[Python] BeautifulSoup을 통한 이미지 블로그 스크래핑하기 — 나무늘보의 개발 블로그

노트북: 첫 번째 노트북
만든 날짜: 2020-10-26 오전 8:28
URL: https://continuous-development.tistory.com/109?category=736681

Python

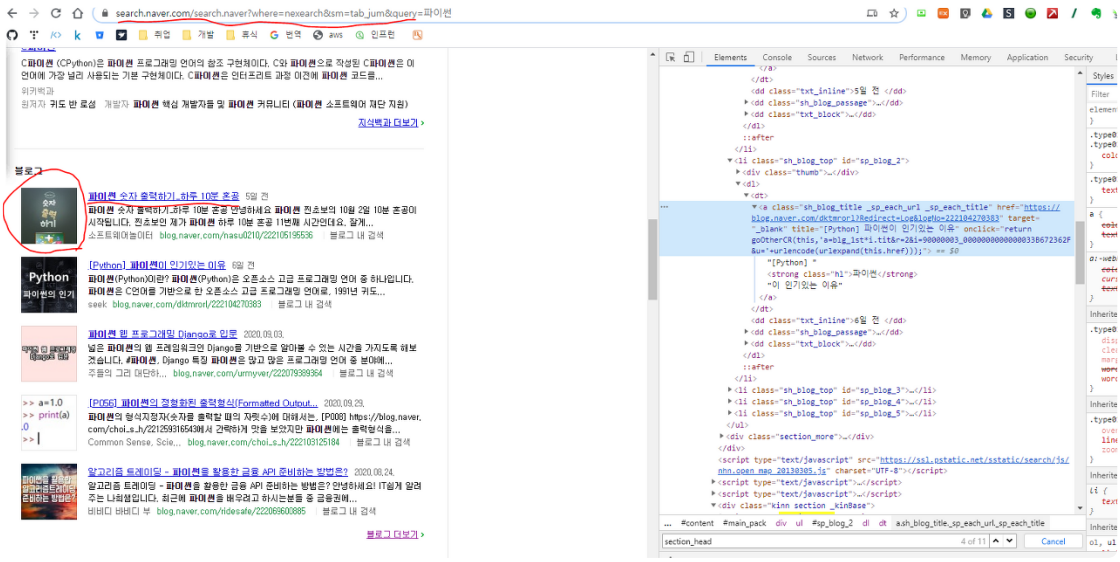
[Python] BeautifulSoup을 통한 이미지 블로그 스크래핑하기

2020. 10. 8. 10:03 수정 삭제 공개

검색 키워드를 이용하여 원하는 제목과 이미지 링크 가져오기

In [3]:
```
from urllib.request import urlopen
from bs4 import BeautifulSoup
from urllib.error import HTTPError
from urllib.error import URLError
import pandas as pd
from urllib.parse import quote_plus
```

기본적인 라이브러리를 넣어주는 부분이다.



위 사이트에서 검색어를 입력해서 블로그의 타이틀과 이미지 값을 받아오는 부분이다.

```
In [16]: base_url = 'https://search.naver.com/search.naver?where=nexearch&sm=tab_jum&query=' # base url 이다. 여기서 뒤에 어떤 검색어를 넣냐에 따라 
         keyword = input('검색어 입력 :') # 여기서 네이버에서 검색하던것 처럼 명령어를 입력해준다.

         try:
             html = urlopen(base_url+quote_plus(keyword)) # quote_plus 를 통해 URL Encoding을 한다.
         except HTTPError as he:
             print('http error')
         except URLError as us:
             print('url error')
         else:
             soup = BeautifulSoup(html.read(),'html.parser')

         검색어 입력 :파이썬
```

해당 검색어를 입력받아서 결과에 대해서 BeautifulSoup 하는 부분이다.

```
In [53]: # sh_blog_title
         titles = soup.find_all('a', {'class' : 'sh_blog_title'})
         titles

Out[53]: [<a class="sh_blog_title _sp_each_url _sp_each_title" href="https://blog.naver.com/nasu0210?Redirect=Log&amp;logNo=222105195536" onclick="r
         eturn goOtherCR(this,'a=blg_1st*i.tit&amp;r=1&amp;i=90000003_0000000000000033B6805410&amp;u='+urlencode(urlexpand(this.href)));" target="_b
         lank" title="파이썬 숫자 출력하기_하루 10분 혼공"><strong class="hl">파이썬</strong> 숫자 출력하기_하루 10분 혼공</a>,
          <a class="sh_blog_title _sp_each_url _sp_each_title" href="https://blog.naver.com/dktmrorl?Redirect=Log&amp;logNo=222104270383" onclick="r
         eturn goOtherCR(this,'a=blg_1st*i.tit&amp;r=2&amp;i=90000003_0000000000000033B672362F&amp;u='+urlencode(urlexpand(this.href)));" target="_b
         lank" title="[Python] 파이썬이 인기있는 이유">[Python] <strong class="hl">파이썬</strong>이 인기있는 이유</a>,
          <a class="sh_blog_title _sp_each_url _sp_each_title" href="https://blog.naver.com/urmyver?Redirect=Log&amp;logNo=222079389364" onclick="re
         turn goOtherCR(this,'a=blg_1st*i.tit&amp;r=3&amp;i=90000003_0000000000000033B4F68EB4&amp;u='+urlencode(urlexpand(this.href)));" target="_bl
         ank" title="파이썬 웹 프로그래밍 Django로 입문"><strong class="hl">파이썬</strong> 웹 프로그래밍 Django로 입문</a>,
          <a class="sh_blog_title _sp_each_url _sp_each_title" href="https://blog.naver.com/choi_s_h?Redirect=Log&amp;logNo=222103125184" onclick="r
         eturn goOtherCR(this,'a=blg_1st*i.tit&amp;r=4&amp;i=90000003_0000000000000033B660BCC0&amp;u='+urlencode(urlexpand(this.href)));" target="_b
         lank" title="[P056] 파이썬의 정형화된 출력형식(Formatted Output of Python)">[P056] <strong class="hl">파이썬</strong>의 정형화된 출력형식(F
         ormatted Output...</a>,
          <a class="sh_blog_title _sp_each_url _sp_each_title" href="https://blog.naver.com/ridesafe?Redirect=Log&amp;logNo=222069600885" onclick="r
         eturn goOtherCR(this,'a=blg_1st*i.tit&amp;r=5&amp;i=90000003_0000000000000033B4613275&amp;u='+urlencode(urlexpand(this.href)));" target="_b
         lank" title="알고리즘 트레이딩 - 파이썬을 활용한 금융 API 준비하는 방법은?">알고리즘 트레이딩 - <strong class="hl">파이썬</strong>을 활용한
         금융 API 준비하는 방법은?</a>]
```

