

Regular Expression

for Beginners

김용원

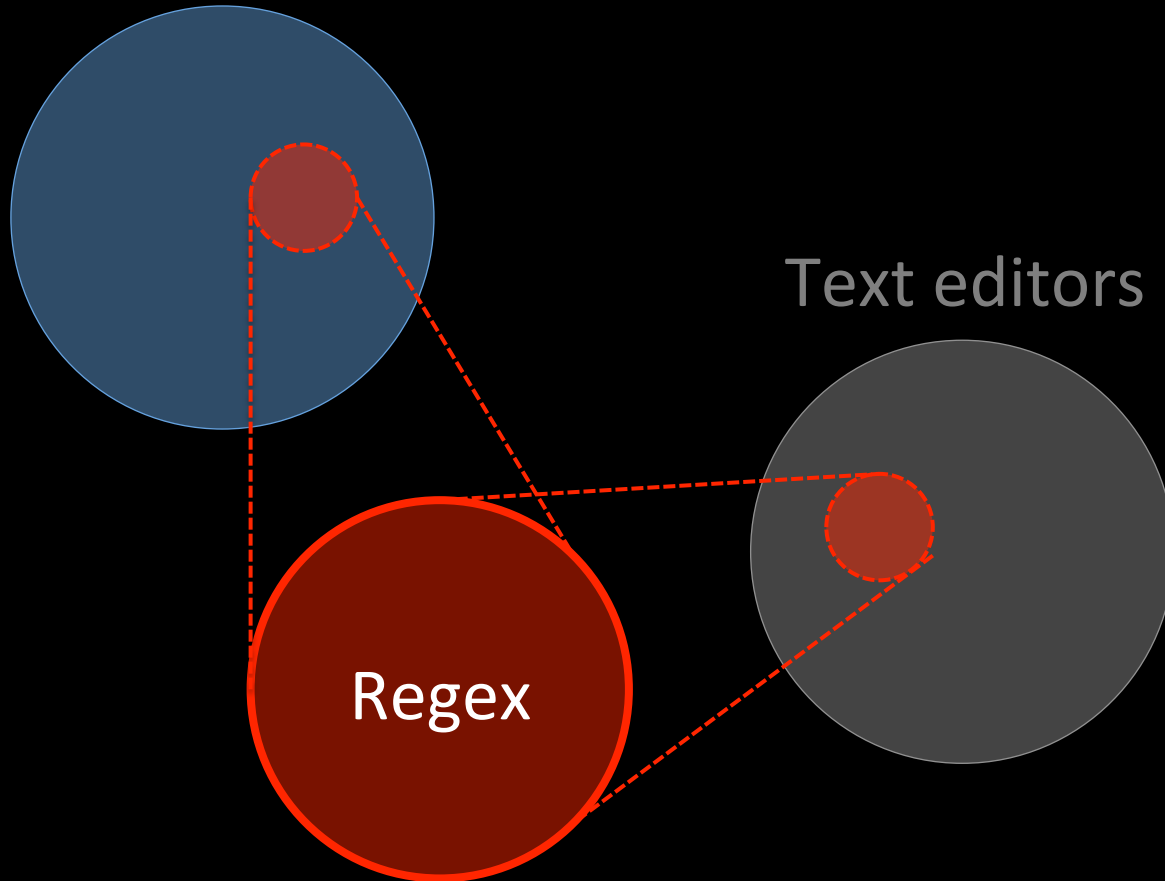
//((-a-zA-Z0-9^\\p{L}\\p{C})\\u00a1-\\uffff@:%_\\.~#?&//=)[2,256])(1)(\\.[a-z]{2,4})(1)(\\.[0-9]*)?(\\[-a-zA-Z0-9\\u00a1-\\uffff(\\[\\]@:%_\\.~#?&//=]*)?[/g

//<script(s|\\S)*?</script>)|(<style(s|\\S)*?</style>)|(<!--(\\s|\\S)*?-->)|(<\\/?(\\s|\\S)*?>)/g

정규 표현식(Regular Expression)

특정한 조건의 텍스트를
찾거나(Find) 바꾸기(Replace)에 쓰는 문자열이며,
정규 표현 언어를 사용해 만든다.

JS, Perl, Java, C ...



