



분석용 데이터

획득 전략

대량 데이터 확보 기법

학습 목표

+ + +

학습 목표

- 특정 메뉴를 선택하는 다양한 방법을 활용할 수 있다.
- 원하는 항목을 선택하여 추출할 수 있다.
- 페이지를 변경하는 다양한 방법을 활용할 수 있다.

학습 내용

- 메뉴 선택과 상세내역 출력 및 페이지 변경하기 - 이론
- 메뉴 선택과 상세내역 출력 및 페이지 변경하기 - 실습

메뉴 선택하기 - 이론

1) 특정 메뉴가 텍스트일 경우



선택하는 메뉴가 텍스트로 되어 있고
해당 메뉴에 하이퍼링크가 존재할 경우
`LINK_TEXT()` 함수를 활용하여 클릭 가능

`LINK_TEXT()`

하이퍼링크가 걸려 있을 때 그 글자를 클릭하라는 명령어

메뉴 선택하기 - 이론

2) 특정 메뉴가 텍스트가 아닐 경우



선택하고 싶은 메뉴가 텍스트가 아닐 경우에는 **LINK_TEXT()** 함수를 사용할 수 없다.

※ 이럴 경우 **XPATH**값을 사용하면 된다.



XPATH값 찾기

오른쪽 HTML에서 마우스 우측 버튼 클릭

→ XPATH 클릭

→ Copy XPATH 선택 후 메모장이나 소스코드에 붙여넣기



XPATH(XML Path Language)란,

W3C의 표준으로 확장 생성 언어 문서의 구조를 통해
경로 위에 지정한 구문을 사용하여 항목을 배치하고
처리하는 방법을 기술하는 언어

- ➡ XML 지시자 언어(XPointer)에 쓰이는 언어
- ➡ XSL 변환(XSLT)과 지시자 언어(Xpointer)에 쓰이는 언어
- ➡ XPATH는 XML 문서의 노드를 정의하기 위하여 경로식을 사용하며, 수학 함수와 기타 확장 가능한 표현들이 있음

정리

- ❖ 메뉴가 **글자**로 되어 있으면, **LINK_TEXT** 함수 사용
- ❖ 메뉴가 **그림**이나 **다른 모양**으로 되어 있으면, **XPATH** 사용

메뉴 선택하기 - 실습

■ 입력한 정보 저장

```
1  # riss.kr 에서 특정 키워드로 논문 / 학술 자료 검색하기
2
3  #Step 1. 필요한 모듈을 로딩합니다
4  from selenium import webdriver
5  from selenium.webdriver.common.by import By
6  from selenium.webdriver.common.keys import Keys
7  from selenium.webdriver.chrome.service import Service
8  import time
9
10 #Step 2. 사용자에게 검색 관련 정보들을 입력 받습니다.
11 print("=" * 100)
12 print(" 이 크롤러는 RISS 사이트의 논문 및 학술자료 수집용 웹크롤러입니다.")
13 print("=" * 100)
14 query_txt = input('1.수집할 자료의 키워드는 무엇입니까? : ')
```

■ 경로를 정확히 입력

```
16 #Step 3. 크롬 드라이버 설치 및 웹 페이지 열기
17 s = Service("c:/py_temp/chromedriver.exe")
18 driver = webdriver.Chrome(service=s)
19
20 url = 'https://www.riss.kr/'
21 driver.get(url)
22 time.sleep(5)
23 driver.maximize_window()
24
25 #Step 4. 자동으로 검색어 입력 후 조회하기
26 element = driver.find_element(By.ID, 'query')
27 driver.find_element(By.ID, 'query').click()
28 element.send_keys(query_txt)
29 element.send_keys("\n")
```

메뉴 선택하기 - 실습

- 주소 앞에 프로토콜 반드시 입력

```
16 #Step 3. 크롬 드라이버 설정 및 웹 페이지 열기
17 s = Service("c:/py_temp/chromedriver.exe")
18 driver = webdriver.Chrome(service=s)
19
20 url = 'https://www.riss.kr/'
21 driver.get(url)
22 time.sleep(5)
23 driver.maximize_window()
24
25 #Step 4. 자동으로 검색어 입력 후 조회하기
26 element = driver.find_element(By.ID, 'query')
27 driver.find_element(By.ID, 'query').click()
28 element.send_keys(query_txt)
29 element.send_keys("\n")
```

