makedirs(img_dir)
34 os.chdir(img_dir)
35
36 #크롬 드라이버를 사용해서 웹 브라우저를 실행합니다.
37 s_time = time.time( )
38
39 s = Service("c:/py_temp/chromedriver.exe")
40 driver = webdriver.Chrome(service=s)
41
42 driver.get('https://www.naver.com')
43 time.sleep(random.randrange(2,5))
```

정리

time.sleep() : 멈추고자 하는 시간(초) 만큼 일시정지

인증정보가 필요 없는 사이트 정보 수집하기

- 2~5초 사이에 랜덤 시간을 뽑아 일시정지

```
#크롬 드라이버를 사용해서 웹 브라우저를 실행합니다.
s_time = time.time( )

s = Service("c:/py_temp/chromedriver.exe")
driver = webdriver.Chrome(service=s)

driver.get('https://www.naver.com')
time.sleep(random.randrange(2,5))

element = driver.find_element(By.ID,"query")
element.send_keys(query_txt)
element.submit()

#이미지 링크를 선택합니다
driver.find_element(By.LINK_TEXT,"이미지").click()

#스크롤 다운 함수 생성 후 화면 이동하기
def scroll_down(driver):
    driver.execute_script("window.scrollTo(0,document.body.scrollHeight);")
    time.sleep(3)
```

- 메뉴나 이모티콘 등

```
for i in img_src :
        img_src1= i.find('img')['src']
        print(img_src1)
        print()
```

```
https://search.pstatic.net/common/?src=http%3A%2F%2Fblogfiles.naver.net%2FMjAyMjExMTlfOSAg%2FMDAxNjY4ODMwMjE1MDk1.PCW99sRAhyOcvirxEu4qZ4r-LUanhM83Dum77Q9gw48g.yFQWWPhFRQncIwCBzPp9it6UHKpLicevwPdVo4tqaKsg.JPEG.milky_way159%2FIMG_9583.JPG&type=a340

https://search.pstatic.net/common/?src=http%3A%2F%2Fblogfiles.naver.net%2FMjAyMjExMTZfOTEg%2FMDAxNjY4NTkzNzc2MjIz.mJdYae-cXI3MCGXQNPLhvE8JkSHV0Kv89GvBk0xv7Ksg.ws6DGLO4Kwplua4fbTjY_Py69szQAn9qOXYdWxrIL-Ag.JPEG.monami5335%2F77DB5D4A-8225-48D7-AF1E-284A351908AE_1_201_a.jpeg&type=a340



https://search.pstatic.net/common/?src=http%3A%2F%2Fblogfiles.naver.net%2FMjAyMjExMjVfNjAg%2FMDAxNjY5MzQzODM4MjY1.1kepJBVBl_B4H9iq49LiTBlMRWJFLXOn2Iu8tsdbbIEg.fIarIpbqVNpBZ0hQWd7JzzTCsfirqrKsCGHfwI7IdK0g.PNG.birdowitchears39%2F20221125_112720.png&type=a340

https://search.pstatic.net/common/?src=http%3A%2F%2Fblogfiles.naver.net%2FMjAyMjEyMDFfMzAw%2FMDAxNjY5ODU5ODc1NjI2.v8DpH133HZHS2rglTT7lL7RlUTELxiMWHcW3QZ6K9IQg.Znye1ZyhwRyinNsBpZnmBzOw2ZrcwaH7lSvM21XxDmMg.PNG.wgeqjripewjr%2F1.PNG&type=a340



https://search.pstatic.net/common/?src=http%3A%2F%2Fblogfiles.naver.net%2FMjAyMjExMjRfMjg4%2FMDAxNjY5Mjc2NTk3MDk4.-tWDTVlJOMo5MuBGtlQcx63C50K8MTEqp7JADVq9R8oq.ArqskvO4RkSij4NxHwwrELRXO2Llobxa63tG25KfiiIg.JPEG.themaltese%2F123_%252813%2529.j
```

```
# 수집 대상 이미지의 URL 정보 추출
for i in img_src :
        img_src1= i.find('img')['src']
```

인증정보가 필요 없는 사이트 정보 수집하기

- 가져올 이미지의 URL 주소

```
# 이미지 다운로드하여 저장하기
for i in range(0,len(img_src2)) :
        try :
                urllib.request.urlretrieve(img_src2[i], str(file_no)+'.jpg')
        except :
                continue

        time.sleep(0.5)
        print("%s 번째 이미지 저장중입니다=======" %file_no)

        file_no += 1

        if file_no > cnt :
            break
```

- 저장할 파일 이름

```
# 이미지 다운로드하여 저장하기
for i in range(0,len(img_src2)) :
        try :
                urllib.request.urlretrieve(img_src2[i], str(file_no)+'.jpg')
        except :
                continue

        time.sleep(0.5)
        print("%s 번째 이미지 저장중입니다=======" %file_no)

        file_no += 1

        if file_no > cnt :
            break
```

인증정보가 필요 없는 사이트 정보 수집하기

- 인코딩 변환

```
# 이미지 이름에 한글이 들어갈 경우 조치방법
for i in range(0,len(img_src2)) :
    file_no += 1
    urllib.request.urlretrieve(urllib.parse.quote(img_src2[i].encode('utf8'), '/:'),\
                               str(file_no)+'.jpg')

    time.sleep(0.5)
    print("%s 번째 이미지 저장중입니다=======" %file_no)

    file_no += 1

    if file_no > cnt :
        break
```

```
# 요약 정보를 출력합니다
e_time = time.time( )
t_time = e_time - s_time

store_cnt = file_no -1

print("=" *70)
print("총 소요시간은 %s 초 입니다 " %round(t_time,1))
print("총 저장 건수는 %s 건 입니다 " %store_cnt)
```

인증정보가 필요한 사이트 정보 수집하기

```
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.chrome.service import Service
import urllib.request
import urllib
import time
import math
import os
import random

#Step 2. 필요한 정보를 입력 받습니다.
print("=" *80)
print(" pixabay 사이트에서 이미지를 검색하여 수집하는 크롤러 입니다 ")
print("=" *80)

query_txt = input('1.크롤링할 이미지의 키워드는 무엇입니까?: ')
cnt = int(input('2.크롤링 할 건수는 몇건입니까?: '))
real_cnt = math.ceil(cnt / 100) # 실제 크롤링 할 페이지 수
f_dir=input('3.파일이 저장될 경로만 쓰세요(예: c:\\py_temp\\ ) : ')
if f_dir =='' :
    f_dir = "c:\\py_temp\\"

print("\n")
print("요청하신 데이터를 수집 중이오니 잠시만 기다려 주세요~~^^")

#Step 3. 파일을 저장할 폴더를 생성합니다
n = time.localtime()
s = '%04d-%02d-%02d-%02d-%02d-%02d' % (n.tm_year, n.tm_mon, n.tm_mday, n.tm_hour, n.tm_min,

_dir+s+'-'+query_txt
(img_dir)
```

주의사항

무료 이미지라도 사용 용도에 따라 저작권 위반이 될 수 있음