

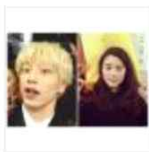
인기검색어 크롤링 과제

저는 아침에 눈을 뜨면 인기 검색어 순위를 먼저 확인해보는 습관이 있습니다. 그런데 이 인기 검색어 순위를 확인해볼 때마다 불편한 점이 있었어요.. 예를 들어, 인기 검색어 순위가 아래와 같이 떠있다 할 때,

2018.07.16.(월)	
14:19:30 기준	
<hr/>	
1	타카하타 미츠키
2	정양
3	황의조
4	서용교
5	gs칼텍스
6	아시안게임
7	의정부고 졸업사진
8	히든싱어 린
9	이강인
10	gs칼텍스 채용

1위에 떠있는 “타카하타 미츠키”라는 것이 왜 인기검색어에 났는지 궁금해져서 클릭을 해보면?

뉴스 관련도순 최신순



[사카구치 켄타로-타카하타 미츠키, '결혼 전제' 열애? "가족에게 소개"](#)

한국일보 | 4분 전 | 네이버뉴스 | [🔗](#)

사카구치 켄타로, 타카하타 미츠키가 결혼 전제 열애 중이라는 보도가 나왔다. 사카구치 켄타로, 타카하타 미츠키 인스타그램 일본 배우 사카구치 켄타로와 타카하타 미츠키가 결혼을 전제로 열애 중이라는 현지...

↳ [타카하타 미츠키, "男 아파트 드나들..."](#) 경남매일신문 | 9분 전

↳ [日 '사귀고 싶은 연예인' 타카하타 미...](#) 뉴스 | 7분 전

[관련뉴스 전체보기 >](#)



[\[록@재팬\] 사카구치 켄타로♥타카하타 미츠키 "결혼 전제 열애" 보도](#)

TV리포트 | 2시간 전 | 네이버뉴스 | [🔗](#)

[TV리포트=박설이 기자] 일본의 유명 배우 사카구치 켄타로와 타카하타 미츠키가 결혼을 전제로 교제 중이라는 보도가 나왔다. 16일 일본 닛칸겐다이 보도에 따르면 두 사람의 열애가 밝혀진 건 1년 반 전으로, 일부...

↳ [사카구치 켄타로-타카하타 미츠키, ...](#) 쿠키뉴스 | 24분 전

↳ [타카하타 미츠키, 사카구치 켄타로와...](#) MBN | 1시간 전 | 네이버뉴스

[관련뉴스 전체보기 >](#)

관심도 없는 열애설이.. 저는 이게 너무 싫습니다. 인기 검색어에 왜 뚝인지 궁금해서 클릭해보면 막상 그렇게 큰 이슈도 아닌.. 그런데 또 궁금은 하니까 1위부터 20위까지 모두 클릭을 다해보고ㅋㅋ 결국 얻는 것은 시간낭비와 허무함..

이번주 과제는 이 시간낭비를 조금이라도 줄여보는 프로그램을 만들어보도록 할게요!

문제 >>

현재의 인기검색어를 뉴스 기사 제목들 (상위 3개만!) 과 함께 출력해주세요.

인기검색어 크롤링 코드는 제공해드립니다. 다음 페이지를 참고해주세요.

(네이버에서는 2020년 1월 17일부로 메인 홈페이지 개편으로 bs4 모듈로는 인기검색어 크롤링이 어려워졌습니다. 따라서 인기검색어 크롤링 부분만 제공해드립니다. 뉴스 검색어 제목 크롤링 부분만 코딩해주시면 됩니다.)

아래는 실행 결과입니다. 확실히 해당 검색어들이 왜 인기검색어에 뚝있는지가 한눈에 보입니다.

