NAVER News Scraping

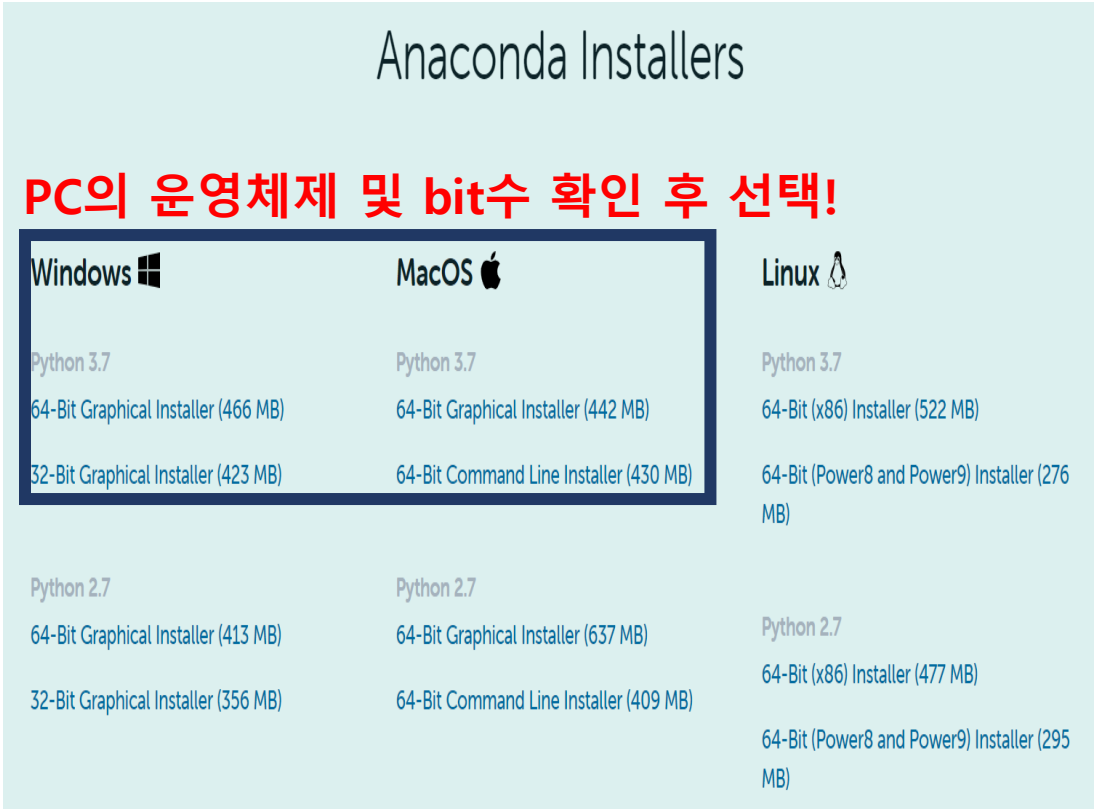
목차

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

- **파이썬 패키지 Requests, BeautifulSoup4 설치**
  - **Anaconda Prompt** 실행
  - **pip install requests** 입력 후 Enter
  - **pip install bs4** 입력 후 Enter



