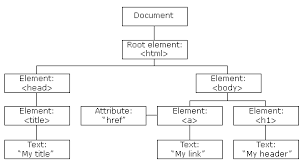
**<파이썬 실전 웹크롤링>**

1. HTML구조(DOM구조)

* DOM : 서버로부터 받은 정보는 텍스트 파일로 되어 있는데 웹브라우저에서 읽기 위해 변환 과정이 거쳐 메모리에 저장되어야 함. 이 과정에서 텍스트 파일을 웹브라우저가 읽을 수 있는 엘리먼트 트리구조로 만드는 것을 DOM이라고 함.
* Selenium()을 이용할 때, 원하는 자료를 파악하기 위해 DOM구조 이해 필요



* Selenium을 이용하여 정보를 크롤링 할 때, DOM구조에 대한 이해 필요.

(Selenium : 웹 브라우저를 이용하여 data를 불러오는 모듈)

* 웹 구조 : HTML, 표현 : CSS, 행위 : Java script

1. 변화 없이 HTML내용이 변하는 웹사이트 구조(동적 웹사이트 구조)

* 동적 웹사이트 구조는 URL변화 없이 웹사이트 내용이 바뀌는 것을 의미함.
* JAVA Script를 이용하여 사용자의 정보에 따라, 사용자 입력, 동작에 의해 HTML문서 내용이 달라져

Ex) web application : 사용자의 행동의 HTML문서 내용 확인 가능. 무신사, OP.GG(게임정보페이지)

1. HTML input 구조

* 사용자에 의해 선택되거나 입력 되는 웹 브라우저 내용은 <input type=”values”> 형태로 작성되고 input 태그로 키보드나 마우스 없이 명령가능
* Values값으로는 아래와 같은 종류의 속성값을 가짐.

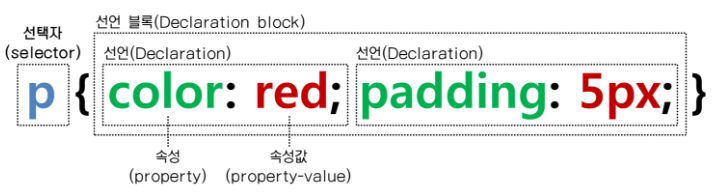
button, checkbox, file, hidden, image, password, radio, reset, submit, text 등

* Drop down메뉴( )

: Select 태그로 내용을 포함하고 있음.

1. Selenium 웹 브라우저 조작

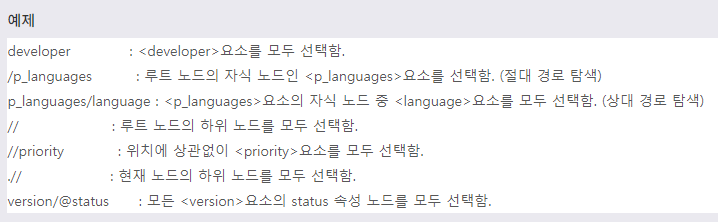
* Selenium으로 웹 브라우저를 열고 각종 버튼과 입력 칸에 접근하여 원하는 정보 탐색 방법
  + 1. driver.find\_element\_by\_id() : 엘리먼트 속성값 id를 사용하여 접근
    2. driver.find\_element\_by\_class\_mane() : 엘리먼튼 속성값 claa를 이용하여 접근
    3. driver.find\_element\_by\_css\_selector() : 엘리먼트의 css\_selector 이용
    4. driver.find\_element\_by\_xpath() : 엘리먼트의 xpath이용
* css\_selector : selctor는 태그가 될수도 있고 속성값이 될 수도 있음.

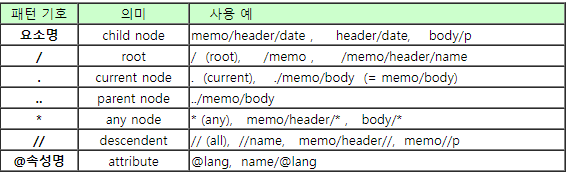


* 변수명. Find(셀렉터 표현식)

|  |  |  |
| --- | --- | --- |
| 셀렉터 표현식 | 예시 | 설명 |
| .class | .intro | class값으로 inrto를 갖는 엘리먼트 |
| #id | #first | Id가 first인 엘리먼트 |
| \* | \* | 모든 엘리먼트 |
| Elelment | P | 모든 p태그 엘리먼트 |
| Element, element | Div, p | 모든 div, p태그 |
| Element element | Div p | Div 안에 있는 p태그 엘리먼트 |
| Element > element | div>p | Div 바로 아래 계층의 P태그 모두 |
| Element + element | Div+p | Div 다음의 같은 계층 P태그 |
| Element ~ element | P~ul | P태그 뒤에 오는 ul 태그 모두 |

* xpath : XML 문서의 특정 요소나 속성에 접근하기 위한 경로를 지정하는 언어





1. 사용 실제 구문

* 참고 자료 selenium1, 2, 3
* Zigbang, musinsa 파일