

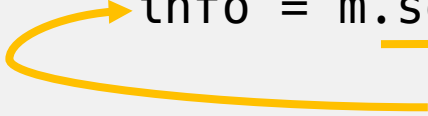
# WEEK 5. 데이터 수집 업그레이드

이성민

# 1. 순서를 활용해서 데이터 선택하기

네이버 영화 데이터 수집 - 선택자가 같을 경우

## ① 리스트 활용

 `info = m.select("dl.info_txt1 dd")`  
데이터를 리스트 형태로 저장  
·인덱싱 방법 활용하기  
`info[0], info[1], info[2]`

해당 데이터 없으면 에러  
(try-except 사용)

## ② 선택자 응용하기

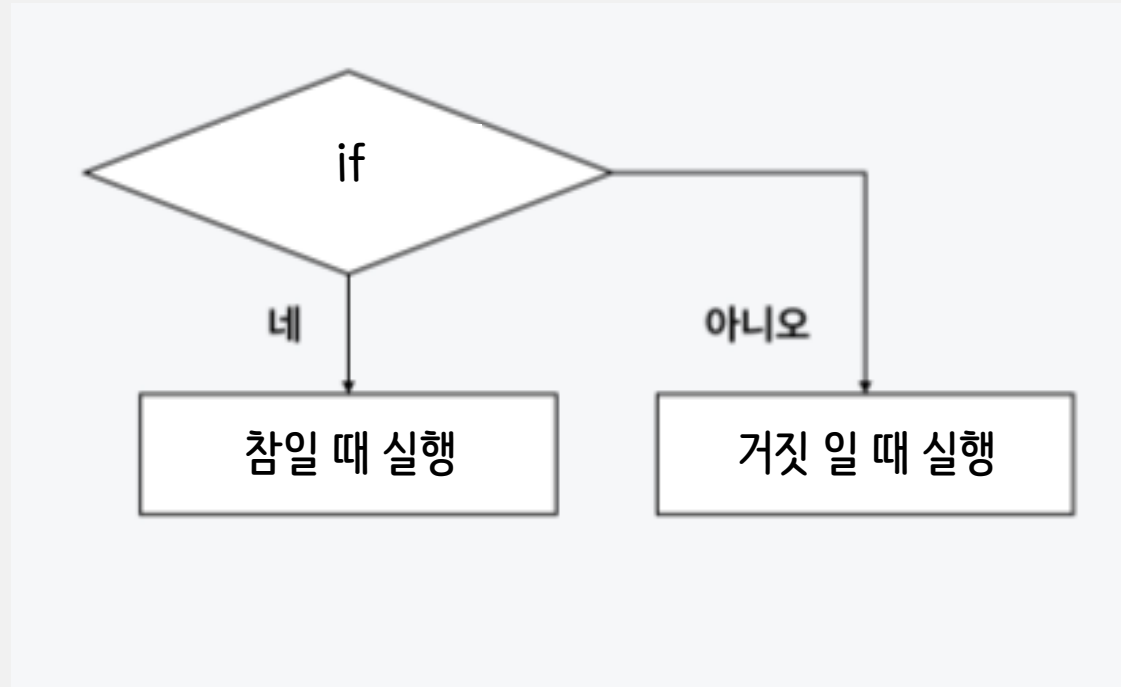
태그이름: `nth-of-type( 순서 )`

클래스, id는 쓸 수 없다

'순서' 번째에 있는 태그 선택 (1부터 시작한다!)

해당 데이터 없어도 에러 X  
(빈 리스트로 들어감)

## 2. 조건에 따라서 데이터 수집하기



if 조건문1 :  
    조건문 1이 참일 때 실행

elif 조건문2 :  
    조건문 1이 거짓이고 2가 참일 때 실행

else:  
    모두 거짓일 때 실행

## 2. 조건에 따라서 데이터 수집하기

### ① 숫자 비교

비교연산자	설명	결과1 (A=3, B=1)	결과2 (A=2, B=2)
A > B	A가 B보다 크다.	True	False
A >= B	A가 B보다 크거나 같다.	False	True
A == B	A와 B가 같다.	False	True
A != B	A와 B가 같지않다.	True	False

### ② 논리 연산

비교연산자	설명	결과1 (A=True, B=True)	결과2 (A=True, B=False)
A and B	A와 B가 둘다 참이다.	True	False
A or B	A 또는 B가 참이다.	True	True
not B	B가 거짓이다.	False	True

비교할 때는 타입 확인! (형변환)

## 2. 조건에 따라서 데이터 수집하기

A in LIST

A가 리스트 안에 있다

A not in LIST

A가 리스트 안에 없다

해당하는 값이 있을 때

A in String

A가 문자열 안에 있다

A not in String

A가 문자열 안에 없다

일부가 포함되어 있을 때  
(for문 활용)

continue

자신이 속해있는 반복문을 다음 반복으로 넘긴다  
(주로 if와 함께 사용)

### 3. 링크 안에서 데이터 수집하기

〈태그 속성=" "〉

태그: a

속성: href

태그: img

속성: src

선택자	<code>BS.attrs["속성이름"]</code>
설명	선택한 코드에서 원하는 속성의 속성값을 저장합니다. *해당하는 속성이 없는 경우 에러가 발생합니다.

```
url = title.attrs["href"]
: title에 들어있는 a 태그의 href 값을 url에 저장
```

### 〈새로운 페이지 안에서 상세 데이터 찾기 (앞과 동일하게 진행)〉

```
raw_each = request.get("앞 주소" + url , headers={'User-Agent' :
            'Mozilla/5.0'})
```

```
html_each = BeautifulSoup(raw_each.text , 'html.parser')
```

## 4. 이미지 데이터 수집하기

```
from urllib.request import urlretrieve
```

파이썬 기본 라이브러리

```
urlretrieve("주소", "파일이름")
```

해당 주소에 저장되어 있는 데이터를 파일 이름으로 저장

```
poster = each_html.select_once("선택자")
```

```
poster_src = poster.attrs["src"]
```

```
urlretrieve(poster_src, "poster/" + title[:2] + ".png")
```