Crawing Study

동적 크롤링 스터디

김하람, 임낙준, 도언

크롤링이란?

웹 페이지를 가져와서 데이터를 추출해 내는 방법.
이렇게 크롤링하는 소프트웨어를 **크롤러(crawler)** 라고 합니다.

크롤링 활용 사례.

- 어떠한 기업에서는 크롤링 담당 부서가 따로 존재
(ex 마케팅 부서의 경우 크롤링을 통해 평점, 후기 등을 쉽게 다룰 수 있게 됨.)

데이터사이언스에는 원하는 데이터를 가져오는 것이 곧 분석력이며 경쟁력!

크롤링 종류

1. 정적 크롤링

정적인 데이터를 수집하는 방법.

- 정적데이터란

변하지 않는 데이터를 의미. 즉 한 페이지 안에서 원하는 정보가 모두 드러날 때 정적 크롤링 사용

크롤링 종류

2. 동적 크롤링

동적인 데이터를 수집하는 방법.

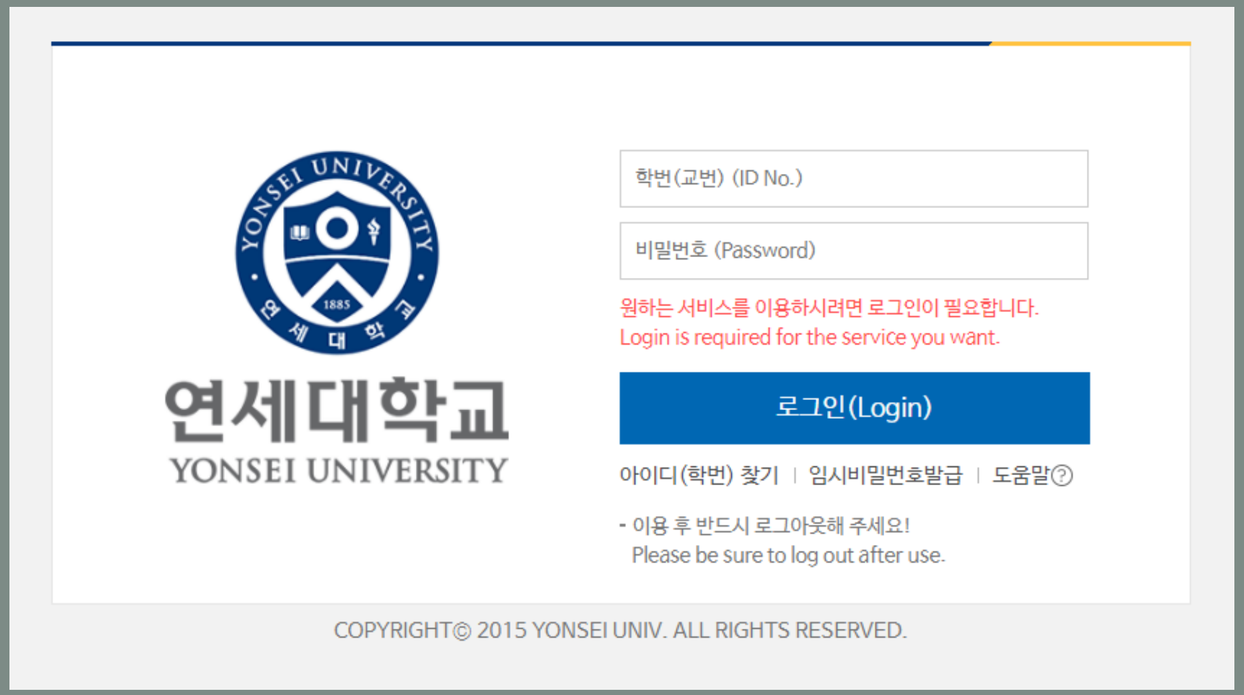
- **동적 데이터란**

입력, 클릭, 로그인 등과 같이 페이지 이동이 있어야 보이는 데이터

복권구매
복권정보
당첨결과
판매점
이벤트
행복공감
QUICK MENU
당첨결과
회차별 당첨번호
당첨결과 > 로또6/45 > 회차별 당첨번호
로또6/45
회차별 당첨번호
내번호 당첨확인
당첨내역
당첨금 지급안내
추첨방송 다시보기
추첨방송 참관신청
연금복권720+
전자복권
회차 바로가기
961
조회
961회 당첨결과
(2021년 05월 01일 추첨)
11 20 29 31 33 42 + 43
당첨번호
보너스

