/*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

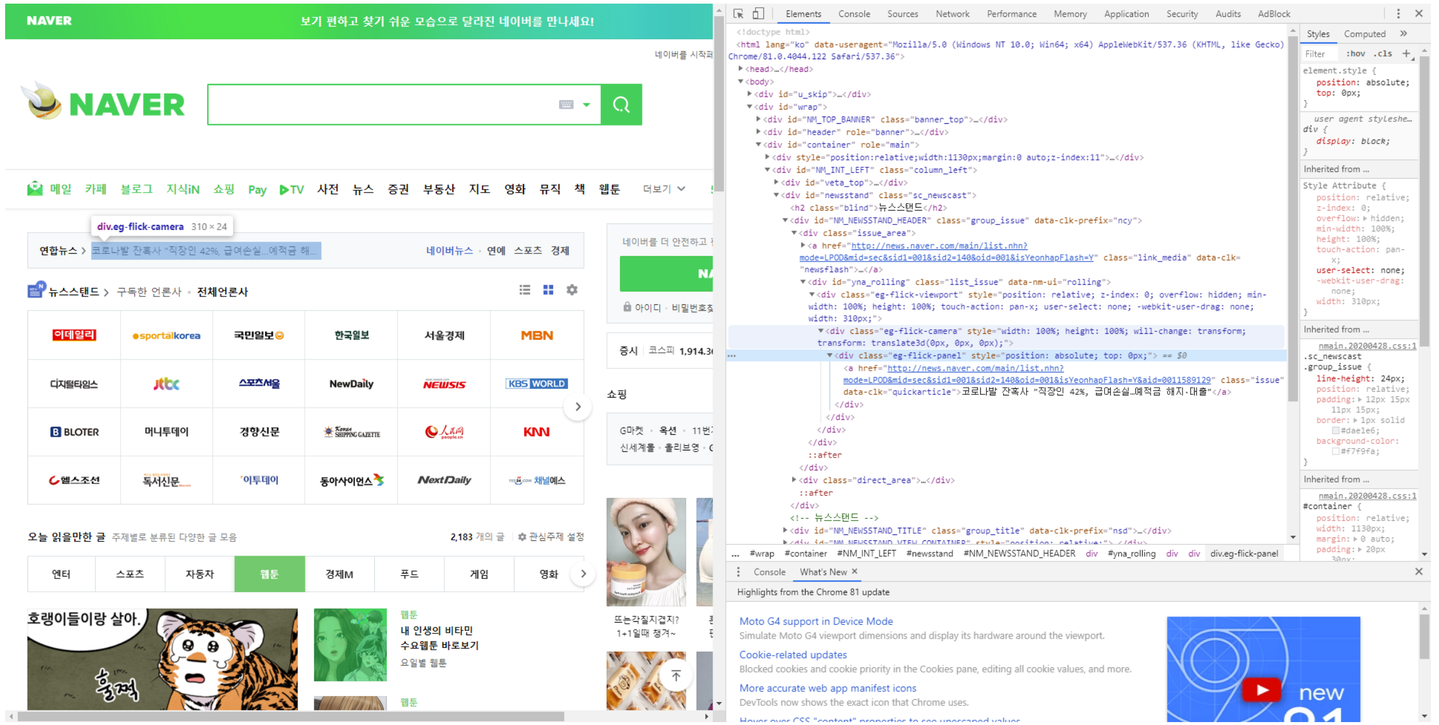
웹 페이지의 HTML을 분석할 수 있습니다.

실전 크롤링

실전 크롤링

배운 내용으로 크롤링 코드를 직접 작성해보도록 하겠습니다.