1 타카하타 미츠키
관련기사 : 日 사카구치 켄타로, 타카하타 미츠키와 열애설 '동갑내기 커플 탄생?'
관련기사 : [룩@재팬] 사카구치 켄타로♥타카하타 미츠키 "결혼 전제 열애" 보도
관련기사 : 타카하타 미츠키, 아시아 여신 '드디어' 터질 게 터졌나? 청순한데 섹시, 시선 강탈해~
2 김우빈
관련기사 : 김우빈, 오늘(16일) 생일...소속사 측 "아무 일 없었다는 듯 돌아오길"
관련기사 : "아무 일 없었다는 듯 돌아오길"...싸이더스HQ, 몽클한 김우빈 생일 축하
관련기사 : '비인두암' 김우빈, 2년째 공백기에도 여전한 관심
3 비인두암
관련기사 : '비인두암' 김우빈, 2년째 공백기에도 여전한 관심
관련기사 : [Oh!쎬 이슈] 김우빈, 비인두암 투병 1년~30번째 생일..쏟아지는 응원(종합)
관련기사 : [이슈Q] '비인두암' 김우빈, 벌써 공백 2년 째... EXID 솔지·문근영, 투병 후 복귀 스타 '눈길'
4 황의조
관련기사 : 황의조 '시즌 11골'...손흥민과 공존도 가능
관련기사 : 김학범 감독이 밝힌 황의조 와일드 카드 선발-이강인 제외 이유
관련기사 : AG축구 김학범 감독 "황의조, 컨디션 좋아 선발한 것"
5 신민아
관련기사 : 신민아 '디바' 7월 크랭크인
관련기사 : 비인두암 김우빈, 신민아와 근황은? "병원서 목격...잘 보살펴 준다고"
관련기사 : 신민아 근황 `변함없는 미모` 마음도 예쁜 김우빈 여친 "내조에 힘써"
6 아시안게임
관련기사 : 인천 김진야, 지난해 'U-20 월드컵' 탈락 아픔 딛고 아시안게임 출격
관련기사 : 조현우, 2018 아시안게임 와일드카드 발탁
관련기사 : 손흥민 따라다니던 '병역 문제', 아시안게임에서 풀릴까(종합)
7 하지원 나이
관련기사 : [EBS 영화-한국영화특선] 천만 영화 '해운대', 한반도를 덮치는 거대한 쓰나미
관련기사 : 첫방 '갈릴레오' 김병만·하지원·닉쿤·김세경, 韓 최초 화성탐사 '첫미션 시작' [종합]
관련기사 : 이서원, 100여 일이 지나도 여전히 싸늘한 시선...연예계 생활 어떻게 될까?
8 캬테
관련기사 : 캬테 명성 무색 부진, 월드컵 결승 교체된 이유
관련기사 : 쑥스러운 캬테, '나도 트로피 만져도 돼?'
관련기사 : 현영민 "캬테, 러시아 월드컵 최고의 선수" (이범의 시선집중)

[1단계]

네이버 인기검색어를 크롤링해주세요.

(인기검색어 크롤링 코드는 제공해드립니다! 아래 코드를 사용해주시면 됩니다. 복사/붙여넣기 하는 과정에서 제대로 복사가 안될 수 있으므로, 직접 따라서 타이핑 쳐주시는 것을 추천드립니다.)

```
import requests # 터미널 창에 pip install requests 를 입력하여 모듈 설치!
json = requests.get('https://www.naver.com/srchrank?frm=main').json()
ranks = json.get("data")

for i in range(20) :
    print("{} ".format(i+1)+ranks[i]["keyword"])
```

[2단계]

그 다음, 뉴스 검색에 각 인기검색어를 검색했을 때 나오는 뉴스 제목 상위 3개를 수집해주세요.

Hint1


해당 검색어의 뉴스 기사제목들을 보려면, 검색어를 네이버 검색창에다가 검색하고, ‘뉴스’ 탭을 누르면 됩니다.

예를 들어, “손흥민”의 뉴스 검색 결과를 보려면? 아래와 같이 뉴스기사들이 모여 있는 뉴스탭으로 이동해야합니다.



우리는 스크레이핑을 위해 웹사이트의 html 소스코드가 필요합니다.urlopen() 함수에 웹사이트의 url 주소를 전달해주면, 그 반환 값으로 웹사이트의 html 소스코드를 가져오죠?

