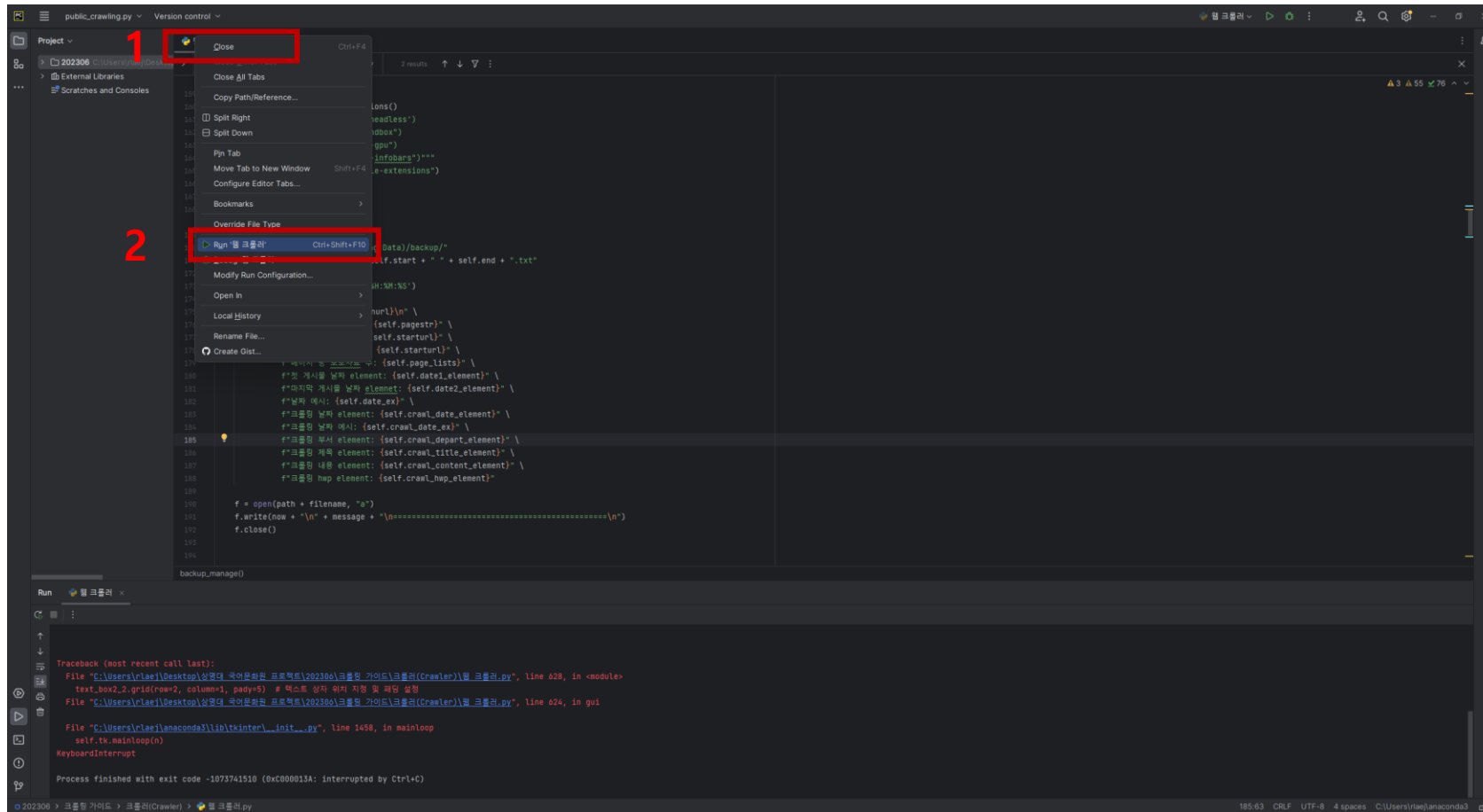


크롤링 가이드 설명서

2023.07.10 작성

크롤링 프로그램 실행

파이참에서 웹 크롤러.py를 실행하고 1번 탭에서 우클릭 후 2번을 클릭하여 실행합니다.



크롤링 프로그램 사용 방법1

보도자료 크롤링

1 기관명 청양군청

2 시작 날짜 20230701

3 끝 날짜 20230710

한글파일 4 ☐ 파일 여부

다음

© SMU KLCC

1. 기관명에는 크롤링을 하려는 기관의 이름을 입력합니다. ex) 청양군청
2. 시작날짜에는 크롤링을 하려는 날짜의 시작 년월일을 입력합니다. ex)20230701
3. 끝 날짜에는 크롤링을 하려는 날짜의 끝 년월일을 입력합니다. ex)20230710
4. 기관의 보도자료 내용이 한글파일을 크롤링하여 받아오는 지 체크합니다.
ex) 청양군청의 경우 한글 파일이 존재하지 않으며 한글파일을 크롤링 할 필요가 없으므로 체크하지 않았습니다.
5. 다음 버튼을 누릅니다.

