NAVER News Scraping

목차

1. 준비사항

2. HTML Tag 이해

- HTML Tag 종류
- HTML Tag 상하 관계 이해

3. NAVER News Scraping

- NAVER News 페이지 Tag 파악
- Python Code 이해 News Scraping
- Python Code 이해 CSV파일 저장
- Python Code 이해 키워드 수정

Web Scraping을 위한 준비

- Google Chrome 설치 및 사용
- https://support.google.com/chrome/answer/95346?co=GENIE.Platform%3DDesktop&hl=ko
- Anaconda 설치(차례로 참고)
- https://www.anaconda.com/products/individual
- https://wikidocs.net/2826



- Anaconda Prompt 실행
- pip install requests 입력 후 Enter
- pip install bs4 입력 후 Enter



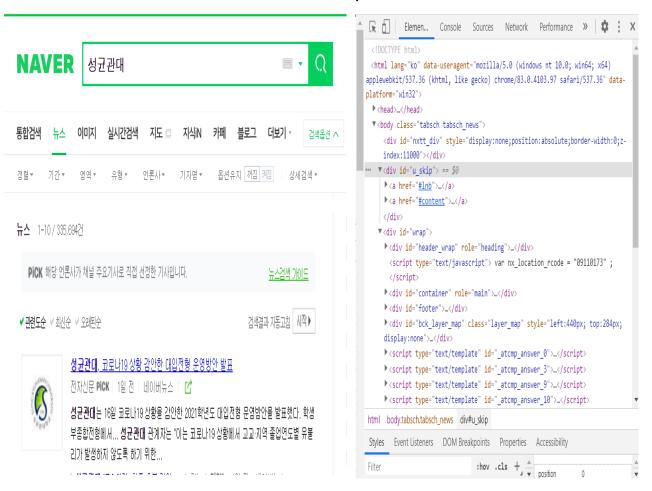
HTML Tag의 종류

태그	설명	
h1, h2, h3, h4, h5, h6	제목(headline)으로 만들기	
br	한줄 띄워쓰기	
b	진하게(bold)	
u	u 밑줄(underline) a 링크 연결하기	
a		

태그	img 이미지 파일 넣기 div 무형의 박스를 넣습니다. input 입력 도구를 만듭니다.	
img		
div		
input		
textarea		

Tag 확인하려면?

- 오른쪽 마우스 -> 검사
- **Ctrl** + **Shift** + **I** (Inspection)



HTML Tag의 ID, Class명

Tag의 ID명: #로 나타낸다 ex) Tag종류#Tag의 ID

Class명: . 로 나타낸다 ex) Tag종류.Tag의 class

예시)
div#u_skip: ID가 items-section인 div 태그 div.items-row: class명이 items-row인 div태그

div#bck_layer 또는 div.layer_map:

ID가 bck_layer 인 div 태그 또는 class명이 layer_map인 div 태그

```
Sources Network
                                                     Performance >>
 <html lang="ko" data-useragent="mozilla/5.0 (windows nt 10.0; win64; x64)</pre>
applewebkit/537.36 (khtml, like gecko) chrome/83.0.4103.97 safari/537.36" data-
platform="win32">
 ▶ <head>...</head>
 ▼<body class="tabsch tabsch news">
     <div id="nxtt_div" style="display:none;position:absolute;border-width:0;z-</pre>
     index:11000"></div>
                                  div#u_skip: ID가 u_skip인 div 태그
    ▼<div id="u skip"> ==
     ▶ <a href="#lnb">...</a>
     ▶ <a href="#content">...</a>
     </div>
   ▼<div id="wrap">
     ▶ <div id="header_wrap" role="heading">...</div>
       <script type="text/javascript"> var nx location rcode = "09110173" ;
       </script>
     ▶ <div id="container" role="main">...</div>
     ▶ <div id="footer">...</div>
     ▶ <div id="bck_layer_map" class="layer_map" style="left:440px; top:284px;
     display:none">...</div>
     \script type="text/template" id="_atcmp_answer_0">...</script>
     ▶ <script type="text/template" id=" atcmp answer 3">...</script>
     ▶ <script type="text/template" id="_atcmp_answer_9">...</script>
     ▶ <script type="text/template" id=" atcmp answer 10">...</script>
html body.tabsch.tabsch_news div#u_skip
Styles Event Listeners DOM Breakpoints Properties Accessibility
                                 :hov .cls + _ position
Filter
```

HTML Tag의 자손, 자식 관계 이해

수집대상인 Tag의 내용을 가져오려면?