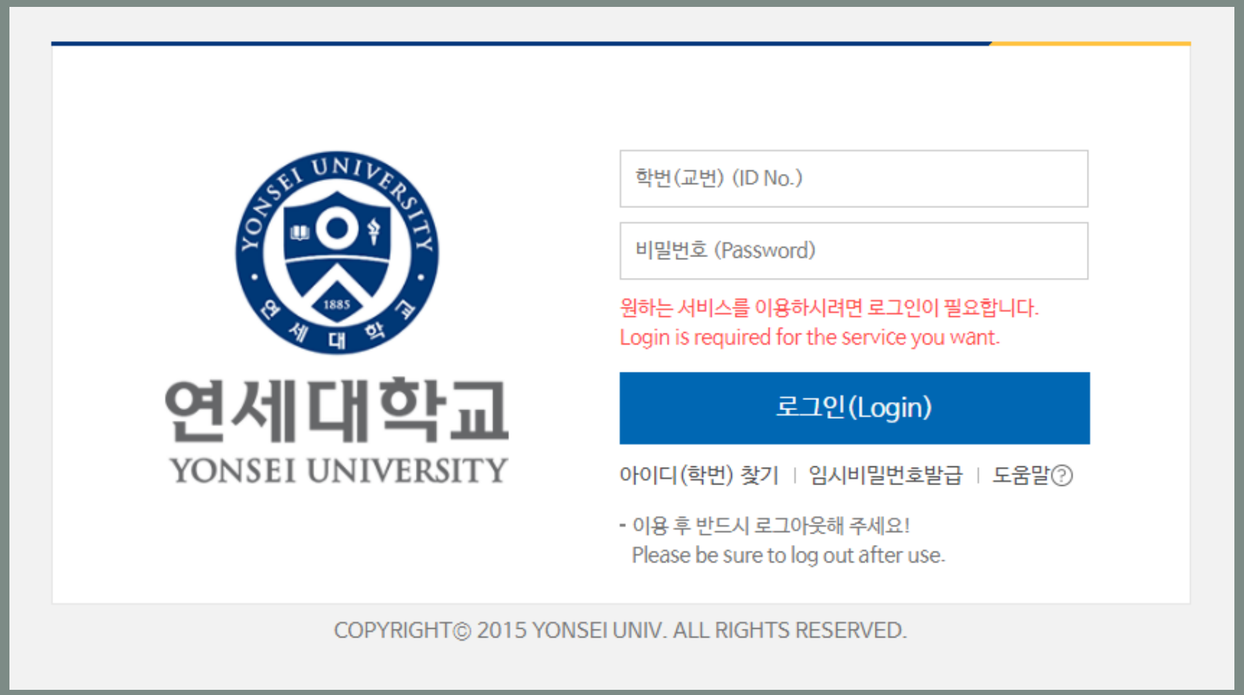
정적 크롤링 예시

복권구매
복권정보
당첨결과
판매점
이벤트
행복공감
QUICKMENU
당첨결과
회차별 당첨번호
당첨결과 > 로또6/45 > 회차별 당첨번호
로또6/45
회차별 당첨번호
내번호 당첨확인
당첨내역
당첨금 지급안내
추첨방송 다시보기
추첨방송 참관신청
연금복권720+
전자복권
회차 바로가기
961
조회
961회 당첨결과
(2021년 05월 01일 추첨)
11 20 29 31 33 42 + 43
당첨번호
보너스



