HTML 기초와 크롤링 실습

HTML Basics and Crawling

M1505.001600 정보모델링기법과 응용

March 27th, 2018

최교윤

kyoyun3@gmail.com

서울대학교 산업공학과

PART 1. HTML & CSS 기초

HTML이란?

- HTML (HyperText Markup Language)
  - WWW(world wide web)에서 볼 수 있는 웹 페이지를 작성하기 위해 사용하는 언어
  - 웹 페이지의 전체 구조와 각 구조 상의 요소들을 구현하는 데 최적화됨
    - 제목, 단락, 목록, 하이퍼링크, 인용 등

- HTML의 핵심 구성 요소
  - 태그(tag)
    - HTML 문서 상에서 정보를 표현하는 방식을 정의하는 요소로서, HTML 문서를 구성하는 문법적 표시
  - 하이퍼링크(hyperlink)
    - HTML 문서 내의 특정 요소와, 해당 문서 내의 다른 요소 혹은 다른 HTML 문서 내의 요소 간의 연결 관계
    - 줄여서 링크(link)라고도 함

HTML의 기본 구성

- HTML 문서의 기본 구조
  - html 태그 영역
    - 하나의 웹 페이지는 기본적으로 <html> 태그로 시작하여 </html> 태그로 끝남
  - head 태그 영역
    - 웹 페이지 화면에 직접적으로 출력되지는 않지만, 웹 브라우저가 알아야 할 중요한 정보를 서술 (메타데이터 등)
  - body 태그 영역
    - 웹 페이지 화면에 직접 출력되는 정보를 서술

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <meta name="author" content="Kilho Kim">
    <meta name="keywords" content="internet">
    <title>Useful Links</title>
  </head>
  <body>
    <a href="http://imlab.snu.ac.kr" target="_blank">Visit IMLab</a>
  </body>
</html>
```

HTML 관련 용어

- 요소 (element)
  - 웹 페이지 화면을 구성하는 기본 단위
  - 대부분의 요소가 시작 태그와 종료 태그로 이루어지며,
    시작 태그만으로 구성되는 요소도 있음
    - e.g.1. 제목 요소: **<title> ~ </title>**
    - e.g.2. 메타데이터 요소: **<meta charset="utf-8">**

- 속성 (attribute)
  - 요소의 부가적인 정보를 명시하는 부분으로, 시작 태그 안에서 명시됨
  - 속성은 **이름="값"** 의 형태로 명시됨
    - e.g. <a **href="http://imlab.snu.ac.kr"**>IMLab
    - href 속성값: http://imlab.snu.ac.kr

주요 HTML 태그 – 텍스트 관련 태그

- **h#** 태그
  - 웹 페이지의 제목(heading)을 표시
    - e.g. <**h1**>Lorem Ipsum</**h1**>

- **p** 태그
  - 웹 페이지 상에서 하나의 문단(paragraph)을 표시
    - e.g. <**p**>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris ac.</**p**>

- **br** 태그
  - 웹 페이지 텍스트 상에서 줄바꿈(line break)을 수행
    - e.g. Suspendisse tincidunt <**br**>condimentum bibendum. Nulla.

(텍스트 출처: http://www.lipsum.com)

주요 HTML 태그 – 리스트 관련 태그

- **ol, li** 태그
  - 순서가 있는 리스트(ordered list) 표시
    - e.g.

```
<ol>
  <li>Donec</li>
  <li>magna</li>
  <li>mauris</li>
</ol>
```

1. Donec
2. magna
3. mauris

- **ul, li** 태그
  - 순서가 없는 리스트(unordered list) 표시
    - e.g.