크롤링 프로그램 사용 방법2

1

보도자료 크롤링

메인 url

page를 나타내는 값

보도자료 처음 url

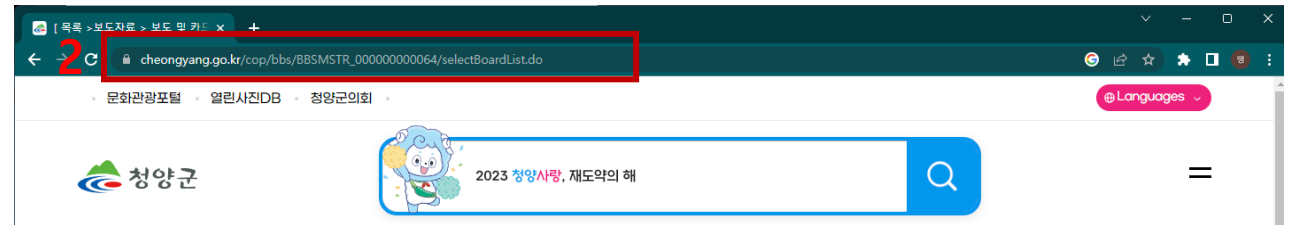
보도자료 마지막 url

페이지 당 보도자료 수

첫 게시물 날짜 element

마지막 게시물 날짜 element

날짜 예시



(2번 탭)보도자료 페이지의 url(링크) 를 복사하여
1번에 붙여넣기로 입력합니다.

크롤링 프로그램 사용 방법2

보도자료 크롤링

메인 url

1 page를 나타내는 값

보도자료 처음 url

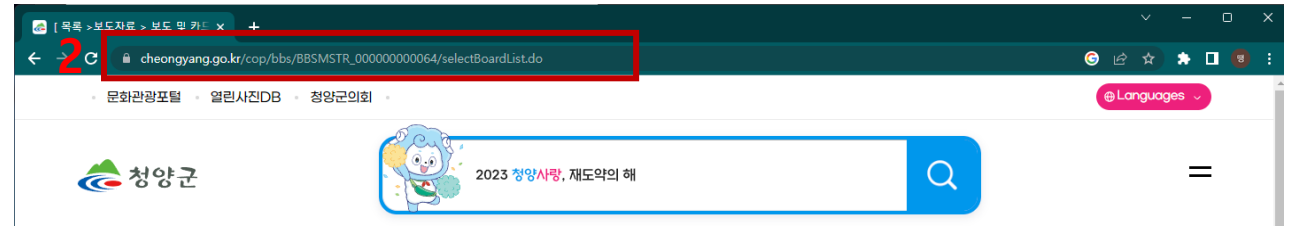
보도자료 마지막 url

페이지 당 보도자료 수

첫 게시물 날짜 element

마지막 게시물 날짜 element

날짜 예시



(2번 탭) 현재 보도자료 페이지는 1페이지 입니다.
그래서 2번 탭의 url에서는 표기가 되어있지 않습니다.

만약 1페이지의 보도자료에서 'page=1' 또는
'pageIndex=1' 과 같이 페이지를 나타내는 값이 포함되어
있다면 6, 7페이지를 무시하고 넘어가면 됩니다.

해당 2번 탭의 예시에는 포함되어 있지 않지만, 만약
Url이

https://www.cheongyang.go.kr/cop/bbs/BBSMSTR_000000000064/selectBoardList.do?pageIndex=1

이와 같다면 'pageIndex'을 page를 나타내는 값(1번 탭)에
입력하면 됩니다.