실전 크롤링



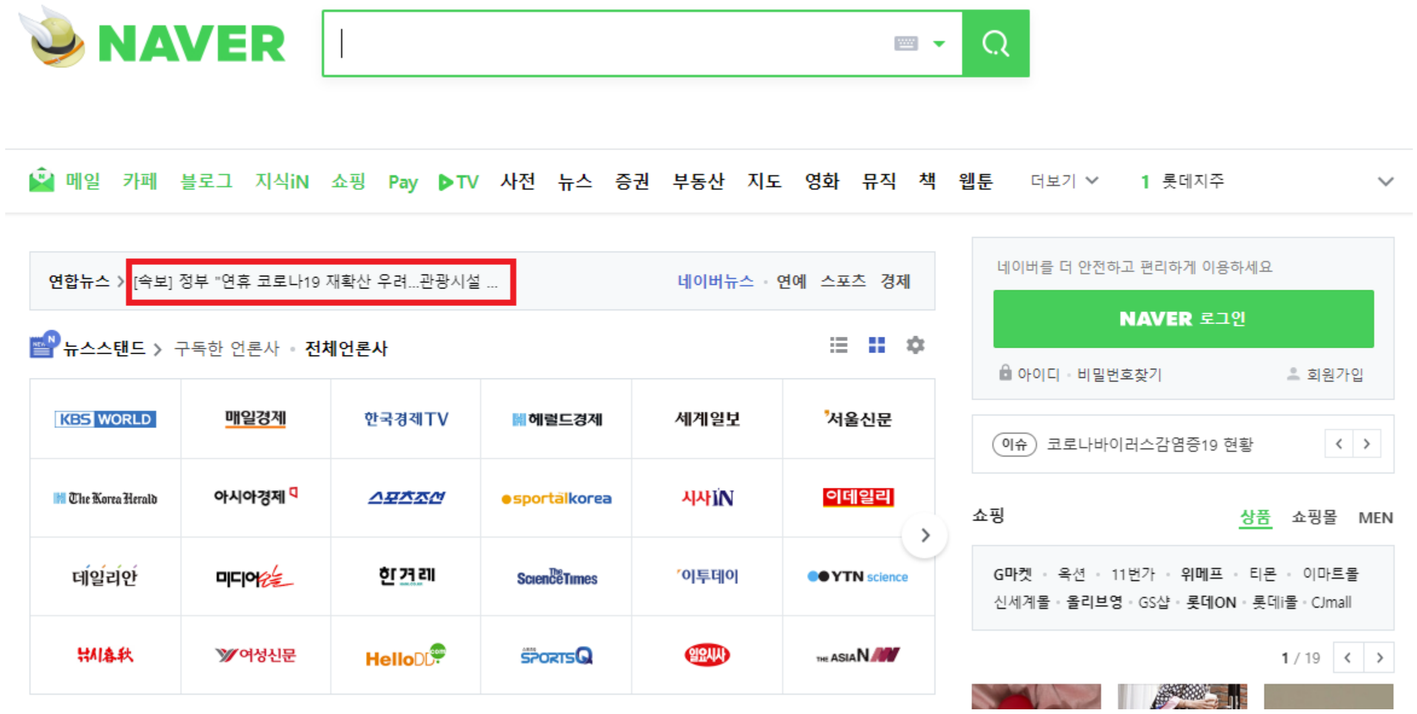
웹 페이지에서 **F12** 버튼을 눌러 개발자 도구를 켤 수 있습니다.

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.