주의사항

- ❖ 느린인터넷상에서접속중명령시에러발생
- ❖ 엘리먼트이름은사이트마다다를수있기에개발자도구를통해탐색후사용

메뉴 선택하기 - 실습

- 대소문자 정확하게 입력

```
15
16 #Step 3. 크롬 드라이버 설정 및 웹 페이지 열기
17 s = Service("c:/py_temp/chromedriver.exe")
18 driver = webdriver.Chrome(service=s)
19
20 url = 'https://www.riss.kr/'
21 driver.get(url)
22 time.sleep(5)
23 driver.maximize_window()
24
25 #Step 4. 자동으로 검색어 입력 후 조회하기
26 element = driver.find_element(By.NAME, 'query')
27 driver.find_element(By.ID, 'query').click()
28 element.send_keys(query_txt)
29 element.send_keys("\n")
```

주의사항

ID도 Name도 없는 사이트일 경우, XPATH값을 이용

상세 항목 추출하기

```
27 driver.find_element(By.ID, 'query').click( )
28 element.send_keys(query_txt)
29 element.send_keys("\n")
```

이 크롤러는 RISS 사이트의 논문 및 학술자료 수집용 웹크롤러입니다.

1. 수집할 자료의 키워드는 무엇입니까? : 전염병

```
1 # Step 6. 학위 논문 메뉴 선택하기
2 # 메뉴선택방법 1 - LINK_TEXT( ) 사용하기
3 driver.find_element(By.LINK_TEXT, '학위논문').click()
4 time.sleep(2)
5
6 # 메뉴 선택방법 2 - XPATH 사용하기
7 #driver.find_element(By.XPATH, '//*[@id="tabMenu"]/ul/li/div/ul/li[3]/a/span').click()
8 #time.sleep(2)
```

The screenshot shows the RISS website interface. At the top, there's a search bar with the keyword '전염병' (Infectious Disease) entered. Below the search bar, there are navigation tabs: '통합검색' (Integrated Search), '국내학술논문' (Domestic Academic Papers), '학위논문' (Thesis Papers), '해외학술논문' (Overseas Academic Papers), '학술지' (Academic Journals), and '단행본' (Monographs). The '학위논문' tab is selected. Below the tabs, there's a section for '검색결과와 좁혀 보기' (Search Results and Narrowing). On the left, there are filters for '원문유무' (Original Document) and '음성지원유무' (Voice Support). The main area displays search results for the keyword '전염병' (Infectious Disease), showing a list of papers with titles, authors, and publication details. The first result is '1 전염병에 대한 기독교 견해와 고찰' (Christian Perspective and Reflection on Infectious Disease) by 이상훈 (Lee Sang-hoon), published in 2021. The second result is '2 전염병 모델의 연대기적 분석과 후속모델에 대한 고찰' (Chronological Analysis of Infectious Disease Models and Reflection on Subsequent Models) by 이상원 (Lee Sang-won), published in 2012.