```
<ul>
  <li>Ut finibus</li>
  <li>varius</li>
  <li>ligula eget</li>
</ul>
```

- Ut finibus
- varius
- ligula eget

주요 HTML 태그 – 링크, 이미지, 주석 태그

- **a** 태그
  - 하이퍼링크(hyperlink) 표시
  - href 속성(hypertext reference)이 반드시 명시됨
    - e.g. <a **href="http://imlab.snu.ac.kr"**>IMLab

- **img** 태그
  - 이미지(image) 표시
  - src 속성(source)이 반드시 명시됨
    - e.g. <img **src="http://codelion.net/lion.jpg"**>

- 주석 태그
  - 웹 페이지에는 표시되지 않으며, HTML 소스 코드 상에서만 표시되는 주석(comment)을 표시
  - 다른 태그와는 다르게, **<!–** 로 열고 **-->** 로 닫음
    - e.g. **<!–** This is a comment. **-->**

주요 HTML 태그 – 표 관련 태그

- **table** 태그
  - 표 요소의 시작과 끝을 나타냄
  - thead 태그, tbody 태그를 포함

- **thead** 태그
  - 표의 머리에 해당하며, 표의 각 열에 대한 제목을 표시
  - tr 태그, th 태그를 포함

- **tbody** 태그
  - 표의 몸통에 해당하며, 표의 실제 내용을 표시
  - tr 태그, td 태그를 포함

- **tr** 태그: 표에 새로운 행을 추가
- **th** 태그: 표에 새로운 열 제목을 추가
- **td** 태그: 표에 새로운 열 데이터를 추가

주요 HTML 태그 – 표 관련 태그

```
<table border="1">
  <thead>
    <tr>
      <th>Number</th>
      <th>First Name</th>
      <th>Last Name</th>
      <th>Student Number</th>
    </tr>
  </thead>
  <tbody>
    <tr>
      <td>1</td>
      <td>Kildong</td>
      <td>Hong</td>
      <td>2018-11111</td>
    </tr>
    <tr>
      <td>2</td>
      <td>Mongryong</td>
      <td>Lee</td>
      <td>2018-11111</td>
    </tr>
    <tr>
      <td>3</td>
      <td>Chunhyang</td>
      <td>Seong</td>
      <td>2018-12321</td>
    </tr>
  </tbody>
</table>
```

Number	First Name	Last Name	Student Number
1	Kildong	Hong	2018-11111
2	Mongryong	Lee	2018-12345
3	Chunhyang	Seong	2018-12321

주요 HTML 태그 – 기타 태그

- **span** 태그
  - 웹 페이지 텍스트의 일부분에 대하여 스타일을 지정하고자 할 때 사용
    - e.g. Sed sit amet ****diam ultricies.

Sed sit amet diam ultricies.

- **div** 태그
  - 웹 페이지 상의 여러 요소들을 하나의 그룹으로 묶어 일괄적으로 스타일을 지정하고자 할 때 사용
    - e.g. div 요소 안의 모든 요소들을 한 번에 가운데 정렬 및 배경색 변경
    - e.g.

```
<div style="background-color: #CCCCCC; text-align: center">
    <h1 style="font-family: Arial Black;">Lorem Ipsum</h1>
    <p>Lorem ipsum dolor sit amet, consectetur
       adipiscing elit. Mauris consequat. </p>
</div>
```

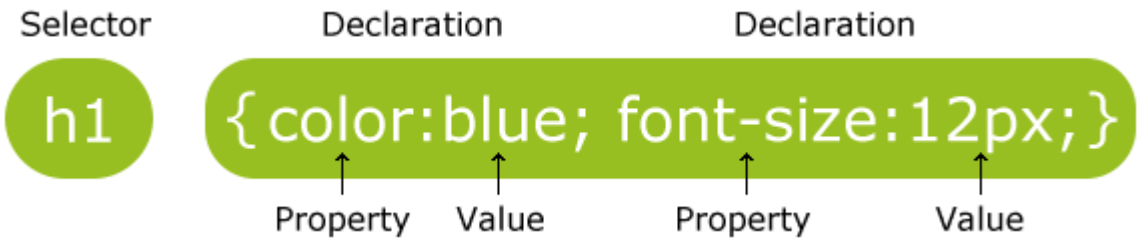
Lorem Ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris consequat.

CSS란?

- CSS (Cascading Style Sheets)
  - HTML 요소에 스타일을 적용하여, 웹 페이지 상에서 어떠한 모양으로 보여질지 서술할 때 사용하는 언어
  - 즉, 웹 페이지를 꾸며줄 때 사용하는 언어

- CSS의 기본 구조
  - 선택자(selector)를 기본 단위로 하며, 이들 각각에 대한 선언(declaration) 부분으로 구성됨
  - 선언 부분 안에는, **속성(property): 값(value);** 이 연속적으로 등장함
    - CSS에서의 속성(property)은 HTML에서의 속성(attribute)와 그 의미가 다름



•(그림 출처: http://www.w3schools.com/css/selector.gif)

HTML에 CSS를 적용하는 방법

- HTML 태그의 style 속성값에 CSS 직접 입력
  - e.g.

