

Chap 8. HTML 기본 문법 배우기

오늘은 웹 크롤러를 만들기 위해서 가장 중요하다고 강조할 수 있는 내용들을 많이 배울 예정입니다. 웹 크롤러란 웹 페이지에 있는 데이터를 가져오는 것이기 때문에 웹 페이지가 어떻게 만들어져 있는지를 잘 알아야 합니다. 그래서 먼저 웹 페이지를 만드는 가장 기본적인 HTML 에 대해서 중요한 부분을 골라서 배웁니다.

웹 크롤러를 만들기 위해서는 이번 장에서 배우는 내용을 반드시 알아야 하니까 두 눈 크게 뜨고 열심히 공부해 주세요~

1) 다양한 HTML 태그의 종류

(1) <html> 태그

이 태그는 HTML 문서의 시작을 알리는 태그입니다. HTML 문서를 시작할 때 반드시 써야 하며 HTML 문서의 마지막 부분에는 </html> 로 마무리하면 됩니다.

그리고 <html lang="ko"> 와 같이 lang 속성을 사용해서 사용될 언어를 지정할 수 있습니다. 주로 많이 지정하는 언어는 아래 표와 같습니다.

코드약어	ko	en	fr	ja	de	zh
언어	한국어	영어	프랑스어	일본어	독일어	중국어

참고로 lang="ko" 와 같이 언어를 지정하는 이유는 아래와 같습니다.

- 검색을 할 때 "한국어 문서"로 검색 범위를 지정할 경우 이 옵션을 보고 판단을 하게 됩니다
- 시각장애인 들을 위해 점자나 음성 인식으로 웹 페이지를 서비스할 때 이 옵션을 보고 해당 언어로 서비스를 합니다.

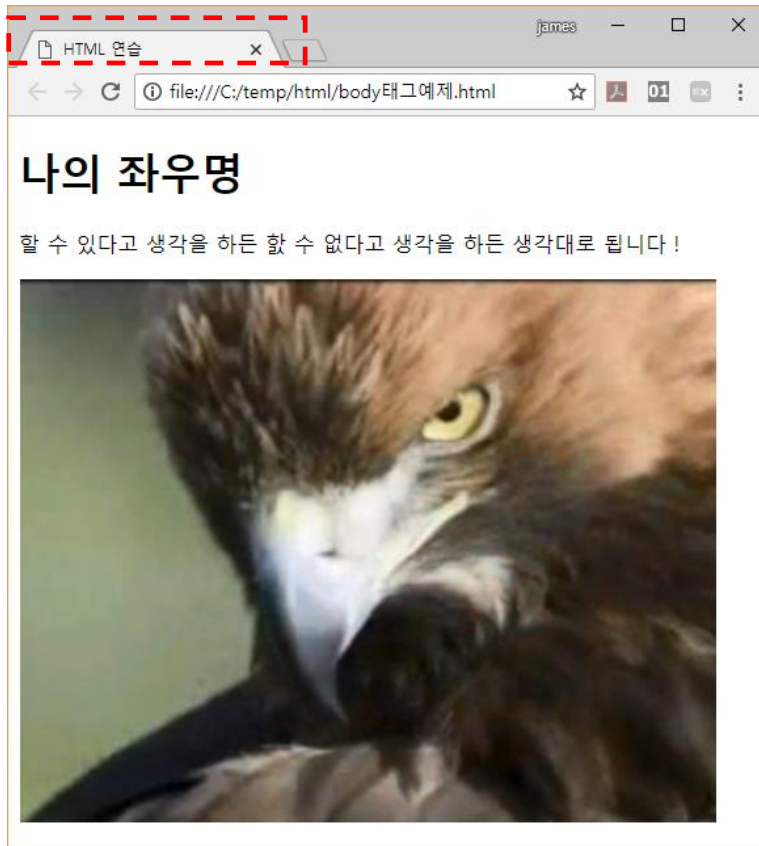
(2) <head> 태그

이 부분에 저장되는 대부분의 정보는 사람을 위해서가 아니라 웹 브라우저를 위해서 사용됩니다. 이 태그안에 사용되는 주요 하위 태그들은 아래와 같습니다.

a) <title> 태그

문법: <title> 제목 </title>

이 태그는 웹 브라우저의 제목 표시줄에 저장될 페이지의 제목을 지정하는 태그입니다. 아래 그림에서 빨간색 네모칸 부분이 <title> 태그입니다.



b) <meta> 태그

문법: <meta charset="utf-8">

이 태그에는 인코딩 방식이나 웹 문서 키워드 등을 웹 브라우저에게 알려 줍니다.
가장 많이 사용하는 기능의 위 문법의 예제처럼 인코딩 방식을 지정하는 것입니다

<head> 태그에서 많이 사용되는 2가지 태그를 살펴 보았습니다.

이 태그들 이외에도 스타일 시트용 태그인 <style> 태그와 <link> 태그 등을 지정할 수 있습니다.

3) <body> 태그

이 부분에는 사람들에게 보여줄 실제 내용들을 지정하는 부분입니다.

아래의 그림에서 빨간 네모 부분이 실제 웹 페이지에 보여주는 <body> 태그 부분입니다.

```

1 <!doctype html>
2 <html lang="ko">
3 <head>
4 <meta charset="utf-8">
5 <title> HTML 연습 </title>
6 </head>
7 <body>
8 <h1> 나의 좌우명 </h1>
9 <p> 할 수 있다고 생각을 하든
10 할 수 없다고 생각을 하든
11 생각대로 됩니다 ! </p>
12 
13 </body>
14 </html>

```

위 그림에서 빨간 네모 박스 안에 보면 <h1> 태그와 <p> 태그, 태그 등이 보이죠?

<body> 태그안에는 실제 사용자들에게 보여줄 정보들이 저장되기 때문에 아주 다양한 태그들이 사용됩니다. 지금부터 자세하게 살펴보겠습니다.

(1) 텍스트와 관련된 주요 태그들

a) <h1> 태그

이 태그는 주로 제목을 표시할 때 많이 사용합니다.

이 태그를 사용하면 다른 글씨들보다 크고 진하게 텍스트가 표시됩니다.

<h1> 에서 n 자리에는 1 ~ 6 까지 숫자가 들어가는데 숫자가 작을수록 글씨는 크게 표시됩니다.

실습을 위해서 아래와 같이 소스코드를 수정했습니다.

아래 코드에서 9번 줄을 추가했습니다.

아래와 같이 수정하고 저장한 후 웹 브라우저에서 열어 볼까요?

```

1  <!doctype html>
2  <html lang="ko">
3  <head>
4      <meta charset="urf-8">
5      <title> HTML 연습 </title>
6  </head>
7  <body>
8      <h1> 나의 좌우명 - h1 크기 </h1>
9      <h5> 꼭 이루자!! - h5 크기 </h5>
10     <p> 할 수 있다고 생각을 하든
11         할 수 없다고 생각을 하든
12         생각대로 됩니다! </p>
13 </body>
14 </html>

```

length : 339 lines : 14 Ln : 14 Col : 9 Sel : 0 | 0 Windows (CR LF) UTF-8 INS



b) <p> 태그 와
태그

이 태그는 'paragraph' 의 약자로 한국어로 번역하면 문단을 지정하는 태그입니다.

문단이라는 것은 문장 1개 또는 여러 개로 구성된 것을 의미합니다.

이 태그로 텍스트를 출력할 경우 웹 브라우저에서 한 줄에 표시할 수 없이 긴 문장은 자동 줄바꿈으로 표시해 줍니다. 그리고
 태그는 강제로 줄바꿈을 시켜주는 태그입니다.