POSIX



(Portable Operating System Interface for uniX)

스페이스 문법 `[:space:]`

PCRE

(Perl Compatible Regular Expressions)

스페이스 문법 `\s`

시작기호

종료기호

문자열 패턴

/Hello\sWorld/gim

메타 문자

플래그

문자 클래스 - Character Classes

.	any character except newline
\w \d \s	word, digit, whitespace
\W \D \S	not word, digit, whitespace
[abc]	any of a, b, or c
[^abc]	not a, b, or c
[a-g]	character between a & g

위치 지정 - Anchors

^abc\$	start / end of the string
\b	word boundary

이스케이프 - Escaped characters

\. * \\	escaped special characters
\t \n \r	tab, linefeed, carriage return

그룹 & 탐색 - Groups & Lookaround

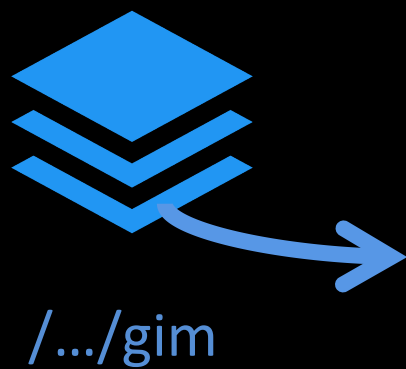
(abc)	capture group
\1	backreference to group #1
(?=abc)	positive lookahead
(?<=abc)	positive lookbehind (not JS)
(?!abc)	negative lookahead
(?<!abc)	negative lookbehind (not JS)

수량자 & 대체 - Quantifiers & Alternation

a* a+ a?	0 or more, 1 or more, 0 or 1
a{5} a{2,}	exactly five, two or more
a{1,3}	between one & three
a+? a{2,}?	match as few as possible (Lazy)
ab cd	match ab or cd

한글 범위 지정하기

[ㄱ- | 가-힝]



어디에 활용할 수 있을까?

- URL 링크로 감싸기(파일 인덱스)
- 다수 페이지에 잘못된 형식의 코드블럭 변경
- CSS클래스 사용빈도 측정
- HTML 소스에서 진짜 HTML 만 남기기
- 다양한 패턴의 이미지 주소 일괄 변경

...