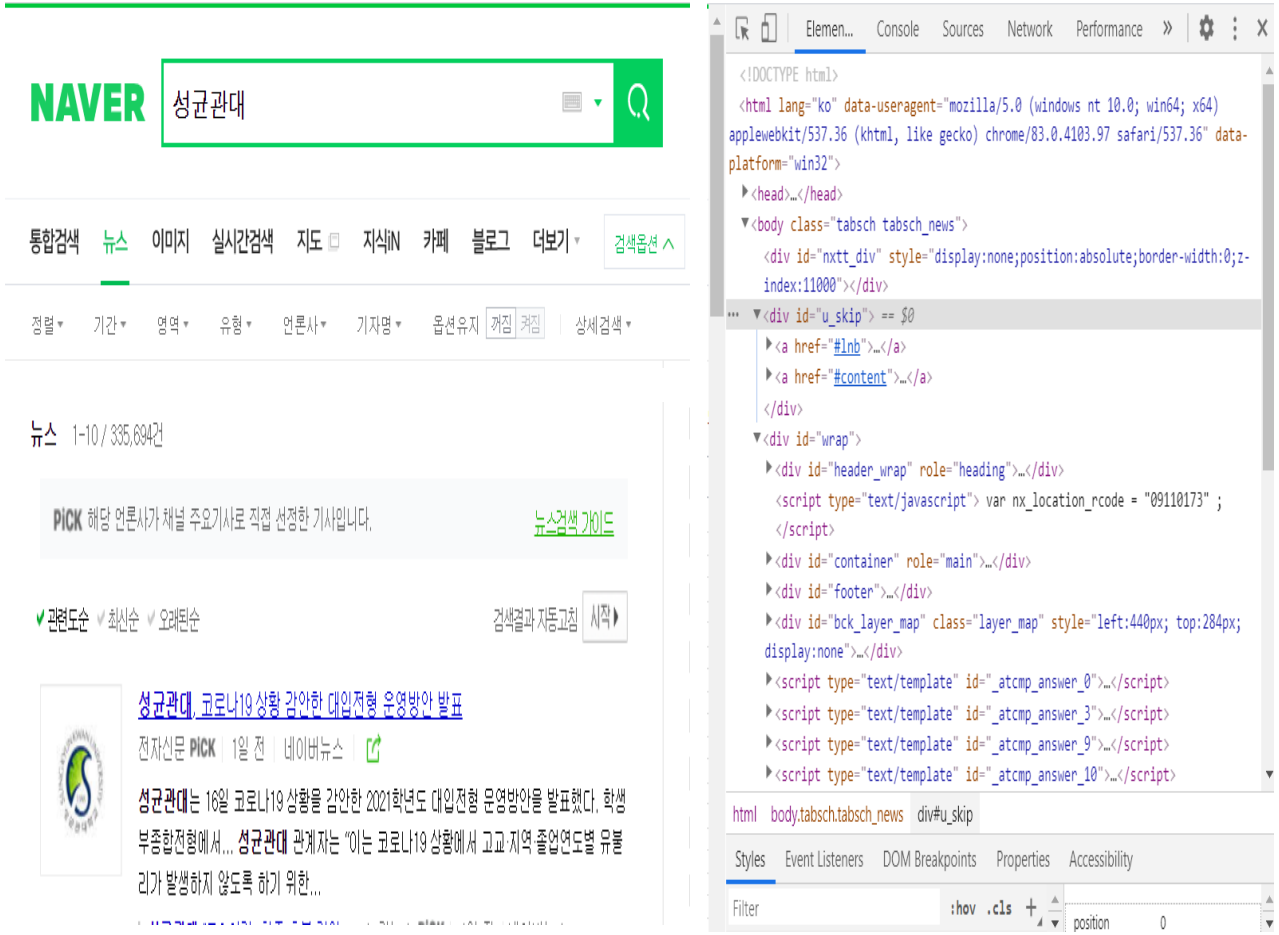
HTML Tag의 종류

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

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

Tag 확인하려면?