아래의 예를 보세요.

```

1 <!doctype html>
2 <html lang="ko">
3 <head>
4     <meta charset="utf-8">
5     <title> HTML 연습 </title>
6 </head>
7 <body>
8     <h1> 나의 좌우명 </h1>
9     <p> 할 수 있다고 생각을 하든
10     할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11     <p> 얼마나 </p>
12     <p> 멋진 삶을 사는가는 <br>
13     스스로 결정하고 만들어가는 것입니다 </p>
14 </body>
15 </html>
  
```

length : 405 lines : 15 Ln : 15 Col : 9 Sel : 0 | 0 Windows (CR LF) UTF-8 INS



c) 태그와 태그 - 글씨를 굵게 표시하기

이 두 가지 태그는 글씨체를 강조해서 표시할 경우 많이 사용하는 태그입니다.

웹 페이지로 보기에 둘 다 동일한 기능이지만 시각 장애인을 위한 음성 번역기에서는 두가지 태그가 다른 역할을 합니다. 음성 번역기에서는 태그가 있을 경우 특정 글자가 강조되었다고 안내를 해 줍니다.

```

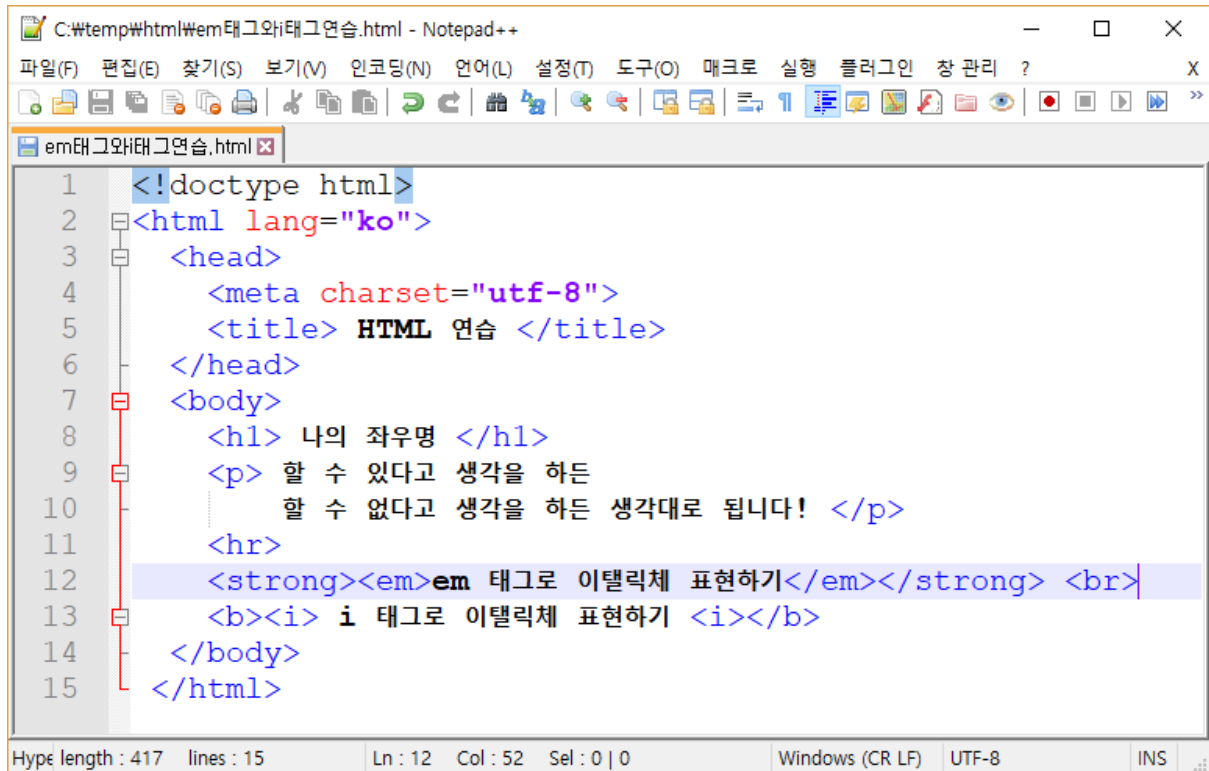
1 <!doctype html>
2 <html lang="ko">
3 <head>
4 <meta charset="utf-8">
5 <title> HTML 연습 </title>
6 </head>
7 <body>
8 <h1> 나의 좌우명 </h1>
9 <p> 할 수 있다고 생각을 하든
10     할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11 <hr>
12 <strong>strong 태그로 강조하기</strong> <br>
13 <b> b 태그로 강조하기 </b>
14 </body>
15 </html>

```



d) 태그와 <i> 태그 - 이탤릭체로 표시하기

이 두 가지 태그는 문장에서 특정 단어나 특정 부분을 강조할 때 이탤릭 체로 표현할 때 많이 사용하는 태그입니다. 아래 예제를 보세요.



```

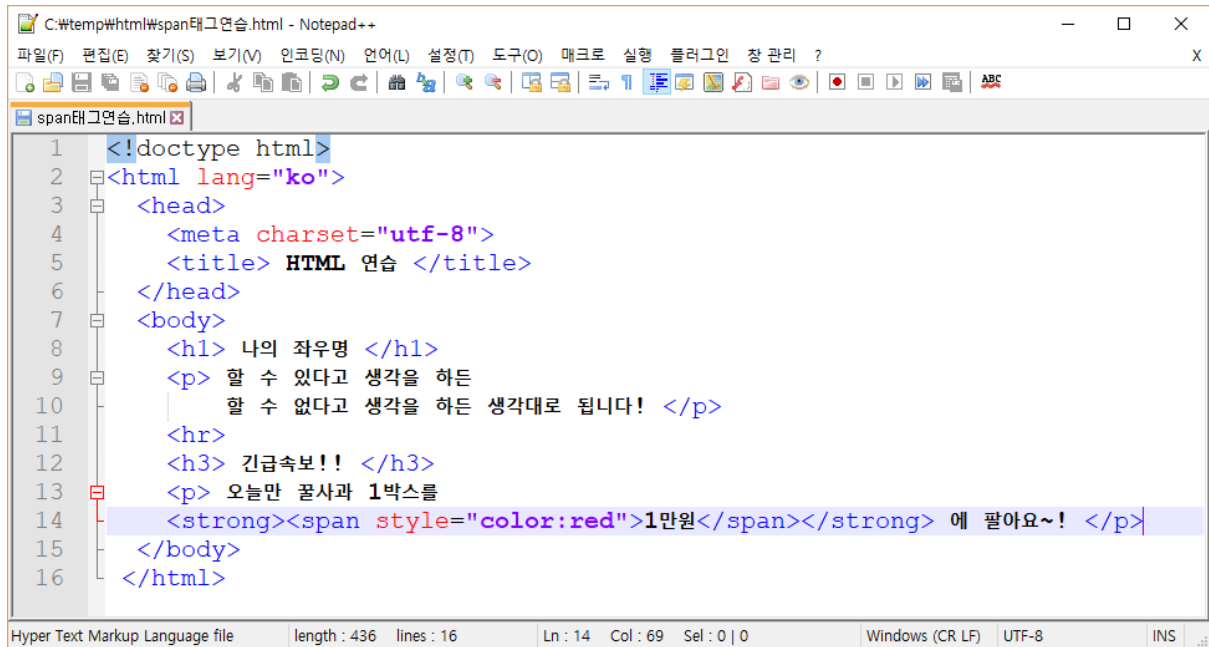
1 <!doctype html>
2 <html lang="ko">
3 <head>
4 <meta charset="utf-8">
5 <title> HTML 연습 </title>
6 </head>
7 <body>
8 <h1> 나의 좌우명 </h1>
9 <p> 할 수 있다고 생각을 하든
10   할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11 <hr>
12 <strong><em>em 태그로 이탤릭체 표현하기</em></strong> <br>
13 <b><i>i 태그로 이탤릭체 표현하기 <i></b>
14 </body>
15 </html>

```



e) 태그 – 특정 영역을 하나로 묶을 때 사용

이 태그는 특정 부분을 하나의 블록으로 묶어서 강조하거나 스타일 시트를 적용할 때 많이 사용하는 태그입니다. 아래 예제를 보세요.