정규식으로 HTML 가공하기

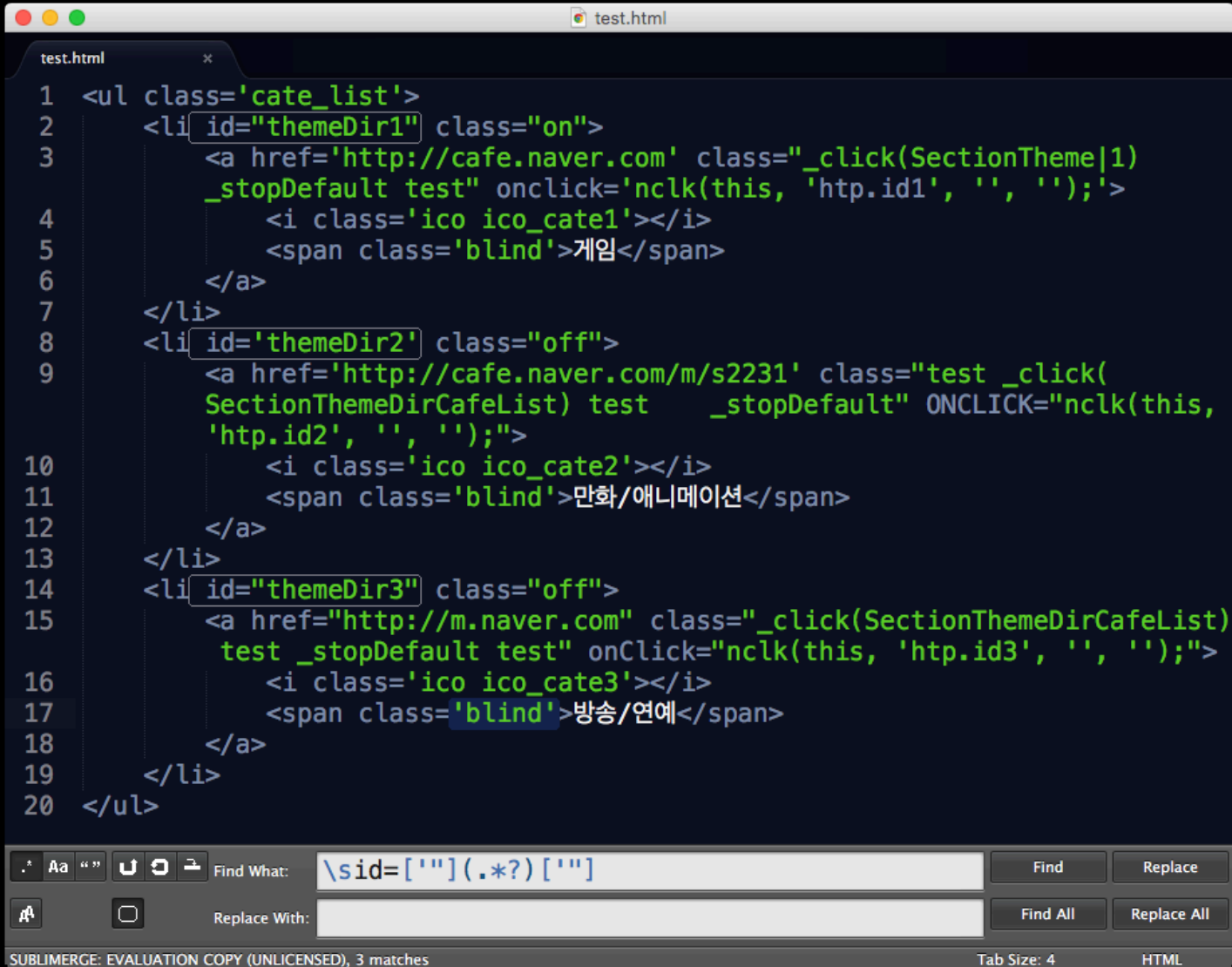
Sublime Text Editor Case

id, href="#", _classname, onclick

```
test.html x
1 <ul class='cate_list'>
2   <li id="themeDir1" class="on">
3     <a href='http://cafe.naver.com' class="_click(SectionTheme|1)
4       _stopDefault test" onclick='nclk(this, 'http.id1', '', '');'>
5       <i class='ico ico_cate1'></i>
6       <span class='blind'>게임</span>
7     </a>
8   </li>
9   <li id='themeDir2' class="off">
10    <a href='http://cafe.naver.com/m/s2231' class="test _click(
11      SectionThemeDirCafeList) test _stopDefault" ONCLICK="nclk(this,
12      'http.id2', '', '');">
13      <i class='ico ico_cate2'></i>
14      <span class='blind'>만화/애니메이션</span>
15    </a>
16  </li>
17  <li id="themeDir3" class="off">
18    <a href="http://m.naver.com" class="_click(SectionThemeDirCafeList)
19      test _stopDefault test" onClicK="nclk(this, 'http.id3', '', '');">
20      <i class='ico ico_cate3'></i>
21      <span class='blind'>방송/연예</span>
22    </a>
23  </li>
24 </ul>
```

SUBLIMERGE: EVALUATION COPY (UNLICENSED), Line 20, Column 6 Tab Size: 4 HTML

아이디 속성/값 제거 : `/\sid=["'](.*)["']/gim`



The screenshot shows a Sublime Text editor window titled 'test.html'. The editor contains an HTML snippet with three list items. The first item has an ID attribute, while the others do not. A search bar at the bottom is active, showing the regex `\sid=["'](.*)["']` and indicating 3 matches. The 'Replace With' field is empty.