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

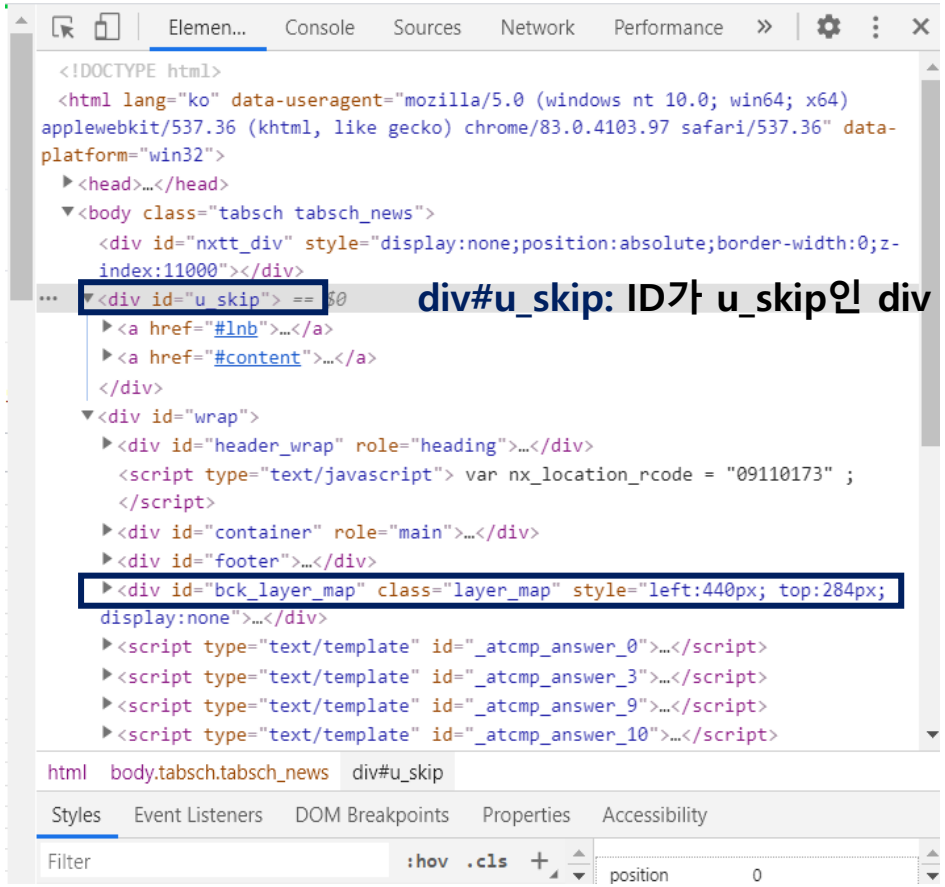


HTML Tag의 ID, Class명

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#u_skip: ID가 u_skip인 div 태그

div#bck_layer 또는 div.layer_map:

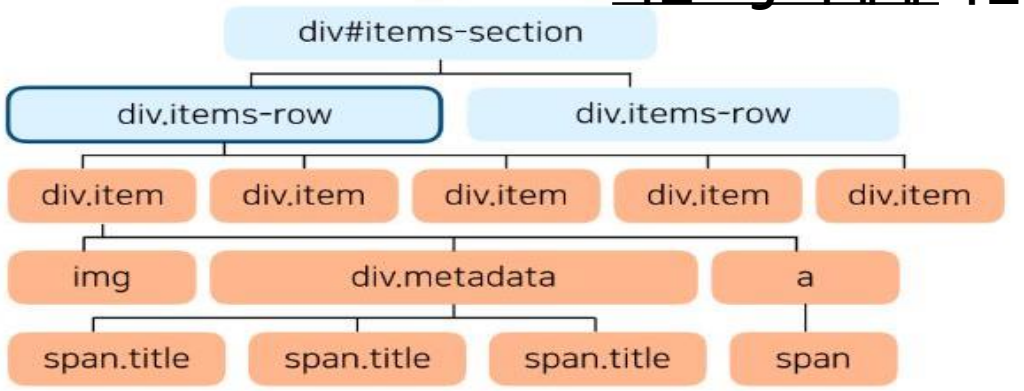
ID가 bck_layer 인 div 태그 또는 class명이 layer_map인 div 태그

HTML Tag의 자손, 자식 관계 이해

수집대상인 Tag의 내용을 가져오려면?

- 상위 Tag에서부터 올바른 경로로 해당 Tag 조회

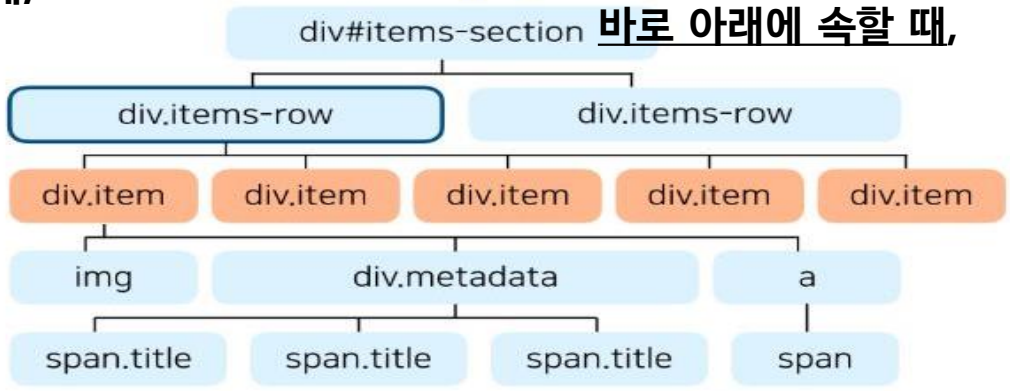
자손 (space) **수집대상인 Tag가 기준Tag 아래에 속할 때,**



자손: 특정 태그에 속해있는 모든 태그

" " 공백으로 연결!