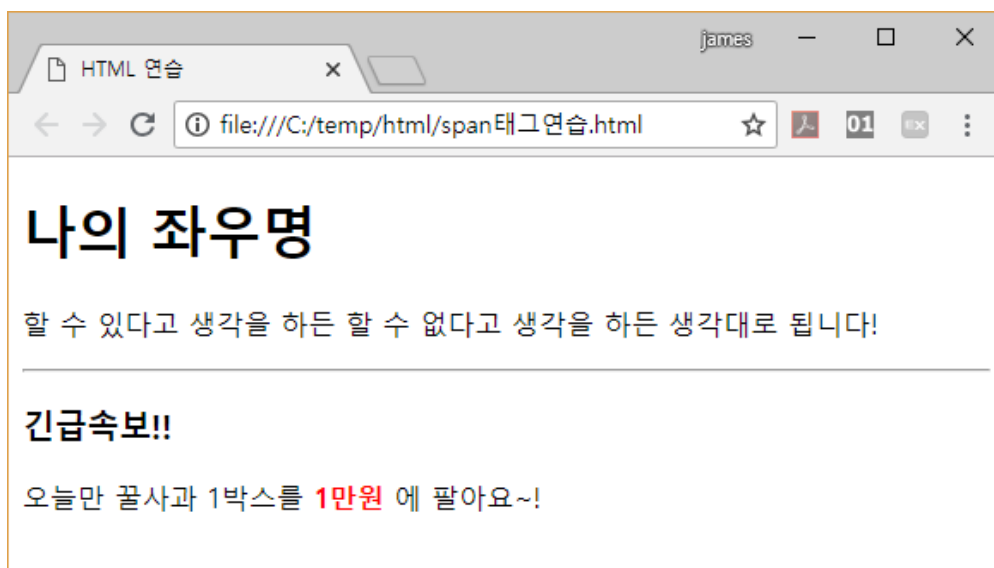


```

1 <!doctype html>
2 <html lang="ko">
3 <head>
4 <meta charset="utf-8">
5 <title> HTML 연습 </title>
6 </head>
7 <body>
8 <h1> 나의 좌우명 </h1>
9 <p> 할 수 있다고 생각을 하든
10     할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11 <hr>
12 <h3> 긴급속보!! </h3>
13 <p> 오늘만 꿀사과 1박스를
14 <strong><span style="color:red">1만원</span></strong> 에 팔아요~! </p>
15 </body>
16 </html>

```

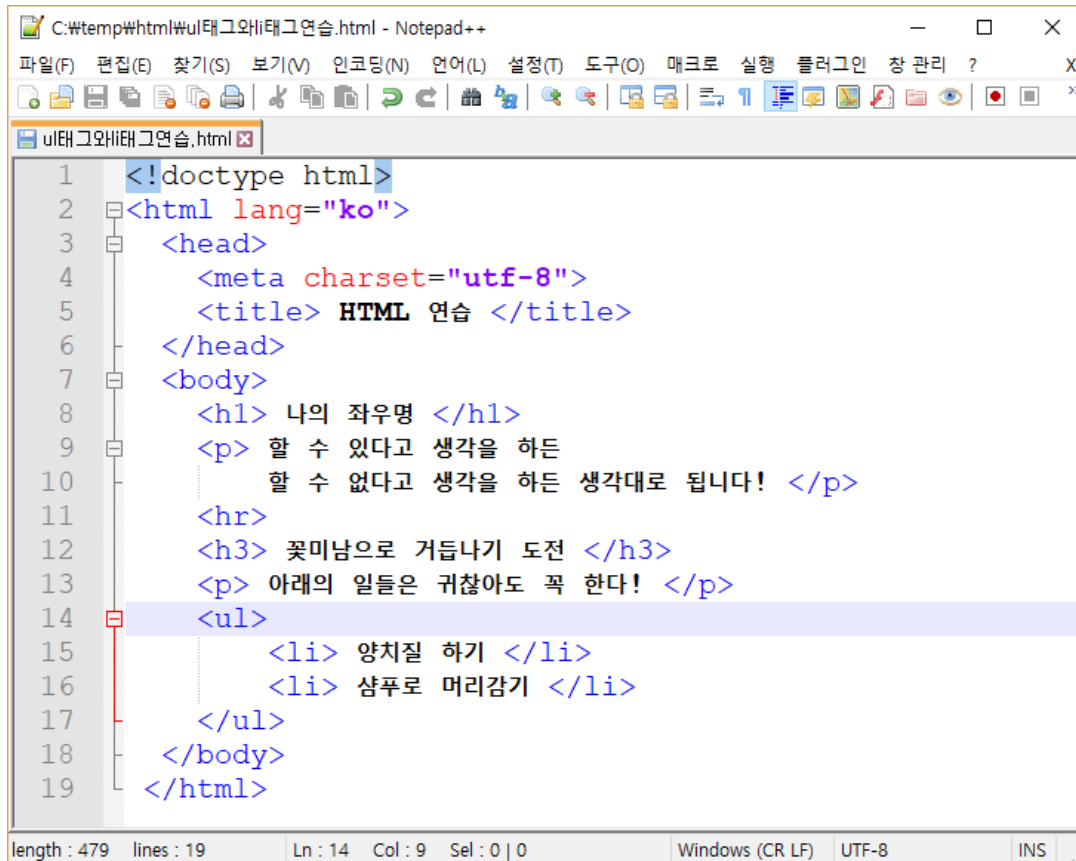
Hyper Text Markup Language file length : 436 lines : 16 Ln : 14 Col : 69 Sel : 0 | 0 Windows (CR LF) UTF-8 INS



(2) 다양한 목록을 만드는 태그들

a) 태그와 태그

 태그는 unordered list 의 약자로 이 태그들은 순서 없는 목록을 만들 때 주로 사용됩니다. 아래의 예제를 보세요.



```

1  <!doctype html>
2  <html lang="ko">
3  <head>
4      <meta charset="utf-8">
5      <title> HTML 연습 </title>
6  </head>
7  <body>
8      <h1> 나의 좌우명 </h1>
9      <p> 할 수 있다고 생각을 하든
10         할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11      <hr>
12      <h3> 꽃미남으로 거듭나기 도전 </h3>
13      <p> 아래의 일들은 귀찮아도 꼭 한다! </p>
14      <ul>
15          <li> 양치질 하기 </li>
16          <li> 샴푸로 머리감기 </li>
17      </ul>
18  </body>
19  </html>
  