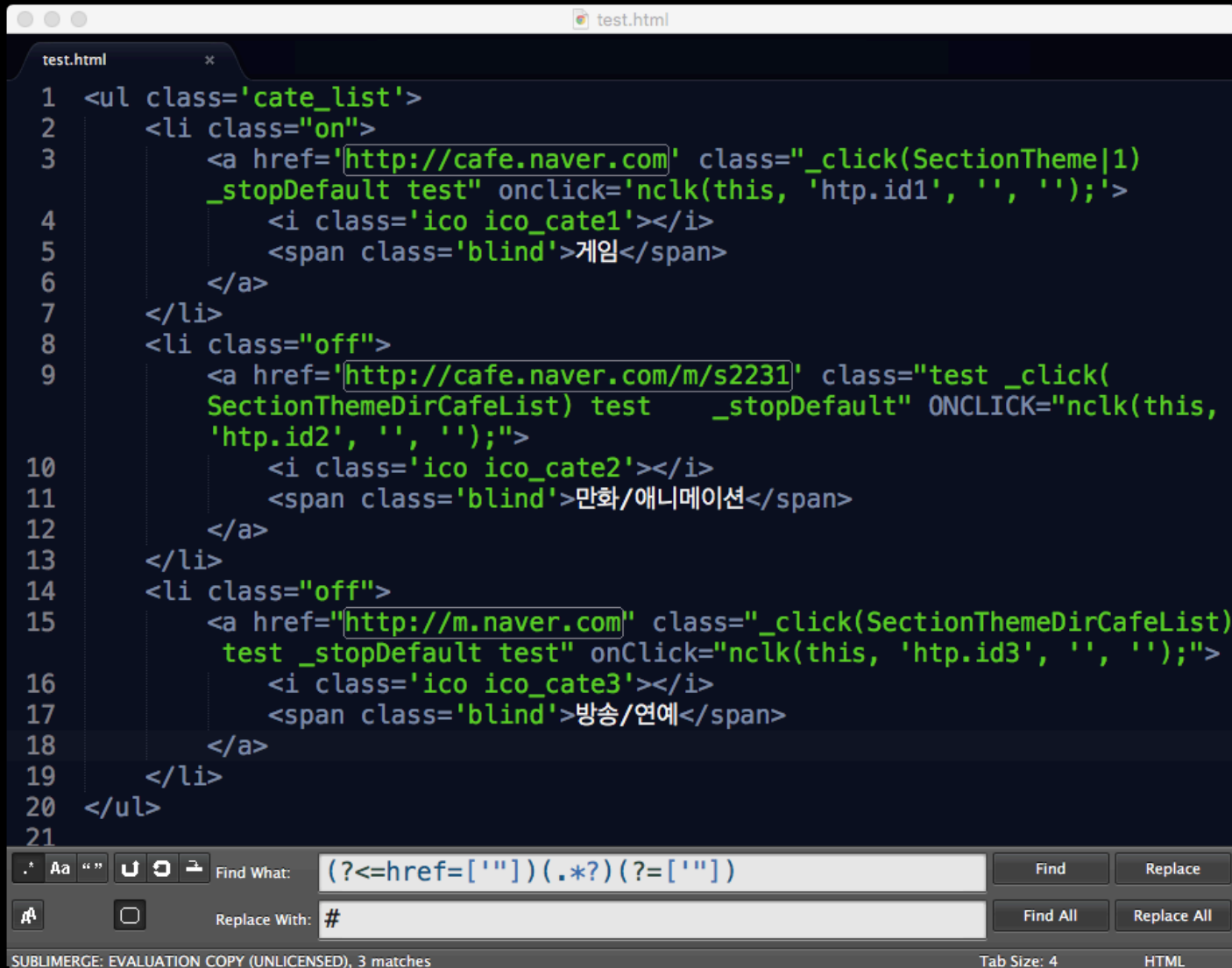
```
1 <ul class='cate_list'>
2   <li id="themeDir1" class="on">
3     <a href='http://cafe.naver.com' class="_click(SectionTheme|1)
4       _stopDefault test" onclick='nclk(this, 'http.id1', '', '');'>
5       <i class='ico ico_cate1'></i>
6       <span class='blind'>게임</span>
7     </a>
8   </li>
9   <li id='themeDir2' class="off">
10    <a href='http://cafe.naver.com/m/s2231' class="test _click(
11      SectionThemeDirCafeList) test _stopDefault" ONCLICK="nclk(this,
12      'http.id2', '', '');">
13      <i class='ico ico_cate2'></i>
14      <span class='blind'>만화/애니메이션</span>
15    </a>
16  </li>
17  <li id="themeDir3" class="off">
18    <a href="http://m.naver.com" class="_click(SectionThemeDirCafeList)
19      test _stopDefault test" onClick="nclk(this, 'http.id3', '', '');">
20      <i class='ico ico_cate3'></i>
21      <span class='blind'>방송/연예</span>
22    </a>
23  </li>
24 </ul>
```