자식 (>) **수집대상 Tag(div.items-row)가 기준Tag(div#items-section에 바로 아래에 속할 때,**



자식: 특정 태그에 직접 속해있는 태그

">" 으로 연결!

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)

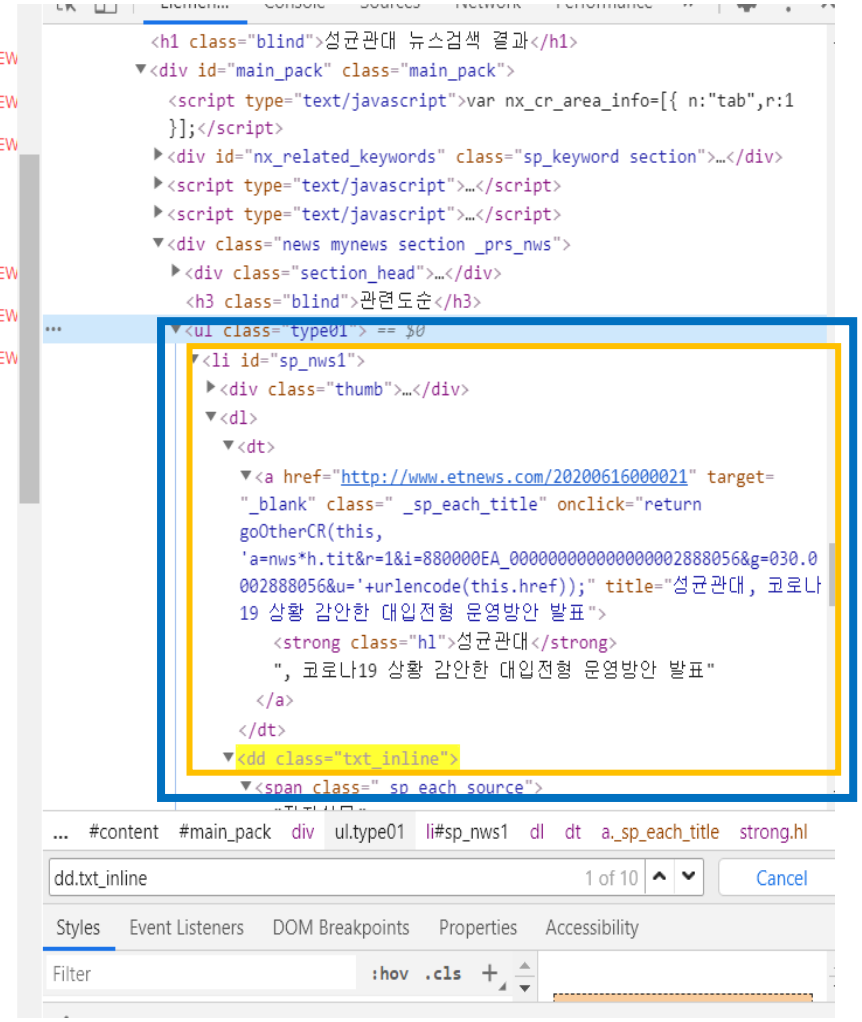
NAVER News Scraping

본격적으로, 특정 keyword에 대한 뉴스기사 Scraping을 해보자

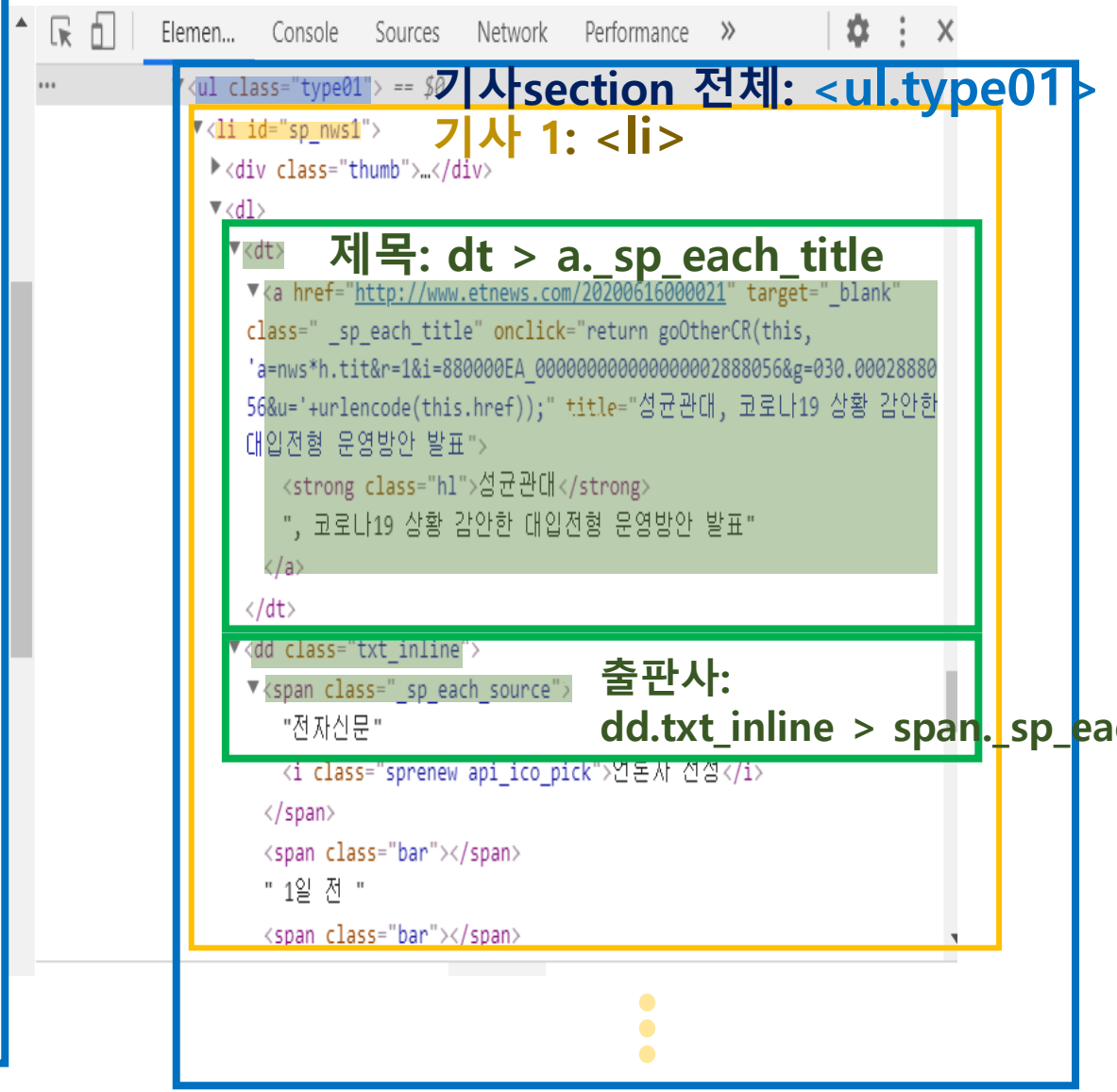
아래 정보들을 수집!

