파이썬
Web Crawling

BeautifulSoup & Selenium

Contents

- Web
- Crawling
- bs4
- selenium

Web

- Web 3대 요소
  - HTML : 구조 (뼈)
  - CSS : 디자인(표현)
  - JavaScript : 동작(행동)

Web

- HTML

  -Hyper Text Markup Language -> HTML

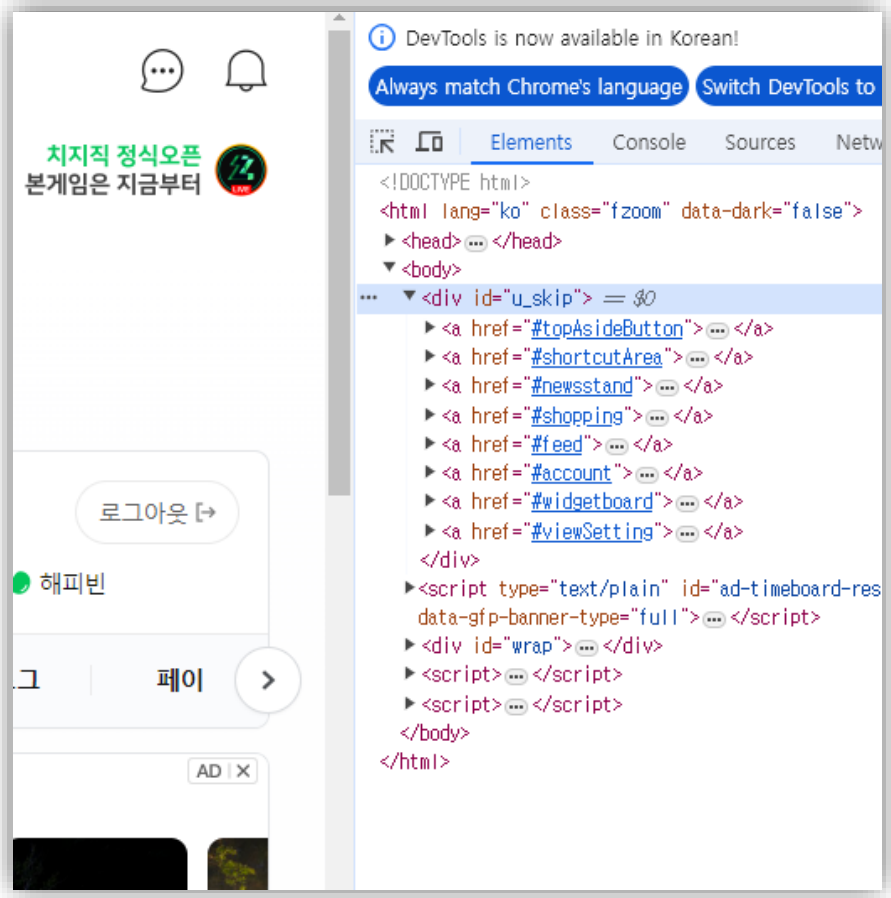
  -태그, 속성, 내용 으로 구성

  -자식, 부모 관계

  예) <tag> text </tag>

  <tag 속성="속성값"> text </tag>

   <a>



Web

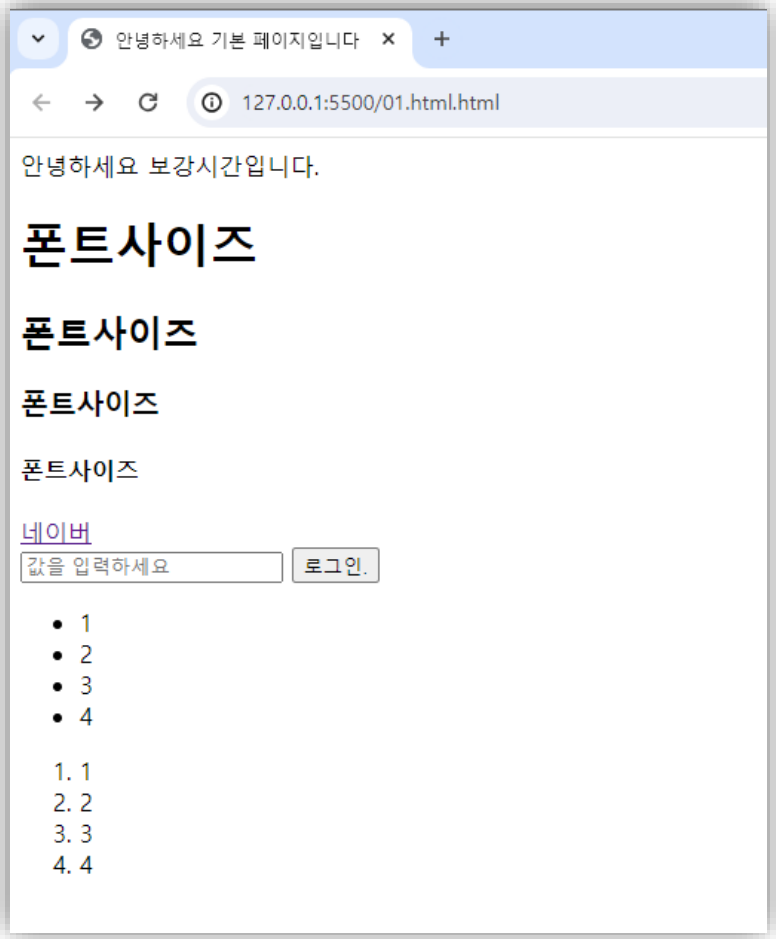
- HTML

```html
<!DOCTYPE html>
<head>
    <title>안녕하세요 기본 페이지입니다</title>
</head>

<body>
    안녕하세요 보강시간입니다.
    <h1>폰트사이즈</h1>
    <h2>폰트사이즈</h2>
    <h3>폰트사이즈</h3>
    <h4>폰트사이즈</h4>

    <a href="https://www.naver.com">네이버</a>
    <div>
        <input type="text" placeholder="값을 입력하세요" />
        <button onclick="alert('로그인')">로그인.</button>
    </div>

    <ul>
        <li>1</li>
        <li>2</li>
        <li>3</li>
        <li>4</li>
    </ul>
    <ol>
        <li>1</li>
        <li>2</li>
        <li>3</li>
        <li>4</li>
    </ol>
</body>
```



Web

- CSS

  -Cascading Style Sheets(CSS) 웹페이지 디자인

  - 선택자, 속성명, 속성값