크롤링 프로그램 사용 방법2

보도자료 크롤링

메인 url

1 page를 나타내는 값

보도자료 처음 url

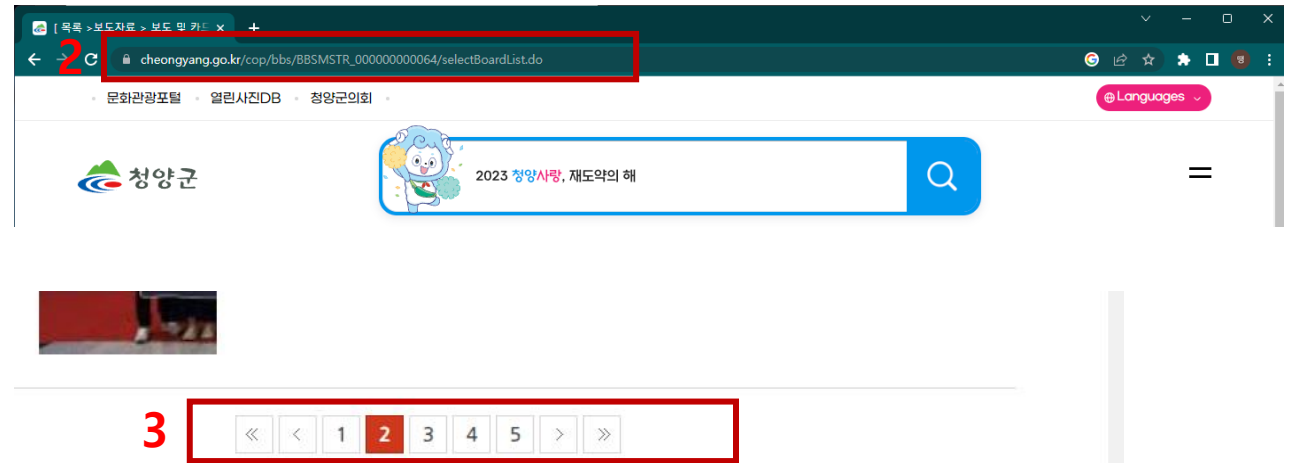
보도자료 마지막 url

페이지 당 보도자료 수

첫 게시물 날짜 element

마지막 게시물 날짜 element

날짜 예시



(3번 탭)보도자료 페이지를 2페이지로 이동하였을 때
(2번 탭)의 url이

https://www.cheongyang.go.kr/cop/bbs/BBSMSTR_000000000064/selectBoardList.do?pageIndex=2

'page=2' 또는 위 처럼 'pageIndex=2' 와 같다면 1번 탭에
page 또는 pageIndex를 입력하면 됩니다.

이처럼 되지 않는다면 다음 페이지를 참고하면 됩니다.
위와 같다면 7 페이지를 무시하고 넘어가면 됩니다.

크롤링 프로그램 사용 방법2

1

4

보도자료 크롤링

메인 url 00064/selectBoardList.do

page를 나타내는 값 pageIndex

보도자료 처음 url

보도자료 마지막 url

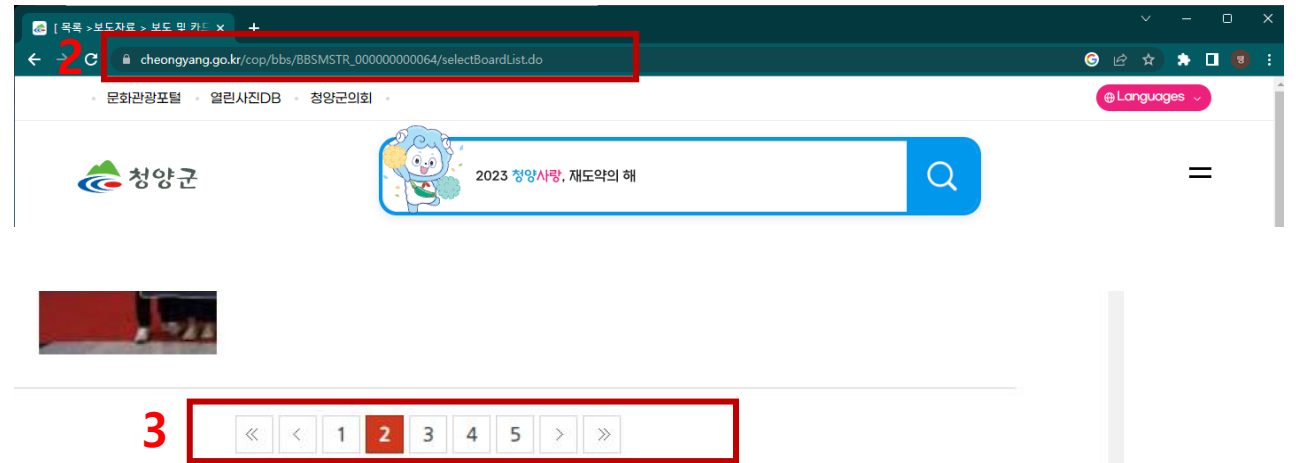
페이지 당 보도자료 수

첫 게시물 날짜 element

마지막 게시물 날짜 element

날짜 예시

이전 다음



(2번 탭)의 url에 아래 링크와 같이

https://www.cheongyang.go.kr/cop/bbs/BBSMSTR_000000000064/selectBoardList.do?pageIndex=2

'page=2' 또는 'pageIndex=2'를 입력했을 때,

3번 탭 처럼 페이지가 2번으로 이동한다면, 그리고 2가 아닌 3, 4, 5로 바뀌서 입력했을 때 페이지가 3, 4, 5로 이동한다면, 4번 탭에 page 또는 pageIndex를 입력하면 됩니다. (위의 예시로는 pageIndex를 입력해야 이동합니다.)

크롤링 프로그램 사용 방법2

1

2

3

메인 url	00064/selectBoardList.do
page를 나타내는 값	pageIndex
첫 게시물 url element	=\"txt\"/div[2]/div[1]/div[2]/a
마지막 게시물 url element	\"txt\"/div[2]/div[10]/div[2]/a
페이지 당 보도자료 수	10
첫 게시물 날짜 element	
마지막 게시물 날짜 element	
날짜 예시	
이전	다음


보도자료 페이지의
첫 게시물과 마지막 게시물로
이동을 위한 url 요소를 파싱하여 알아내야 합니다

이 방법에 대한 설명이 다음 페이지부터 적혀있습니다.