- 상위 Tag에서부터 올바른 경로로 해당 Tag 조회



div#items-section div.item (자손관계 이용) div#items-section img (자손관계 이용) div#items-section span (자손관계 이용) div.items-row>div.item(자식관계 이용) div#items-section>div.items-row (자식관계 이용)

NAVER News Scraping

본격적으로, 특정 keyword에 대한 뉴스기사 Scraping을 해보자

아래 정보들을 수집!

- 뉴스기사 제목
- 출판언론사 이름

NAVER 기사 Tag 구조 파악

NAVER

성균관대



3 북한군 총참모부

8 국제표준도 선점



<h1 class="blind">성균관대 뉴스검색 결과</h1> NEW ▼ <div id="main pack" class="main pack"> <script type="text/javascript">var nx cr area info=[{ n:"tab",r:1 NEW }];</script> NEW ▶ <div id="nx related keywords" class="sp keyword section">...</div> <script type="text/javascript">...</script> <script type="text/javascript">...</script> ▼<div class="news mynews section _prs_nws"> NEW ▶ <div class="section head">...</div> <h3 class="blind">관련도순</h3> NEW ▼ == \$0 NEW // ▶ <div class="thumb">...</div> ▼<d1> ▼ <strong class="hl">성균관대 ", 코로나19 상황 감안한 대입전형 문영방안 발표" </dt> ▼<dd class="txt inline": ▼ ... #content #main_pack div ul.type01 li#sp_nws1 dl dt a._sp_each_title dd.txt_inline 1 of 10 ^ V Cancel Styles Event Listeners DOM Breakpoints Properties Accessibility :hov .cls + 📥 Filter

NAVER 기사 Tag 구조 파악



성균관대 코로나19 상황 감안한 대입전형 운영방안 발표

5자신분 MCK | 1일 전 | 데이버뉴스 | 🕜

성균관대는 16일 코로나19 상황을 감안한 2021학년도 대입전형 운영방안을 발표했다. 학성 부종합전형에서... 성균관대 관계자는 "이는 코로나19 상황에서 고교·지역·졸업연도별 유불 리가 발생하지 않도록 하기 위한...

- └ 성균관대 "고3 여건, 학종 충분 감안… 노컷뉴스 PICK | 1일 전 | 네이버뉴스
- 느`무리는 연세대와 다르다`…**성균관대** … 매일경제 **PICK** | 1일 전 | 네이버뉴스
- **└ 성균관대** "학생부종합 고3 여건 감안··· 연합뉴스 | 1일 전 | 네이버뉴스
- ▶ 성균관대, 코로나19 상황 감안해 대··· YTN | 1일 전 | 네이버뉴스

관련뉴스 24건 전체보기》



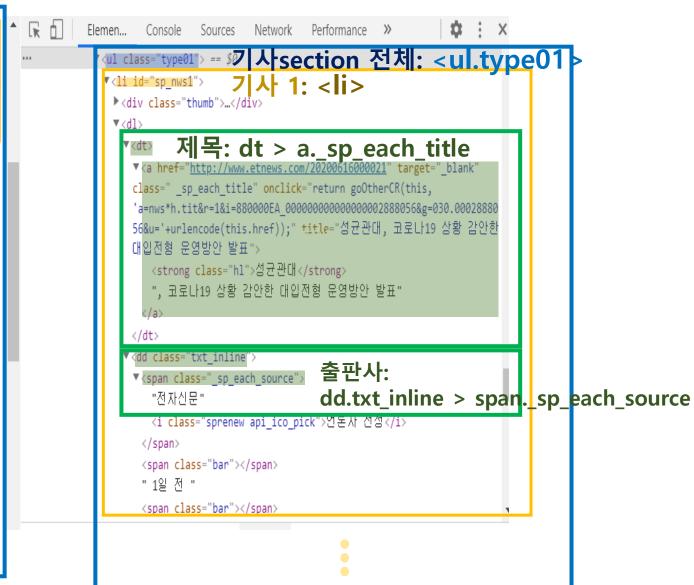
경희·서강·**성균관대**, 수시 논술 비교과 영역 '만점' 준다

서울신문 PICK | 🔳 8면1단 | 7시간 전 | 네이버뉴스 | 🗹

일부 **대학**은 면접을 온라인으로 진행한다. 16일 경희대, 서강대, **성균관대**, 이화여대(가나 다순)는 코로나19에 따른 2021학년도 입학전형 변경 계획을 발표했다. 이들 **대학**은 학교생활기록부와 자기소개서 등을 종합해...

