/*elice*/

파이썬 크롤링

크롤링의 기초

김경민 선생님

크롤링이란?

크롤링이란?

웹 페이지에서 필요한 데이터를 **추출**해내는 작업

크롤링을 하는 프로그램 : 크롤러

크롤링이란?

웹 페이지는 정보를 **HTML** 문서로 표현합니다.

크롤링을 위해 필요한 것

웹 페이지의 **HTML**을 얻기 위해

requests 라이브러리를,

가져온 **HTML**을 분석하기 위해

BeautifulSoup 라이브러리를 사용합니다.

BeautifulSoup

BeautifulSoup 라이브러리

HTML, XML, JSON 등 파일의 **구문을 분석**하는 모듈

웹 페이지를 표현하는 **HTML**을 분석하기 위해 사용합니다.

BeautifulSoup 라이브러리

```
soup = BeautifulSoup(open("index.html"), "html.parser")
```

HTML 파일로 **BeautifulSoup** 객체를 만들 수 있습니다.

변수 이름은 관습적으로 **soup** 라고 짓습니다.

BeautifulSoup 라이브러리

```
soup = BeautifulSoup(open("index.html"), "html.parser")
```

“html.parser”의 의미는, BeautifulSoup 객체에게

“HTML을 분석해라” 라고 알려주는 의미입니다.

BeautifulSoup 라이브러리

```
soup.find("p")          # 처음 등장하는 태그 찾기
soup.find_all("p")      # 모든 태그 찾기
```

find, find_all 메소드를 이용하여

HTML 태그를 추출할 수 있습니다.

BeautifulSoup 라이브러리

```
soup.find("p")          # 처음 등장하는 태그 찾기
soup.find_all("p")      # 모든 태그 찾기
```

find는 추출한 **HTML 태그 하나**를,

find_all은 HTML 태그를 여러 개 담고 있는 **리스트**를 얻습니다.

BeautifulSoup 라이브러리

예시 코드

```
print(soup.find("p"))
print(soup.find_all("p"))
```

출력 결과

```
<p></p>
[<p></p>, <p></p>, ... , <p></p>]
```

BeautifulSoup 라이브러리

```
<!DOCTYPE html>
...
<body>
  <div class="cheshire">
    <p>Don't crawl this.</p>
  </div>
  <div class="elice">
    <p>Hello, Python Crawling!</p>
  </div>
</body>
```

div 태그 중, **클래스**가

elice인 것만 추출하려면

어떻게 해야 할까요?

BeautifulSoup 라이브러리

```
soup.find("div")
soup.find("div", class_="elice")
```

class_ 매개변수에 값을 저장함으로써

특정 클래스를 가진 태그를 추출할 수 있습니다.

BeautifulSoup 라이브러리

```
soup.find("div", class_="elice").find("p")
```

find로 얻은 결과도 **BeautifulSoup 객체**입니다.

따라서 find를 한 결과에 또 find를 적용할 수 있습니다.

위 코드는 **div 태그 안에 있는 p 태그**를 추출합니다.

BeautifulSoup 라이브러리

```
soup.find("div", class_="elice").find("p").get_text()
```

BeautifulSoup 객체에 **get_text** 메소드를 적용하면

태그가 갖고 있는 텍스트를 얻을 수 있습니다.

BeautifulSoup 라이브러리

예시 코드

```
print(soup.find("p"))
print(soup.find("p").get_text())
```

출력 결과

```
<p>Hello, Python Crawling!</p>
Hello, Crawling!
```

BeautifulSoup 라이브러리

```
soup.find("div")
soup.find("div", id="elice")
```

특정 id의 값을 추출하고자 하는 경우에는

id 매개변수의 값을 지정할 수 있습니다.

Requests

requests 라이브러리

Python에서 HTTP 요청을 보낼 수 있는 모듈

HTTP 요청이란?

GET 요청 : 정보를 **조회**하기 위한 요청

(예 : 네이버 홈페이지에 접속한다. 구글에 키워드를 검색한다.)

POST 요청 : 정보를 **생성, 변경**하기 위한 요청

(예 : 웹 사이트에 로그인한다. 메일을 삭제한다.)

HTTP 요청이란?

본 과목에서는 **GET** 요청만 사용합니다.

requests 라이브러리

```
url = "https://www.google.com"
result = requests.get(url)
```

지정한 **URL**로 **GET** 요청을 보냈고,

서버에서는 요청을 받아 처리한 후

result 변수에 **응답**을 보냅니다.

requests 라이브러리

```
print(result.status_code)
print(result.text)
```

응답의 **status_code**로는 요청의 결과를 알 수 있습니다.

만약 요청이 성공했다면

text로 해당 웹 사이트의 **HTML**을 얻을 수 있습니다.

두 라이브러리 조합하기

```
soup = BeautifulSoup(result.text, "html.parser")
```

requests와 **BeautifulSoup**를 조합하여

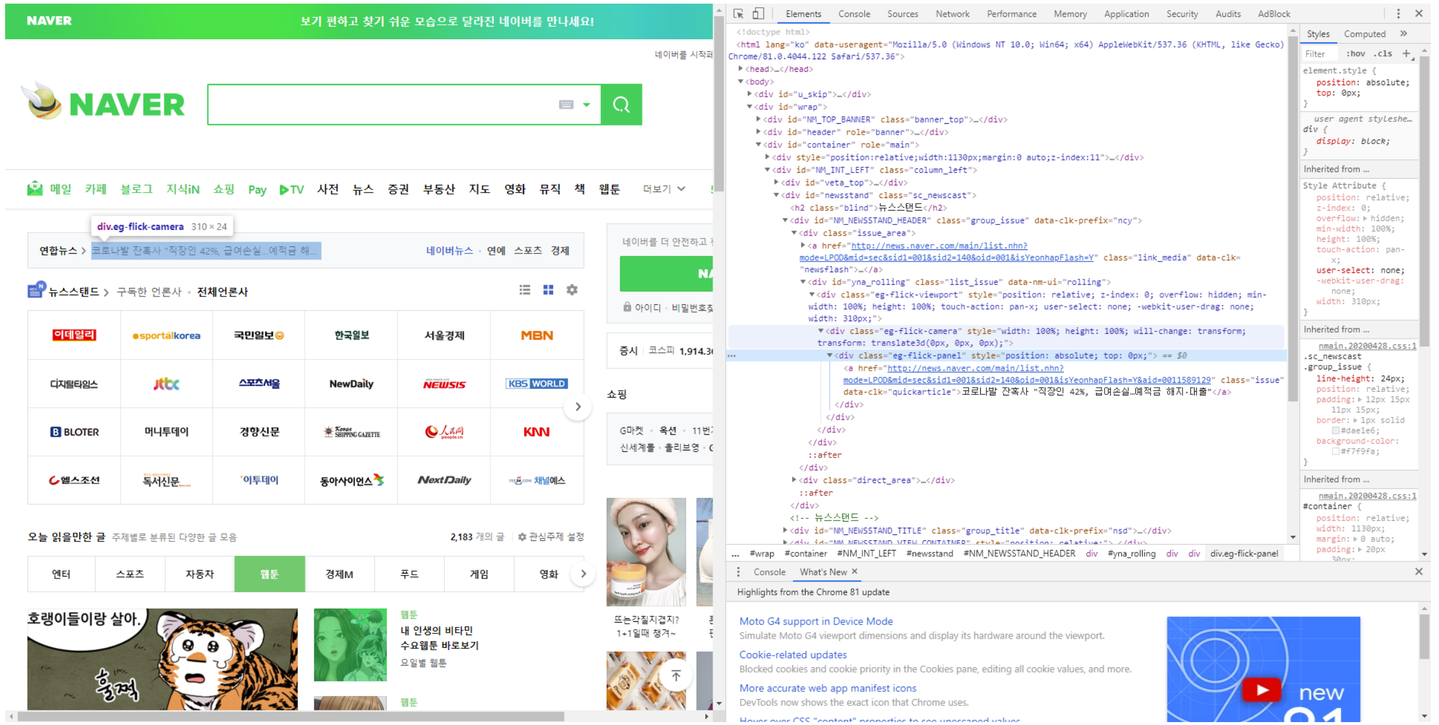
웹 페이지의 HTML을 분석할 수 있습니다.

실전 크롤링

실전 크롤링

배운 내용으로 크롤링 코드를 직접 작성해보도록 하겠습니다.