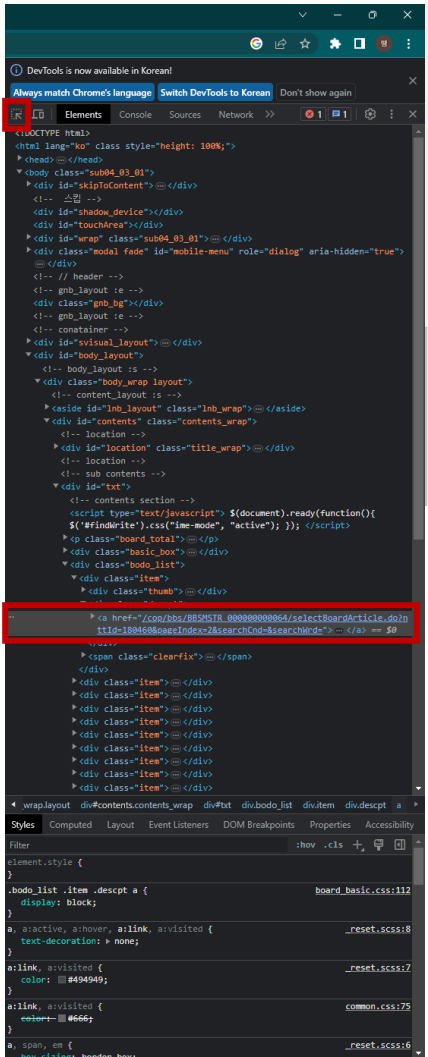
보도자료 페이지에 게시물이 총 몇 개가 있는 지를 세고
3번 탭에 입력합니다.