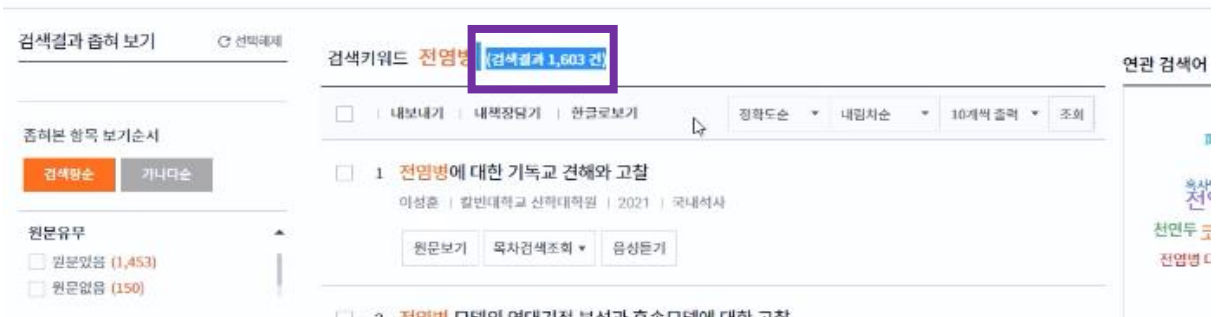
상세 항목 추출하기

총 검색 건수 추출하기

```

1 #Step 7.Beautiful Soup 로 본문 내용만 추출하기
2 from bs4 import BeautifulSoup
3 html_1 = driver.page_source
4 soup_1 = BeautifulSoup(html_1, 'html.parser')
5
6 # 총 검색 건수를 보여주고 수집할 건수 입력받기
7 import math
8 total_cnt = soup_1.find('div', 'searchBox pd').find('span', 'num').get_text()
9 print('검색하신 키워드 %s (으)로 총 %s 건의 학위논문이 검색되었습니다.' %(query,
10 total_cnt))
11 collect_cnt = int(input('이 중에서 몇 건을 수집하시겠습니까?: '))
12 collect_page_cnt = math.ceil(collect_cnt / 10)
13 print('%s 건의 데이터를 수집하기 위해 %s 페이지의 게시물을 조회합니다.' %(collect_cnt, collect_page_cnt))
14 print('=' * 80)
15
16 #Step 8. 각 항목별로 데이터를 추출하여 리스트에 저장하기
17 no2 = [ ] #번호 저장
18 title2 = [ ] #논문제목 저장
19 author2 = [ ] #논문저자 저장
20 company2 = [ ] #소속기관 저장
21 date2 = [ ] #발표년도 저장
22 hak2 = [ ] #학위정보 저장
23 no = 1
24 for a in range(1, collect_page_cnt + 1) :
25

```



메뉴 선택과 상세내역 출력 및 페이지 변경하기 - 실습

상세 항목 추출하기

개발자 도구 F12

RISS 검색 - 학위논문

검색키워드: 전염병 (검색결과 1,603 건)

1 전염병에 대한 기독교 견해와 고찰
이성훈 | 칼빈대학교 신학대학원 | 2021 | 국내석사

2 전염병 모델의 연대기적 분석과 후속모델에 대한 고찰
이성훈 | 高麗大學校 大學院 | 2012 | 국내석사

3 전염병과 현대의 상황에서 메디컬 처치의 운영에 관한 연구
이성훈 | 칼빈대학교 신학대학원 | 2021 | 국내석사

4 전염병 위험자가 관광행동의도에 미치는 영향
김해진 | 세종대학교 대학원 | 2021 | 국내석사

DevTools is now available in Korean!

Always match Chrome's language | Switch DevTools to Korean | Don't show again

Elements | Console | Sources | Network

```
<!-- SearchKeyword -->
<!-- 검색결과 리스트 -->
<div class="srchResultW">
  <!-- 상단결과영역 -->
  <!-- 정렬영역 -->
  <div class="srchResultTop">...</div>
  <!-- 정렬영역 -->
  <div class="srchResultListW">
    <ul>
      <li>
        <span style="display:none;">...</span>
        <div class="markW"> </div>
        <span class="num wd45">...</span>
        <div class="cont ml60">
          <p class="title"> == $0
            <a href="/search/detail/DetailView.do?p_mat_type=be54d9b...
              control_no=53df85fac8ad6dc3ffe0bdc3ef48d419&keyword=전염
              <span class="highlight">전염병</span>
            </a>
          </p>
          <p class="etc">...</p>
          <div style="display:none;">RANK : 27772927</div>
          <div class="btnW">...</div>
        </div>
      </li>
    </ul>
  </div>
</div>
```

로그인

검색도움말 | 최근 검색어

검색 | 상세검색

다국어임력

단행본

내림차순 | 10개씩 출력

DevTools is now available in Korean!