```

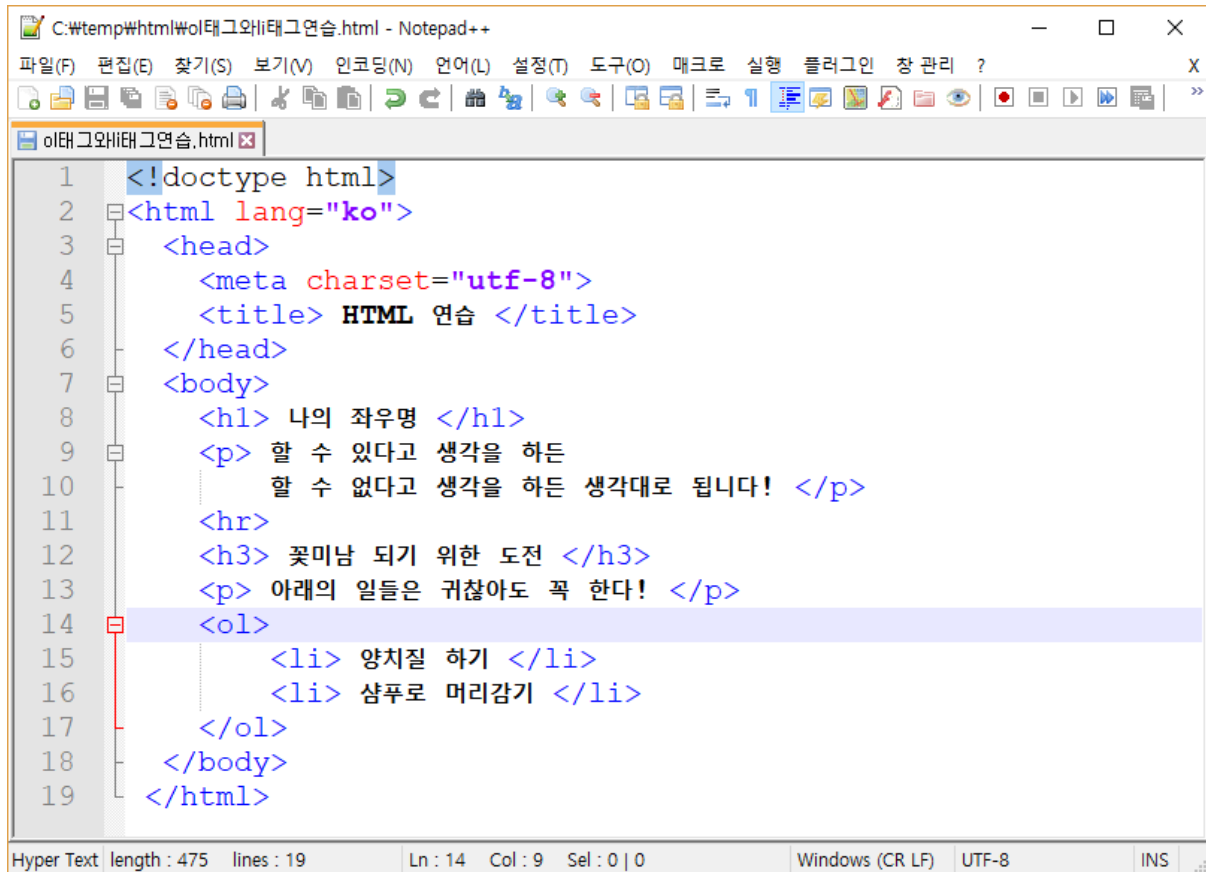
length : 479 lines : 19 Ln : 14 Col : 9 Sel : 0 | 0 Windows (CR LF) UTF-8 INS



b) 태그와 태그

 태그는 ordered list 의 약자로 순서가 필요한 목록을 만드는 태그입니다.

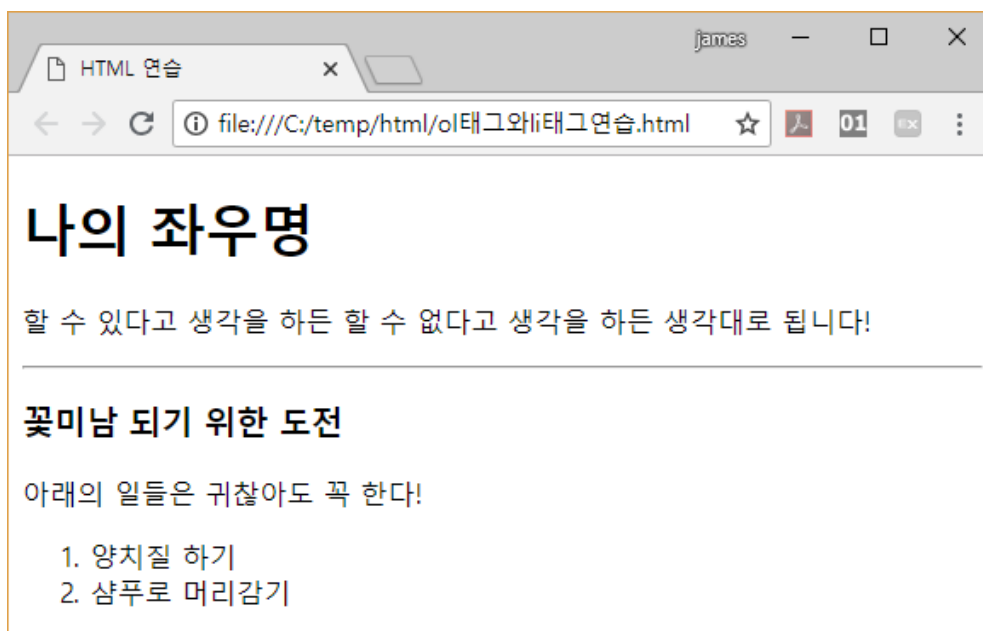
아래 예제를 보세요.



```

1  <!doctype html>
2  <html lang="ko">
3  <head>
4      <meta charset="utf-8">
5      <title> HTML 연습 </title>
6  </head>
7  <body>
8      <h1> 나의 좌우명 </h1>
9      <p> 할 수 있다고 생각을 하든
10         할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11      <hr>
12      <h3> 꽃미남 되기 위한 도전 </h3>
13      <p> 아래의 일들은 귀찮아도 꼭 한다! </p>
14      <ol>
15          <li> 양치질 하기 </li>
16          <li> 샴푸로 머리감기 </li>
17      </ol>
18  </body>
19 </html>

```



앞에서 살펴본대로 태그에서 기본값은 숫자로 번호가 매겨집니다.
 그런데 옵션을 변경할 경우 다양한 형태로 순서를 지정할 수 있어요.
 옵션은 아래 표와 같습니다.

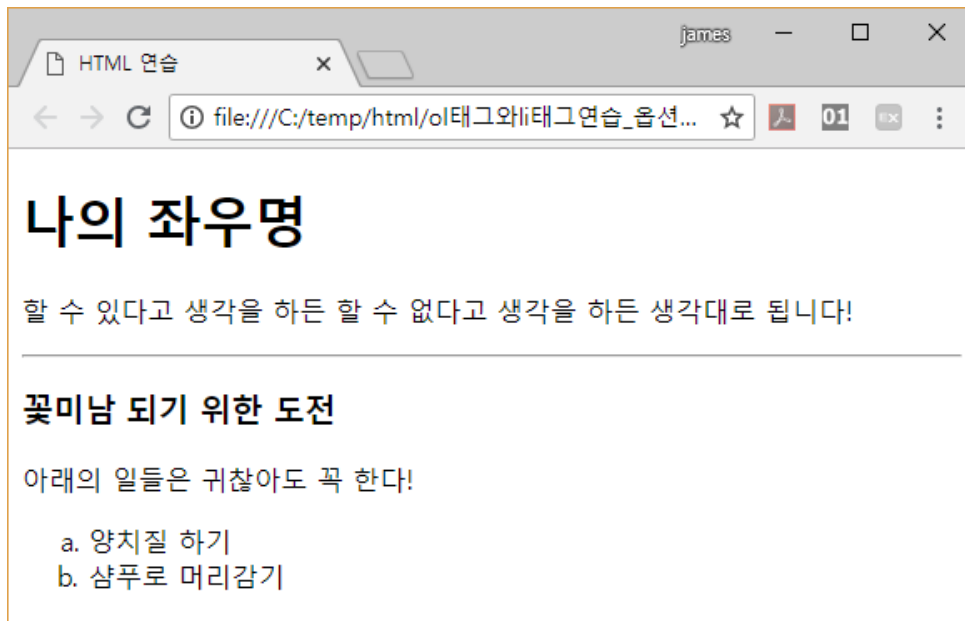
속성 옵션	표시값
1	숫자(옵션지정 없을 경우 기본적용)
a	영문 소문자
A	영문 대문자
i	로마숫자 소문자
I	로마숫자 대문자
start	시작되는 번호를 지정할 수 있음
reversed	번호를 역순으로 출력함

위 표에 있는 옵션들을 사용하여 목록을 만들어 보겠습니다.
 아래 예제들을 보세요.

```

1 <!doctype html>
2 <html lang="ko">
3   <head>
4     <meta charset="utf-8">
5     <title> HTML 연습 </title>
6   </head>
7   <body>
8     <h1> 나의 좌우명 </h1>
9     <p> 할 수 있다고 생각을 하든
10      할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11     <hr>
12     <h3> 꽃미남 되기 위한 도전 </h3>
13     <p> 아래의 일들은 귀찮아도 꼭 한다! </p>
14     <ol type="a">
15       <li> 양치질 하기 </li>
16       <li> 샴푸로 머리감기 </li>
17     </ol>
18   </body>
19 </html>
  
```

위 그림에서 14번 행에 type="a" 라는 부분 보이죠?
 저렇게 옵션을 지정할 경우 아래 그림처럼 목록 순서를 표시하는 부분이 변경됩니다.