크롤링 프로그램 사용 방법2

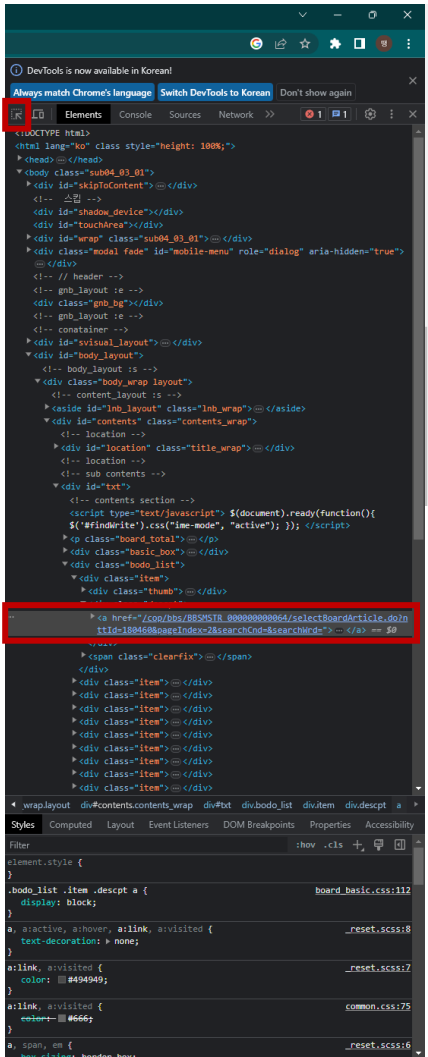
1



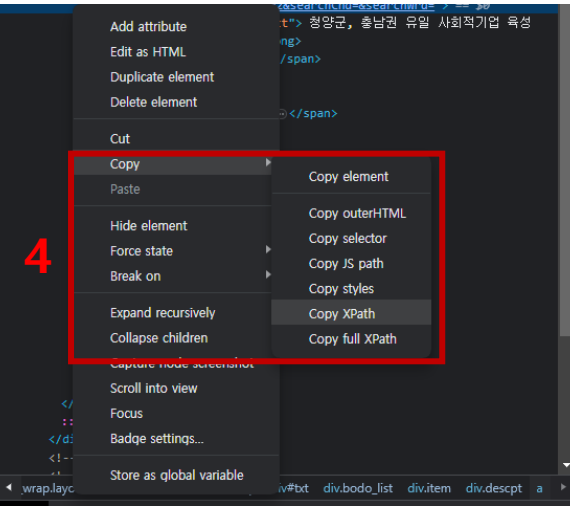
2



3



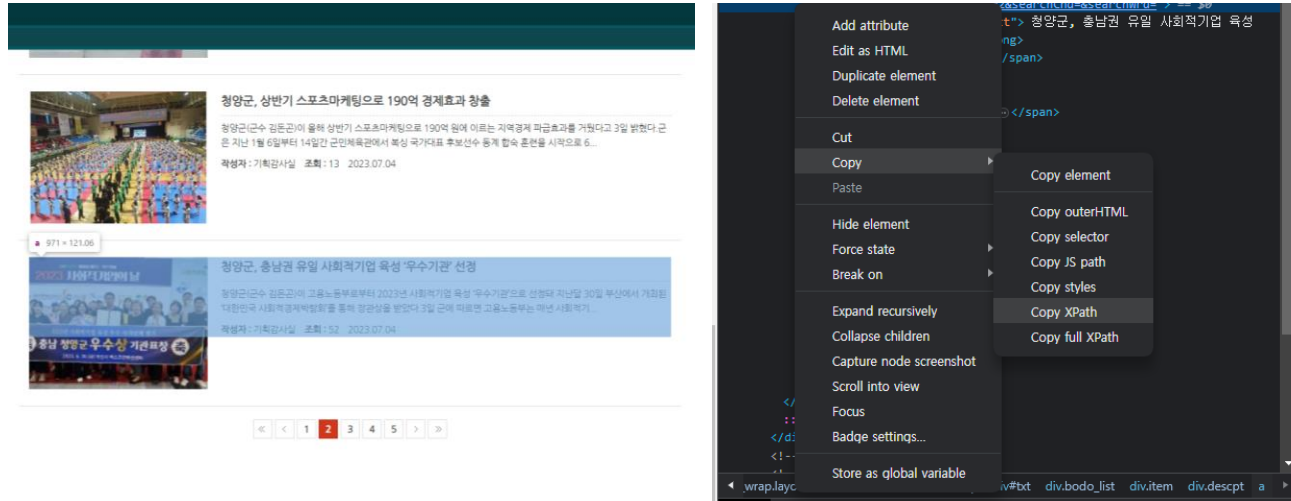
4



첫 게시물 url element를 파싱하기 위해
[처음 보도자료 웹페이지에서 F12키를 눌러
2번 탭을 선택합니다.

마우스를 첫 게시물 (1번 탭)을 클릭합니다.
3번 탭에 '<a href' 라고 표기되어 있는 부분을
클릭하였을 때, 1번 탭 부분이 사진과 같이 파란 박스가
생기면 됩니다.
이 3번 탭 부분을 우클릭하고 4번 탭 부분의 Copy-> Copy Xpath를
클릭하고] 8번 페이지의 1번에 붙여넣기로 입력합니다.

크롤링 프로그램 사용 방법2



마지막 게시물 url element를 파싱하기 위해
마지막 보도자료 웹페이지에서 F12키를 누른 후 9번 페이지의 [] 대괄호의
방법을 똑같이 진행합니다
8번 페이지의 2번에 붙여넣기로 입력합니다.

크롤링 프로그램 사용 방법2

메인 url	00064/selectBoardList.do
page를 나타내는 값	pageIndex
첫 게시물 url element	= "txt"/div[2]/div[1]/div[2]/a
마지막 게시물 url element	"txt"/div[2]/div[10]/div[2]/a
페이지 당 보도자료 수	10
1 첫 게시물 날짜 element	div[2]/a/span/span/span[3]
2 마지막 게시물 날짜 element	div[2]/a/span/span/span[3]
3 날짜 예시	2023.07.05
이전 4 다음	

보도자료 페이지의
첫 게시물과 마지막 게시물의
날짜 요소를 파싱하여 알아내야 합니다

이 방법에 대한 설명이 다음 페이지부터 적혀있습니다.



청양군 암 환자 의료비 지원 뭐가 있나?