- ►경희·서강·**성균관대**도 "고3 배려 전… 머니투데이 | 3시간 전 |네이버뉴스
- **└ 성균관대**, 코로나19 상황에 대응한 2··· 유교신문 │ 5시간 전

- <u>고려대·**성균관대**, 매년 30억 지원 ICT인재양성 사업 선정</u> IT조선 | 22시간 전 | 🗹



NAVER 기사 Tag 구조 파악

기사section 전체 1 ~ 10: <ul.type01>

기사 1: < li >	dt > asp_each_title Title	dd.txt_inline > spansp_each_source Publisher
기사 2: >	dt > asp_each_title Title	dd.txt_inline > spansp_each_source Publisher
기사 9: < li >	dt > asp_each_title Title	dd.txt_inline > spansp_each_source Publisher
기사 10: >	dt > a. sp each title Title	dd.txt_inline > spansp_each_source Publisher

Code

```
import requests
                                                       Line1~2: Scraping에 필요한 package를 import한다
from bs4 import BeautifulSoup
url='https://search.naver.com/search.naver?&where=news&query=성균관대
                                                       Line1: url 주소에 접속을 요청, 데이터를 가져온다
raw=requests.get(url)
                                                       Line2: 가져온 html 소스코드를 Tag기준으로 구분한다
html=BeautifulSoup(raw.text,'html.parser')
article=html.select('ul.type01 > li')
                                                       Line3: <ul.type01>의 아들인 여러 li> Tag내용을 가져온다
for i in article:
   title=i.select one('dt > a._sp_each_title').text.strip().replace(',','_')
   pub=i.select_one('dd.txt_inline > span._sp_each_source').text.strip().replace(',','_')
                                                       i.select_one('원하는 Tag).text
   print(title)
                                                       - Title: 한 개의 에서 해당 Tag의 text를 가져온다
   print(pub)
                                                       - pub: 한 개의 에서 해당 Tag의 text를 가져온다
   print('='*60)
성균관대 코로나19 상황 감안한 대입전형 문영방안 발표
                                                       .strip().replace(',' , '_')
전자신문언론사 선정
                                                       - Text의 불필요한 공백을 제거하고 내용 속 COMMA를
경희·서강·성균관대 수시 논술 비교과 영역 '만점' 준다
                                                       UNDERBAR(' ')로 대체한다
서울신문언론사 선정
                                                       - CSV 저장 시 구분기호인 COMMA 존재로 인한 문제를 방지
고려대·성균관대 매년 30억 지원 ICT인재양성 사업 선정
TT조선
LG 서울대학교 AI연구팀과 AI 생태계 확장
FETV
```

Code - CSV file로 저장

```
import requests
from bs4 import BeautifulSoup
f=open('naverarticle.csv','w')
f.write('기사제목, 언론사\n')
url='https://search.naver.com/search.naver?&where=news&querv=성균관대
raw=requests.get(url)
html=BeautifulSoup(raw.text,'html.parser')
article=html.select('ul.type01 > li')
for i in article:
    title=i.select one('dt > a. sp each title').text.strip().replace(',',' ')
    pub=i.select one('dd.txt inline > span. sp each source').text.strip().replace(',',' ')
   print(title)
    print(pub)
   print('='*60)
    f.write(title + ',' + pub + '\n')
f.close()
```

 csv: comma separated value

콤마로 구분되는 값들의 모음 값들의 열구분(가로 데이터 구분)은 **콤마(,)** 값들의 행구분(세로 데이터 구분)은 <mark>엔터(\n)</mark>

- 1. 파일이름을 지정하여("naverarticle.csv") 저장할 CSV파일 생성
- 2. Column 명 입력

데이터를 , 와 \ n 으로 구분해주어야 함.