Always match Chrome's language | Switch DevTools to Korean | Don't show again

Elements | Console | Sources | Network

```
<!-- SearchKeyword -->
<!-- 검색결과 리스트 -->
<div class="srchResultW">
  <!-- 상단결과영역 -->
  <!-- 정렬영역 -->
  <div class="srchResultTop">...</div>
  <!-- 정렬영역 -->
  <div class="srchResultListW">
    <ul>
      <li>
        <span style="display:none;">...</span>
        <div class="markW"> </div>
        <span class="num wd45">...</span>
        <div class="cont ml60">
          <p class="title"> == $0
            <a href="/search/detail/DetailView.do?p_mat_type=be54d9b...
              control_no=53df85fac8ad6dc3ffe0bdc3ef48d419&keyword=전염
              <span class="highlight">전염병</span>
            </a>
          </p>
          <p class="etc">...</p>
          <div style="display:none;">RANK : 27772927</div>
          <div class="btnW">...</div>
        </div>
      </li>
    </ul>
  </div>
</div>
```

메뉴 선택과 상세내역 출력 및 페이지 변경하기 - 실습

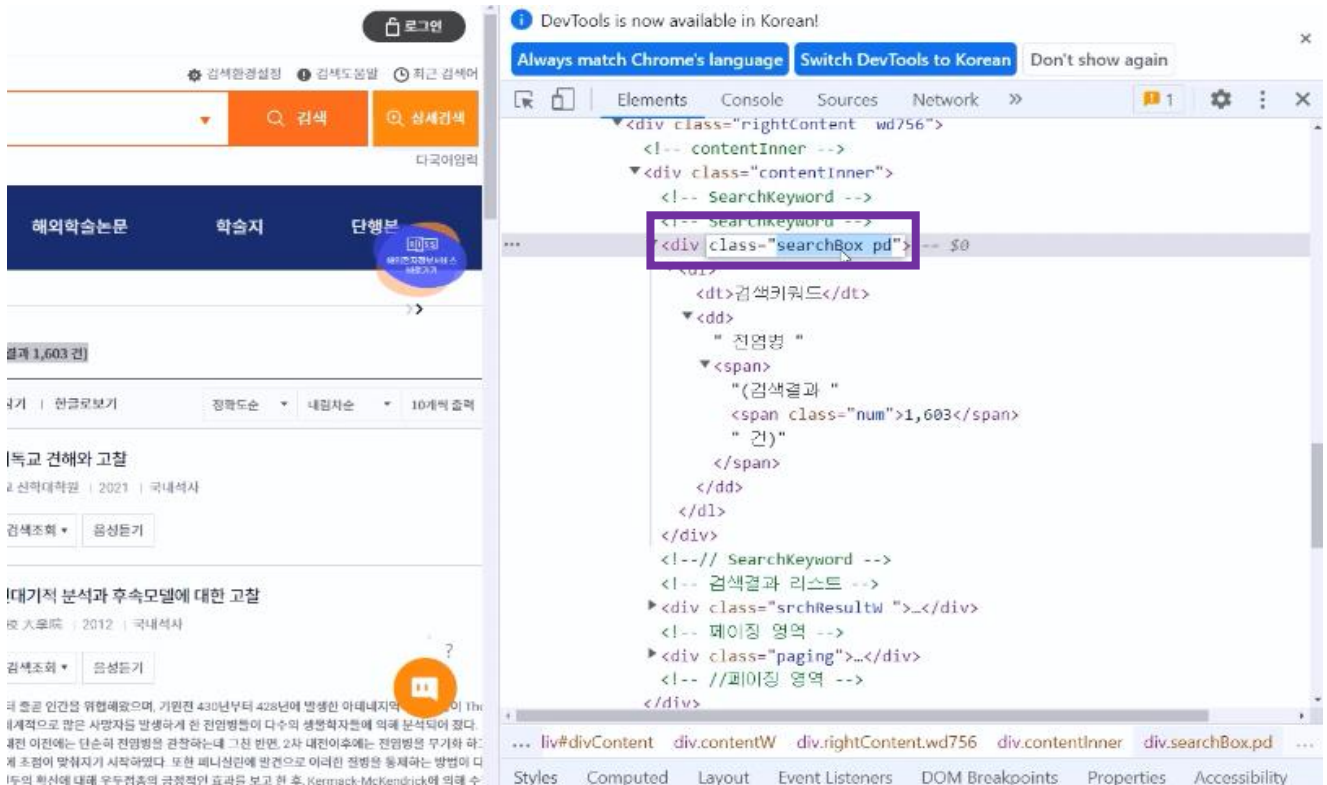
상세 항목 추출하기

```
1 #Step 7.Beautiful Soup 로 본문 내용만 추출하기
2 from bs4 import BeautifulSoup
3 html_1 = driver.page_source
4 soup_1 = BeautifulSoup(html_1, 'html.parser')
5
6 # 총 검색 건수를 보여주고 수집할 건수 입력받기
7 import math
8 total_cnt = soup_1.find('div', 'searchBox_pd').find('span', 'num').get_text()
9 print('검색하신 키워드 %s (으)로 총 %s 건의 학위논문이 검색되었습니다.' % (query_, total_cnt))
10 collect_cnt = int(input('이 중에서 몇 건을 수집하시겠습니까?: '))
11 collect_page_cnt = math.ceil(collect_cnt / 10)
12 print('%s 건의 데이터를 수집하기 위해 %s 페이지의 게시물을 조회합니다.' % (collect_cnt, collect_page_cnt))
13 print('=' * 80)
14
```