  예) h1{color : red;}

```
h1 { color : red; }
```

선택자 **속성명** **속성값**

-> ***모든 h1 태그의 글자 색을 빨간색으로 바꿔라***

Web

- 선택자 (selector)

  1) 태그 선택자 : 태그 이름으로 선택하는 것

```
<body>
    <h1>제목_h1</h1>
    <p>본문내용</p>
</body>
```

```
h1{color: red;}
p{color: blue;}
```

Web

- 선택자 (selector)

2) 클래스 선택자 : 클래스 속성 값으로 선택하는 것, 태그에 별명을 주는 것

- 그룹을 지정해주는 것 -> ***css 앞에 .***

```
<body>
    <h1>제목_h1</h1>
    <p class="cls1">본문내용1</p>
    <p class="cls1">본문내용2</p>
    <p class="cls2">본문내용3</p>
    <p class="cls1">본문내용4</p>
</body>
```

```
.cls1{color: red;}
.cls2{color: blue;}
```

```
.cls1 {
    color: red;
}
.cls2 {
    color: blue;
}
```

Web

- 선택자 (selector)

  3)아이디(id) 선택자 : 아이디 속성 값으로 선택하는 것

  -태그에 별명을 주는 것. ***-> css 앞에 #***

```
<body>
    <h1>제목_h1</h1>
    <p id="id1">본문내용1</p>
    <p>본문내용2</p>
    <p>본문내용3</p>
</body>
```

```
#id1{color: red;}
```

Web

- 선택자 (selector)

4)자식 선택자 : 바로 아래 자식태그를 선택하는 것, 원하는 태그의 별명이 없을 때 사용

```
<body>
    <div>
        <p>p1</p>
        <p>p2</p>
    </div>
</body>
```

```
div > p {color: red;}
```