- \ n을 쓸 때 따옴표로 씌워주기!

꼭 f.close()로 닫아주기!

Code - CSV file로 저장

```
import requests
from bs4 import BeautifulSoup
f=open('naverarticle.csv','w')
f.write('기사제목, 언론사\n')
url='https://search.naver.com/search.naver?&where=news&query=성균관대'
raw=requests.get(url)
html=BeautifulSoup(raw.text,'html.parser')
article=html.select('ul.type01 > li')
for i in article:
   title=i.select_one('dt > a._sp_each_title').text.strip().replace(',','_')
   pub=i.select one('dd.txt inline > span. sp each source').text.strip().replace(',',' ')
   print(title)
   print(pub)
   print('='*60)
   f.write(title + ',' + pub + '\n')
f.close()
성균관대 코로나19 상황 감안한 대입전형 운영방안 발표
전자신문언론사 선정
경희·서강·성균관대 수시 논술 비교과 영역 '만점' 준다
서울신문언론사 선정
고려대·성균관대_ 매년 30억 지원 ICT인재양성 사업 선정
LG 서울대학교 AI연구팀과 AI 생태계 확장
LG전자 성균관대와 협력한 '제조 인공지능 리더' 교육과정 마쳐
비즈니스포스트
_____
성균관대 LG전자와 'AI 스마트팩토리 전문가' 양성
파이낸셜뉴스
```

저장된 CSV파일 ↓

1	A	В
1	기사제목	언론사
2	성균관대_ 코로나19 상황 감안한 대입전형 운영방안 발표	전자신문언론사 선정
3	경희·서강·성균관대_ 수시 논술 비교과 영역 '만점' 준다	서울신문언론사 선정
4	고려대·성균관대_ 매년 30억 지원 ICT인재양성 사업 선정	IT조선
5	LG_ 서울대학교 AI연구팀과 AI 생태계 확장	FETV
6	LG전자_ 성균관대와 협력한 '제조 인공지능 리더' 교육과정 마쳐	비즈니스포스트
7	성균관대_ LG전자와 'AI 스마트팩토리 전문가' 양성	파이낸셜뉴스
8	조선대학교_ 'AI중심 산업융합 집적단지 조성사업' 3개 연구팀 선정	교수신문
9	LG전자_ 성균관대와 협력 인공지능-빅데이터 전문가 육성 박차	베이비타임즈
10	최저학력 기준 낮추고_ 어학시험 폐지서울대·성균관대 고3 구제책 확정	이투데이
4.4	대차기 ㅇ기이 나치 ㅂ저채이 되트 이쁘에 서그라며 나라며	어깎ㄷㅆ여ㅋ!! 저저

Code – keyword 수정

```
import requests
from bs4 import BeautifulSoup
f=open('naverarticle.csv','w')
f.write('기사제목, 언론사\n')
url='https://search.naver.com/search.naver?&where=news&querv=섬균관
raw=requests.get(url)
html=BeautifulSoup(raw.text,'html.parser')
article=html.select('ul.type01 > li')
for i in article:
   title=i.select one('dt > a. sp each title').text.strip().replace(',','')
   pub=i.select one('dd.txt inline > span. sp each source').text.strip().replace(',','')
   print(title)
   print(pub)
   print('='*60)
   f.write(title + ',' + pub + '\n')
f.close()
성균관대 코로나19 상황 감안한 대입전형 문영방안 발표
전자신문언론사 선정
경희·서강·성균관대 수시 논술 비교과 영역 '만점' 준다
서울신문언론사 선정
고려대·성균관대 매년 30억 지원 ICT인재양성 사업 선정
IT조선
LG 서울대학교 AI연구팀과 AI 생태계 확장
FETV
LG전자 성균관대와 협력한 '제조 인공지능 리더' 교육과정 마쳐
비즈니스포스트
성균관대_ LG전자와 'AI 스마트팩토리 전문가' 양성
파이낸셜뉴스
```

Query 값을 수정하게 되면 다른 키워드에 대한 News 수집 가능!

Summary

- 수집을 페이지의 URL 파악
- 수집대상의 HTML Tag
- 파이썬 코드로 데이터 수집 실시
- 수집한 데이터를 CSV로 저장

수고하셨습니다!