그럼 현재의 URL 주소를 확인해볼게요.

 https://search.naver.com/search.naver?where=news&sm=tab_jum&query=손흥민

맨 마지막에 ‘query=손흥민’ 이라는 부분에 주목해주세요. ‘손흥민’이라는 검색어가 query 라는 변수에 전달되는 형태입니다. 그럼 내가 원하는 검색어를 저기 ‘query=’ 뒤에다가 붙혀주기만 하면 될까요?

한번 아래와 같이 코딩해볼게요. ‘query=’ 뒤에다가 내가 원하는 검색어를 덧붙힌 후 urlopen() 함수로 웹사이트를 열게 했습니다.

```
1 import urllib.request as req
2 from bs4 import BeautifulSoup
3
4 url = "https://search.naver.com/search.naver?where=news&sm=tab_jum&query="
5 keyword = "치킨"
6 url_result = url + keyword
7 code = req.urlopen(url_result)
```

결과呢?

```

Traceback (most recent call last):
  File "C:/Users/Admin/PycharmProjects/Taling_11_Saturday/test.py", line 7, in <module>
    code = req.urlopen(url_result)
  File "C:/Users/Admin/AppData/Local/Programs/Python/Python36/lib\urllib\request.py", line 223, in urlopen
    return opener.open(url, data, timeout)
  File "C:/Users/Admin/AppData/Local/Programs/Python/Python36/lib\urllib\request.py", line 526, in open
    response = self._open(req, data)
  File "C:/Users/Admin/AppData/Local/Programs/Python/Python36/lib\urllib\request.py", line 544, in _open
    '_open', req)
  File "C:/Users/Admin/AppData/Local/Programs/Python/Python36/lib\urllib\request.py", line 504, in _call_chain
    result = func(*args)
  File "C:/Users/Admin/AppData/Local/Programs/Python/Python36/lib\urllib\request.py", line 1361, in https_open
    context=self._context, check_hostname=self._check_hostname)
  File "C:/Users/Admin/AppData/Local/Programs/Python/Python36/lib\urllib\request.py", line 1318, in do_open
    encode_chunked=req.has_header('Transfer-encoding'))
  File "C:/Users/Admin/AppData/Local/Programs/Python/Python36/lib\http\client.py", line 1239, in request
    self._send_request(method, url, body, headers, encode_chunked)
  File "C:/Users/Admin/AppData/Local/Programs/Python/Python36/lib\http\client.py", line 1250, in _send_request
    self.putrequest(method, url, **skips)
  File "C:/Users/Admin/AppData/Local/Programs/Python/Python36/lib\http\client.py", line 1117, in putrequest
    self._output(request.encode('ascii'))
UnicodeEncodeError: 'ascii' codec can't encode characters in position 77-78: ordinal not in range(128)

```

에러가 납니다. 그 이유는 URL 주소에 사실 한글이 들어갈 수 없습니다.

“엥? 근데 크롬 창에서는 URL 주소에 한글이 들어가 있는데요?”

그건 크롬 브라우저의 ‘친절함’이라 볼 수 있죠. 원래는 한글이 있어서는 안되는데 이쁘게 한글로 보이게 만들어 놓은 것입니다. 크롬창의 URL주소를 복사해서 여기에다가 붙혀볼까요?

https://search.naver.com/search.naver?where=news&sm=tab_jum&query=%EC%86%90%ED%9D%A5%EB%AF%BC

‘query=’ 뒤에 있어야할 ‘손흥민’은 온데간데 없고 이상한 문자들만 남아있습니다. 이렇게 한글은 URL 주소에 쓰여질 때, 특수한 문자들로 ‘변환’되어야 합니다. 따라서 아래와 같이 코딩하면 됩니다.

```

1 import urllib.request as req
2 from bs4 import BeautifulSoup
3 import urllib.parse as par
4
5 url = "https://search.naver.com/search.naver?where=news&sm=tab_jum&query="
6 keyword = par.quote("치킨") # 한글을 특수한 문자로 변환
7 url_result = url + keyword
8 print(url_result)