나머지 옵션들도 방법이 동일하니 꼭 확인해 보세요~

c) 중첩된 목록 만들기

이번에는 앞에서 배웠던 태그들을 활용해서 실제 웹사이트에서 많이 보는 중첩 목록을 만들겠습니다.

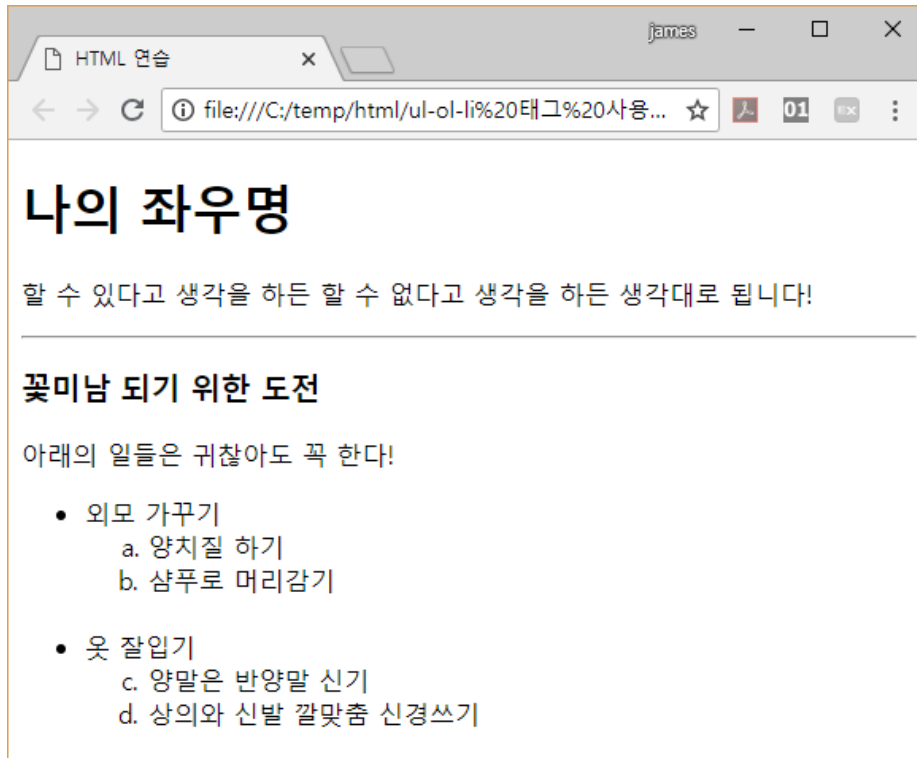
```

1 <!doctype html>
2 <html lang="ko">
3   <head>
4     <meta charset="utf-8">
5     <title> HTML 연습 </title>
6   </head>
7   <body>
8     <h1> 나의 좌우명 </h1>
9     <p> 할 수 있다고 생각을 하든
10    | 할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11    <hr>
12    <h3> 꽃미남 되기 위한 도전 </h3>
13    <p> 아래의 일들은 귀찮아도 꼭 한다! </p>
14    <ul>
15      <li> 외모 가꾸기
16        <ol type="a">
17          <li> 양치질 하기
18          <li> 샴푸로 머리감기
19        </ol>
20      </li> <br>
21      <li> 옷 잘입기
22        <ol type="a" start="3">
23          <li> 양말은 반양말 신기
24          <li> 상의와 신발 깔맞춤 신경쓰기
25        </ol>
26      </li>
27    </ul>
28  </body>
29 </html>

```

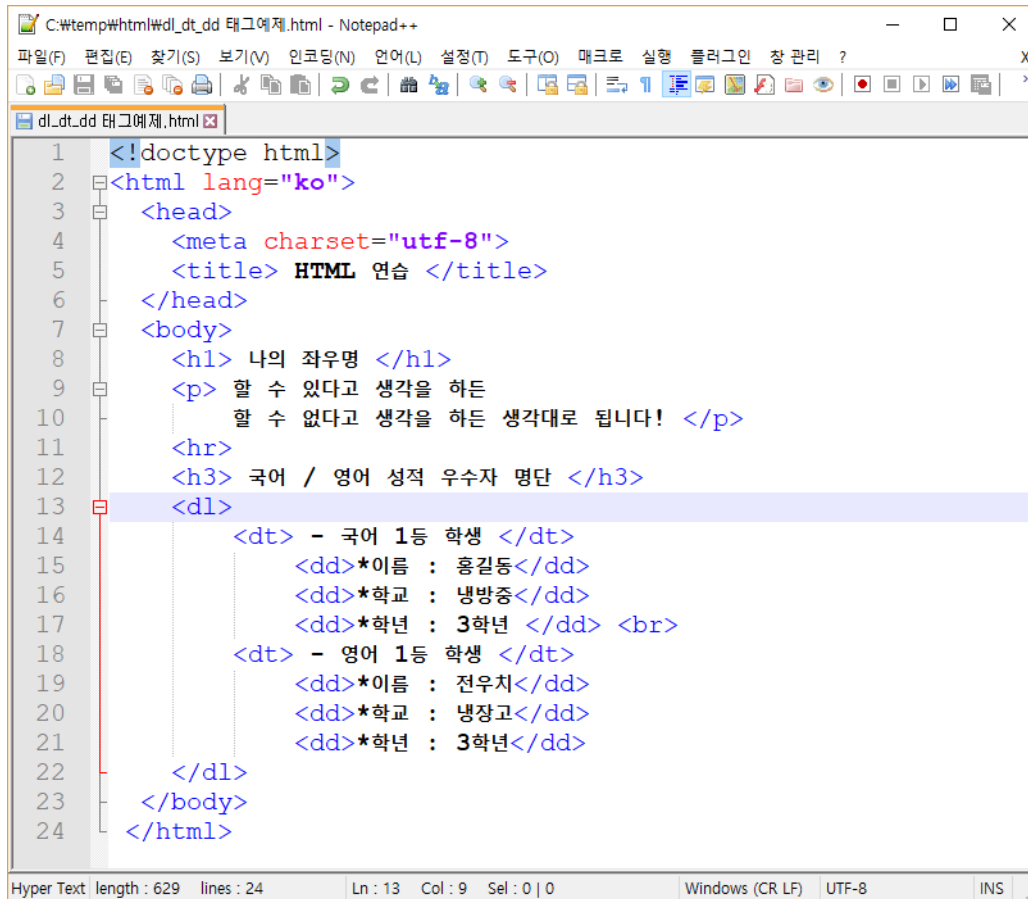
위 그림에서 22번 줄에 start="3" 옵션은 옷 잘입기 옵션의 시작되는 번호를 알파벳(type="a") 에서 3번째 글자인 "c" 로 지정한다는 의미 입니다. 아래 그림을 보면 어떤 의미인지 이해하실 거예요.