```
<html>
  <head>
  </head>
  <body>
    <p style="color: orange;">Lorem Ipsum</p>
  </body>
</html>
```

- HTML 파일 안의 style 태그에 CSS 입력
  - e.g.

```
<html>
  <head>
    <style>
        p {
            color: orange;
        }
    </style>
  </head>
  <body>
    <p>Lorem Ipsum</p>
  </body>
</html>
```

HTML에 CSS를 적용하는 방법

- 외부에 CSS 파일을 만들고, link 태그를 사용하여 이를 불러오기
  - e.g. html 파일

```
<html>
  <head>
    <link rel="stylesheet" type="text/css" href="orangestyle.css">
  </head>
  <body>
    <p>Lorem Ipsum</p>
  </body>
</html>
```

  - css 파일

```
p {
  color: orange;
}
```

CSS 선택자

- 요소 선택자
  - 특정 요소의 HTML 태그 이름을 선택자로 사용함
    - e.g.

```
p {
  color: orange;
}
```

- **id 선택자**
  - 특정한 요소 하나에 대해서만 스타일을 적용하고자 할 경우, 해당 요소의 id를 정의하고 이를 선택자로 사용함
  - 하나의 HTML 요소에 대하여 반드시 고유한 id값을 지정해야 함
    - HTML 문서 상의 서로 다른 두 개의 요소에 대하여 동일한 id값을 지정할 수 없음
    - e.g. HTML 파일

```
<p>Lorem ipsum dolor sit amet,
  <span id="strong">consectetur</span>
  adipiscing elit. Mauris consequat. </p>
```

CSS 파일

```
#strong {
  color: red;
}
```

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris consequat.

CSS 선택자

- class 선택자
  - 특정한 요소 여러 개에 대해서 동일한 스타일을 일괄적으로 적용하고자 할 경우, 해당 요소들의 class를 정의하고 이를 선택자로 사용함
  - 하나의 class 값이 여러 개의 HTML 요소에 대하여 동시에 지정될 수 있음
    - e.g. HTML 파일

```
<p>Lorem ipsum dolor <span class="weak">sit</span> amet,
   <span id="strong">consectetur</span>
   adipiscing <span class="weak">elit</span>. Mauris consequat. </p>
```

    - CSS 파일