종별	학정번호-분반(-실습)	학점	교과목명	담당교수	강의시간	강의실	유의사항	국외교환학...
교기	YCA1007-01-00	.5	비대면-채플(C)	이대성		채플	ⓑ①ⓒ	O
교기	YCA1007-02-00	.5	비대면-채플(C)	이대성		채플	ⓑ①②ⓑⓒ	O
교기	YCA1009-01-00	.5	비대면-예배채플(A)	정미현		채플	ⓑ①ⓒ	O
교기	YCA1101-12-00	3	비대면-기독교와세계문화	김학철	월15,토0/수8	동영상콘텐츠/실시간온라인	ⓖⓑ	X
교기	YCA1102-13-00	3	비대면-기독교와현대사회	한인철	화14,15/목4	동영상콘텐츠/실시간온라인	ⓖⓑ	X
교기	YCA1103-08-00	3	비대면-성서와기독교	조재국	목7/토4,5	원B101/동영상콘텐츠	ⓗ①ⓑ	X
교기	YCB1101-58-00	3	비대면-글쓰기	김선혜	월3,4/수4	실시간온라인/백S408	①①ⓑ	X
교기	YCB1101-59-00	3	비대면-글쓰기	송태욱	월5,6/수6	실시간온라인/백S408	①①ⓑ	X
교기	YCB1101-60-00	3	비대면-글쓰기	전은주	화1/목2,3	백S408/실시간온라인	①①ⓑ	X
교기	YCB1101-61-00	3	비대면-글쓰기	하신애	화5,6/목4	실시간온라인/백S408	①①ⓑ	X
교기	YCB1101-62-00	3	비대면-글쓰기	전은주	화7/목8,9	백S408/실시간온라인	①①ⓑ	X

동적 크롤링 예시



종별	학정번호-분반(-실습)	학점	교과목명	담당교수	강의시간	강의실	유의사항	국외교환학...
교기	YCA1007-01-00	.5	비대면-채플(C)	이대성		채플	ⓑ①ⓒ	O
교기	YCA1007-02-00	.5	비대면-채플(C)	이대성		채플	ⓑ①②ⓑⓒ	O
교기	YCA1009-01-00	.5	비대면-예배채플(A)	정미현		채플	ⓑ①ⓒ	O
교기	YCA1101-12-00	3	비대면-기독교와세계문화	김학철	월15,토0/수8	동영상콘텐즈/실시간온라인	ⓖⓑ	X
교기	YCA1102-13-00	3	비대면-기독교와현대사회	한인철	화14,15/목4	동영상콘텐즈/실시간온라인	ⓖⓑ	X
교기	YCA1103-08-00	3	비대면-성서와기독교	조재국	목7/토4,5	원B101/동영상콘텐즈	ⓗ①ⓑ	X
교기	YCB1101-58-00	3	비대면-글쓰기	김선혜	월3,4/수4	실시간온라인/백S408	①①ⓑ	X
교기	YCB1101-59-00	3	비대면-글쓰기	송태욱	월5,6/수6	실시간온라인/백S408	①①ⓑ	X
교기	YCB1101-60-00	3	비대면-글쓰기	전은주	화1/목2,3	백S408/실시간온라인	①①ⓑ	X
교기	YCB1101-61-00	3	비대면-글쓰기	하신애	화5,6/목4	실시간온라인/백S408	①①ⓑ	X
교기	YCB1101-62-00	3	비대면-글쓰기	전은주	화7/목8,9	백S408/실시간온라인	①①ⓑ	X

+) 메일함에 있는 메일 제목데이터를 수집하고 싶다고 생각을 해보자. 그렇기 위해서는 로그인과정을 거친 후 메일함에 들어가야 하는 동적인 과정이 필요

정적 크롤링 vs 동적 크롤링

	정적 크롤링	동적 크롤링
연속성	주소를 통해 단발적으로 접근	브라우저를 사용하여 연속적으로 접근
수집 능력	수집 데이터의 한계가 존재	수집 데이터의 한계가 없음
속도	빠름	느림
라이브러리	requests, BeautifulSoup	selenium, chromedriver

HTML이란?

웹페이지를 이루고 있는 구성요소.
크롤링이란 웹 데이터 중 특정 데이터를 컴퓨터에게 요청하는 것
웹에서 'F12'를 누르면 보이는 수많은 데이터 중,
원하는 데이터를 구별하기 위해 HTML 구조 공부가 크롤링 공부에 선행!