실전 크롤링



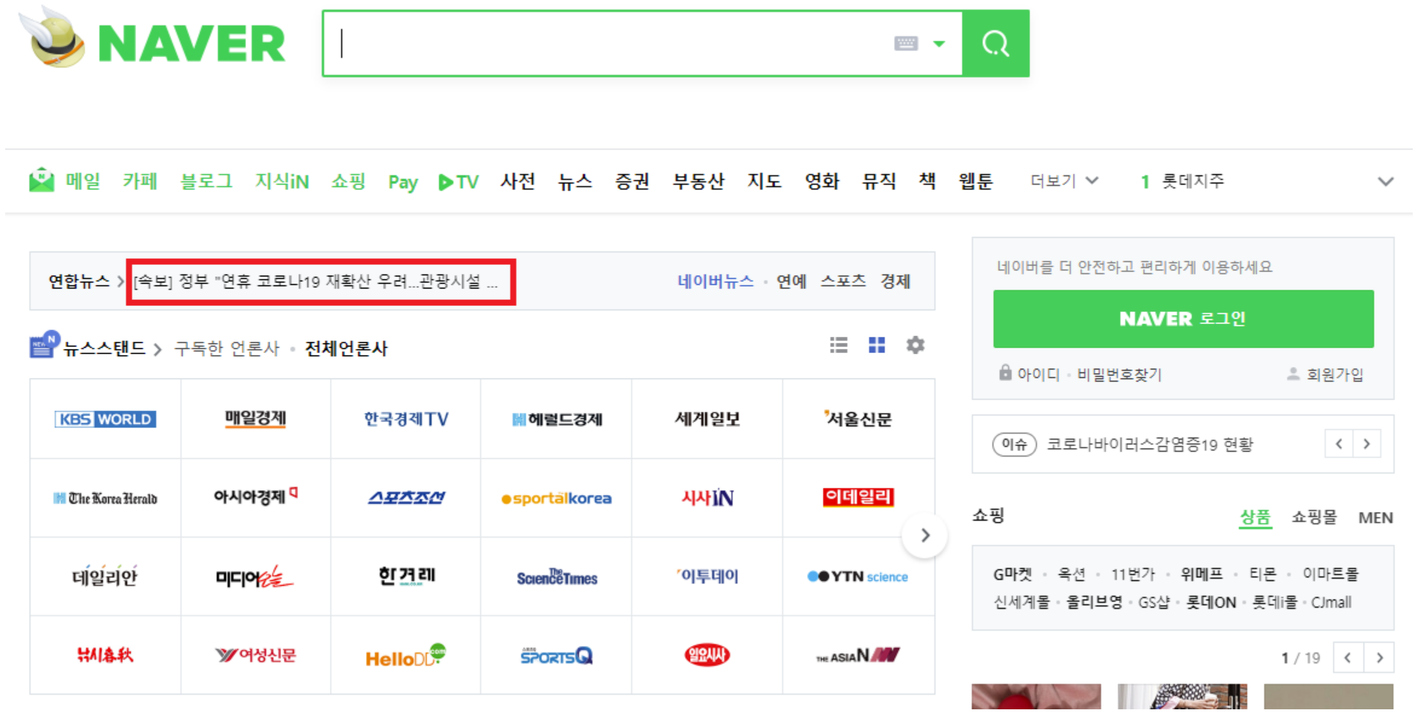
웹 페이지에서 **F12** 버튼을 눌러 개발자 도구를 켤 수 있습니다.

실전 크롤링

검색을 원하는 요소에 **오른쪽 마우스**를 클릭하고,

‘검사’ 를 눌러 개발자 도구를 켤 수도 있습니다.

네이버 헤드뉴스 찾기



네이버의 헤드뉴스 부분을 크롤링해보려고 합니다.

네이버 헤드뉴스 찾기

```
['분양가 상한제 시행 3개월 연기…"개포주공1단지 총회 미뤄야"', '대구서 17세 청소년 숨져... 보건당국 "사후 검체 검사 중"', '한사랑요양 75명등 대구 요양병원 5곳 88명 확진…집단감염 빈발', '문대통령 "정부 힘만으론 부족"…코로나극복 \'범국가연대\' 강조', '오래 쓰는 나노 마스크 첫 상용화 추진…마스크 부족 해결될까', "분당제생병원장 접촉한 복지부 차관 '예방적 자가격리'", "보호·부양의무 외면한 가족 상속 막는 '구하라법' 입법 청원", '내일 때아닌 태풍급 강풍 분다…"선별진료소 등 관리 유의해야"', "코로나추경 '위기가구'에 2천억 투입…2인가구 月77만원 지원", '가수 최종훈, 불법촬영 인정…"이제라도 처벌받게 돼 홀가분"']
```

해당 부분을 포함하는 태그와 클래스를 참조하여

정보를 크롤링하세요.

연합뉴스 속보 기사 제목 추출하기

네이버에 있는 연합뉴스 속보 기사들의 제목을

크롤링하려고 합니다.

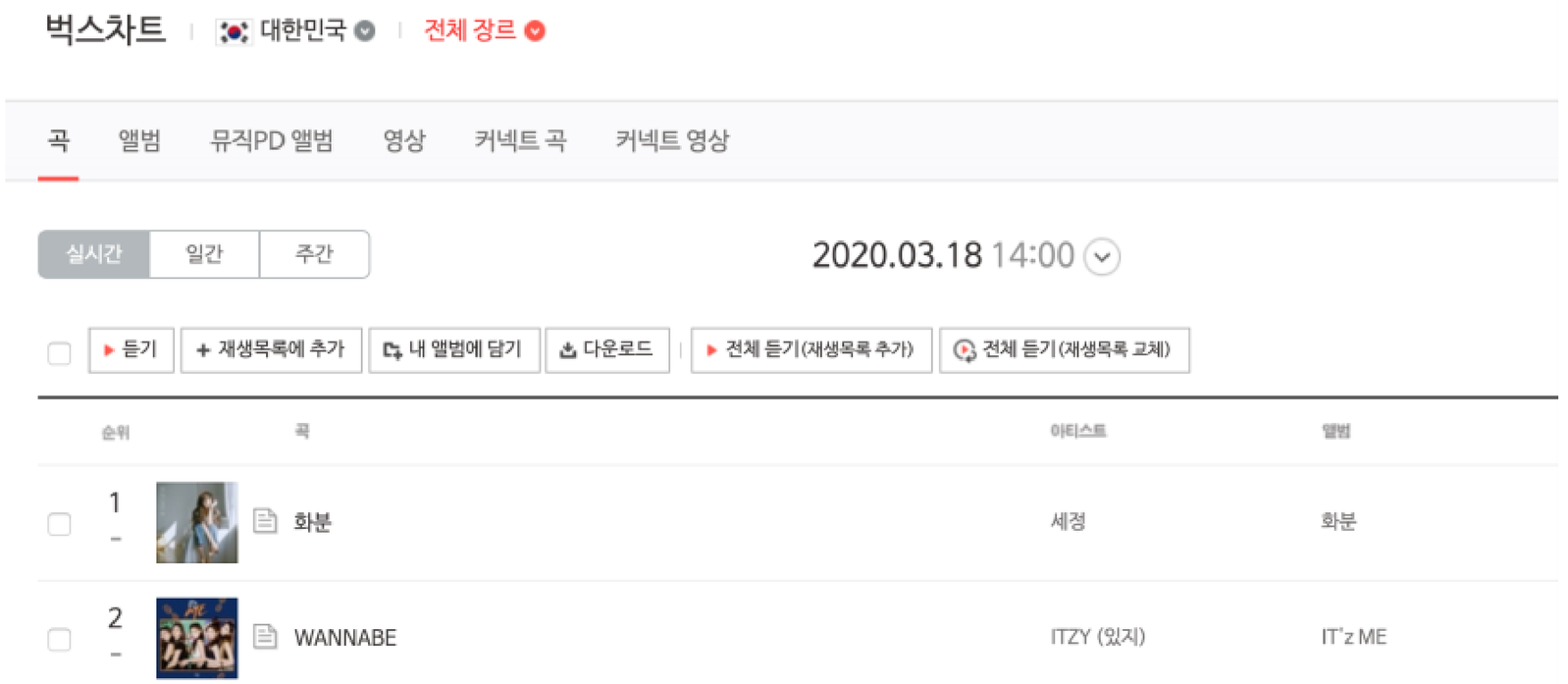
연합뉴스 속보 기사 제목 추출하기

```
['\n', '분양가 상한제 시행 3개월 연기…"개포주공1단지 총회 미뤄야"', '코로나에 발목 잡힌 상한제…한숨 돌린 재개발·재건축 조합', '서울 아파트 공시가격 14.75% 상승, 13년만에 최대…강남 25.57%', "래미안대치팰리스 전용 84㎡ 공시가격 15억→21억원…'41%' 급등", '서울 서초동 트라움하우스5차, 15년째 제일 비싼 공동주택', '30억 이상 아파트 공시가 30%↑…종부세 편입대상 41.8% 늘어', '고가·다주택자 보유세 상한까지 오른다…집값 하락 본격화될까', '"서울 한강 이남 아파트값 3.3㎡당 4천만원 돌파"', '대구서 17세 청소년 숨져... 보건당국 "사후 검체 검사 중"', '한사랑요양 75명등 대구 요양병원 5곳 88명 확진…집단감염 빈발', '문대통령 "정부 힘만으론 부족"…코로나극복 \'범국가연대\' 강조', '오래 쓰는 나노 마스크 첫 상용화 추진…마스크 부족 해결될까', "분당제생병원장 접촉한 복지부 차관 '예방적 자가격리'", "보호·부양의무 외면한 가족 상속 막는 '구하라법' 입법 청원", '내일 때아닌 태풍급 강풍 분다…"선별진료소 등 관리 유의해야"', "코로나추경 '위기가구'에 2천억 투입…2인가구 月77만원 지원", '가수 최종훈, 불법촬영 인정…"이제라도 처벌받게 돼 홀가분"']
```

연합뉴스 속보 기사의 제목은 find 함수로 찾은 요소 안에서

또 find를 사용해야 할 수도 있습니다.