 태그와 태그, 태그는 목록을 만들 때 아주 많이 사용하는 태그이므로 꼭 기억해 주세요~~



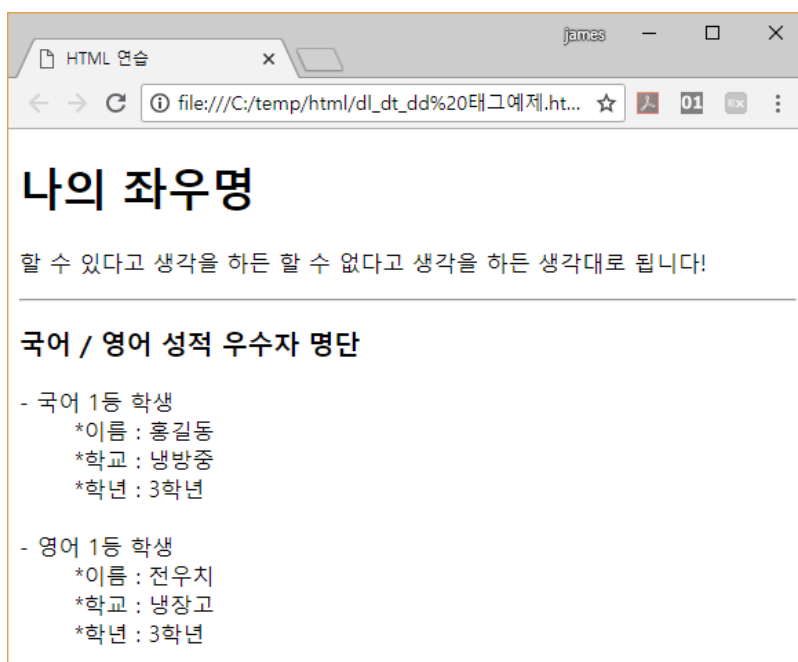
d) <dl> 태그와 <dt> 태그와 <dd> 태그로 목록 만들기

<dl> 태그는 description list 의 약자로 특정 항목과 설명을 한 세트로 묶어서 표시하는 목록을 만들 때 사용합니다. 아래의 예를 보세요.



```

1 <!doctype html>
2 <html lang="ko">
3 <head>
4 <meta charset="utf-8">
5 <title> HTML 연습 </title>
6 </head>
7 <body>
8 <h1> 나의 좌우명 </h1>
9 <p> 할 수 있다고 생각을 하든
10 할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11 <hr>
12 <h3> 국어 / 영어 성적 우수자 명단 </h3>
13 <dl>
14 <dt> - 국어 1등 학생 </dt>
15 <dd>*이름 : 홍길동</dd>
16 <dd>*학교 : 냉방중</dd>
17 <dd>*학년 : 3학년 </dd> <br>
18 <dt> - 영어 1등 학생 </dt>
19 <dd>*이름 : 전우치</dd>
20 <dd>*학교 : 냉장고</dd>
21 <dd>*학년 : 3학년</dd>
22 </dl>
23 </body>
24 </html>
  
```



(3) 표를 만드는 태그들

웹 페이지 중에서 표를 포함하고 있는 페이지들이 아주 많습니다. 나중에 크롤링 할 때 표 자체에 있는 내용을 크롤링 해야 할 경우가 아주 많기 때문에 표를 구성하는 여러가지 태그들을 잘 알아야 합니다.

a) <table>, <tr>, <td> 태그를 활용하여 기본적인 표 만들기

표를 만들 때 가장 먼저 사용하는 태그는 <table> 태그입니다.

기본 문법은 아래와 같습니다.

```

1  <!doctype html>
2  <html lang="ko">
3  <head>
4      <meta charset="utf-8">
5      <title> HTML 연습 </title>
6  </head>
7  <body>
8      <h1> 나의 좌우명 </h1>
9      <p> 할 수 있다고 생각을 하든
10     할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11     <hr>
12     <h3> 표 만들기 연습 </h3>
13     <table>
14     <tr>
15         <td> 내용 1 </td>
16         <td> 내용 2 </td>
17         <td> 내용 3 </td>
18     </tr>
19     <tr>
20         <td> 내용 4 </td>
21         <td> 내용 5 </td>
22         <td> 내용 6 </td>
23     </tr>
24     </table>
25 </body>
26 </html>

```




그런데 위 그림을 보면 표 부분에 테두리 선이 없어서 보기가 불편하죠? 이럴 땐 table 태그의 옵션 중 border 옵션을 사용하면 됩니다. 아래 그림을 보세요.

```

1 <!doctype html>
2 <html lang="ko">
3   <head>
4     <meta charset="utf-8">
5     <title> HTML 연습 </title>
6   </head>
7   <body>
8     <h1> 나의 좌우명 </h1>
9     <p> 할 수 있다고 생각을 하든
10      할 수 없다고 생각을 하든 생각대로 됩니다! </p>
11     <hr>
12     <h3> 표 만들기 연습 </h3>
13     <table border="1">
14       <tr>
15         <td> 내용 1 </td>
16         <td> 내용 2 </td>
17         <td> 내용 3 </td>
18       </tr>
19       <tr>
20         <td> 내용 4 </td>
21         <td> 내용 5 </td>
22         <td> 내용 6 </td>
23       </tr>
24     </table>
25   </body>
26 </html>

```

Hyper Text length : 616 lines : 26 Ln : 13 Col : 23 Sel : 0 | 0 Windows (CR LF) UTF-8 INS



b) <thead> , <tbody>, <tfoot> 태그로 표 만들기

앞에서 표를 만드는 기본적인 태그를 보았습니다.

그런데 표를 만들 때 컬럼 이름 부분과 내용부분과 마지막 행 부분을 따로 만들어서 관리하는 경우도 종종 있습니다. 특히 마지막 행에 요약된 집계 내용이 들어갈 경우나 내용 부분이 아주 많은 경우에는 컬럼 이름부분과 내용 , 요약 부분을 별도로 만들어서 관리하는 경우가 많죠.

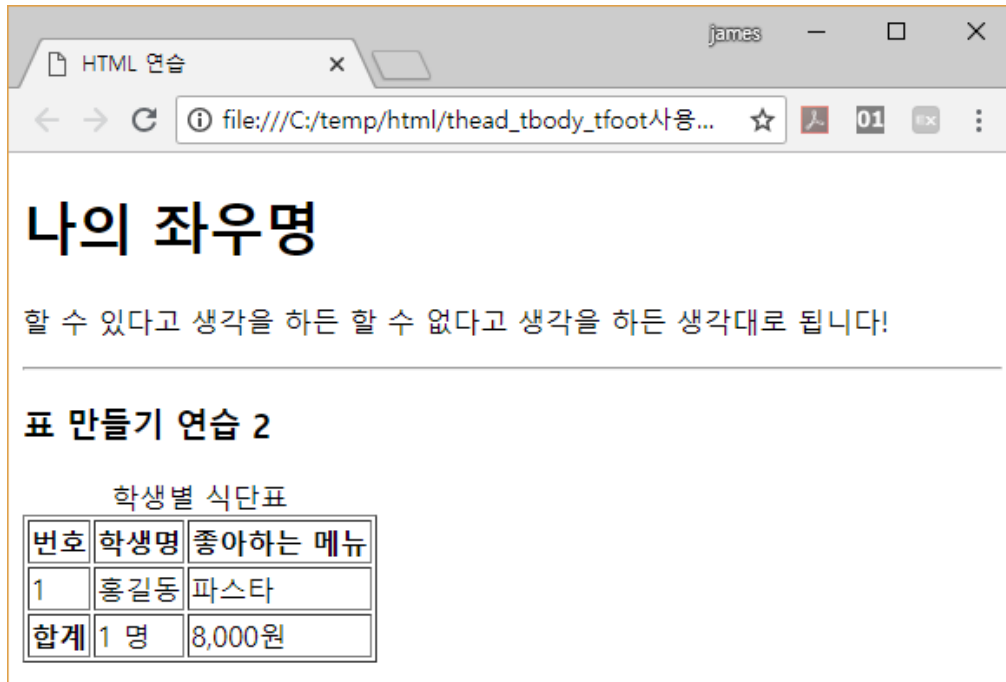
이 때 사용하는 태그가 컬럼 이름을 관리하는 <thead>, 내용 부분을 생성하는 <tbody>, 마지막의 요약 부분을 관리하는 <tfoot> 태그 입니다.