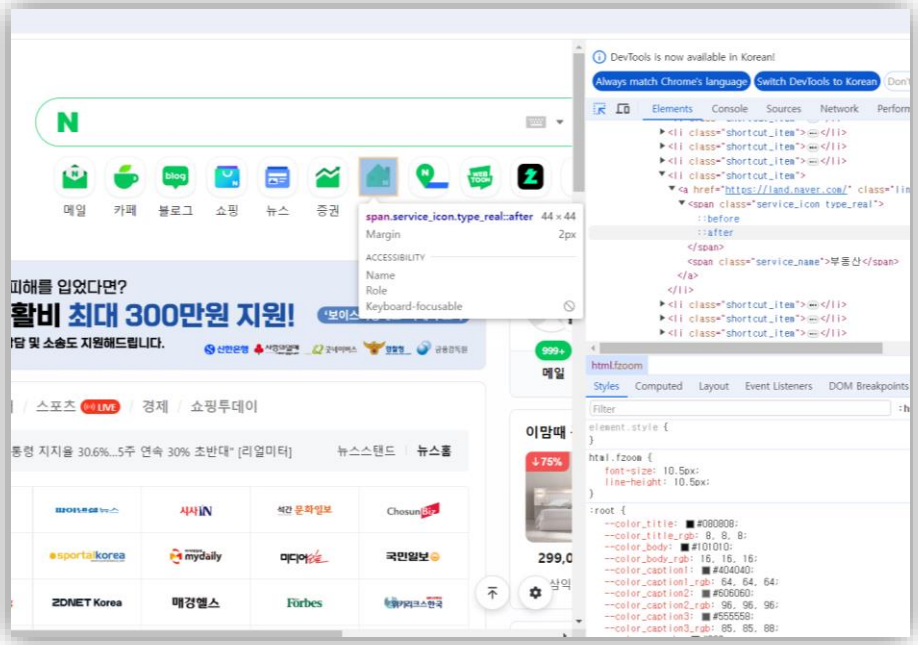
```
<body>
    <div class="cls1">
        <p>p1</p>
        <p>p2</p>
    </div>
</body>
```

```
.cls1 > p {color: red;}
```

