select

업데이트
2020.06.16. 19:16
참여저자
성빈, 도리도리m, 뷥븩탄 외 1명 상세보기
댓글
바로가기

책갈피
https://cafe.naver.com/nameyee/book5105233/15687 주소복사
Jsoup의 select 메소드를 이용하여, HTML 문서 내의 필요한 element만을 선택할 수 있습니다.
기본적으로, get/post/execute() 메소드로 Request를 날린 이후에 select를 하게 됩니다.
select메소드는 elements 객체를 반환합니다.

여기에서는 https://osu.ppy.sh/rankings/osu/performance 페이지를 기준으로 select를 실습해 봅니다.
soup = org.jsoup.Jsoup.connect('https://osu.ppy.sh/rankings/osu/performance').get() 을 실행해두었다 가정하고 작성합니다.

1. Select 기초

※ 아래 나열되는 선택방법은 서로 조합하여 사용이 가능합니다.

1) tag 선택 : 'tag'
soup.select('table')을 실행해 봅시다.
<table class="ranking-page-table">이라는 element와 모든 하위 태그가 select 되었습니다.

2) id 선택 : '#id'
soup.select('#scores')를 실행해 봅시다.
 element가 select되었습니다.

3) class 선택 : '.class'
soup.select('.flag_country')를 실행해 봅시다.

...
등 여러 국가의 국기가 담긴 span element들이 select되었습니다.

4) 속성 선택 : '[속성명]'
속성(attribute)을 이용해서도 선택할 수 있습니다.
soup.select('[data-orig-title]')을 실행해 봅시다.
 등 role이라는 속성값을 가진 태그 2개가 select 되었습니다.

5) 속성의 값으로 찾기 : '[속성명=속성값]'
속성명, 속성값의 조합으로 찾는 방법도 존재합니다.
soup.select('[data-user-id="9224078"]')를 실행하면,
 FlyingTuna
이렇게 특정 속성과 속성값이 존재하는 element를 찾습니다.

6) 속성값의 일부로만 찾기
속성값의 일부(시작, 끝, 포함, 정규식)로 select를 할 수도 있습니다.
아래와 같은 형태로 select를 합니다.
.select('[속성명^=시작값]')
.select('[속성명$=끝값]')
.select('[속성명*=포함값]')

예를들어 page-table 이라는 class 값을 포함하는 태그를 select하려면,
.select('[class*=page-table]') 과 같이 표현합니다.

7) pseudo selector
pseudo selector는 여러 조건들을 이용하여 select하기 위한 방법으로,
위의 셀렉터 뒤에 콜론(:)을 붙여 사용합니다.
많이 쓰이는 몇가지만 소개합니다.
:eq(n) : 동일 태그 중 n번째 태그를 선택합니다.
:has(selector) : selector 조건을 하위에 두고있는 태그를 선택합니다.
:contains(abcd) : 태그 내 abcd라는 텍스트를 포함하는 태그를 선택합니다.
:not(selector) : selector 조건에 해당되지 않는 태그만 선택합니다.

예를 들어 body tag 안에 div 태그가 4개가 있는 경우, .select('body > div:eq(4)')로 셀렉트를 하면 마지막 div가 select 됩니다.

8) Tag의 상속구조를 이용하여 찾기

html tag의 구조를 이용하여 찾는 방법입니다
soup.select('body > div.warp > em')과 같이 실행하면,
html 태그 내 body 안에 div class = "warp" 안에 em을 얻어올수 있습니다.

2. Select된 element의 값 얻어오기

1) text(), wholeText(), toString()

text()는 select된 태그 내 텍스트(string)를 반환합니다. 이 때, 더블스페이스는 스페이스 한개로 치환되며, 불필요한 공백문자, 라인브레이크를 삭제하여 순수 문자열만 반환합니다. 모든 하위 태그는 사라집니다.

wholeText() 역시 select된 태그 내 텍스트(string)를 반환합니다. 다만, 더블스페이스를 치환하지 않습니다. text()와 마찬가지로 모든 하위태그는 사라집니다. wholeText()는 elements 객체에서는 사용이 불가능하며, 아래에서 설명할 .get() 등을 이용하여 명시적으로 한개의 element만 가져올 경우 사용이 가능합니다.

toString()은 가져오려는 값 중 태그가 존재하는 경우, 태그를 살리기위해 사용하며, 아래와 같이 사용합니다.

```
elem = soup.select('셀렉트할조건');
elem.outputSettings().prettyPrint(false);
elem.toString();
```

2) size(), get()

select 시 여러개의 element가 반환될 수 있습니다.
이 때 반환된 element의 갯수는 .size()로 확인할 수 있으며,
n번째의 element를 얻고자 하는 경우 .get(n)으로 특정 태그만 선택할 수 있습니다.

이를 이용하여 for문을 사용하면, 반복되는 구조를 손쉽게 처리할 수 있습니다.
아래의 예제를 이용하여 실습해 봅니다.

```
elems = soup.select('.ranking-page-table__row');
res = "";
size = elems.size();
for (i = 0;i<size;i++){
currentElem = elems.get(i)
res = res + [currentElem.select('a').get(1).text(),
currentElem.select('td').get(2).text(),
currentElem.select('td').get(3).text(),
currentElem.select('td').get(4).text()].join('\n') + '\n\n';
}
```

위와 같이 실행하면, 예제 사이트에서
순위별 사용자명, 정확도, 플레이횟수, 퍼포먼스 점수를 나열하게 됩니다.

참조
https://jsoup.org/cookbook/extracting-data/selector-syntax

 뷥븨탄

- 2020.04.08. 22:16

신고

BennyK Get은 url뒤에 붙고 post는 안붙어요 그리고 execute는 모르겟네요

- BennyK
- 2020.04.08. 22:16

신고

뷥븨탄 흠... 억떡계

-  뷥븨탄
- 2020.04.08. 22:19

신고

BennyK cute는 아는데 execute는 모르겠네요

- BennyK
- 2020.04.08. 22:20

신고

뷥븨탄 아 제가 cute하죠 ㅠㅠ

-  뷥븨탄
- 2020.04.08. 22:20

신고

BennyK ?

-  뷥븨탄
- 2020.04.08. 22:20

신고

참고로 1번의 7번 제가함

-  뷥븨탄
- 2020.04.09. 21:28

신고

7번이 아니라 8번으로 바뀌었네

-

▲ top

이전글 parse [1]

다음글 예외문 [1]

도미 doami2005 2020.04.08.

아두이노 하는 사람 2020.04.12.