Find What: `\sid=["'](.*)["']` Find Replace

Replace With: Find All Replace All

SUBLIMERGE: EVALUATION COPY (UNLICENSED), 3 matches Tab Size: 4 HTML

href 값 #으로 변경 : `/(?<=href=['"])(.*?)(?=['"])/gim`



The screenshot shows a Sublime Text editor window titled "test.html". The editor contains an HTML document with a list of links. A search and replace operation is in progress. The search pattern is `(?<=href=['"])(.*?)(?=['"])/gim` and the replacement text is `#`. The search has found 3 matches in the href attributes of the links.

```
1 <ul class='cate_list'>
2   <li class="on">
3     <a href='http://cafe.naver.com' class="_click(SectionTheme|1)
4       <i class='ico ico_cate1'></i>
5       <span class='blind'>게임</span>
6     </a>
7   </li>
8   <li class="off">
9     <a href='http://cafe.naver.com/m/s2231' class="test _click(
10      SectionThemeDirCafeList) test _stopDefault" ONCLICK="nclk(this,
11      'http.id2', '', '');">
12      <i class='ico ico_cate2'></i>
13      <span class='blind'>만화/애니메이션</span>
14    </a>
15  </li>
16  <li class="off">
17    <a href="http://m.naver.com" class="_click(SectionThemeDirCafeList)
18      test _stopDefault test" onClick="nclk(this, 'http.id3', '', '');">
19      <i class='ico ico_cate3'></i>
20      <span class='blind'>방송/연예</span>
21    </a>
22  </li>
23 </ul>
```

