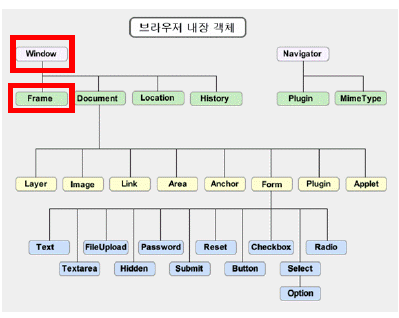
<WEEK 4>

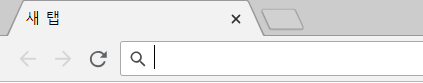
**파이썬을 활용한 실전 웹 크롤링**

* HTML 계층 학습
* HEADLESS BROWSER

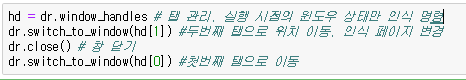
1. HTML계층
2. HTML구조



* Window : 크롬창의 탭. 새창이 열리거나 여러창을 열어서 크롤링 해야하는 경우 사용

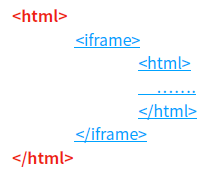


* 팝업, 새창이 생겼을 때, 크롤링을 방해하게 되는데 window를 이용하여 탭 이동 가능
* Index는 [0]부터 시작
* 실습예제 “popup\_handling”



1. Frame

* HTML 안에 또 다른 HTML문서를 넣을 때, 이용



* 크롤링을 할 때, 가장 바깥의 HTML문서만을 명령으로 인식하고 찾기 때문에 HTML내 HTML문서의 내용을 찾기 위해서 Frame에 대한 이해 필요
* 실습예제 “NAVER café”

1. HEADLESS BROSER

* 작업이 진행되는 과정을 보여주는 창 없이 명령 실행하는 모드

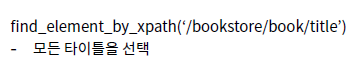
|  |  |  |
| --- | --- | --- |
|  | Browser | Headless browser |
| User interface | Graphic User Interface  시각적 사용자 인터페이스  Ex) 다운로드 버튼 | Command-Line Interface  커맨드 창을 이용한 인터페이스 |
| 조작 방법 | 브라우저를 마우스로 조작 | 콘솔상의 코드 입력 |
| 실제 예시 | 브라우저 웹서핑 시 GUI를 통해 웹브라우저 통제 | 코드로 크롬드라이버에 입력을 줄 경우, CLI를 통해서 브라우저 통제 |
| 장점 | selenium코드에 의한 진행과정 확인 가능 | Rendering 작업 가상으로 처리하여 속도 빨라 |
| 단점 | 브라우저 상에 시각적으로 표현하는 rendering작업으로 느림 | 예상치 못했던 오류 시, 코드를 다시 실행해서 원인을 찾아야 함. |

* 따라서 Headless 모드는 완성된 코드로 작업을 수행할 때, 이용
* 종류 : Phantom.js, Chrome, Firefox
* Phantome.JS 더 이상 오류 수정 되지 않음.
* groups.google.com/forum/#!forum/chromedriver-users에서 관련 오류정보 습득이나 문의 가능
* Chrome driver option : 자동 로그인, 사용자가 설정한 옵션들을 추가하여 검색환경 변경하는 옵션

1. X-path

* Selector의 한 종류.
* HTML계층 구조를 이용하여 같은 구조의 계층이 여러 가지로 구성되어 있어 하나의 내용만을 불러오고 싶을 때, 용이함.
* 같은 계층이 있을 때, index값을 지정하여 불러오기 가능. Index는 [1]부터 시작.

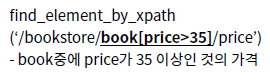




>>> Everyday Italian, Harry Potter, XQuery Kick Start, Learning XML 라인이 webelement형태로 출력



>>>Everyday Italian 라인 web element형태로 출력

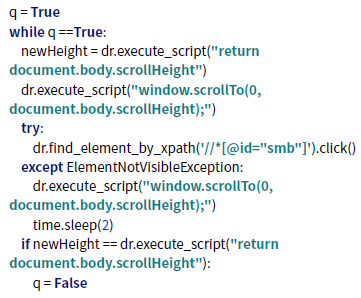


>>> XQuery Kick Start, Learning XML 라인이 web element형태로 출력

* Book태그 안에 price 태그 값의 범위에 대한 표현 가능

1. 웹 스크롤 이동하여 정보 불러오기

* Lazy load처럼 스크롤을 이용해야 정보를 불러와 보여주는 경우(쇼핑몰, 구글 이미지 등) 스크롤 내리는 명령 필요
* 아래의 코드는 구글 이미지를 예시로 함.



**현재 보이는 페이지의 전체 스크롤 높이**

* newHeight는 화면에 보이는 구성의 스크롤 높이를 할당.(최하단의 위치 의미)
* (“Window.scrollTO(0, document.body.scrollHeight);”

현재 창에서 스크롤이 있는 위치(0)부터 현재 창의 최하단까지 스크롤 이동

* Try/except 코드는 예외 상황 조건 부여

코드 내용 : 버튼이 생성될 때, 버튼을 누르거나 아니면 스크롤을 계속 내린다.

* If문 : 반환되는 스크롤의 위치가 처음 저장했던 newHeight과 같은 값이 되면 반복문 종료