아래의 예를 보면 금방 이해할 거예요.

```

1  <!doctype html>
2  <html lang="ko">
3  <head>
4      <meta charset="utf-8">
5      <title> HTML 연습 </title>
6  </head>
7  <body>
8      <h1> 나의 좌우명 </h1>
9      <p> 할 수 있다고 생각을 하든 할 수 없다고 생각을 하든 생각대로 됩니다! </p>
10     <hr>
11     <h3> 표 만들기 연습 2 </h3>
12     <table border="1">
13         <caption> 학생별 식단표 </caption>
14         <thead>
15             <tr>
16                 <th> 번호 </th>
17                 <th> 학생명 </th>
18                 <th> 좋아하는 메뉴 </th>
19             </tr>
20         </thead>
21         <tbody>
22             <tr>
23                 <td> 1 </td>
24                 <td> 홍길동 </td>
25                 <td> 파스타 </td>
26             </tr>
27         </tbody>
28         <tfoot>
29             <tr>
30                 <th> 합계 </th>
31                 <td> 1 명</td>
32                 <td> 8,000원 </td>
33             </tr>
34         </tfoot>
35     </table>
36 </body>
37 </html>

```



표를 만들 때 자주 사용하는 태그이니 꼭 기억해주세요~

(4) 본문에 그림 넣기

이번에는 웹 페이지에 이미지를 한번 띄워 보도록 할까요???

이미지 태그는 아래와 같이 사용합니다.

```

```

```

1 <!doctype html>
2 <html lang="ko">
3   <head>
4     <meta charset="utf-8">
5     <title> HTML 연습 </title>
6   </head>
7   <body>
8     <h1> 나의 좌우명 </h1>
9     <p> 할 수 있다고 생각을 하든 할 수 없다고 생각을 하든 생각대로 됩니다! </p>
10    <hr>
11    <p> 멋진 솔개 </p>
12    
13  </body>
14 </html>

```

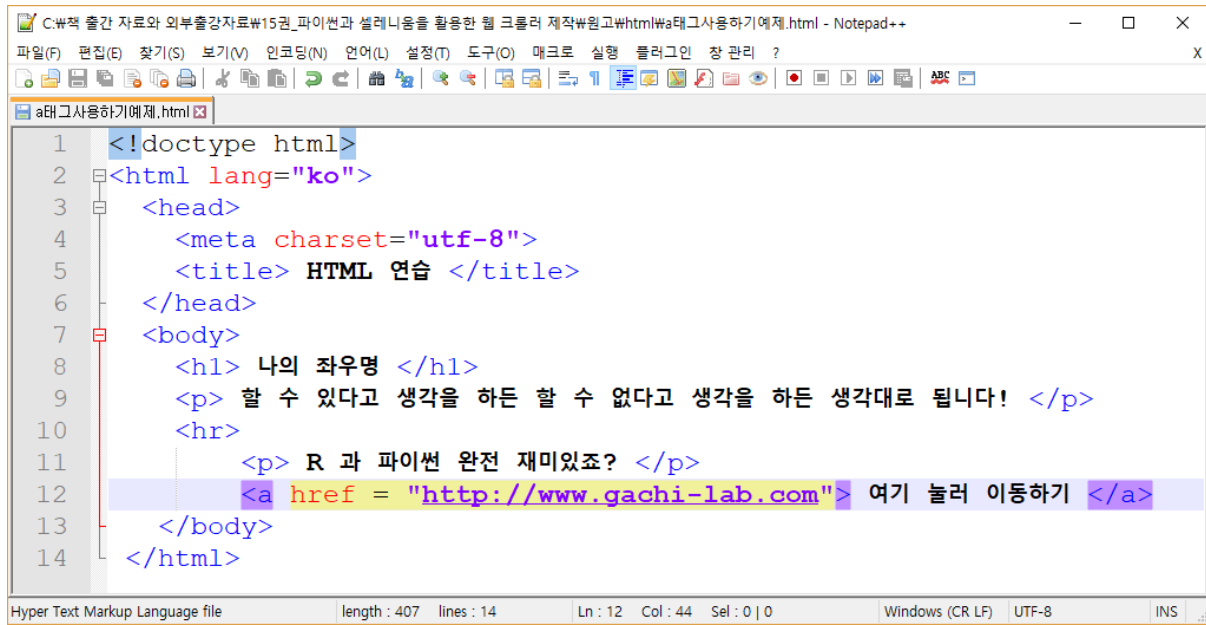


(5) 본문에 하이퍼링크 넣기

인터넷 홈페이지의 특징 중 한 가지는 홈페이지에서 어떤 부분을 누르면 다른 페이지로 연결되는 것입니다. 이런 기능을 하이퍼링크라고 합니다.

 태그를 사용하는데 아주 많이 사용하는 기능인데 아주 쉽습니다.

소스 코드를 아래와 같이 변경하고 저장 한 후 실행 하세요

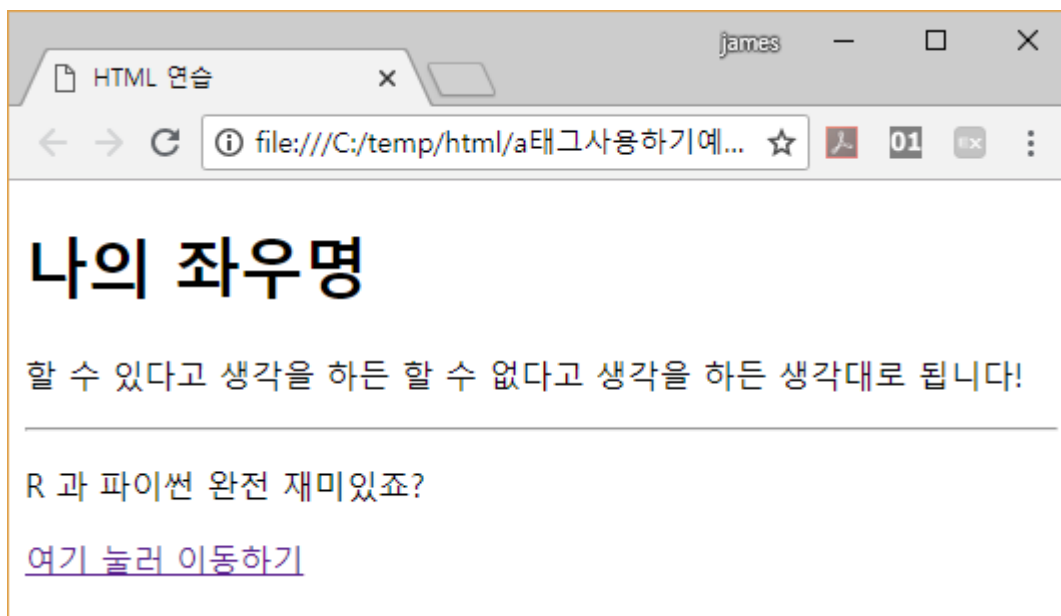


```

1 <!doctype html>
2 <html lang="ko">
3 <head>
4 <meta charset="utf-8">
5 <title> HTML 연습 </title>
6 </head>
7 <body>
8 <h1> 나의 좌우명 </h1>
9 <p> 할 수 있다고 생각을 하든 할 수 없다고 생각을 하든 생각대로 됩니다! </p>
10 <hr>
11 <p> R 과 파이썬 완전 재미있죠? </p>
12 <a href = "http://www.gachi-lab.com"> 여기 눌러 이동하기 </a>
13 </body>
14 </html>