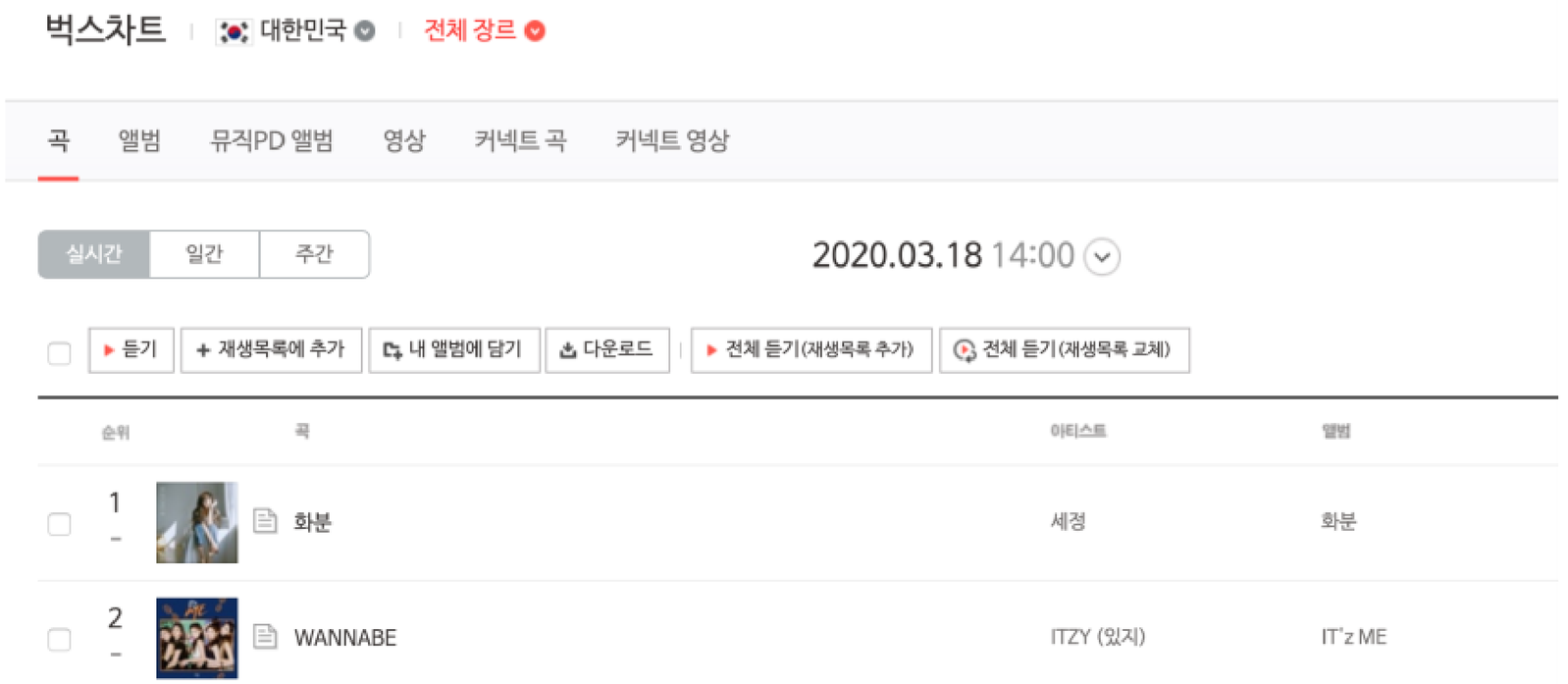
bugs 실시간 음원차트 순위 추출하기



음원 사이트 bugs의 실시간 차트를 크롤링하여

높은 순위부터 차례대로 곡명을 출력하려고 합니다.

bugs 실시간 음원차트 순위 추출하기



음원 사이트 bugs의 실시간 차트를 크롤링하여

높은 순위부터 차례대로 곡명을 출력하려고 합니다.

bugs 실시간 음원차트 순위 추출하기

```
['\n화분\n', '\nWANNABE\n', '\n시작\n', '\n어떻게 지내\n', '\n돌덩이\n', '\n아무노래\n', '\n그때 그 아인\n', '\n흔들리는 꽃들 속에서 네 샴푸향이 느껴진거야\n', '\nFIESTA\n', '\n문득\n', '\n마음을 드려요\n', '\nPhysical (feat. 화사)\n', '\nTo Die For\n', '\nON\n', '\nBlueming\n', '\nPsycho\n', '\nMETEOR\n', '\nSquare (2017)\n', '\nBirthday\n', '\n찐이야\n', '\n바빠서 (Feat. 헤이즈)\n', '\nManiac\n', "\nDon't Start Now\n", '\n둘만의 세상으로 가\n', '\n다시 난, 여기\n', '\nPainkiller\n', '\nHIP\n', '\nMemories\n', '\nLove poem\n', '\n사랑의 인사\n', '\n솔직히 지친다\n', '\n어떻게 이별까지 사랑하겠어, 널 사랑하는 거지\n', '\n늦은 밤 너의 집 앞 골목길에서\n', '\n영웅 (英雄; Kick It)\n', '\n2002\n', '\nChanges\n', '\nNerdy Love (Feat. 백예린)\n', '\nSay\n', '\n작은 것들을 위한 시 (Boy With Luv) (Feat. Halsey)\n', '\n막걸리 한잔\n', '\n모든 날, 모든 순간 (Every day, Every Moment)\n', '\nBlack Swan\n', '\n어느 60대 노부부이야기\n', '\n오늘도 빛나는 너에게 (To You My Light) (Feat.이라온)\n', '\n반만\n', '\n사계 (Four Seasons)\n', '\n안녕\n', '\n보
```

아마 텍스트를 그냥 출력하면

위와 같이 개행문자가 끼어 있을 것입니다.

bugs 실시간 음원차트 순위 추출하기

```
['화분', 'WANNABE', '시작', '어떻게 지내', '돌덩이', '아무노래', '그때 그 아인', '흔들리는 꽃들 속에서 네 샴푸향이 느껴진거야', 'FIESTA', '문득', '마음을 드려요', 'Physical (feat. 화사)', 'To Die For', 'ON', 'Blueming', 'Psycho', 'METEOR', 'Square (2017)', 'Birthday', '찐이야', '바빠서 (Feat. 헤이즈)', 'Maniac', "Don't Start Now", '둘만의 세상으로 가', '다시 난, 여기', 'Painkiller', 'HIP', 'Memories', 'Love poem', '사랑의 인사', '솔직히 지친다', '어떻게 이별까지 사랑하겠어, 널 사랑하는 거지', '늦은 밤 너의 집 앞 골목길에서', '영웅 (英雄; Kick It)', '2002', 'Changes', 'Nerdy Love (Feat. 백예린)', 'Say', '작은 것들을 위한 시 (Boy With Luv) (Feat. Halsey)', '막걸리 한잔', '모든 날, 모든 순간 (Every day, Every Moment)', 'Black Swan', '어느 60대 노부부이야기', '오늘도 빛나는 너에게 (To You My Light) (Feat.이라온)', '반만', '사계 (Four Seasons)', '안녕', '보라빛 엽서', 'ROXANNE', 'Sweet Night', '추억으로 가는 당신', 'SKYLINE', 'bad guy', '너를 사랑하고 있어', '00:00 (Zero O'Clock)', '배신자', '또 사랑에 속다', '18세 순이', 'Know Me Too Well', '진또배기', '나의 오
```

실습에서는 위와 같이 필요 없는 문자는

모두 필터링하려고 합니다.

영화 후기 수집하기

네이버 영화 페이지에 있는 영화평의 제목을

수집하여 출력해봅니다.

영화 후기 수집하기

```
['다들 구라치지마세요 재미1도 없는 영화', '닐 암스트롱 vs 스티븐 시걸', '미국의 성공과 암스트롱의 성공, 그 이면의 이야기', '진짜 영화 제대로 보지도않고, 볼줄도 모르는 사람이 안타깝다', '[영화] 퍼스트맨 (2018), "우주보단 인간, 노래보단 적막, 라라랜드보단 위플래쉬"', '"적막을 감성과 감정으로 채운 영화" 퍼스트맨 짧은 리뷰 스포없음', '데미언 셔젤은 왜 ‘닐 암스트롱’을 선택했는가 (in 2018 부산국제영화제)', '<퍼스트맨(First Man , 2018)> 달에 첫 발을 내딛기까지.. 닐 암스트롱의 전기 영화 (데이미언 셔젤 감독, 라이언 고슬링, 클레어 포이, 실화 영화)', '천재감독의 3번째 영화!!!', '정말 많은걸 느끼게 한듯']
```

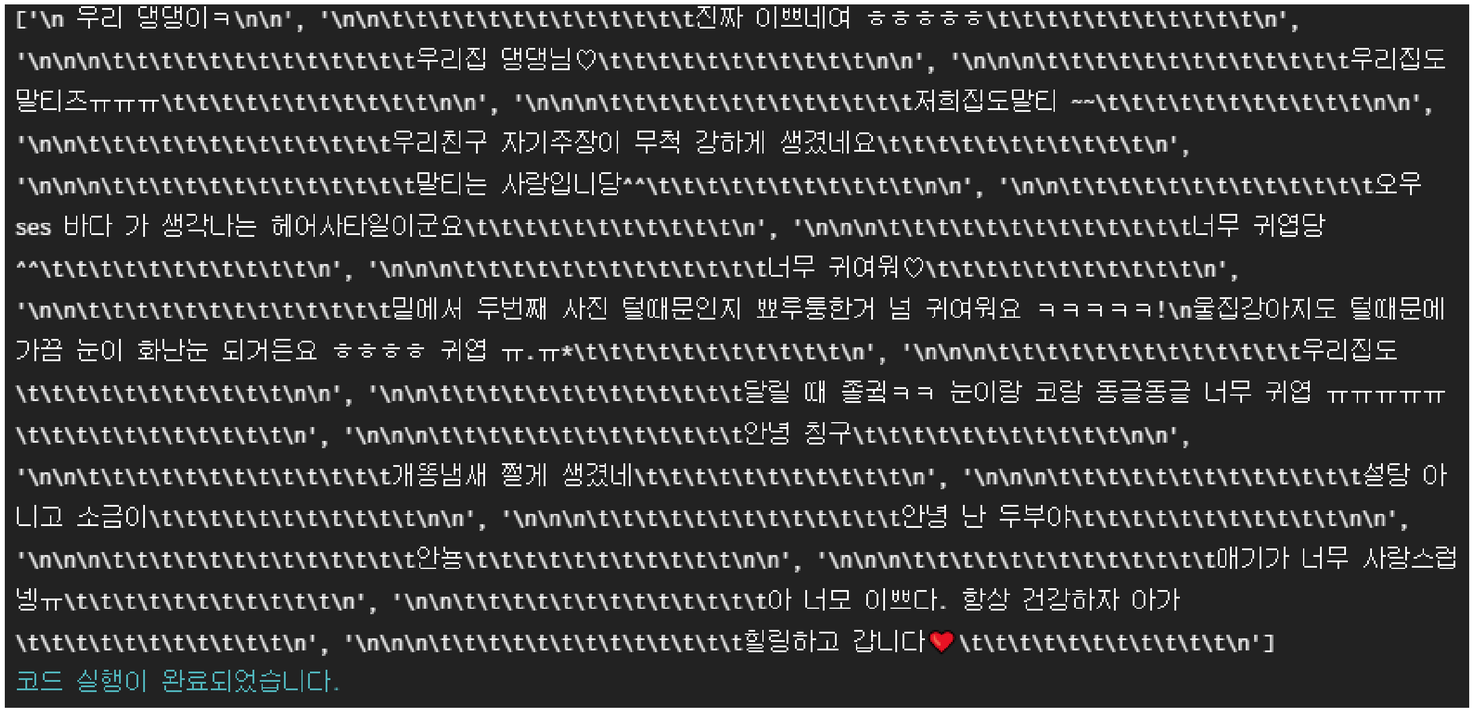
경우에 따라서 find 함수를 3중으로 사용해야 할 수도 있으니

데이터를 담고 있는 태그를 잘 확인하세요.