주의사항

searchBoxpd:대소문자,공백개수 등 정확히 입력해야함

- 더블 클릭 → ctrl + c 눌러 복사 + 붙여넣기



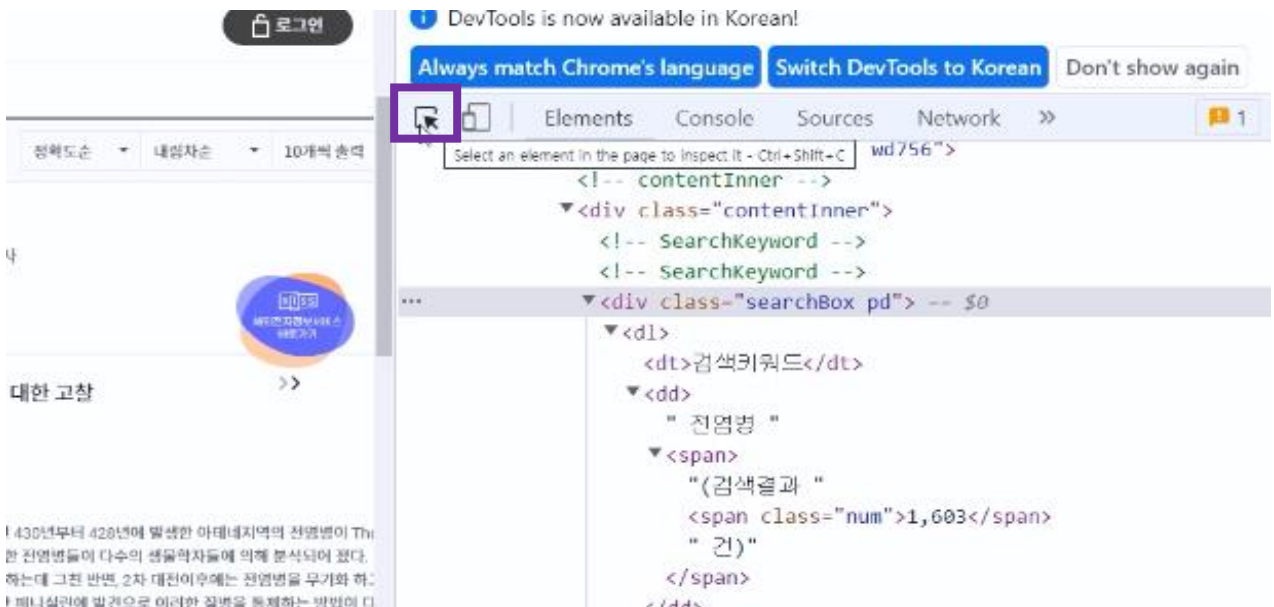
메뉴 선택과 상세내역 출력 및 페이지 변경하기 - 실습

상세 항목 추출하기

```
1 #Step 7.Beautiful Soup 로 본문 내용만 추출하기
2 from bs4 import BeautifulSoup
3 html_1 = driver.page_source
4 soup_1 = BeautifulSoup(html_1, 'html.parser')
5
6 # 총 검색 건수를 보여주고 수집할 건수 입력받기
7 import math
8 total_cnt = soup_1.find('div', 'searchBox pd').find('span', 'num').get_text()
9 print('검색하신 키워드 %s (으)로 총 %s 건의 학위논문이 검색되었습니다' %(query_txt, total_cnt))
10 collect_cnt = int(input('이 중에서 몇 건을 수집하시겠습니까?: '))
11 collect_page_cnt = math.ceil(collect_cnt / 10)
12 print('총 %s건의 데이터를 수집할 수 있습니다.' %(collect_cnt, collect_page_cnt))
13
14 #Step 8. 각 항목별로 데이터를 추출하여 리스트에 저장하기
15 no2 = [ ] #번호 저장
16 title2 = [ ] #논문제목 저장
17 author2 = [ ] #논문저자 저장
18 company2 = [ ] #소속기관 저장
19 date2 = [ ] #발표년도 저장
20 hak2 = [ ] #학위정보 저장
21 no = 1
22
23
24 for a in range(1, collect_page_cnt + 1) :
25
26     html_2 = driver.page_source
27     soup_2 = BeautifulSoup(html_2, 'html.parser')
28
29     content_2 = soup_2.find('div', 'srchResultListW').find_all('li')
30
```

ceil()

주어진숫자보다크고가장가까운 정수를 뽑아내는 함수





상세 항목 추출하기

■ 논문 제목 추출하기

```
24 for a in range(1, collect_page_cnt + 1) :
25
26     html_2 = driver.page_source
27     soup_2 = BeautifulSoup(html_2, 'html.parser')
28
29     content_2 = soup_2.find('div', 'srchResultListW').find_all('li')
30
31     for i in content_2 :
32         #1. 논문제목 있을 경우만
33         try :
34             title = i.find('div', 'cont').find('p', 'title').get_text()
35         except :
36             continue
37         else :
38             print('1. 번호:', no)
39             no2.append(no)
40
41             print('2. 논문제목:', title)
42             title2.append(title)
43
44             # 저자 정보
45             all_data = i.find('p', 'etc').find_all('span')
46             try :
47                 author=all_data[0].get_text().strip()
48             except :
49                 author = '저자 정보가 없습니다'
50             print("3. 저자정보 : %s" %author)
```