HTML 구조

```html
<div>
    <div>
        <h4><a href="#">경제</a></h4>
        <span>
            <a href="#">일반</a>
            <a href="#">금융</a>
            <a href="#">생활경제</a>
        </span>
    </div>
    <div>
        <dl>
            <a>
                <img src="http://imgnews.naver.net/image/upload/item/2018/
                            182203604_%25C1%25A6%25B8%25F1-%25BE%25F8%25C0
                            ?type=f270_166" width="300" height="190"> <br/


            <a href="#">[한미 금리차 확대 전문가 진단] "外人 자금 이탈땐 증시 출

        </dl>
        <div>
            <ul>
```

HTML의 작성 방식 – 태그의 구조

<태그>내용물</태그>

HTML 태그의 기본 구조입니다.

```
<span>X심 봉지라면 4종 기획 세트</span>
```

HTML의 작성 방식- 중첩 태그

```
<태그1>
  <태그2>
    내용물
  </태그2>
</태그1>
```

```html
<div>
    <h4><a>경제</a></h4>
    <h4>
        <a>일반</a>
    </h4>
</div>
```

HTML 태그의 문법

```
<div>
    <div>
        <div>
            <img>
            </img>
            <div>
                <span>
                    14U380-EU1TK
                </span>
                <span>
                    최저 389,000원
                </span>
                <span>
                    상품평 599
                </span>
            </div>
        </div>
        <div>
            ...
        </div>
    ...
...
```

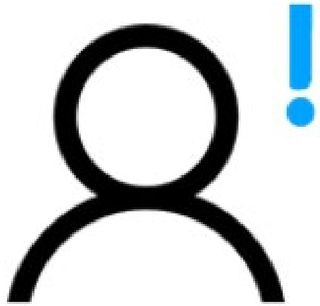
똑같은 div인가?

똑같은 span인가?

div, img, span...

HTML 태그의 문법 – 태그의 이름표, 선택자

```html
<div id="items-section">
    <div class="items-row">
        <div class="item">
            <img>
            </img>
            <div class="metadata">
                <span class="title">
                    14U380-EU1TK
                </span>
                <span class="price">
                    최저 389,000원
                </span>
                <span class="comments">
                    상품평 599
                </span>
            </div>
        </div>
    </div>
    <div class="items-row">
        ...
    </div>
</div>
```



컴퓨터 위주의 태그 구성에 인간이 확인할 별명을 추가

태그의 별명, 선택자의 종류

ID, class: 가장 많이 쓰이는 선택자
한 ID는 전체 HTML 문서에 하나만 존재(like 학번)
한 태그가 여러 개의 클래스 가질 수 있음