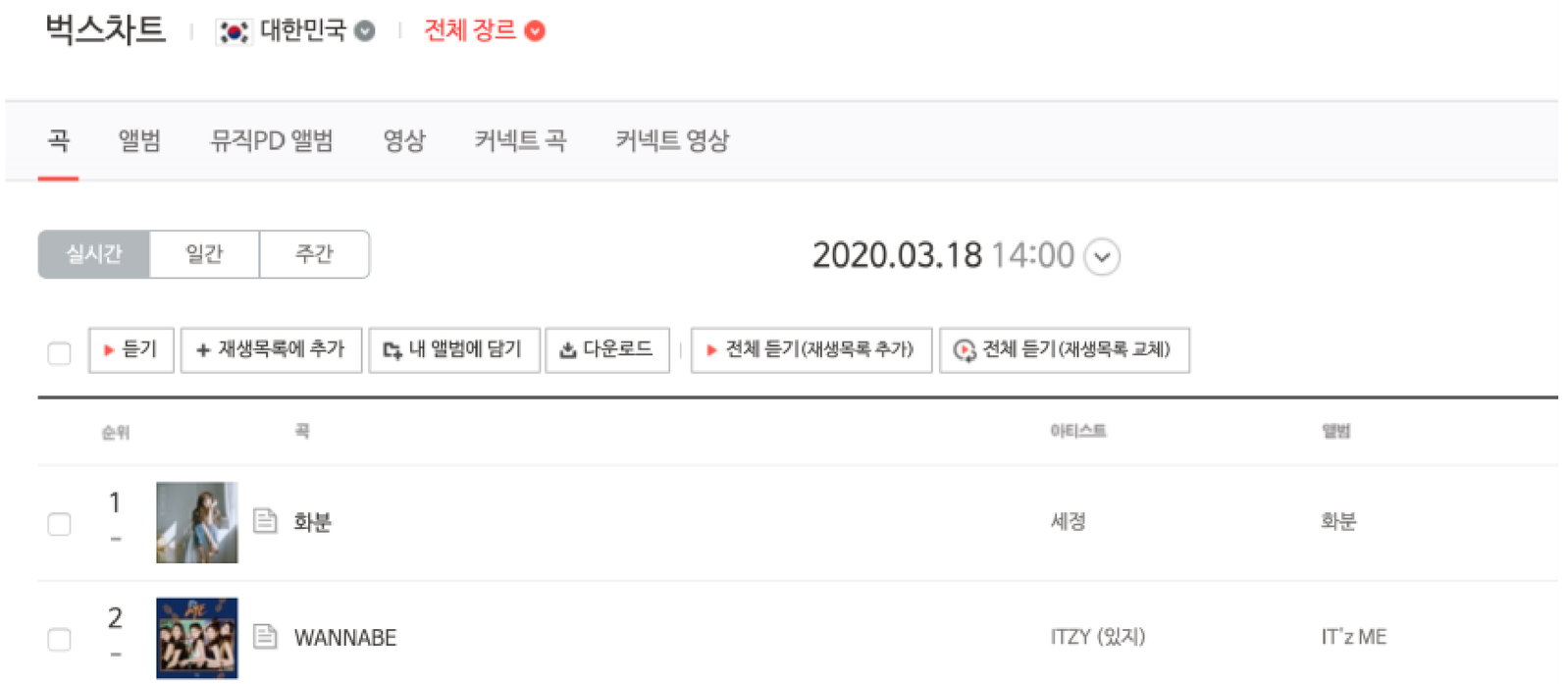
연합뉴스 속보 기사 제목 추출하기

```
['\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.

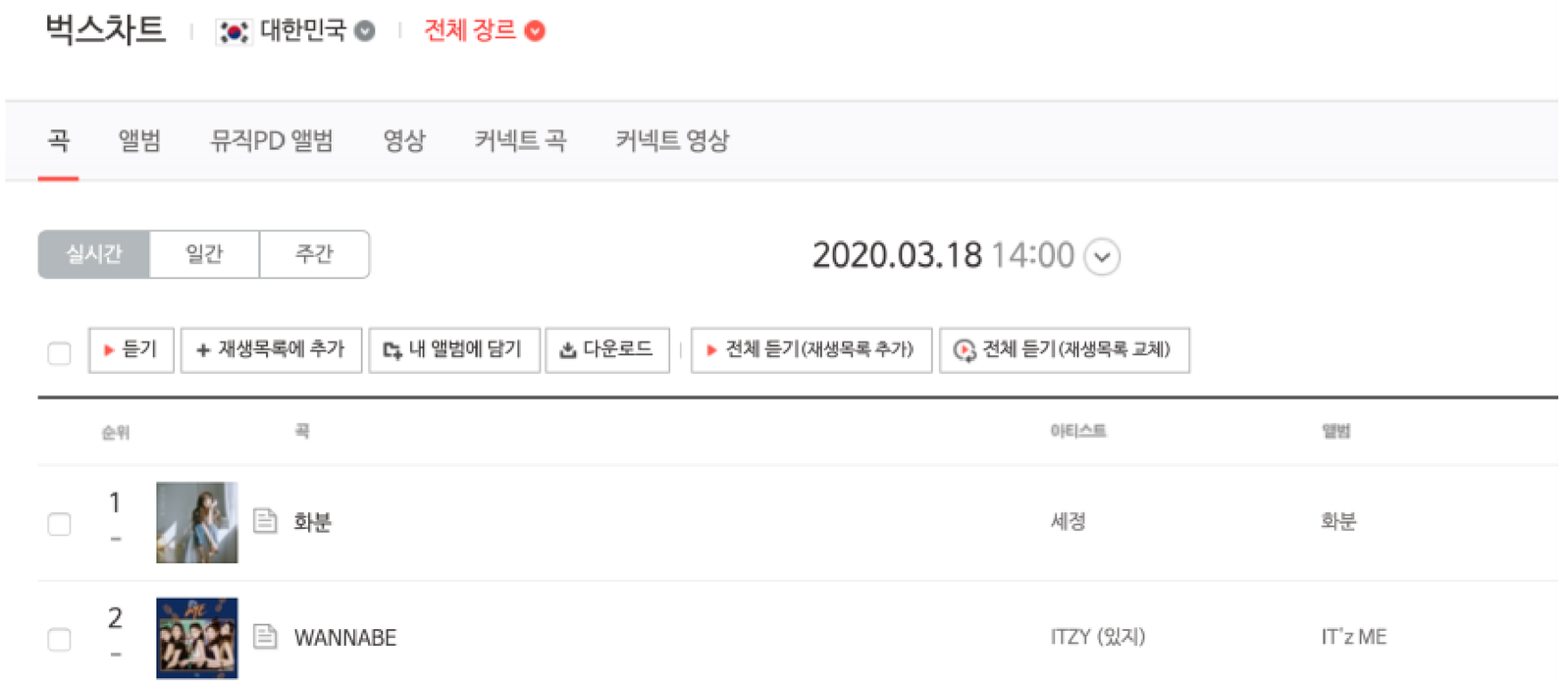
bugs 실시간 음원차트 순위 추출하기



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']
```