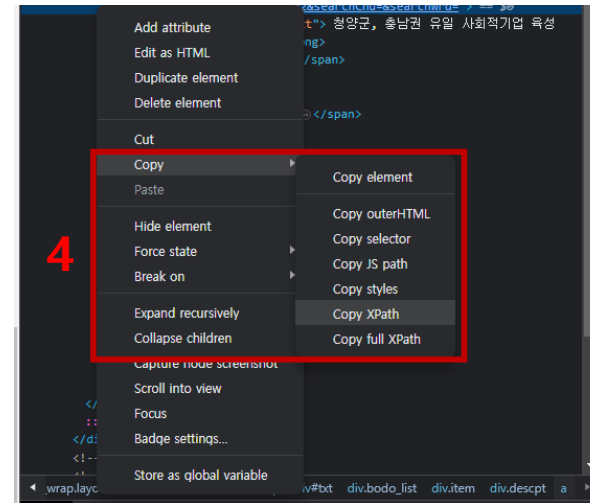
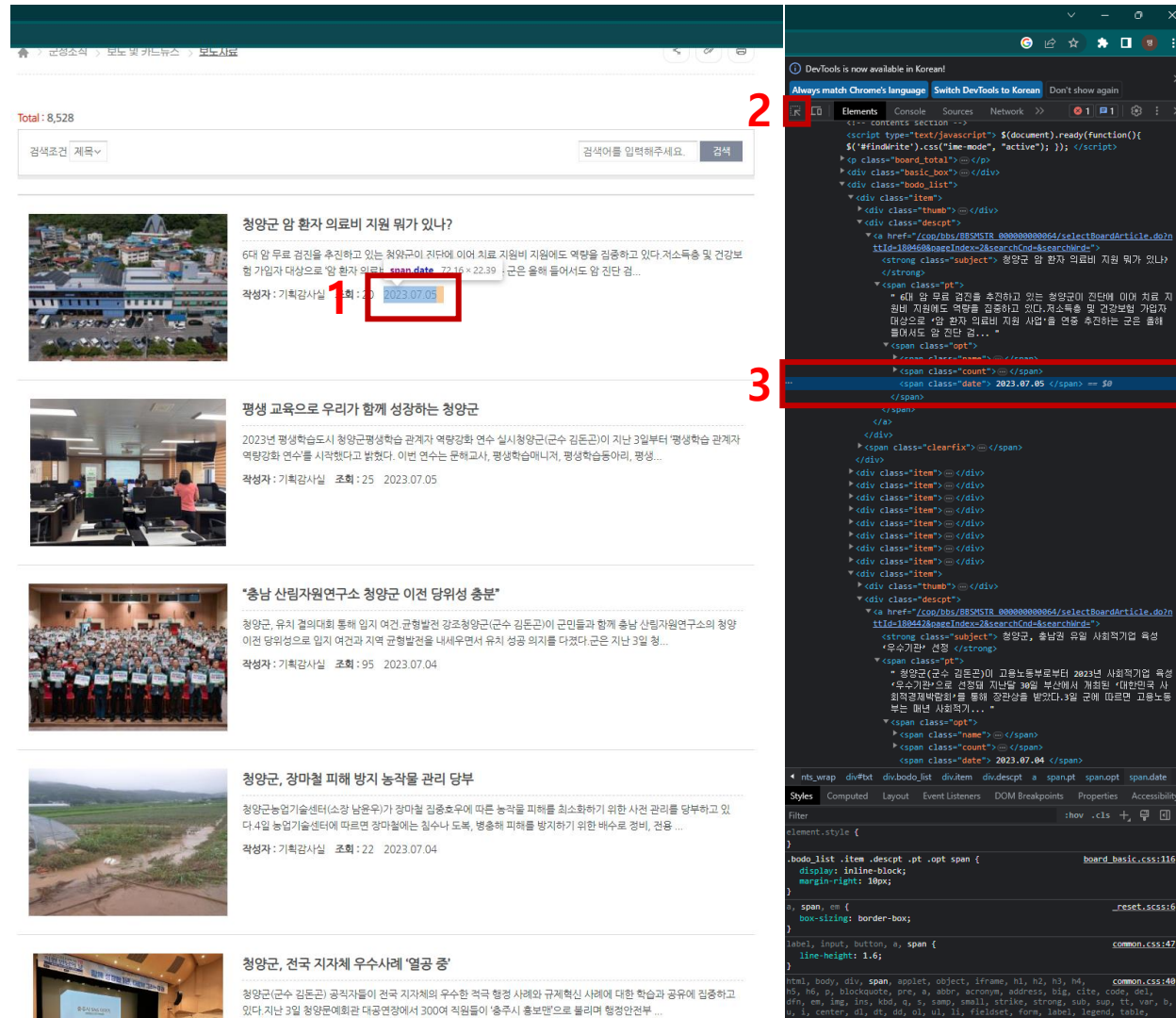
6대 암 무료 검진을 추진하고 있는 청양군이 진단에 이어 치료 지원비 지원에도 역량을 집중하고 있다. 저소득층 및 건강보험 가입자 대상으로 '암 환자 의료비 지원 사업'을 연중 추진하는 군은 올해 들어서도 암 진단 검...

작성자 : 기획감사실 조회 : 20 2023.07.05

보도자료 페이지에 게시물의 날짜를 3번 탭에 입력하면 됩니다.

위 사진의 예시는 2023.07.05 입니다.