ID와 클래스

ID : 4-김민재
클래스 : 2학년 5반, 축구부, 남학생

ID : 6-정우영
클래스 : 2학년 5반, 남학생

ID : 3-강혜원
클래스 : 2학년 5반, 여학생

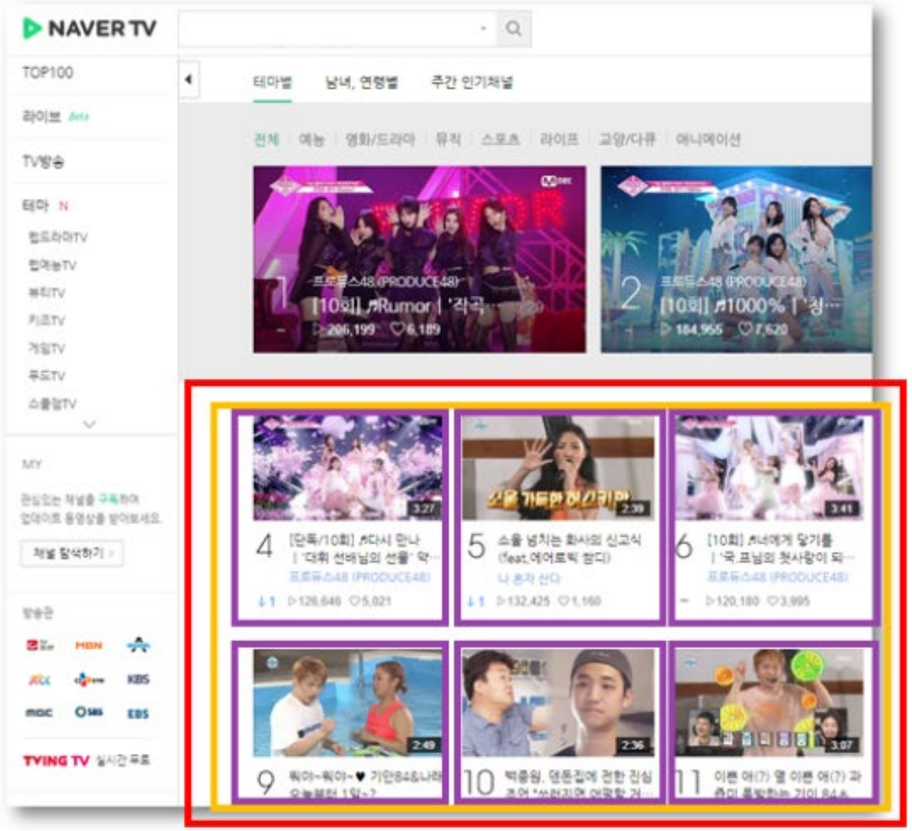
ID : 35-사토 미나미
클래스 : 2학년 5반, 댄스부, 여학생

선택자 경로

: HTML요소

우리가 할 일은 원하는 요소의 위치를 컴퓨터에게 알려주는 것!

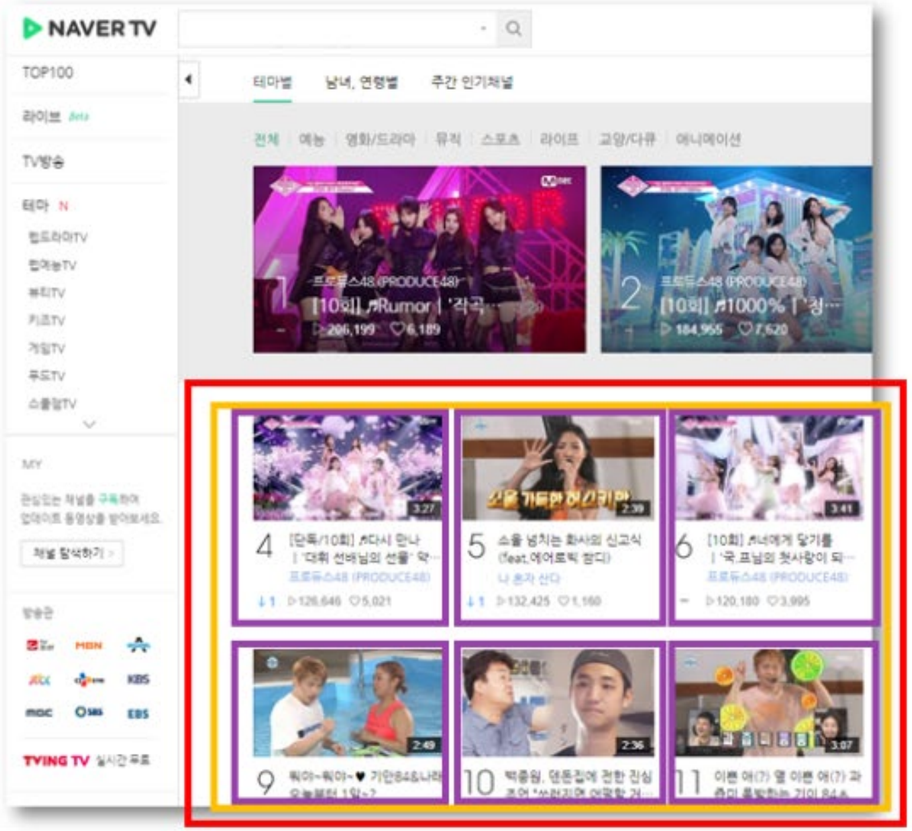
이를 “**선택자 경로** " 라고 함




```
<div id="content">
   ...
        <div class="cds_area">
          <div class="cds"></div>
          <div class="cds"></div>
          <div class="cds"></div>
          <div class="cds"></div>
          <div class="cds"></div>
          <div class="cds"></div>
        </div>
   ...
</div>
```