Crawling

- 웹크롤링

  - 웹 페이지 내 데이터를 추출



Crawling

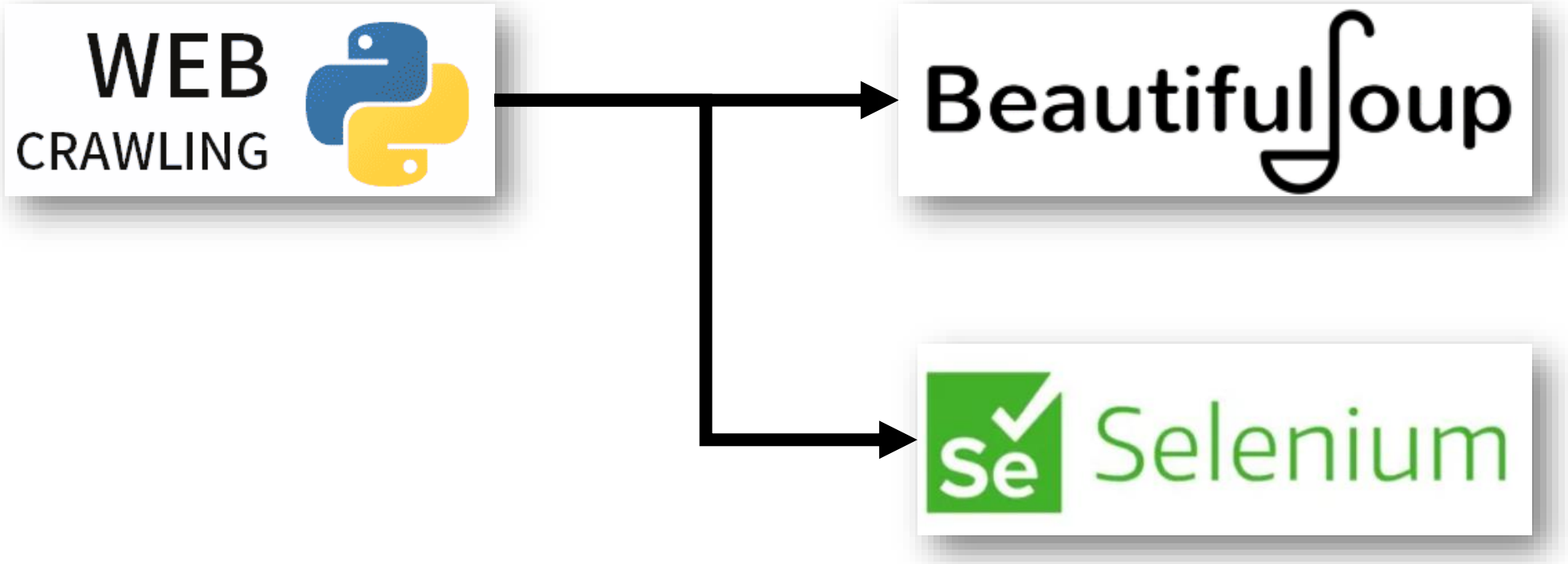
- 크롤링 VS 스크래핑

Crawling

- Process

Crawling

- Python Crawling Library



Crawling

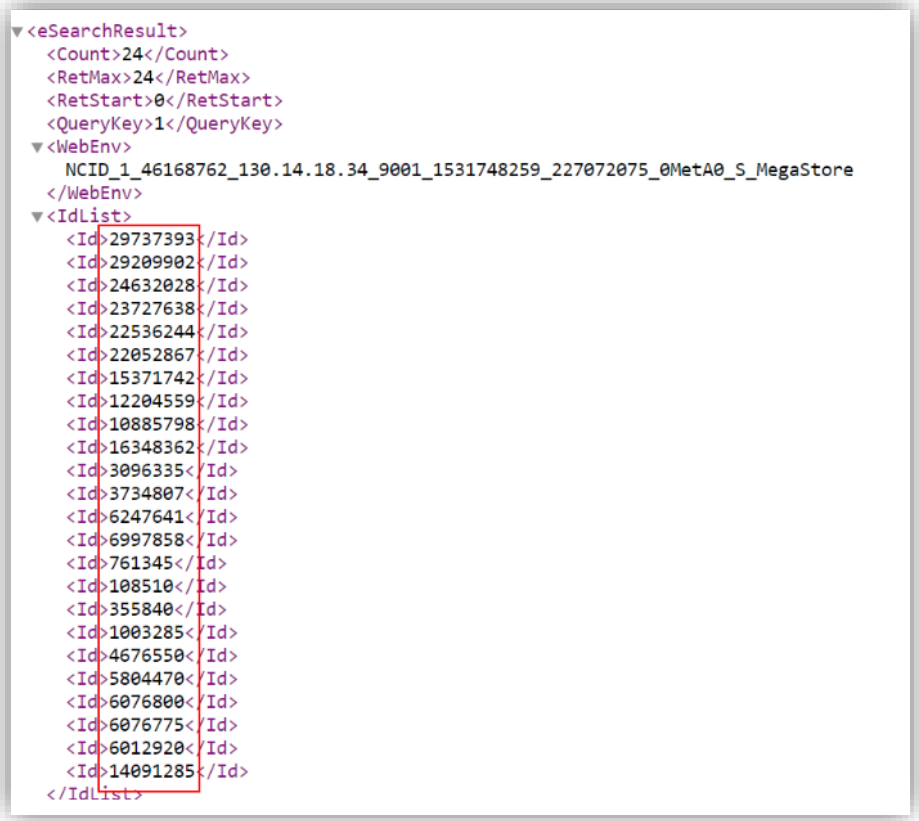
- BeautifulSoup4

  대표적인 파이썬 정적 웹 크롤링 라이브러리

  - 멈춰있는(정적) 페이지의 html을

    requests와 함께 이용하여 사용

  - 수집속도가 빠르지만 범용성이 떨어짐



Crawling

- Selenium

  대표적인 파이썬 동적 웹 크롤링 라이브러리

  - 계속 움직이는 페이지를 다루는 패키지
  - Chromdriver와 함께 사용되며,

    직관적으로 crawling 가능
  - 수집속도가 느리지만 범용성이 높음