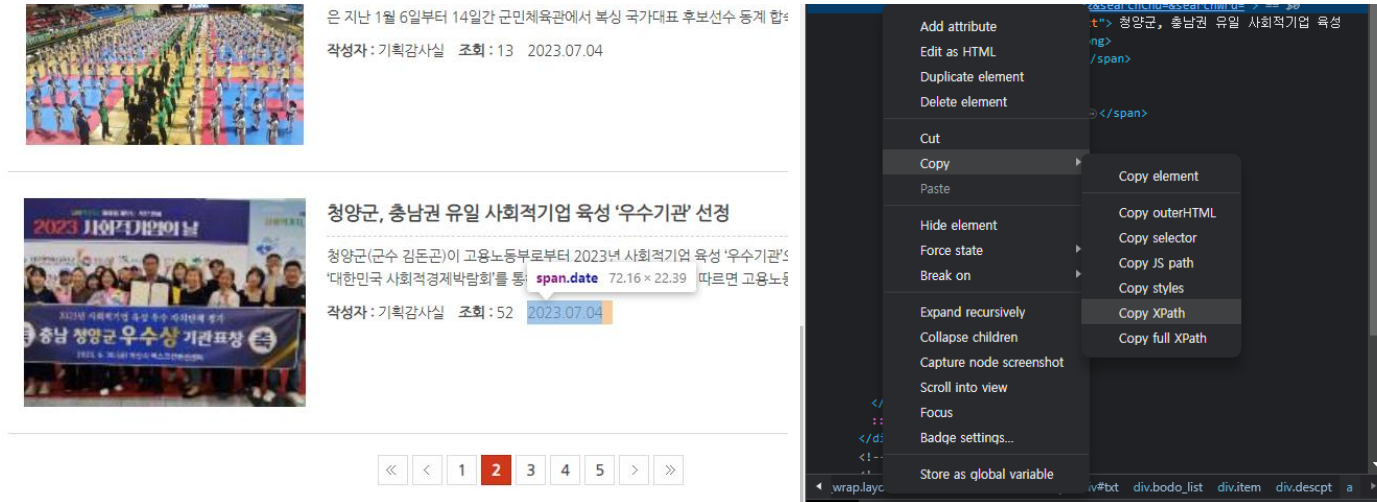
크롤링 프로그램 사용 방법2



첫 게시물 날짜 element를 파싱하기 위해
[처음 보도자료 웹페이지에서 F12키를 눌러
2번 탭을 선택합니다.

마우스를 첫 게시물 (1번 탭)을 클릭합니다.
3번 탭에 날짜 정보가 있는 부분을
클릭하였을 때, 1번 탭 부분이 사진과 같이 파란 박스가
생기면 됩니다.
이 3번 탭 부분을 우클릭하고 4번 탭 부분의 Copy-> Copy Xpath를
클릭하고] 11번 페이지의 1번에 붙여넣기로 입력합니다.

크롤링 프로그램 사용 방법2



마지막 게시물 날짜 element를 파싱하기 위해
마지막 보도자료 웹페이지에서 F12키를 누른 후 12번 페이지의 []
대괄호의 방법을 똑같이 진행합니다
11번 페이지의 2번에 붙여넣기로 입력합니다.

그 다음 11번 페이지의 4번 탭 '다음' 버튼을 누릅니다.

크롤링 프로그램 사용 방법3

1 크롤링 날짜 element `iv[1]/table/tbody/tr[2]/td[2]`

2 크롤링 날짜 예시 `2023-07-05`

3 크롤링 부서 element `iv[1]/table/tbody/tr[2]/td[1]`

4 크롤링 제목 element `"/div[1]/table/tbody/tr[1]/td`

5 크롤링 내용 element `"/div[1]/table/tbody/tr[5]/td`

6 한글파일 element

미전

7 크롤링 시작

보도자료 게시물의
날짜 요소, 부서, 제목, 내용을 파싱하여 알아내야 합니다

만약 3번 페이지에서 한글 파일 여부에서 체크 해제를
했다면, 6번 탭은 뜨지 않습니다.

이 방법에 대한 설명이 다음 페이지부터 적혀있습니다.

크롤링 프로그램 사용 방법3

제목 1 청양군 암 환자 의료비 지원 뭐가 있나?

작성자 2 기획감사실

등록일 3 2023-07-05

조회 22

첨부 KakaoTalk_20220620_085506779_03.jpg [7.833 mbyte] 바로보기



4

6대 암 무료 검진을 추진하고 있는 청양군이 진단에 이어 치료 지원비 지원에도 역량을 집중하고 있다.

저소득층 및 건강보험 가입자 대상으로 암 환자 의료비 지원 사업을 연중 추진하는 군은 올해 들어서도 암 진단 검사 비용, 암 치료 비용 및 약제비를 지원한다.

지원 대상은 청양군 주소지를 둔 암 진단받은 의료급여수급권자와 건강보험 가입자다. 소아·성인 의료급여수급권자는 전체 암, 건강보험 가입자는 국가 암 검진 대상인 5대 암과 폐암 진단자이다.

단, 건강보험 가입자는 2021년 6월 30일까지 5대 암(위암, 대장암, 간암, 유방암, 자궁경부암) 국가암검진을 받고 진단된 암 환자 또는 폐암을 진단받은 자여야 한다.

2023년 건강보험료 기준으로 지역가입자 6만 2,500원 이하, 직장가입자 11만 7,000원 이하 충족하여야 하며 연 200만원씩 3년 연속 지원된다. 또, 의료급여수급권자, 차상위계층은 모든 암종 지원이 가능하며 검진 인력 또한 필요 없이 연 300만원씩 3년 연속 지원된다.

자세한 사항은 청양보건의료원 방문보건팀(940-4580)으로 문의하면 된다.

