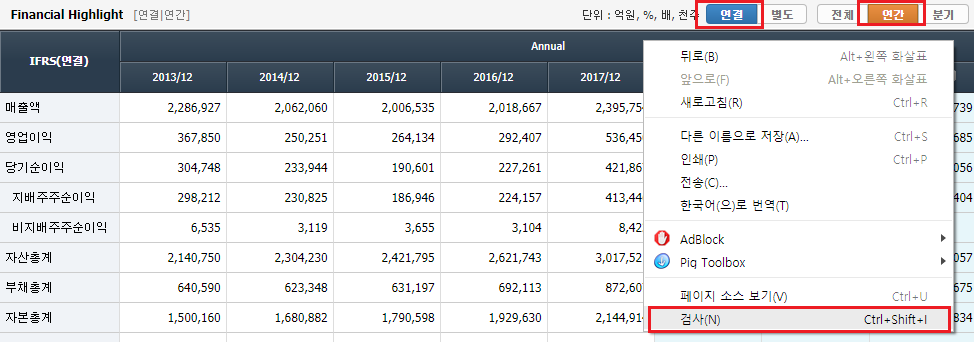
페이지 내에서 특정값 혹은 테이블만 크롤링 할 경우 XPath를 이용[[1]](#endnote-1)하는 것이 매우 유용할 때가 있습니다. XPath란 XML 중 특정 값의 태그나 속성을 찾기 쉽게 만든 주소라 생각하면 됩니다. 예를 들어 R 프로그램이 저장된 곳을 윈도우 탐색기를 이용하여 이용할 경우 **C:\Program Files\R\R-3.4.1** 형태의 주소를 보이며, 이는 윈도우의 path 문법입니다. XML 역시 이와 동일한 개념의 XPath가 존재합니다.

XPath를 찾는 방법은 매우 간단합니다. Financial Highlight 부분에서 [연결]과 [연간] 탭을 선택한 후, 크롤링하고자 하는 테이블위에 마우스를 둔 채 우클릭을 한 후 가장 하단의 검사버튼을 누르면 개발자도구 화면이 열립니다.

**[그림] 크롤링 하고자 하는 부분의 검색과정**



그 후 개발자도구 화면의 HTML 문서에서 마우스를 위아래로 움직이면, 해당 문서가 웹페이지에서 나타내는 영역이 음영처리가 됩니다.

**[그림] HTML 문서 내 Financial Highlight 영역**



해당 문서에서는 **highlight\_D\_Y** ID 하부의 table 태그 부분이 Financial Highlight 테이블을 나타내는 부분입니다. 해당 부분에는 총 6개의 ID가 존재하며, 구체적 항목은 다음과 같습니다.

|  |  |
| --- | --- |
| **ID** | **구분** |
| highlight\_D\_A | 연결 / 전체 |
| highlight\_D\_Y | 연결 / 연간 |
| highlight\_D\_Q | 연결 / 분기 |
| highlight\_B\_A | 별도 / 전체 |
| highlight\_B\_Y | 별도 / 연간 |
| highlight\_B\_Q | 별도 / 분기 |

개발자도구 화면의 <table class="us\_table\_ty1 h\_fix zigbg\_no"> 부분에 마우스를 둔 채 우클릭 후 Copy → Copy XPath를 선택하면, 클립보드에 자동으로 XPath가 복사됩니다. 복사된 내용을 확인하면, 해당 부분의 XPath는 다음과 같습니다.

**[그림] 개발자도구 화면에서 XPath의 복사 방법**



|  |
| --- |
| //\*[@id="highlight\_D\_Y"]/table |

해당 XPath를 이용하여 요약 재무제표를 크롤링하도록 합니다. read\_html() 함수를 통해 해당 페이지의 HTML 정보를 불러온 후, html\_nodes() 함수를 사용하며, xpath 항목에 위에서 구한 값을 입력해 줍니다. 그 후 html\_table() 함수를 이용하여 표 형태의 데이터 프레임을 찾아 data\_table 변수에 저장합니다. 셀병합 등으로 모든 행의 열 개수가 일치하지 않을 경우 html\_table() 함수는 오류가 발생하며, fill = TRUE 변수를 추가해주어 병합된 셀들을 해제하여 줍니다. 또한 크롤링 단계 앞뒤로 Sys.setlocale() 함수를 추가해주어 인코딩 문제를 방지해 줍니다.

1. read\_html() 함수와 html\_table() 함수를 이용할 경우 바로 테이블 형태를 수집할 수 있지만, 각 종목의 페이지별 테이블의 개수가 다르므로 해당 방법은 바람직하지 않습니다. [↑](#endnote-ref-1)