- **뉴스기사 제목**
- **출판언론사 이름**

NAVER 기사 Tag 구조 파악

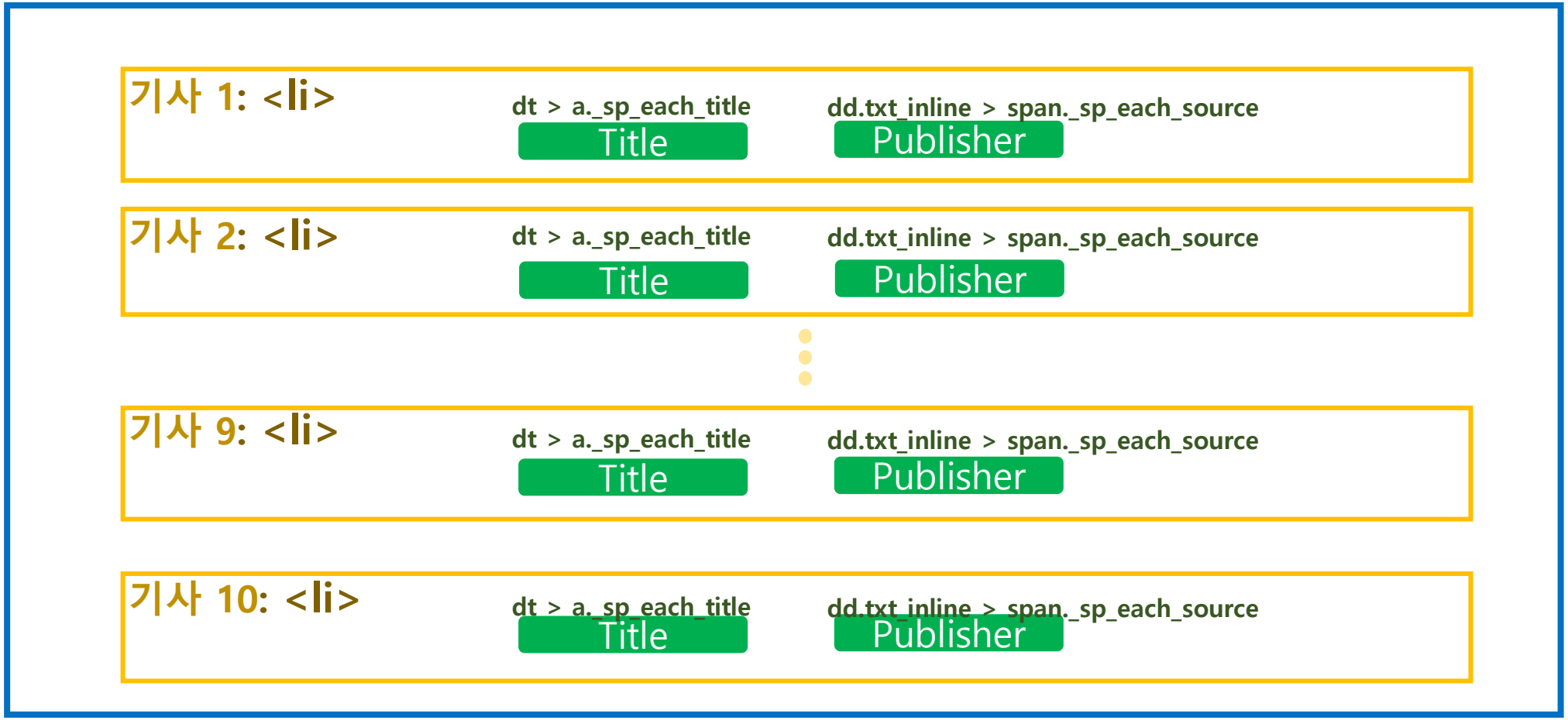


NAVER 기사 Tag 구조 파악



NAVER 기사 Tag 구조 파악

기사section 전체 1 ~ 10: <ul.type01>



Code

```
import requests
from bs4 import BeautifulSoup

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)
```

```
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
IT조선
============================================================
LG_ 서울대학교 AI연구팀과 AI 생태계 확장
FETV
============================================================
```

Line1~2: Scraping에 필요한 package를 import한다

Line1: url 주소에 접속을 요청, 데이터를 가져온다
Line2: 가져온 html 소스코드를 Tag기준으로 구분한다
Line3: <ul.type01>의 아들인 여러 <li> Tag내용을 가져온다

i.select_one('원하는 Tag).text

- Title: 한 개의 에서 해당 Tag의 text를 가져온다
- pub: 한 개의 에서 해당 Tag의 text를 가져온다

.strip().replace(',' , '_')

- Text의 불필요한 공백을 제거하고 내용 속 COMMA를 UNDERBAR('_')로 대체한다
- CSV 저장 시 구분기호인 COMMA 존재로 인한 문제를 방지

Code – CSV file로 저장

csv: comma separated value
콤마로 구분되는 값들의 모음
값들의 열구분(가로 데이터 구분)은 **콤마(,)**
값들의 행구분(세로 데이터 구분)은 **엔터(\n)**

```
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
IT조선
============================================================
LG_ 서울대학교 AI연구팀과 AI 생태계 확장
FETV
============================================================
LG전자_ 성균관대와 협력한 '제조 인공지능 리더' 교육과정 마쳐
비즈니스포스트
============================================================
성균관대_ LG전자와 'AI 스마트팩토리 전문가' 양성
파이낸셜뉴스
============================================================
```

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

```python
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
IT조선
============================================================
LG_ 서울대학교 AI연구팀과 AI 생태계 확장
FETV
============================================================
LG전자_ 성균관대와 협력한 '제조 인공지능 리더' 교육과정 마쳐
비즈니스포스트
============================================================
성균관대_ LG전자와 'AI 스마트팩토리 전문가' 양성
파이낸셜뉴스
============================================================
```

저장된 CSV파일 ↓

	A	B
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 폐지...서울대・성균관대 고3 구제책 확정	이투데이

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&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
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!