선택자 경로

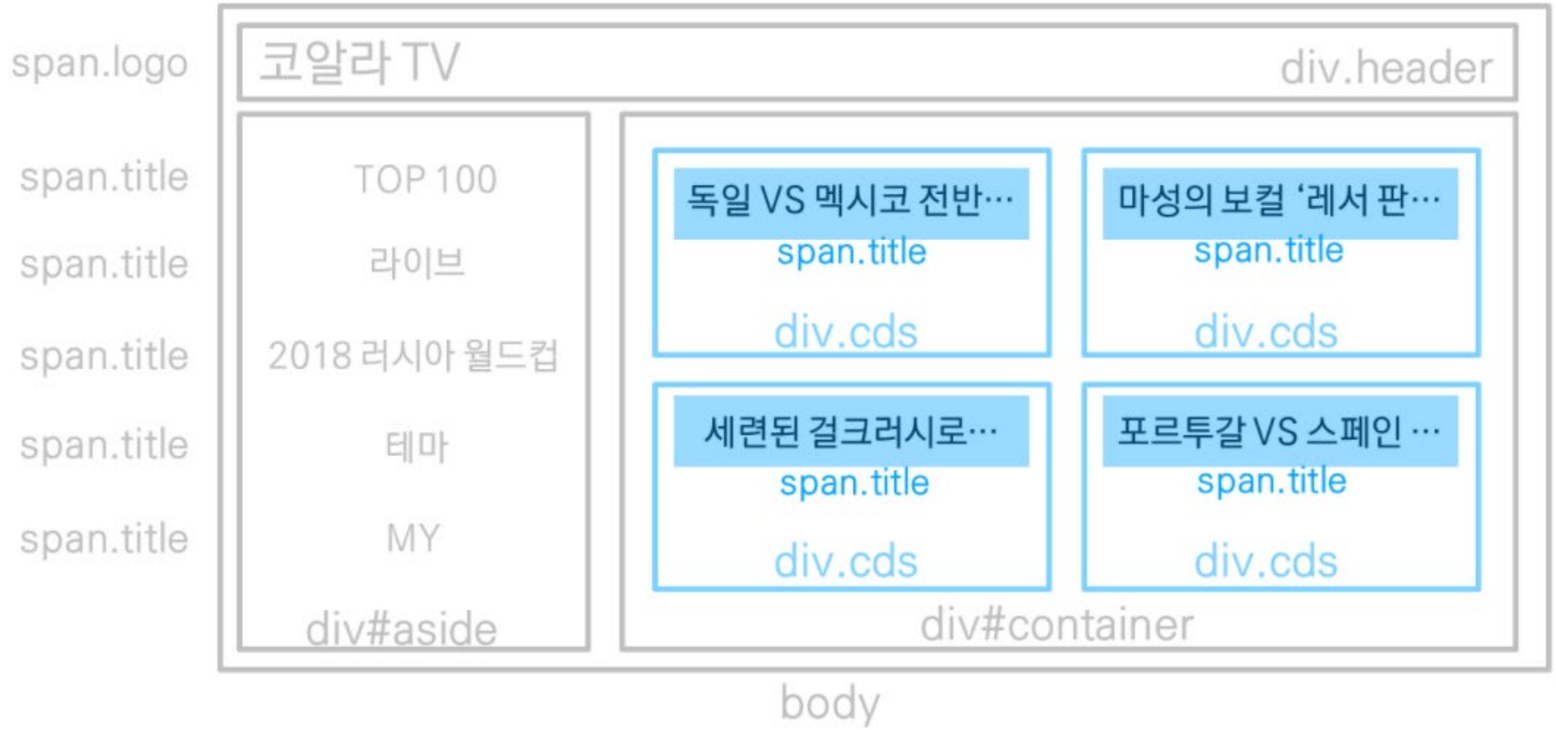
div#content > div.cds area > div.cds
Or
Div#content div.cds




```
<div id="content">
   ...
        <div class="cds_area">
          <div class="cds"></div>
          <div class="cds"></div>
          <div class="cds"></div>
          <div class="cds"></div>
          <div class="cds"></div>
          <div class="cds"></div>
        </div>
   ...
</div>
```