The screenshot shows a web browser displaying search results for '전염병' (Infectious Disease). The search results are listed in a table with columns for '번호' (Number), '제목' (Title), '저자' (Author), and '연도' (Year). The first result is '1. 전염병에 대한 기독교 견해와 고찰' (Christian View and Reflection on Infectious Disease) by '이성훈' (Lee Seong-hoon) from '전원대학교 신학대학원' (Jeonwon University Theological Seminary) in 2021. The second result is '2. 전염병 모델의 연대기적 분석과 후속모델에 대한 고찰' (Chronological Analysis of Infectious Disease Models and Reflection on Subsequent Models) by '이성원' (Lee Seong-won) from '고려대학교 대학원' (Korea University Graduate School) in 2012. The third result is '3. 전염병과 팬데믹 상황에서 의료적 처치의 운영에 관한 연구' (Study on the Operation of Medical Treatment in the Context of Infectious Disease and Pandemic) by '이재훈' (Lee Jae-hoon) from '칼빈대학교 신학대학원' (Calvin University Theological Seminary) in 2021. The fourth result is '4. 전염병 위험지각이 관광행동의도에 미치는 영향' (The Effect of Perceived Risk of Infectious Disease on the Intention to Engage in Tourist Behavior).

The Chrome DevTools Elements panel is open, showing the HTML structure of the search results. The 'srchResultListW' container contains a list of search results, each with a 'div' containing the title, author, and year. The 'div' with class 'srchResultListW' is highlighted, and the 'div' with class 'srchResultTop' is also highlighted.

상세 항목 추출하기

■ 제목이 없는 경우 예외 처리

```

24 for a in range(1, collect_page_cnt + 1) :
25
26     html_2 = driver.page_source
27     soup_2 = BeautifulSoup(html_2, 'html.parser')
28
29     content_2 = soup_2.find('div', 'srchResultListW').find_all('li')
30
31     for i in content_2 :
32         #1. 논문제목 있을 경우만
33         try :
34             title = i.find('div', 'cont').find('p', 'title').get_text()
35         except :
36             continue
37         else :
38             print('1.번호:',no)
39             no2.append(no)
40
41             print('2.논문제목:', title)
42             title2.append(title)
43
44             # 저자 정보
45             all_data = i.find('p', 'etc').find_all('span')
46             try :
47                 author=all_data[0].get_text().strip()
48             except :
49                 author = '저자 정보가 없습니다'
50                 print("3.저자정보 : %s" %author)
51                 author2.append(author)