soup에서 find_all 명령어를 통해 해당하는 부분을 전부 가져오는데 a 태그 안에 있는 class의 sh_blog_title을 가져온다.
find_all로 가져올 경우 , 를 기준으로 인덱스로 들어간다.

```
In [64]: title_list =[]
         link_list = []
         for i in titles:
             print(i.attrs['title'])
             print(i.attrs['href'])
             print()
             title_list.append(i.attrs['title'])
             link_list.append(i.attrs['href'])
```

```
파이썬 숫자 출력하기_하루 10분 혼공
https://blog.naver.com/nasu0210?Redirect=Log&logNo=222105195536

[Python] 파이썬이 인기있는 이유
https://blog.naver.com/dktmrorl?Redirect=Log&logNo=222104270383

파이썬 웹 프로그래밍 Django로 입문
https://blog.naver.com/urmyver?Redirect=Log&logNo=222079389364

[P056] 파이썬의 정형화된 출력형식(Formatted Output of Python)
https://blog.naver.com/choi_s_h?Redirect=Log&logNo=222103125184

알고리즘 트레이딩 - 파이썬을 활용한 금융 API 준비하는 방법은?
https://blog.naver.com/ridesafe?Redirect=Log&logNo=222069600885
```

그 뒤 list형태로 된 것을 for 문을 이용해 값을 하나씩 뽑아온다.
또한 attrs를 이용해 해당 태그를 뽑아온다. 그 후 그 값들을 list에 넣어준다.

```
In [65]: import pandas as pd
         python_df = pd.DataFrame({'title':title_list,'link':link_list})
```

```
In [66]: python_df
```

Out[66]:

	title	link
0	파이썬 숫자 출력하기_하루 10분 혼공	https://blog.naver.com/nasu0210?Redirect=Log&l...
1	[Python] 파이썬이 인기있는 이유	https://blog.naver.com/dktmrorl?Redirect=Log&l...
2	파이썬 웹 프로그래밍 Django로 입문	https://blog.naver.com/urmyver?Redirect=Log&lo...
3	[P056] 파이썬의 정형화된 출력형식(Formatted Output of Python)	https://blog.naver.com/choi_s_h?Redirect=Log&l...
4	알고리즘 트레이딩 - 파이썬을 활용한 금융 API 준비하는 방법은?	https://blog.naver.com/ridesafe?Redirect=Log&l...

리스트로 빼낸 값을 pandas의 DataFrame으로 넣어준다.

```
In [67]: python_df.to_csv('.python_df.csv',mode='w',encoding='utf-8',index=False)
         print('success')
```

```
success
```

데이터 프레임으로 만든 값을 csv 로 저장한다. mode w는 저장한다는 뜻이고 encoding 은 저장 방식 , index는 행의 인덱스를 같이 보낼지를 결정하는 건데 False로 안 쓴다고 명시한다.

```
In [68]: df = pd.read_csv('.python_df.csv',encoding='utf-8')
         df
```

Out[68]:

	title	link
0	파이썬 숫자 출력하기_하루 10분 혼공	https://blog.naver.com/nasu0210?Redirect=Log&l...
1	[Python] 파이썬이 인기있는 이유	https://blog.naver.com/dktmrorl?Redirect=Log&l...
2	파이썬 웹 프로그래밍 Django로 입문	https://blog.naver.com/urmyver?Redirect=Log&lo...
3	[P056] 파이썬의 정형화된 출력형식(Formatted Output of Python)	https://blog.naver.com/choi_s_h?Redirect=Log&l...
4	알고리즘 트레이딩 - 파이썬을 활용한 금융 API 준비하는 방법은?	https://blog.naver.com/ridesafe?Redirect=Log&l...

이 값을 다시 가져와서 확인해 본다.

'Python' 카테고리의 다른 글

[Python] python 에서 Seleium을 통한 동적 크롤링 - 1

[Python] python 에서 Seleium을 설치 방법

[Python] BeautifulSoup을 통한 이미지 블로그 스크래핑하기

[Python] BeautifulSoup을 통한 이미지 스크래핑 하기

[python] 영화 리뷰에 대한 자연어 처리분석/ 감성분석하기 feat. 스크래핑

[python] BeautifulSoup를 통한 영화리뷰 scraping 하기

python 스크래핑 검색어를 통한 스크래핑 블로그 스크래핑

나무늘보스

혼자 끄적끄적하는 블로그 입니다.