김상경 원장은 "올 상반기 건강보험료자 12명, 의료보험료 9명에게 의료비 2,200여만 원을 지원했다"면서 "암 진단 후 치료까지 연속적 지원으로 치료 접근성이 향상되어 암 환자와 가족 모두의 행복하고 건강한 삶에 도움이 되길 바란다"라고 전했다.

(사진 없음) 보건의료원 방문보건팀(940-4580)

기획감사실(가) 작성한 청양군 암 환자 의료비 지원 뭐가 있나? 저작물은 공공누리 "출처표시+상업적 이용금지+변경금지" 조건에 따라 이용할 수 있습니다.

5

```
<!DOCTYPE html>
<html lang="ko" style="height: 100%;>
<head>
</head>
<body class="sub04_03_01">
  <div id="skipToContent">
  </div>
  <!-- 스크립트 -->
  <div id="shadow_device">
  </div>
  <div id="touchArea">
  </div>
  <div id="wrap" class="sub04_03_01">
  </div>
  <div class="modal fade" id="mobile-menu" role="dialog" aria-hidden="true">
  </div>
  <!-- // header -->
  <!-- gnb_layout -->
  <div class="gnb_bg">
  </div>
  <!-- gnb_layout -->
  <!-- container -->
  <div id="svisual_layout">
  </div>
  <!-- body_layout -->
  <div class="body_wrap layout">
    <!-- content_layout -->
    <div id="lbn_layout" class="lbn_wrap">
    </div>
    <div id="contents" class="contents_wrap">
      <!-- location -->
      <div id="location" class="title_wrap">
      </div>
      <!-- location -->
      <!-- sub contents -->
      <div id="txt">
        <!-- contents section -->
        <div class="gnb_content">

```

6

Copy element

Copy XPath

앞 서 보도자료 게시물의 url element와 날짜 element를 파싱했던 방법 처럼

이번에는 해당 기관 보도자료의 게시물들 중에서 하나를 클릭한 후 이동한 페이지에서 파싱을 진행합니다.

제목(1), 부서(2), 날짜(3), 내용(4)에 대해서 보도자료 게시물 페이지에서 F12키를 눌러 5번 탭을 선택합니다.

마우스로 1,2,3,4 탭을 클릭합니다. 이에 대한 부분들을 우클릭하여 6번 탭 부분의 Copy-> Copy XPath를 클릭하고 제목은 14페이지의 4번 탭에, 부서는 14페이지의 3번 탭에, 날짜는 14페이지의 1번 탭에, 내용은 14페이지의 5번 탭에 각 element(요소)를 입력하면 됩니다.

14페이지의 2번 탭의 경우에는 현재 페이지의 3번 탭의 날짜가 '2023-07-05' 이므로 이에 대한 정보를 입력하면 됩니다.

크롤링 프로그램 사용 방법3



이전까지의 크롤링은 한글 파일이 존재하지 않거나 한글 파일의 내용을 크롤링 할 필요가 없었기에 시행했던 방법입니다 그렇기에 예시 기관으로 청양군청을 선정하여 진행했습니다.

이번에는 예시 기관으로 행안부의 경우입니다.

보도자료 게시물의 내용을 한글파일로 부터 가져오는 것이기에

14페이지의 5번 탭에는 아무 element나 넣어도 상관은 없습니다.

다만 14페이지의 6번 탭에는 현재 페이지의 1번탭에 해당하는 부분처럼 .hwp 또는 .hwpX 확장자를 가진 부분의 element(요소)를 파싱하면 됩니다.

이 부분의 요소는 보통 8페이지에서 url element를 파싱했던 것 처럼 ' <a href' 의 부분을 Copy-> Copy Xpath 하여 14페이지의 6번 탭에 붙여넣기 입력을 한 후 크롤링 시작을 누르면 크롤링이 시작됩니다.

추가 내용

모든 기관의 웹페이지를 이 프로그램을 통해서 크롤링 할 수는 없습니다.

각 웹페이지별로 다양한 변수가 존재하여 크롤링을 의도적으로 할 수 없게 또는 어렵게 하도록 제작되어 있습니다.

이를 조금이나마 해결하기 위해선 실질적으로 코딩을 진행하여야 하는데, 이를 교수님이 쉽게 하실 수 있도록 조금 더 연구하여 프로그램 내에 탑재해야 할 것 같습니다.

추가적으로, 크롤링 시작 버튼을 눌러 진행하였지만 오류가 발생할 수도 있습니다.

이를 조금 더 쉽게 알아내고 조금이나마 직관적으로 보고 판단하실 수 있도록 크롤링 데이터 폴더 내에 backup 폴더에는 프로그램 실행을 위해 입력한 element 파싱 정보 및 크롤링 정보가 텍스트 파일로 존재할 것입니다.

또한 error 폴더에는 프로그램 동작 중에 발생한 오류가 어느 부분에서 발생했는 지에 대한 정보가 텍스트 파일로 존재 할 것입니다.

크롤링을 진행하다가 잘 안 되는 기관 혹은 어려운 부분은 저에게 말씀해주세요.