Find What: `(?<=href=['"])(.*?)(?=['"])/gim`

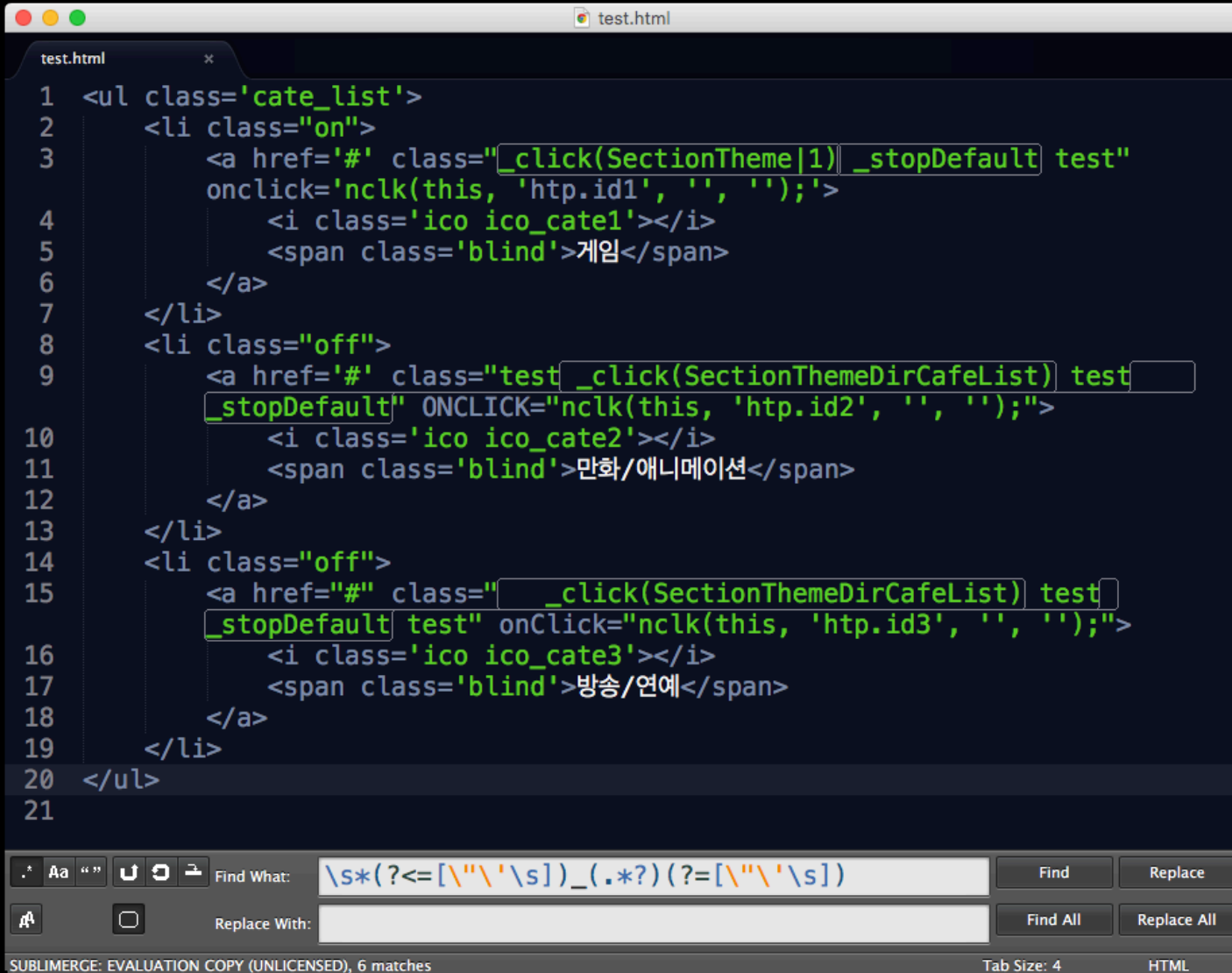
Replace With: `#`

Find Replace Find All Replace All

SUBLIMERGE: EVALUATION COPY (UNLICENSED), 3 matches

Tab Size: 4 HTML

클래스 제거 : `/\s*(?<=[\"'\s])(.*?)(?=[\"'\s])/gim`



The screenshot shows a Sublime Text editor window titled "test.html". The editor contains HTML code for a list of items. A search and replace dialog is open at the bottom. The search pattern is `/\s*(?<=[\"'\s])_(.*?)(?=[\"'\s])/gim`, which is designed to remove underscores and the text between them from class names in the HTML. The replace field is empty. The status bar at the bottom indicates "6 matches".

```
1 <ul class='cate_list'>
2   <li class="on">
3     <a href='#' class="_click(SectionTheme|1)_stopDefault test"
4       onclick='nclk(this, 'http.id1', '', '');'>
5       <i class='ico ico_cate1'></i>
6       <span class='blind'>게임</span>
7     </a>
8   </li>
9   <li class="off">
10    <a href='#' class="test_click(SectionThemeDirCafeList) test
11      _stopDefault" ONCLICK="nclk(this, 'http.id2', '', '');">
12      <i class='ico ico_cate2'></i>
13      <span class='blind'>만화/애니메이션</span>
14    </a>
15  </li>
16  <li class="off">
17    <a href="#" class="_click(SectionThemeDirCafeList) test
18      _stopDefault test" onClick="nclk(this, 'http.id3', '', '');">
19      <i class='ico ico_cate3'></i>
20      <span class='blind'>방송/연예</span>
21    </a>
22  </li>
23 </ul>
```