```

Hyper Text Markup Language file length : 407 lines : 14 Ln : 12 Col : 44 Sel : 0 | 0 Windows (CR LF) UTF-8 INS



대부분의 홈페이지에서 소스코드를 보면 지금 배운 href= 라는 구문이 아주 많이 나오는데 이 링크를 따라가서 데이터를 가져와야 하는 경우도 아주 많기 때문에 꼭 이해해야 합니다~

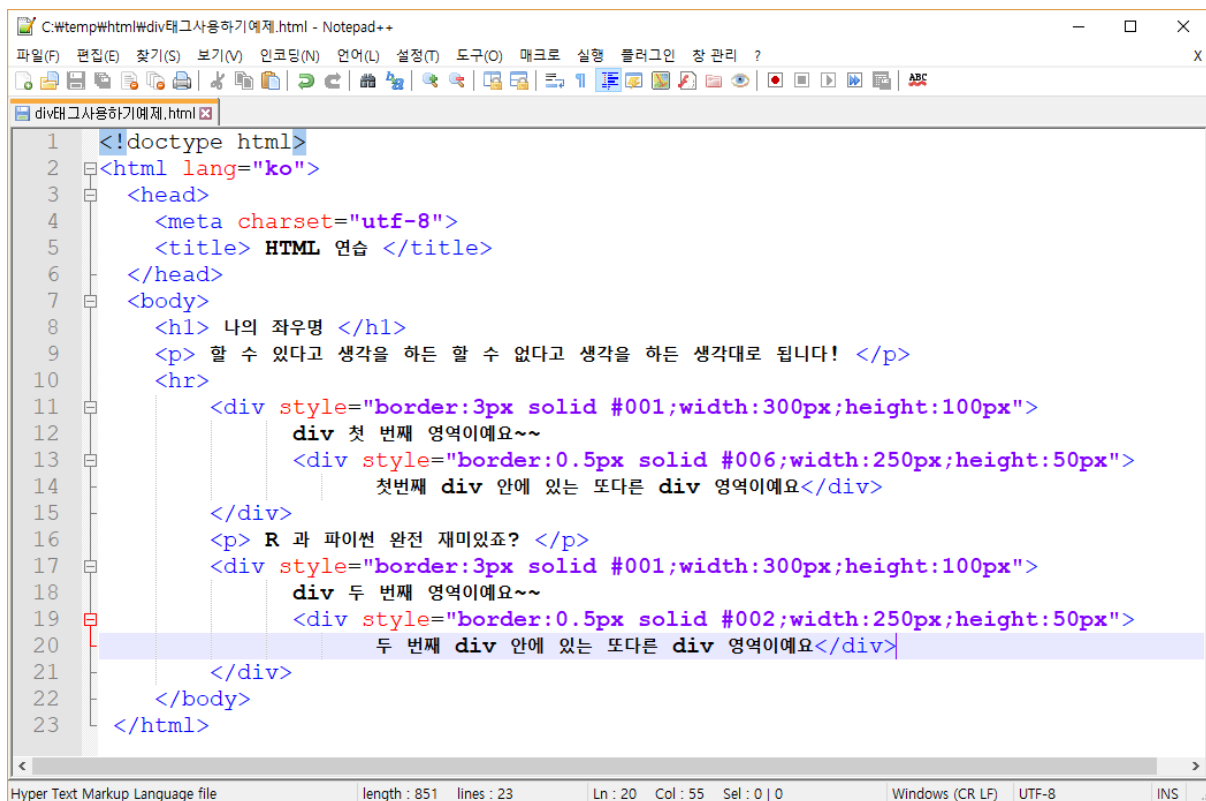
(6) 본문 안에 영역을 지정해서 구분하기 - div 태그 사용하기

웹 페이지 안에 아주 많은 내용이 들어가죠?

그런데 이 내용들이 보일 위치를 지정하고 싶을 경우 있잖아요.

그 때 사용하는 것이 div 태그 입니다.

아래의 예로 살펴 보겠습니다.

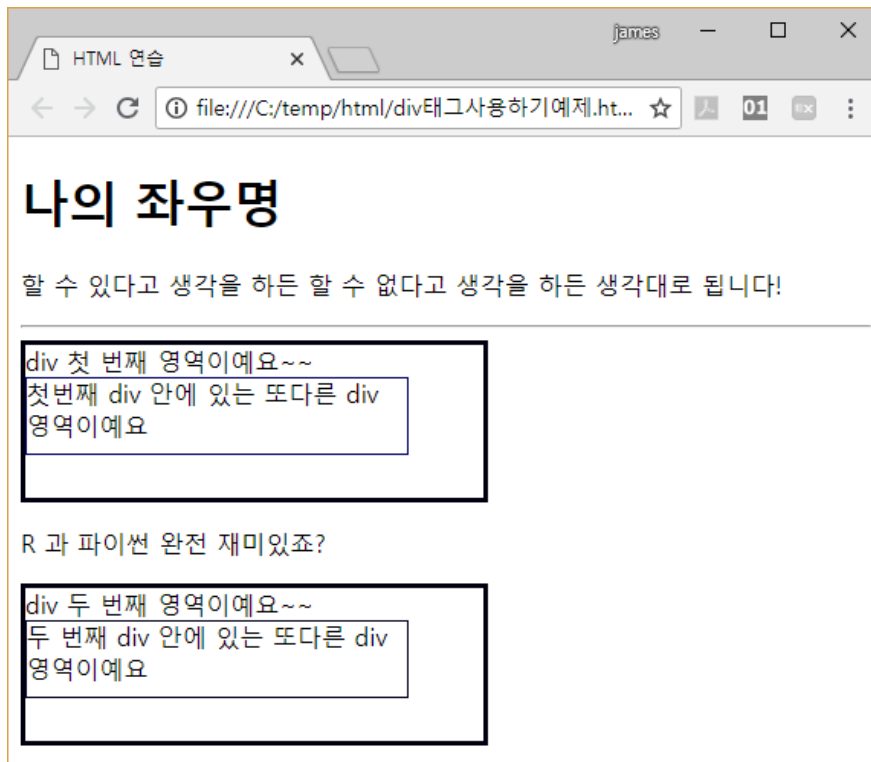


```

1  <!doctype html>
2  <html lang="ko">
3  <head>
4    <meta charset="utf-8">
5    <title> HTML 연습 </title>
6  </head>
7  <body>
8    <h1> 나의 좌우명 </h1>
9    <p> 할 수 있다고 생각을 하든 할 수 없다고 생각을 하든 생각대로 됩니다! </p>
10   <hr>
11   <div style="border:3px solid #001;width:300px;height:100px">
12     <div 첫 번째 영역이에요~~
13       <div style="border:0.5px solid #006;width:250px;height:50px">
14         첫번째 div 안에 있는 또다른 div 영역이에요</div>
15     </div>
16     <p> R 과 파이썬 완전 재미있죠? </p>
17   <div style="border:3px solid #001;width:300px;height:100px">
18     <div 두 번째 영역이에요~~
19       <div style="border:0.5px solid #002;width:250px;height:50px">
20         두 번째 div 안에 있는 또다른 div 영역이에요</div>
21     </div>
22   </div>
23 </body>
</html>

```

위 코드를 저장하고 웹 브라우저로 실행하면 아래와 같이 나옵니다.



위 그림을 보면 웹 페이지에 영역이 정해져 있는 것 보이죠?

지금까지 다양한 HTML 언어 관련 내용들을 살펴 보았습니다.

사실 훨씬 더 많은 내용이 있지만 이 책은 HTML 전공 책이 아니고 웹 크롤러 만들기 위한 목적이라서 그 목적과 관련된 부분만 살펴 보았습니다.

만약 HTML 에 대한 더 자세한 내용을 원하시는 분들은 HTML 관련된 전문 서적을 참고하시길 바랍니다.