커뮤니티 댓글 수집하기

커뮤니티 사이트의 댓글을 수집하여 출력하고자 합니다.

커뮤니티 댓글 수집하기



```
['\n 우리 댕댕이ㅋ\n\n', '\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t진짜 이쁘네여 ㅎㅎㅎㅎㅎ\t\t\t\t\t\t\t\t\t\t\n',
'\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t우리집 댕댕님♡\t\t\t\t\t\t\t\t\t\t\n\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t우리집도
말티즈ㅠㅠㅠ\t\t\t\t\t\t\t\t\t\t\t\n\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t저희집도말티 ~~\t\t\t\t\t\t\t\t\t\t\n\n',
'\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t우리친구 자기주장이 무척 강하게 생겼네요\t\t\t\t\t\t\t\t\t\t\n',
'\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t말티는 사랑입니당^^\t\t\t\t\t\t\t\t\t\t\n\n', '\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t오우
ses 바다 가 생각나는 헤어사타일이군요\t\t\t\t\t\t\t\t\t\t\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t너무 귀엽당
^^\t\t\t\t\t\t\t\t\t\t\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t너무 귀여워♡\t\t\t\t\t\t\t\t\t\t\n',
'\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t밀에서 두번째 사진 털때문인지 뾰루퉁한거 넘 귀여워요 ㅋㅋㅋㅋㅋ!\n물집강아지도 털때문에
가끔 눈이 화난눈 되거든요 ㅎㅎㅎㅎ 귀엽 ㅠ.ㅠ*\t\t\t\t\t\t\t\t\t\t\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t우리집도
\t\t\t\t\t\t\t\t\t\t\n\n', '\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t달릴 때 졸귘ㅋㅋ 눈이랑 코랑 똥글똥글 너무 귀엽 ㅠㅠㅠㅠㅠ
\t\t\t\t\t\t\t\t\t\t\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t안녕 침구\t\t\t\t\t\t\t\t\t\t\n\n',
'\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t개똥냄새 쩔게 생겼네\t\t\t\t\t\t\t\t\t\t\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t설탕 아
니고 소금이\t\t\t\t\t\t\t\t\t\t\n\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t안녕 난 두부야\t\t\t\t\t\t\t\t\t\t\n\n',
'\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t안뇽\t\t\t\t\t\t\t\t\t\t\n\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t애기가 너무 사랑스럽
넹ㅠ\t\t\t\t\t\t\t\t\t\t\n', '\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t아 너모 이쁘다. 항상 건강하자 아가
\t\t\t\t\t\t\t\t\t\t\n', '\n\n\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t힐링하고 갑니다❤️\t\t\t\t\t\t\t\t\t\t\n']
코드 실행이 완료되었습니다.
```

크롤링해온 데이터에는 개행 문자와 탭 문자가 끼어 있습니다.

커뮤니티 댓글 수집하기

```
[' 우리 댕댕이ㅋ', '진짜 이쁘네여 ㅎㅎㅎㅎㅎ', '우리집 댕댕님♡', '우리집도 말티즈ㅠㅠㅠ',
'저희집도말티 ~~', '우리친구 자기주장이 무척 강하게 생겼네요', '말티는 사랑입니당^^', '오우
ses 바다 가 생각나는 헤어사타일이군요', '너무 귀엽당^^', '너무 귀여워♡', '밑에서 두번째 사
진 덜때문인지 뾰루퉁한거 넘 귀여워요 ㅋㅋㅋㅋㅋ!둘집강아지도 덜때문에 가끔 눈이 화난눈 되거
든요 ㅎㅎㅎㅎ 귀엽 ㅠ.ㅠ*', '우리집도', '달릴 때 졸귑ㅋㅋ 눈이랑 코랑 동글동글 너무 귀엽 ㅠ
ㅠㅠㅠㅠ', '안녕 칭구', '개똥냄새 쩔게 생겼네', '설탕 아니고 소금이', '안녕 난 두부야', '안
뇽', '애기가 너무 사랑스럽넹ㅠ', '아 너모 이쁘다. 항상 건강하자 아가', '힐링하고 갑니다❤️']
코드 실행이 완료되었습니다.
```

bugs 실습때와 마찬가지로, 필요 없는 문자는

모두 필터링하려고 합니다.


/*elice*/
contact@elice.io

/*elice*/

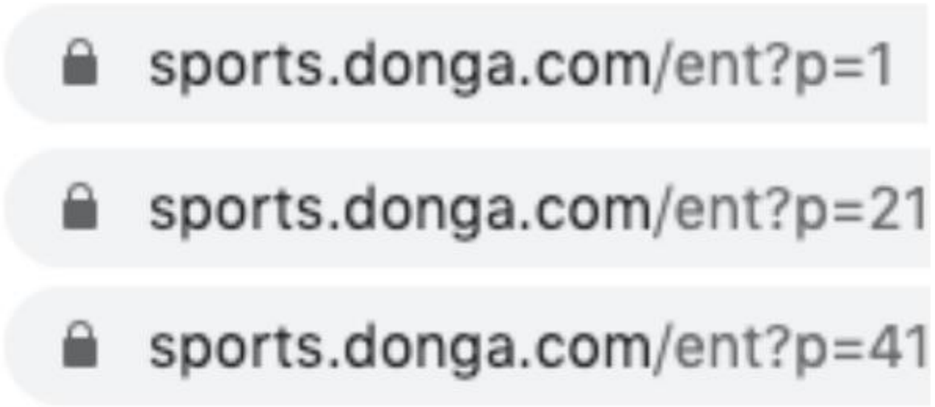
파이썬 크롤링

여러 페이지 크롤링하기

김경민 선생님

Query

Query



이 뉴스 웹사이트는 각 페이지의 URL에서 **p=(숫자)** 부분이

20씩 증가하고 있는 규칙이 있습니다.

이 사이트에서 여러 페이지를 크롤링하려면 어떻게 해야 할까요?

Query

```
for i in range(0, 5) :
    url = "http://sports.donga.com/Enter?p="+str((i*20+1))
    ...
```

쉬운 방법으로는 URL을 **문자열 연산**으로 처리하여

새로운 URL을 얻는 것입니다.

Query

하지만, URL의 **query(쿼리)**를 이용하면

이 작업을 더 효과적으로 할 수 있습니다.

Query

웹 서버에 GET 요청을 보낼 때

조건에 맞는 정보를 표현하기 위한 변수

예) **번호**가 1번인 학생을 보여줘라

전체 기사 중 **페이지**가 21인 기사들을 보여줘라

Query

google에 ‘elice’를 검색한 결과입니다.

q라는 변수에 elice라는 값이 담겨,

전체 데이터 중 elice라는 키워드로 검색한 결과만을 보여줍니다.

Query

네이버 영화 서비스에서 특정 영화를 클릭하면,

code라는 변수에 영화 코드가 담겨

해당 영화에 대한 정보를 보여줍니다.

requests 라이브러리

```
url = "https://www.google.com/search"
result = requests.get(url, params = {'q':'elice'})
```

requests의 get 메소드로 GET 요청을 보낼 때

params 매개변수에 **딕셔너리**를 전달함으로서

쿼리를 지정할 수 있습니다.

requests 라이브러리

```
code = ... # 영화 코드에 대한 정보를 얻는다.
result = requests.get(url, params = {'movie':code})
```

전체 영화 데이터에서 영화 코드에 대한 정보를 찾고,

다시 requests를 이용하여 특정 영화에 대한 정보를

얻는 요청을 할 수 있습니다.

Tag Attribute

태그와 속성

```
<div class="elice" id="title">제목</div>
 태그   속성          속성
```

HTML에는 여러 종류의 **태그**와,

태그에 특정 기능이나 유형을 적용하는 **속성**이 있습니다.

태그와 속성

```
div = soup.find("div")
print(div.attrs)
```

어떤 태그의 속성이 무엇이 있는지 확인할 때는

attrs 멤버변수를 출력합니다.

태그와 속성

```
print(div['class'])
```

attrs 딕셔너리의 키로 인덱싱하여,

태그의 속성에 접근할 수 있습니다.

href 속성

```
<a href="https...">기사 제목</a>
```

a 태그는 하이퍼링크를 걸어주는 태그로써

이동할 URL을 href 속성에 담고 있습니다.

href 속성

```
a = soup.find("a")

href_url = a["href"]
```

위와 같이 href 속성을 이용하여

웹 페이지에 존재하는 하이퍼링크의 URL을 얻을 수 있습니다.

Children, Name

Children, Name

웹 사이트의 구조가 복잡한 경우

다양한 옵션을 적용해야 할 수도 있습니다.

children은 어떤 태그가 **포함하고 있는 태그**를,

name은 어떤 태그의 **이름**을 의미하는 속성입니다.

Children

```
<div>
  <span>span1</span>
  <span>span2</span>
  <p>p tag</p>
  <img ... />
</div>
```

옆의 div 태그는

여러 태그들을 갖고 있습니다.

Children

beautifulsoup의 **children** 속성으로

어떤 태그가 **포함하고 있는 태그**들도 조회할 수 있습니다.

Children

```
soup.find("div").children
#span, p, img 태그를 갖는 리스트를 얻습니다.
```

위의 코드는 어떤 div 태그를 찾고,

그 div 태그에 **포함된 태그들의 리스트**를 얻는 코드입니다.