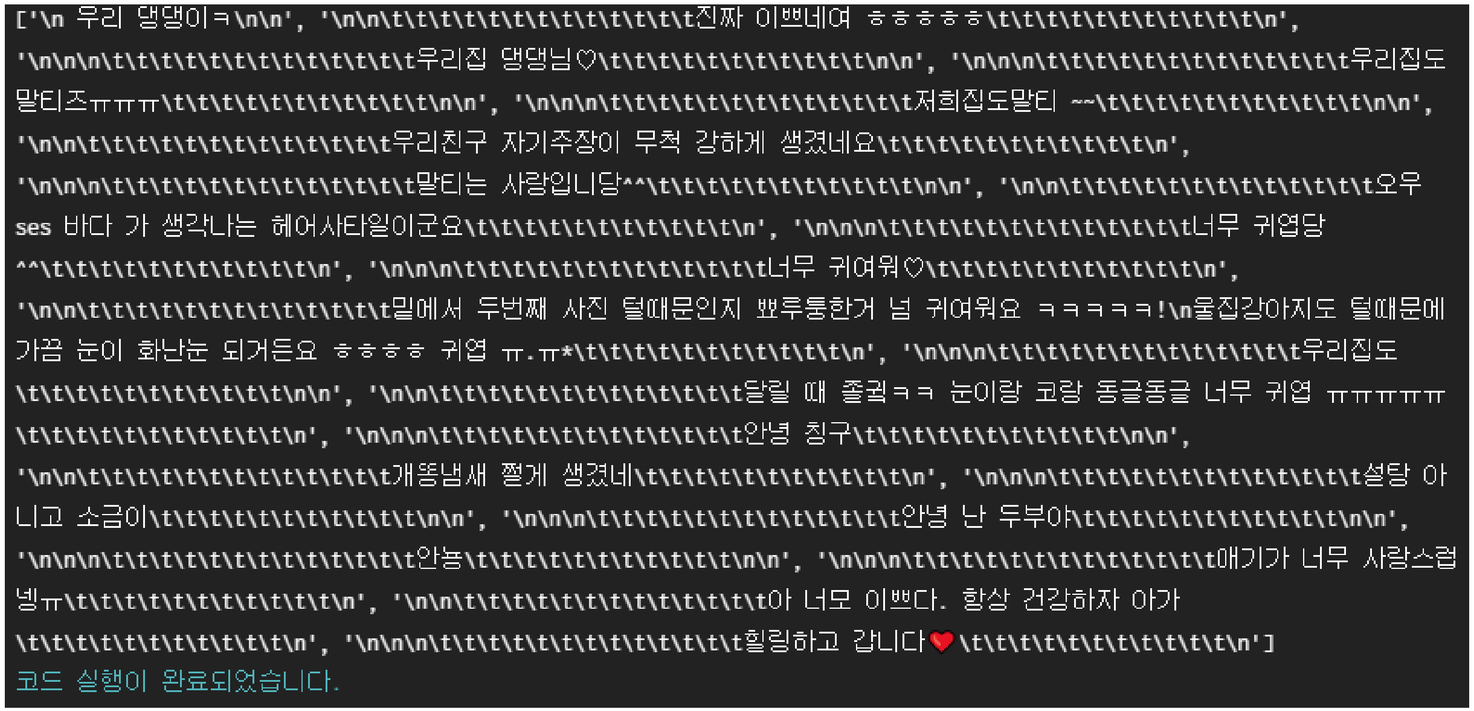
경우에 따라서 find 함수를 3중으로 사용해야 할 수도 있으니

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