Find What: `/\s*(?<=[\"'\s])_(.*?)(?=[\"'\s])/gim` Find Replace

Replace With: Find All Replace All

SUBLIMERGE: EVALUATION COPY (UNLICENSED), 6 matches Tab Size: 4 HTML

_클래스 제거 후 스페이스 제거 : `/(?<=class=[\\"'\"])\s+/gim`

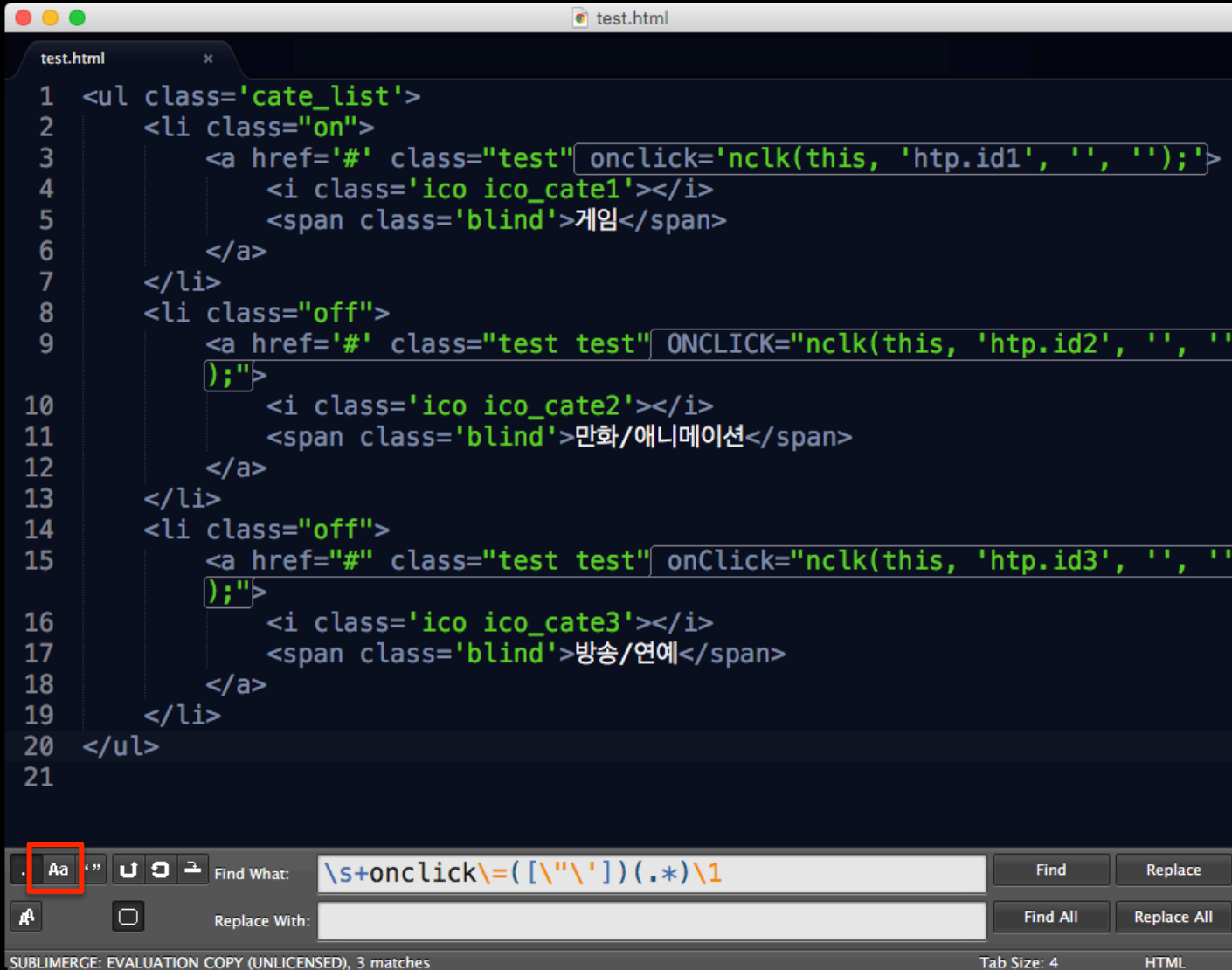
```
1 <ul class='cate_list'>
2   <li class="on">
3     <a href='#' class=" test" onclick='nclk(this, 'http.id1', '', '');>
4       <i class='ico ico_cate1'></i>
5       <span class='blind'>게임</span>
6     </a>
7   </li>
8   <li class="off">
9     <a href='#' class="test test" ONCLICK="nclk(this, 'http.id2', '', ''
10      );">
11       <i class='ico ico_cate2'></i>
12       <span class='blind'>만화/애니메이션</span>
13     </a>
14   </li>
15   <li class="off">
16     <a href="#" class=" test test" onClick="nclk(this, 'http.id3', '',
17      ');">
18       <i class='ico ico_cate3'></i>
19       <span class='blind'>방송/연예</span>
20     </a>
21   </li>
22 </ul>
```

Find What: `(?<=class=[\\"'\"])\s+` Find Replace

Replace With: Find All Replace All

SUBLIMERGE: EVALUATION COPY (UNLICENSED), 2 matches Tab Size: 4 HTML

onclick 이벤트 속성/값 제거 : `/\s+onclick\=([\\"'])(.*)\1/gim`



The screenshot shows a Sublime Text editor window titled 'test.html'. The editor contains an HTML document with a list of categories. The first category is 'on' (game), the second is 'off' (cartoon/animation), and the third is 'off' (broadcast/entertainment). Each category has an icon and a name. The 'on' category has an icon class 'ico_ico_cate1' and a name '게임'. The 'off' categories have icon classes 'ico_ico_cate2' and 'ico_ico_cate3' and names '만화/애니메이션' and '방송/연예' respectively. The 'on' category's link has an 'onclick' attribute, while the 'off' categories have an 'ONCLICK' attribute. A search bar at the bottom of the editor shows the regex pattern `/\s+onclick\=([\\"'])(.*)\1/gim` in the 'Find What' field. The 'Replace With' field is empty. The search results show 3 matches. The status bar at the bottom indicates 'SUBLIMERGE: EVALUATION COPY (UNLICENSED), 3 matches' and 'Tab Size: 4 HTML'.

```
1 <ul class='cate_list'>
2   <li class="on">
3     <a href='#' class="test" onclick='nclk(this, 'http.id1', '', '');>
4       <i class='ico_ico_cate1'></i>
5       <span class='blind'>게임</span>
6     </a>
7   </li>
8   <li class="off">
9     <a href='#' class="test test" ONCLICK="nclk(this, 'http.id2', '', ''
10      );">
11       <i class='ico_ico_cate2'></i>
12       <span class='blind'>만화/애니메이션</span>
13     </a>
14   </li>
15   <li class="off">
16     <a href="#" class="test test" onClicK="nclk(this, 'http.id3', '', ''
17      );">
18       <i class='ico_ico_cate3'></i>
19       <span class='blind'>방송/연예</span>
20     </a>
21   </li>
22 </ul>
```