Name

```
children = soup.find("div").children
for child in children :
    print(child.name)
    # span, span, p, img가 각각 출력됩니다.
```

어떤 태그의 이름을 알고 싶다면 name 속성을 이용할 수 있습니다.

태그가 존재하지 않는 경우 None 값을 얻습니다.

실전 크롤링

실전 크롤링

배운 내용으로 크롤링 코드를 직접 작성해보도록 하겠습니다.

여러 페이지의 기사 제목 수집하기

'소녀의 세계' 권현빈, 강렬한 첫 등장...아린과 훈훈 케미

'소녀의 세계' 권현빈, 강렬한 첫 등장...아린과 훈훈 케미 '소녀의 세계' 권현빈이 오마이걸 아린과 귀여운 케미를 선보였다. 권현빈은 지난 22일 공개된 tvN D 웹드라마 '소녀의 세계' 1화 에필로그에 ...

2020-04-23 09:43:00

[TV북마크] '기막힌 유산' 강세정X신정윤, 심상치 않은 재회...박...

[TV북마크] '기막힌 유산' 강세정X신정윤, 심상치 않은 재회...박인환 재산정리 어제(22일) KBS1 새 저녁 일일드라마 '기막힌 유산'(연출 김형일 극본 김경희) 3회에서는 자식들에게 증여를 준비하는 박...

2020-04-23 09:43:00

동아스포츠의 연예부 기사의

제목 부분을 크롤링해보려고 합니다.

여러 페이지의 기사 제목 수집하기

[DA:차트] 조정석 '아로하', 16주차 가온차트서 2관왕 영예

배우 조정석의 '아로하'가 16주차 가온차트에서 2관왕을 차지했다. 가온차트를 운영하는 사단법인 한국음악콘텐츠협회는, "16주차(2020.04.12~2020.04.18) 디지털차트, 스트리밍차트에서 슬기로운 의...

2020-04-23 09:28:00

1 2 3 4 5 6 7 8 9 10 >

사이트의 하단에서 버튼을 눌러 페이지를 이동할 수 있습니다.

1페이지부터 5페이지까지의 기사 제목을 크롤링하려고 합니다.

여러 페이지의 기사 제목 수집하기

https://sports.donga.com/ent?p=1

이 사이트는 URL의 쿼리 부분에서 p의 값에 따라 페이지가 결정됩니다.

한 페이지에 기사가 20개씩 존재하므로

p=1이면 1페이지, p=21이면 2페이지와 같은 식입니다.

여러 페이지의 기사 제목 수집하기

URL을 문자열의 덧셈 연산으로 만드실 수도 있지만,

requests.get 함수의 **params 매개변수**로

쿼리 변수를 추가할 수 있습니다.

여러 페이지의 기사 제목 수집하기

```
['나띠, 데뷔 스케줄러 공개…5월7일 쇼케이스 'M2 방송'', '[DA:차트] '너를 만나'→'우리 만남이'…폴킴, 연속 음원차트 1위', ''컴백' 공원소녀, 4色 오버뷰 영상 공개…전곡 하이라이트', '원어스, '쉽게 쓰여진 노래' 성공적 마무리 "'로드 투 킹덤'에서 봐요"', ''소녀의 세계' 권현빈, 강렬한 첫 등장…아린과 훈훈 케미', '[TV북마크] '기막힌 유산' 강세정X신정윤, 심상치 않은 재회…박인환 재산정…', '신원호 PD "'슬의생' 2막…5인방 중심 주변 서사 깊어질 것"', '1TEAM(원팀), '덕분에 챌리지' 동참…코로나19 의료진 응원', '블랙핑크, 레이디가가와 콜라보…피처링 참여 [공식]', ''본 어게인' 장기용·진세연·이수혁, 새 운명 시작…전생 악연 풀릴까', ''가족입니다' 메인 포스터…때론 타인보다 낯선 가족 이야기', '리밋, '날말' MV 티저 공개…'찐엔딩' 주인공 재회', ''모던 패밀리' 김영옥 "신구·이순재 활동 보면서 연기 자극 받아"', '[DA:할리우드] 샤를리즈테론 기부, 코로나19 관련 100만달러 쾌척', ''끼리끼리' 인교진→하승진, 24일 '굿모닝FM' 출격 (ft.장성규) [공식…', ''뭉쳐야 찬다' 윤성빈 vs 모태범, 허벅지 힘겨루기 한 판', ''사냥의 시간', 오늘 공개…안재홍, 타투+탈색 '파격 변신'', ''컴백' 파나틱스, 6色 '바비걸' 콘셉트컷 공개', '[DA:차트] 조정석 '아로하', 16주차 가온차트서 2관왕 영예', ''엠카' 측 "갓세븐·솔라·에이프릴, 오늘 컴백 무대 최초 공개" [공식]', ''엠카' 측 "갓세븐·솔라·에이프릴, 오늘 컴백 무대 최초 공개" [공식]', '넷플릭스X연상호 '지옥' 제작 확정…'송곳' 작가 공동집필 [공식]', ''나 혼자 산다' 송승헌, 무지개 회원 입성…10년차 자취일상 최초 공개', '오인택 결혼, 9월 승무원 출신 예비 신부와 백년가약 [공식]', ''쌍갑포차' 이준혁, 머리부터 발끝까지 올 화이트…新 저승사자 탄생', ''29일 컴백' NCT DREAM, 오늘(23일) 트랙 비디오 공개', ''미스터트롯' 이대
```

크롤링한 결과는 다음과 같습니다.

각 기사의 href 수집

‘소녀의 세계’ 권현빈, 강렬한 첫 등장...아린과 훈훈 케미

‘소녀의 세계’ 권현빈, 강렬한 첫 등장...아린과 훈훈 케미 ‘소녀의 세계’ 권현빈이 오마이걸 아린과 귀여운 케미를 선보였다. 권현빈은 지난 22일 공개된 tvN D 웹드라마 ‘소녀의 세계’ 1화 에필로그에 ...

2020-04-23 09:43:00

[TV북마크] ‘기막힌 유산’ 강세정X신정윤, 심상치 않은 재회...박...

[TV북마크] ‘기막힌 유산’ 강세정X신정윤, 심상치 않은 재회...박인환 재산정리 어제(22일) KBS1 새 저녁 일일드라마 ‘기막힌 유산’(연출 김형일 극본 김경희) 3회에서는 자식들에게 증여를 준비하는 박...

2020-04-23 09:43:00

이번에는 각 기사로 이동할 수 있는 href를 수집해봅시다.

각 기사의 href 수집

href는 a 태그의 속성으로 존재하며

크롤링된 a 태그에 접근하여 얻을 수 있습니다.

각 기사의 href 수집

```
['https://sports.donga.com/ent/article/all/20200423/100773540/1', 'https://sports.donga.com/ent/article/all/20200423/100773507/1',
'https://sports.donga.com/ent/article/all/20200423/100773479/1', 'https://sports.donga.com/ent/article/all/20200423/100773425/1',
'https://sports.donga.com/ent/article/all/20200423/100773430/1', 'https://sports.donga.com/ent/article/all/20200423/100773413/1',
'https://sports.donga.com/ent/article/all/20200423/100773388/1', 'https://sports.donga.com/ent/article/all/20200423/100773334/1',
'https://sports.donga.com/ent/article/all/20200423/100773314/1', 'https://sports.donga.com/ent/article/all/20200423/100773308/1',
'https://sports.donga.com/ent/article/all/20200423/100773312/1', 'https://sports.donga.com/ent/article/all/20200423/100773280/1',
'https://sports.donga.com/ent/article/all/20200423/100773265/1', 'https://sports.donga.com/ent/article/all/20200423/100773258/1',
'https://sports.donga.com/ent/article/all/20200423/100773216/1', 'https://sports.donga.com/ent/article/all/20200423/100773200/1',
'https://sports.donga.com/ent/article/all/20200423/100773088/1', 'https://sports.donga.com/ent/article/all/20200423/100773129/1',
'https://sports.donga.com/ent/article/all/20200423/100773094/1', 'https://sports.donga.com/ent/article/all/20200423/100773059/1']
코드 실행이 완료되었습니다.
```

크롤링한 결과는 다음과 같습니다.

네이트 최신뉴스 href 수집하기

네이트의 최신뉴스 목록에서

각 기사의 href를 수집해보겠습니다.

네이트 최신뉴스 href 수집하기

```
['http://news.nate.com/view/20200423n11258?mid=n0100', 'http://news.nate.com/view/20200423n11257?mid=n0100',
'http://news.nate.com/view/20200423n11256?mid=n0100', 'http://news.nate.com/view/20200423n11255?mid=n0100',
'http://news.nate.com/view/20200423n11249?mid=n0100', 'http://news.nate.com/view/20200423n11253?mid=n0100',
'http://news.nate.com/view/20200423n11252?mid=n0100', 'http://news.nate.com/view/20200423n11251?mid=n0100',
'http://news.nate.com/view/20200423n11250?mid=n0100', 'http://news.nate.com/view/20200423n09470?mid=n0100',
'http://news.nate.com/view/20200423n11247?mid=n0100', 'http://news.nate.com/view/20200423n11221?mid=n0100',
'http://news.nate.com/view/20200423n11245?mid=n0100', 'http://news.nate.com/view/20200423n11244?mid=n0100',
'http://news.nate.com/view/20200423n11243?mid=n0100', 'http://news.nate.com/view/20200423n11242?mid=n0100',
'http://news.nate.com/view/20200423n06012?mid=n0100', 'http://news.nate.com/view/20200423n11240?mid=n0100',
'http://news.nate.com/view/20200423n11239?mid=n0100', 'http://news.nate.com/view/20200423n11238?mid=n0100']
코드 실행이 완료되었습니다.
```

크롤링한 결과는 다음과 같습니다.