```

드 전염병 (검색결과 1,603 건)

내보내기 | 내책장읽기 | 한글로보기 | 정책도순 | 내일차순 | 10개씩 출력

전염병에 대한 기독교 견해와 고찰
 이상원 | 칼빈대학교 신학대학원 | 2021 | 국내석사
 원문보기 | 목차검색조회 | 음성듣기

전염병 모델의 연대기적 분석과 후속모델에 대한 고찰
 이상원 | 高麗大學校 大學院 | 2012 | 국내석사
 원문보기 | 목차검색조회 | 음성듣기

전염병은 고대 시대부터 출근 인간을 위협해왔으며, 기원전 430년부터 426년에 발생한 아테네지역의 전염병이 Th의 의해 보고된 이후, 전 세계적으로 많은 사망자를 발생하게 한 전염병들이 다수의 생물학자들에 의해 분석되어 왔다. 대한 연구가 2차 세계대전 이전에는 단순히 전염병을 관찰하는데 그친 반면, 2차 대전이후에는 전염병을 무기와 하. 어떻게 통제할 것인지에 초점이 맞춰지기 시작하였다. 또한 제나실린에 발견으로 이러한 질병을 통제하는 방법이 디 으며, Bernoulli가 천연두의 확산에 대해 우두접종의 긍정적인 효과를 보고 한 후, Kermack-McKendrick에 의해 수

전염병과 팬데믹 상황에서 메디컬 처치의 운영에 관한 연구
 이재훈 | 칼빈대학교 신학대학원 | 2021 | 국내석사

Always match Chrome's language | Switch DevTools to Korean

Elements | Console | Sources | Network

```

<!-- SearchKeyword -->
<!-- 검색결과 리스트 -->
<div class="srchResultW">
  <!-- 상단정렬영역 -->
  <!-- 정렬영역 -->
  <div class="srchResultTop">...</div>
  <!-- //정렬영역 -->
  <div class="srchResultListW">
    <ul>
      <li>
        <span style="display:none;">...</span>
        <div class="markw"> </div>
        <span class="num wd45">...</span>
        <div class="cont ml60">
          <p class="title"> == $0
            <a href="/search/detail/Deta
              control_no=53df85fac8ad6dc3f
                <span class="highlight">전
                  "에 대한 기독교 견해와 고찰"
                </a>

```


상세 항목 추출하기

- no = 추출한 건수
collect_cnt = 사용자가 요구한 건수

```
//
78         # 학위정보
79         try :
80             hak = all_data[3].get_text().strip()
81         except :
82             hak='학위정보가 없습니다'
83             hak2.append(hak)
84             print("6. 학위정보 : %s" %hak)
85         else :
86             hak2.append(hak)
87             print("6. 학위정보 : %s" %hak)
88
89         no += 1
90         print("\n")
91
92         if no > collect_cnt :
93             break
94
95         time.sleep(1)          # 페이지 변경 전 1초 대기
96
97         a += 1
98         b = str(a)
99
100        try :
101            driver.find_element(By.LINK_TEXT , '%s' %b).click()
102        except :
103            driver.find_element(By.LINK_TEXT, ('다음 페이지로')).click()
104
105        print("요청하신 작업이 모두 완료되었습니다")
106
```

페이지 변경하기

```
79         try :
80             hak = all_data[3].get_text().strip()
81         except :
82             hak='학위정보가 없습니다'
83             hak2.append(hak)
84             print("6.학위정보 : %s" %hak)
85         else :
86             hak2.append(hak)
87             print("6.학위정보 : %s" %hak)
88
89         no += 1
90         print("\n")
91
92         if no > collect_cnt :
93             break
94
95         time.sleep(1)          # 페이지 변경 전 1초 대기
96
97         a += 1
98         b = str(a)
99
100        try :
101            driver.find_element(By.LINK_TEXT, '%s' %b).click()
102        except :
103            driver.find_element(By.LINK_TEXT, ('다음 페이지로')).click()
104
105    print("요청하신 작업이 모두 완료되었습니다")
106
```