Find What: `/\s+onclick\=([\\"'])(.*)\1/gim`

Replace With:

Find All Replace All

SUBLIMERGE: EVALUATION COPY (UNLICENSED), 3 matches Tab Size: 4 HTML

최종 코드

```
test.html x
1 <ul class='cate_list'>
2   <li class="on">
3     <a href='#' class="test">
4       <i class='ico ico_cate1'></i>
5       <span class='blind'>게임</span>
6     </a>
7   </li>
8   <li class="off">
9     <a href='#' class="test test">
10      <i class='ico ico_cate2'></i>
11      <span class='blind'>만화/애니메이션</span>
12    </a>
13  </li>
14  <li class="off">
15    <a href="#" class="test test">
16      <i class='ico ico_cate3'></i>
17      <span class='blind'>방송/연예</span>
18    </a>
19  </li>
20 </ul>
21
```

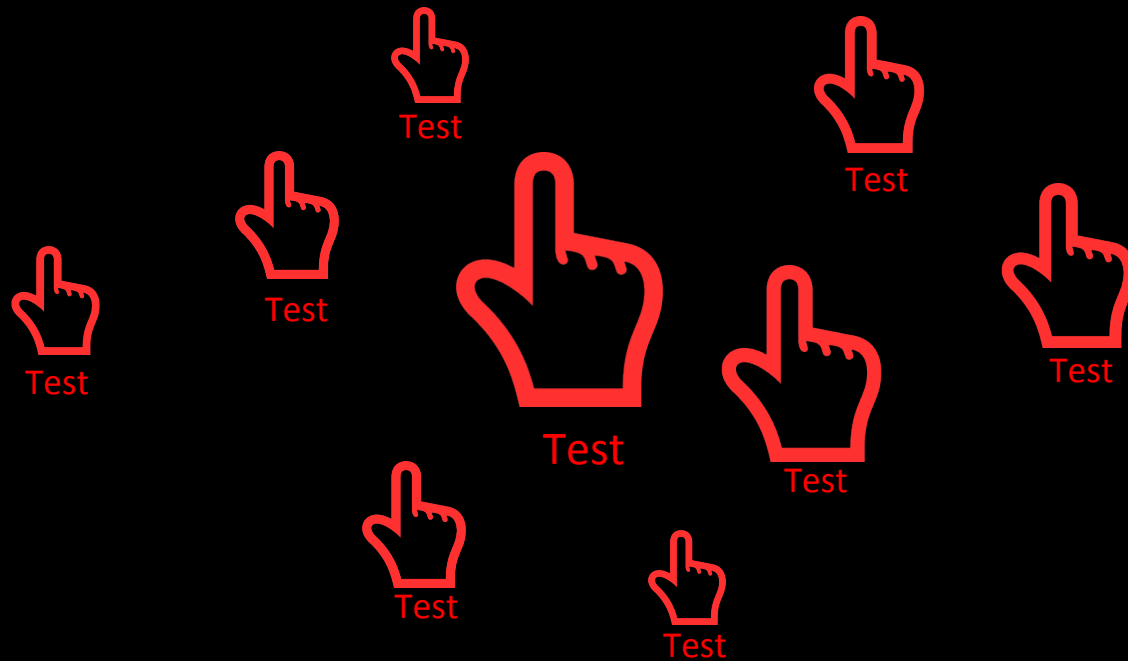
Sublime Merge: EVALUATION COPY (UNLICENSED), Line 21, Column 1

Tab Size: 4

HTML

Perfect

전문가가 되려면 어떻게 해야 하나?



일치하는 문자열을 찾는 거 보다 일치하지 않는 문자열을 찾는게 더 어렵다.

참고 도구들

- <http://www.regexper.com/>
표현식을 시각화해주는 도구
- <http://www.regexr.com/>
정규 표현식에 대한 도움말과 각종 사례들을 보여줌
- <http://zvon.org/comp/r/tut-Regexp.html#Pages~Contents>
정규 표현식 tutorials를 제공함

/Thank\sYou/gim