이전 실습과 비슷한 내용이니 연습해보세요.

sbs 뉴스 최신 기사 목록의 내용 수집하기

sbs 뉴스의 최신 기사 목록의 href를 추출하고

href로 접근할 수 있는 기사들의 내용을 추출해봅니다.

sbs 뉴스 최신 기사 목록의 내용 수집하기

각 기사의 href 주소를 얻는 get_href 함수와

기사의 내용을 얻는 crawling 함수를 각각

올바르게 구현해주세요.

sbs 뉴스 최신 기사 목록의 내용 수집하기

오거돈 부산시장이 사퇴를 결정했습니다.부산시는 "오 시장이 23일 오전 11시 기자회견을 열어 공식적으로 사퇴 의사를 밝힐 예정"이라고 밝혔습니다.오 시장은 일신상의 사유를 들어 사퇴 의사를 밝힌 것으로 알려졌습니다.오 시장은 최근 건강 이상설이 나왔습니다.21대 총선 하루 전인 14일 연가를 냈고, 선거 당일인 15일도 비공개 투표를 했습니다.이후에도 부산시청으로 출근은 했지만, 외부활동을 하지 않았습니다.오 시장 한 측근은 "일신상 이유로 사퇴하는 것은 맞지만 자세한 내용은 알 수 없다"고 말했습니다.부산지역 정가와 시청 안팎에서는 오 시장 사퇴 사유를 두고 다른 배경이 있는 것 아닌가 하는 얘기들이 나오고 있습니다.(사진=부산시 제공, 연합뉴스)
중국의 코로나19 확산이 진정된 가운데 시진핑 국가주석이 마스크를 쓴 채 초등학교를 방문해 강한 체력을 강조하고 나섰다.23일 중국중앙TV 등에 따르면 시진핑 주석은 지난 21일 산시(陝西)성 안강시(安康市) 평리현(平利縣) 시찰에서 마스크를 착용하고 초등학교 5학년 교실에 들어가 코로나19 속 학생들의 수업 상황 등을 살펴봤다.시진핑 주석은 이들 학생에게 학습 및 생활 상황을 물어보면서 "요새 아이들은 대부분 안경을 쓰고 있다"면서 "이게 은근히 걱정되는 부분"이라고 말했다.시 주석은 "체력 단련 부족으로 아이들의 신체 및 건강 수준도 낮아졌다"면서 "아이들은 밝은 정신과 강건한 체력이 있어야 하며 강건한 체력이란 바로 신체를 튼튼히 하는 것"이라고 언급했다.시진핑 주석은 이날 한 이주민의 집도 방문해 마스크를 쓴 채 주민들과 소파에 같이 앉아 담소를 나눴다.이 자리에서 "거주가 안정되려면 안정된 취업이 있어야 한다"면서

크롤링한 결과는 다음과 같습니다.

다양한 섹션의 속보 기사 href 추출하기

네이버 뉴스 속보 페이지에는 여러 섹션이 있습니다.

다양한 섹션의 속보 기사 href 추출하기

https://news.naver.com/main/list.nhn?**sid1=100**

URL의 쿼리 부분에서 sid1의 값에 따라 섹션이 결정됩니다.

어떤 섹션을 크롤링할지는 input 함수로 입력하세요.

다양한 섹션의 속보 기사 href 추출하기

쿼리와 함께 get 요청을 담고 있는 **requests 객체**를 반환하는

get_request 함수와,

섹션별로 나뉘어진 목록에 있는 기사들의 **href를 추출**하는

get_href 함수를 올바르게 구현하세요.

다양한 섹션의 속보 기사 href 추출하기

```
"정치", "경제", "사회", "생활", "세계", "과학" 중 하나를 입력하세요.
  > 정치
['https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=277&aid=0004666877',
'https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=277&aid=0004666877',
'https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=047&aid=0002267478',
'https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=047&aid=0002267478',
'https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=018&aid=0004626432',
'https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=018&aid=0004626432',
'https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=003&aid=0009831311',
'https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=003&aid=0009831311',
'https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=001&aid=0011567544',
```

크롤링한 결과는 다음과 같습니다.

다양한 섹션의 속보 기사 내용 추출하기

이번에는 섹션별 기사의 **내용**을 추출해보도록 하겠습니다.

다양한 섹션의 속보 기사 내용 추출하기

기사의 href에서 내용을 추출할 때, **본문 영역**은 **태그**가 없습니다.

따라서, 기사 본문을 싸고 있는 **div** 태그의 **children**을 얻고,

각 children의 **name이 None**인 요소만 추출해야 합니다.

다양한 섹션의 속보 기사 내용 추출하기

그 뿐만 아니라 HTML 문서에 적혀있는 주석도

걸림돌이 될 수 있으므로 어려울 수 있는 실습입니다.

어려움을 겪으신다면 정답 코드가 담긴 solution.py 파일을

확인해보시거나, 튜터에게 문의해주세요.

다양한 섹션의 속보 기사 내용 추출하기

```
"정치", "경제", "사회", "생활", "세계", "과학" 중 하나를 입력하세요.
  > 세계
['KOSDAQ 643.79 UP 8.63 points (close)(END) ', 'KOSPI 1,914.73 UP 18.58 points (close)(END) ', '[서울=뉴시스] 김예진 기자 = 23일 일본 도쿄증시에서 닛케이225지수(닛케이평균주가)는 전일 대비 291.49 포인트, 1.52% 상승하며 1만 9429.44에 장을 마감했다. 4 거래일 만에 상승 마감했다.JPX 닛케이 인덱스 400지수는 전일 대비 157.30 포인트, 1.25% 오른 1만 2782.45에 거래를 마쳤다. 토픽스(TOPIX)지수는 전일 대비 19.08 포인트, 1.36% 상승한 1425.98에 시장을 마무리했다.니혼게이자이 신문에 따르면 미국에서 경제 활동 재개 기대가 고조되고 국제유가가 상승하자 투자자의 심리가 개선됐다. 해외 투자자들에 의해 일본 증시가 상승했다. aci27@newsis.com<© 공감언론 뉴시스통신사. 무단전재-재배포 금지> ', '[서울=뉴시스] 김예진 기자 = 23일 일본 도쿄증시에서 닛케이225지수(닛케이평균주가)는 전일 대비 291.49 포인트, 1.52% 상승하며 1만 9429.44에 장을 마감했다. 4 거래일 만에 상승 마감했다.JPX 닛케이 인덱스 400지수는 전일 대비 157.30 포인트, 1.25% 오른 1만 2782.45에 거래를 마쳤다. 토픽스(TOPIX)지수는 전일 대비 19.08 포인트, 1.36% 상승한 1425.98에
```

크롤링한 결과는 다음과 같습니다.

특정 영화 리뷰 추출하기

네이버 영화 페이지의 영화 리뷰의 제목을 크롤링하겠습니다.

이번에는 정보를 얻고자 하는 영화의 제목이 입력으로 주어지고,

해당 영화에 대한 리뷰 결과를 보여주어야 합니다.

특정 영화 리뷰 추출하기

이를 위해 get_url, get_href, crawling

세 개의 함수를 올바르게 구현해주셔야 합니다.

특정 영화 리뷰 추출하기

get_url은 영화 제목을 입력받고,

해당 제목을 **검색하였을 때 나오는 URL**을 반환해야 합니다.

requests.get 메소드의 params 매개변수를 이용해도 되지만,

이번에는 문자열의 결합을 이용하는 편이 더 간편하기 때문에

get_url은 **문자열의 결합**을 이용하여 URL을 만듭니다.

특정 영화 리뷰 추출하기

'스타워즈'에 대한 영화 통합검색결과 입니다.

◉ 전체 ○ 영화 ○ 영화인 ○ 영화제 ○ 영화사 ○ 극장 ○ 이미지 ○ 동영상 ○ 배역

영화 (1-5 / 26건) : **정확도순** · 제작년도순

스타워즈: 라스트 제다이 (Star Wars: The Last Jedi)

★★★☆☆ **6.44** (참여 7110명)

액션, 모험, 판타지, SF | 미국 | 152분 | 2017

감독 : 라이언 존슨 | 출연 : 데이지 리들리, 마크 해밀, 오스카 아이삭, 아담 드라이버, 캐리 피셔, 존 보예가

다운로드

스타워즈 에피소드 3 - 시스의 복수 (Star Wars: Episode III - Revenge Of The Sith)

★★★★½ **9.09** (참여 3024명)

SF, 모험, 액션 | 미국 | 139분 | 2005

감독 : 조지 루카스 | 출연 : 이완 맥그리거, 나탈리 포트만, 헤이든 크리스텐슨, 이언 맥디어미드, 프랭크 오즈

다운로드

네이버 영화 페이지에서 키워드를 검색했을 때 화면입니다.

get_url 함수는 이 때의 URL을 반환하면 됩니다.

특정 영화 리뷰 추출하기

get_href는 검색 결과,

가장 위에 있는 영화의 href를 반환합니다.

특정 영화 리뷰 추출하기

마지막으로 crawling 함수를 구현하여

get_href에서 얻은 영화의 href로 접근하고,

해당 영화의 리뷰 목록을 크롤링하세요.

특정 영화 리뷰 추출하기

```
영화 제목을 입력하세요.
  > 스타워즈
['루카스 없는 스타워즈 = 영혼 없는 라이트 세이버 (스포주의)', '시리즈의 정점을 찍다 <스타워즈:라스트 제다이>', '갈수록 재미가 없어져...', '(스포)최고의 소재로 최악의 영화를 만들다', '(스포,분노) 이제부터 이 영화는 포스닥이 입니다.', "[영화 분석(스포일러有)] '모두'를 위한 '마지막' 제다이", '역대 최악의 스타워즈 영화.', '스타워즈를 3류 영화로 전락시킨 쓰레기', "한마디로 '개망작' 이라 생각합니다. (스포 많아요. 영화 보신분만 리뷰 보세요...)", '스타워즈: 라스트 제다이 분노해서 쓸수밖에 없었던 리뷰']
코드 실행이 완료되었습니다.
```

크롤링한 결과는 다음과 같습니다.