선택자 경로 지정 – 자식 선택자 '〉'

자식 선택자 내 바로 아래 자식 태그만 검색

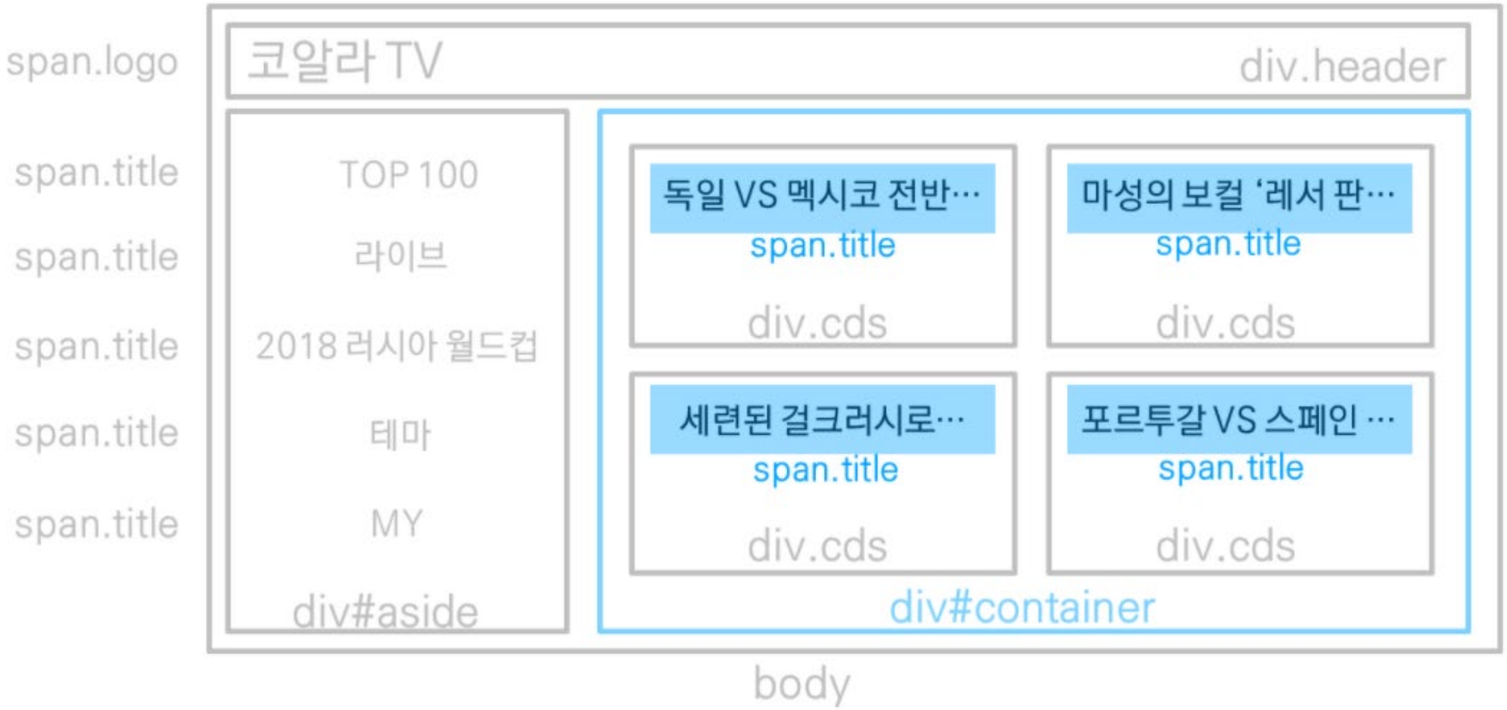


div.cds 〉 span.title

선택자 경로 지정 – 자손 선택자 ‘ ’(공백)

자손 선택자

내부에 존재하는 모든 태그를 검색
(자식, 그 자식의 자식, 그 자식의 자식의 자식..)



* div#container > span.title
이었다면 아무것도 찾아내지 못함

div#container span.title

에브리타임 html 구조 살펴보기

THANK YOU