```

urllib.parse 라는 모듈을 불러와서, quote() 라는 함수를 사용하였습니다. URL 주소의 가장 마지막 부분에 들어가야 할 “치킨” 부분을 quote()라는 함수로 한글을 특수한 문자로 변환한 다음, URL 의 맨 마지막 부분에 합쳐줬죠. 아래는 최종 전체 URL 주소를 출력한 결과입니다.

```

https://search.naver.com/search.naver?where=news&sm=tab_jum&query=%EC%B9%98%ED%82%A8
Process finished with exit code 0

```


자 이제 완벽한 URL 주소를 얻었으니 이걸 urlopen() 함수에만 넣기만 하면 우리가 원하는 곳의 HTML 주소를 얻을 수 있겠죠?? 아래의 코드를 추가해줍니다.

```
1 import urllib.request as req
2     from bs4 import BeautifulSoup
3 import urllib.parse as par
4
5 url = "https://search.naver.com/search.naver?where=news&sm=tab_jum&query="
6 keyword = par.quote("치킨") # 한글을 특수한 문자로 변환
7 url_result = url + keyword
8 code = req.urlopen(url_result)
```

Hint2

우리가 여태 꺼내오려던 것들은 모두 요소의 ‘내용’부분 이었잖아요. 그래서 ‘.string’ 이라는 속성을 사용해 ‘내용’부분만 쏙 빼왔어요. 그런데 만약 우리가 원하는 부분이 ‘내용’이 아닌 ‘속성값’이라면?? ‘내용’이 아닌 ‘속성값’을 빼오는 방법을 알려드릴게요.

우선 HTML을 분석시킨 후에 내가 원하는 요소를 가져와 볼게요.
제가 원하는 요소는 아래와 같습니다.

```
▶ <a data-clk="top.mkhome" href="http://help.naver.com/support/alias/contents2/
naverhome/naverhome_1.naver" class="al_favorite">...</a> == $0
```

속성이 뭐가 많죠..? 정리해보자면 속성명과 속성값은 이렇습니다.

```
data-clk = "top.mkhome"
href = "http://help.naver.com/(중략)/naverhome_1.naver"
class = "al_favorite"
```

```
from bs4 import BeautifulSoup
import urllib.request as req

url="https://www.naver.com"
res=req.urlopen(url)

soup=BeautifulSoup(res, "html.parser")

result = soup.select_one("div.area_links > a")
```

여기서 이제 이 요소의 ‘내용’을 뽑고 싶다면

```
result.string
```

이라고 하면 됐었습니다.

그러면 이렇게 한번 해볼게요.

```
result.attrs
```

어떤 결과가 나오는지 출력해보겠습니다.

```
from bs4 import BeautifulSoup
import urllib.request as req

url="https://www.naver.com"
res=req.urlopen(url)

soup=BeautifulSoup(res, "html.parser")

result = soup.select_one("div.area_links > a")
print(result.attrs)
```

아래가 실행결과입니다.

```
{'data-clk': 'top.mkhome', 'href': 'http://help.naver.com/support/alias/contents2/naverhome/naverhome_1.naver', 'class': ['al_favorite']}
```

오 뭔가 눈치 채셨나요.

요소의 string 속성에는 요소의 ‘내용’이 들어있던 것처럼

요소의 attrs 속성에는 요소의 ‘속성명과 속성값들’이 “딕셔너리 자료형” 형태로 들어가 있네요.

그러면 여기서 ‘data-clk’의 속성값을 뽑고 싶다면? 그저 딕셔너리 자료형에서 인덱싱을 하듯 쓰면 되겠죠? 이렇게요

```
result.attrs['data-clk']
```

다시 또 예를 들어, ‘href’의 속성값을 뽑고 싶다면?

```
result.attrs['href']
```

라고 쓰면 될겁니다.

위 힌트는 이 과제를 하는데 유용하게 쓰일 것이고, 후에 크롤링에 대해서 배울 때에도 유용하게 쓰일 거예요!