/*elice*/
contact@elice.io

/*elice*/

파이썬 크롤링

API를 이용한 크롤링

김경민 선생님

API

API란?

API(Application Programming Interface)는

어떤 **프로그램**과 또 다른 **프로그램**을 연결해주는 **매개체**입니다.

컴퓨터를 다루기 위해 **마우스**와 **키보드**를 이용하는 것처럼

API는 프로그램 사이를 연결해주는 역할을 합니다.

API란?

예를 들어 **지도 데이터**를 이용하여

맛집 찾기 웹 서비스를 제작하려면 어떻게 해야 할까요?

API란?

보통의 일반인들에게는 지도 데이터를 **갖고 있지 않고,**

이를 **수집**하는 것도 매우 어렵습니다.

그렇다고 **공개된 데이터**를 그대로 사용하는 것도 어렵습니다.

API란?

Google이 갖고 있는 지도 데이터를 공개하였다고 가정해봅시다.

API란?

그러나 **원본** 데이터는 너무 방대하기도 하고,

호환성 등의 문제도 있어 **쉽게 사용할 수 없습니다.**

마치 키보드와 마우스가 없는 컴퓨터를 사용하는 것과 같습니다.

API란?

그래서 Google은 지도 데이터를 응용하여 사용할 수 있도록

Google Map API라는 **매개체**를 사용자들에게 **제공**합니다.

API란?

위 사진은 daum 증권 사이트입니다.

여러 기업들의 주가 정보를 **API**를 거쳐 받아온 후 표시하고 있습니다.

API란?

방금 보신 daum 증권 사이트와 같이

API를 이용해 정보를 가져오는 웹 사이트가 꽤 있습니다.

이런 경우 정보가 HTML에 **처음부터 존재하지 않고,**

정보를 **API로부터 불러오고 나서** HTML에 존재하게 됩니다.

API란?

따라서 daum 증권 사이트에서는

BeautifulSoup를 이용하여 주가 데이터를 크롤링할 수 없습니다.

웹 사이트를 처음 로드할 때 HTML 문서에는

주가 데이터가 존재하지 않기 때문입니다.

API란?

보통 API를 이용하여 데이터를 불러오는 경우는

데이터가 '동적'으로 변화하는 일이 많아

실시간으로 값을 불러와야 하는 경우입니다.

기업의 주가도 하나의 예시입니다.

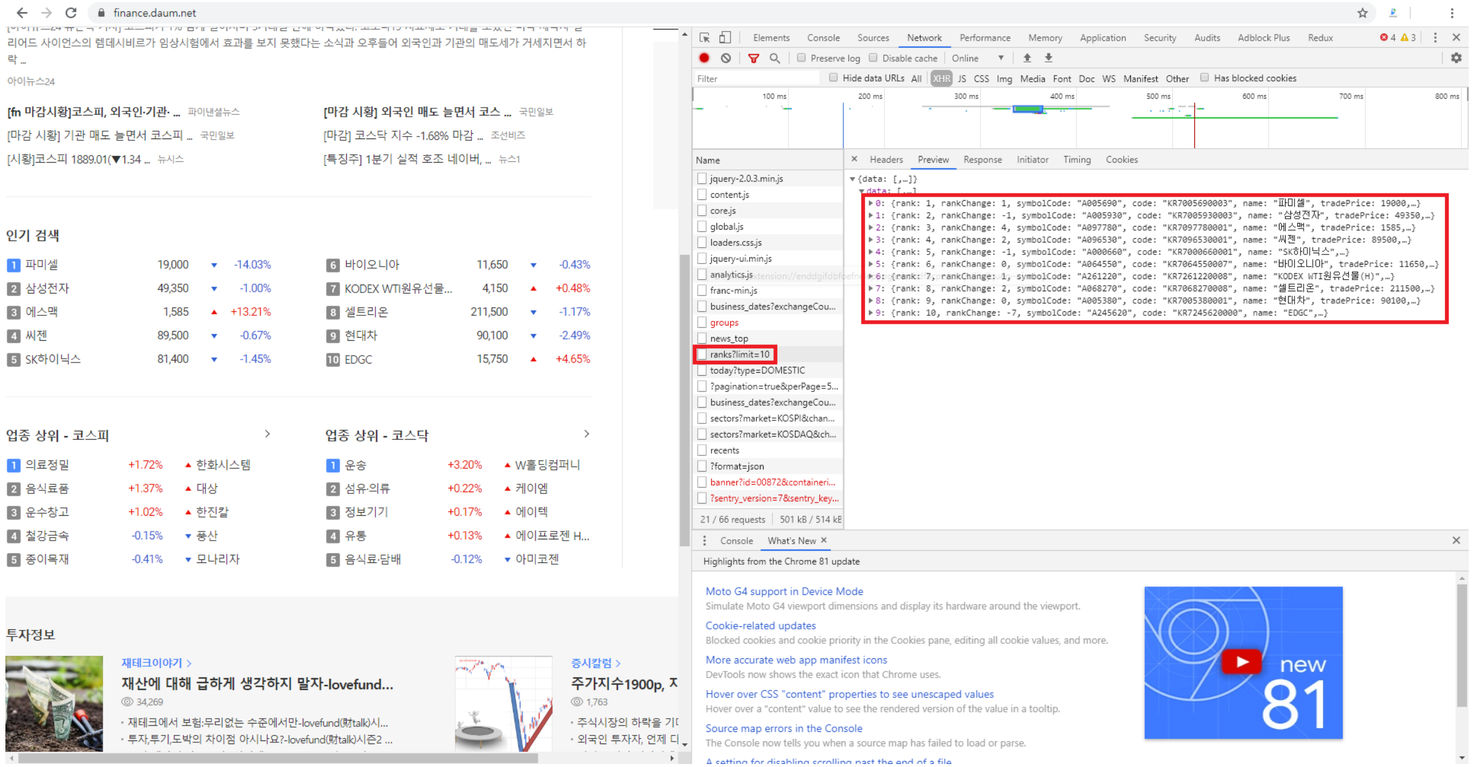
API란?

이럴 땐 daum 증권 사이트에서

주가 정보를 요청하는 **API에 접근**하여

어떤 정보를 전달해주고 있는지 접근하면 됩니다.

API란?



크롬 개발자 도구의 **Network** 탭에서

웹사이트가 데이터를 요청하는 API를 볼 수 있습니다.

API란?

```
url = "http://finance.daum.net/api/search/ranks?limit=10"
req = requests.get(url) # JSON 데이터
```

API의 URL에 GET 요청을 보내면 **JSON 데이터**를 얻을 수 있습니다.

JSON은 **key**와 **value**를 저장하는, 딕셔너리 꼴의 데이터 형식입니다.

API란?

몇몇 웹 사이트들은 크롤러 등을 통한 기계적인 접근을 막고 있습니다.

이를 우회하기 위해 requests.get 메소드에

"headers" 매개변수를 지정해주셔야 합니다.

API란?

‘헤더’란 HTTP 상에서 클라이언트와 서버가

요청 또는 응답을 보낼 때 전송하는 **부가적인 정보**를 의미합니다.

실습에서 headers에 사용할 옵션을 제공하고 있습니다.

API란?

```
custom_header = {
    'referer’ : ...
    'user-agent’ : ...  }
```

referer와 **user-agent** 옵션을 지정하고 있는데,

referer는 **이전 웹 페이지의 주소**를 의미하고

user-agent는 이용자의 여러 가지 **사양**을 의미합니다.

/*elice*/

[실습1]

필요한 정보를 담고 있는 API에 접근

/*elice*/

[실습2]

네이버 실시간 검색어 크롤링

프로젝트 – 음식점 리뷰 크롤링하기

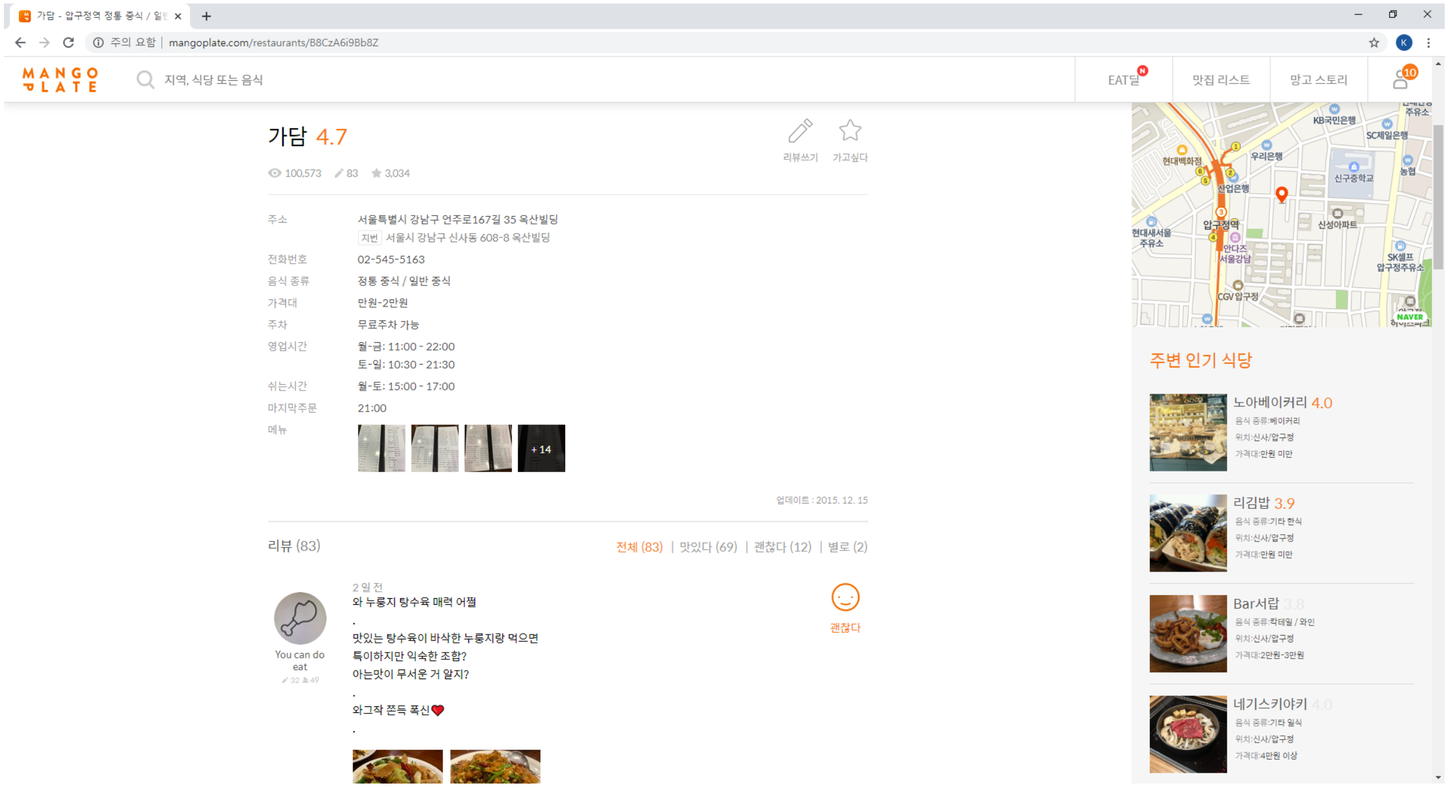
음식점 리뷰 크롤링

음식점을 소개하고 추천하는 서비스를 제공하는 웹 사이트인

‘망고플레이트’의 데이터로 음식점 리뷰를

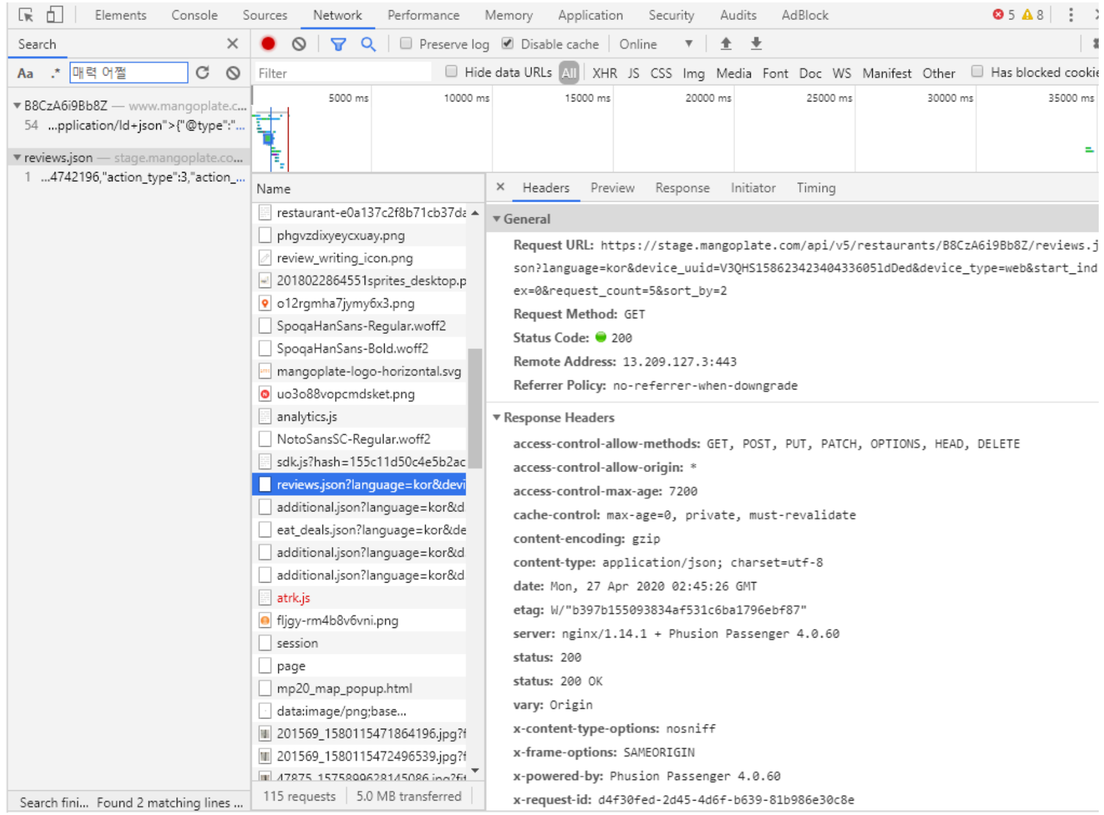
크롤링하는 실습을 만들어보겠습니다.

음식점 리뷰 크롤링



음식점 페이지에 들어가면, 하단에 리뷰가 있습니다.

음식점 리뷰 크롤링



개발자 도구를 통해 리뷰를 불러오는 곳을 조사할 수 있습니다.

음식점 리뷰 크롤링

리뷰를 불러오는 API의 URL에 접근하여

어떤 음식점의 URL이 주어졌을 때 **해당 음식점의 리뷰**를

모두 불러오는 코드를 작성해보세요.

음식점 리뷰 크롤링

그리고 **특정 키워드를 검색**하였을 때

나타나는 음식점들의 href으로 접근하여

여러 음식점들의 리뷰를 불러오는 코드도 작성해보세요.

/*elice*/

[실습3]

음식점 리뷰 크롤링

/*elice*/

[실습4]

음식점 href 링크 크롤링

/*elice*/

[실습5]

검색 결과 음식점 리뷰 크롤링


/*elice*/
contact@elice.io

/*elice*/

파이썬 크롤링

워드 클라우드 프로젝트

김경민 선생님

워드클라우드

워드클라우드란?

데이터에서 단어 빈도를 분석하여 시각화하는 기법

워드클라우드 준비

워드클라우드를 그리기 위해서 **텍스트 데이터**가 필요합니다.

네이버 **뉴스 기사**의 내용의 텍스트 데이터로

워드클라우드를 그려보도록 하겠습니다.

영어 문장 나누기

워드클라우드의 각 단어는 **빈도**에 따라 크기가 결정됩니다.

크기가 큰 단어일수록 **빈도**가 높습니다.

영어 문장의 경우, 공백을 기준으로 나누어

각각의 단어를 얻을 수 있습니다.

영어 문장 나누기

영어로 이루어진 텍스트 데이터가 주어집니다.

텍스트 데이터를 **공백**을 기준으로 나누어 빈도를 조사하고,

이를 바탕으로 워드클라우드를 그려보도록 하겠습니다.

/*elice*/

[실습1]

영어 문장 나누기

/*elice*/

[실습2]

워드클라우드 출력하기

네이버 뉴스 기사 워드클라우드

본격적으로 네이버 뉴스 기사의

워드클라우드를 그려보도록 하겠습니다.

이전 장의 실습에서 활용했던 코드로

네이버 뉴스 기사의 내용을 크롤링하세요.

네이버 뉴스 기사 워드클라우드

원하는 기사의 URL을 입력하시고,

워드클라우드를 출력해보신 후 출력된 모습을 관찰해보세요.

/*elice*/

[실습3]

네이버 뉴스 기사 내용 크롤링하기

/*elice*/

[실습4]

네이버 뉴스 기사 워드클라우드 출력하기

이전 실습의 문제점

이전 실습에서 그렸던 워드클라우드의 문제점은

단어에 **어미**와 **조사**가 붙어 분석이 왜곡되는 것입니다.

예를 들어 '대통령이' 와 '대통령은'은 둘 다

대통령이라는 공통된 키워드로 집계되어야 합니다.

형태소 추출

이를 추출하기 위해 한국어 단어에 붙는

어미와 조사를 제거하고, 단어의 어근만 집계되도록 하는

형태소 추출 과정이 필요합니다.

형태소 추출

이 과정에서 한국어 자연어 처리 라이브러리인

mecab을 사용합니다.

/*elice*/

[실습5]

형태소 추출하기

/*elice*/

[실습6]

형태소를 추출한 워드클라우드 출력하기

여러 개의 기사 내용 크롤링하기

하나의 기사만으로는 단어의 빈도수를 파악하기 어려울 수 있습니다.

기사의 분량, 기자의 성향 등 여러 요인이 반영되기 때문입니다.

여러 개의 기사 내용 크롤링하기

따라서 공통된 주제에 대한 **여러 기사**의 텍스트 데이터를 같이 분석하면

효과적인 워드클라우드를 출력할 수 있습니다.

여러 개의 기사 내용 크롤링하기

네이버 뉴스 페이지는 관련된 주제의

여러 기사를 묶어서 보여주고 있습니다.

여러 개의 기사 내용 크롤링하기

각각의 분야(정치, 경제, 사회, 생활, 세계, 과학)에 대해

페이지 최상단에 보이는 **주제에 해당하는 기사**들의

텍스트 데이터로 워드클라우드를 출력해봅시다.

/*elice*/

[실습7]

여러 개의 기사 내용 크롤링하기

/*elice*/

[실습8]

여러 개의 기사 내용으로 워드클라우드 출력하기

더 많은 기사 내용 크롤링하기

이전 실습으로 각 주제마다 **3~4개 기사**의

텍스트 데이터를 크롤링 할 수 있게 되었습니다.

이 상태에서, **더 많은 기사**의 내용을 크롤링하면

텍스트 데이터를 풍부하게 만들 수 있습니다.

더 많은 기사 내용 크롤링하기

기사 페이지에서 더 많은 기사를 확인할 수 있습니다.

더 많은 기사 내용 크롤링하기

세부 페이지에서 더 많은 기사에 각각 접근하실 수 있습니다.

/*elice*/

[실습9]

더 많은 기사 내용 크롤링하기

/*elice*/

[실습10]

더 많은 기사로 워드클라우드 출력하기

CREDIT

코스 매니저
김경민

콘텐츠 제작자
김경민

강사
김경민

감수자
장석준, 강윤수

디자이너
박주연


/*elice*/
contact@elice.io