```
#strong {
color: red;
}
.weak {
color: gray;
}
```

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris consequat.

참조 링크

- 코드라이언 (CODELION)
  - http://codelion.net
  - HTML/CSS, AWS EC2, Vim, Ruby on Rails 등의 기초 강의 사이트
  - 1주차 5강 ~ 2주차 7강 참조

- W3Schools
  - http://www.w3schools.com
  - HTML, CSS, JavaScript 등의 기초에 대한 백과사전식 서술 사이트
  - HTML의 모든 태그 및 속성(이름-값)을 탐색할 수 있음

- Codecademy HTML & CSS 강좌
  - https://www.codecademy.com/learn/web
  - 웹 사이트 상에서 즉각적인 실습 가능

PART 2. 크롤링 실습

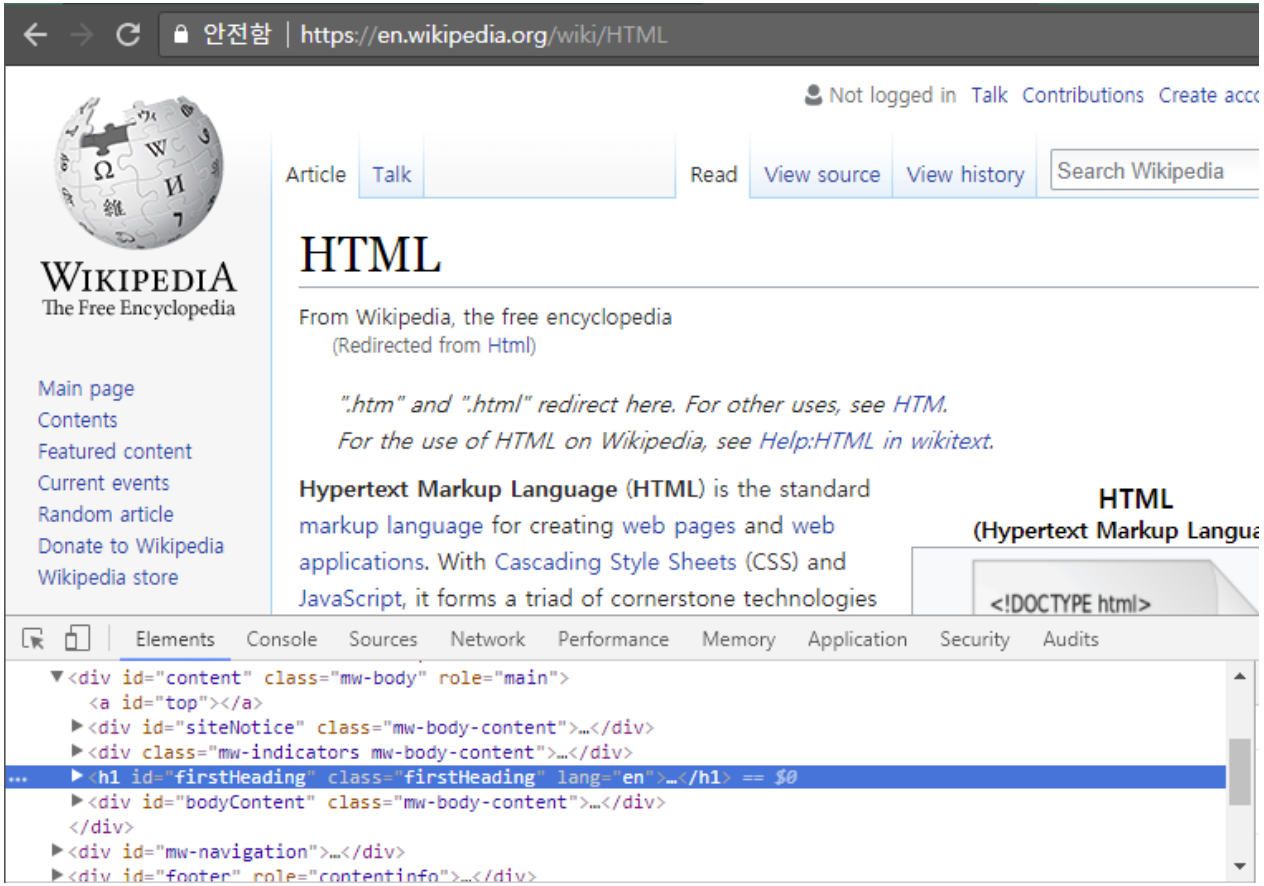
BeautifulSoup 소개

- BeautifulSoup이란?
  - HTML 문서에서의 파싱(parsing)을 위한 python 라이브러리
  - 웹 페이지 URL, HTML파일, 혹은 문자열 등으로부터 파싱 가능
  - **CSS 선택자**를 사용하여 필요한 데이터 포착 및 추출 가능

- 설치
  - Anaconda 설치 시 기본 설치
  - 설치 안 되어있을 경우 cmd창에 다음 명령어 입력
  - conda install beautifulsoup4

BeautifulSoup 예시 코드

- Wikipedia ‘HTML’ 페이지 parsing
  - https://en.wikipedia.org/wiki/HTML
  - 페이지 제목(h1 요소) parsing



BeautifulSoup 예시 코드

- Wikipedia ‘HTML’ 페이지 parsing
  - https://en.wikipedia.org/wiki/HTML
  - 페이지 제목(h1 요소) parsing

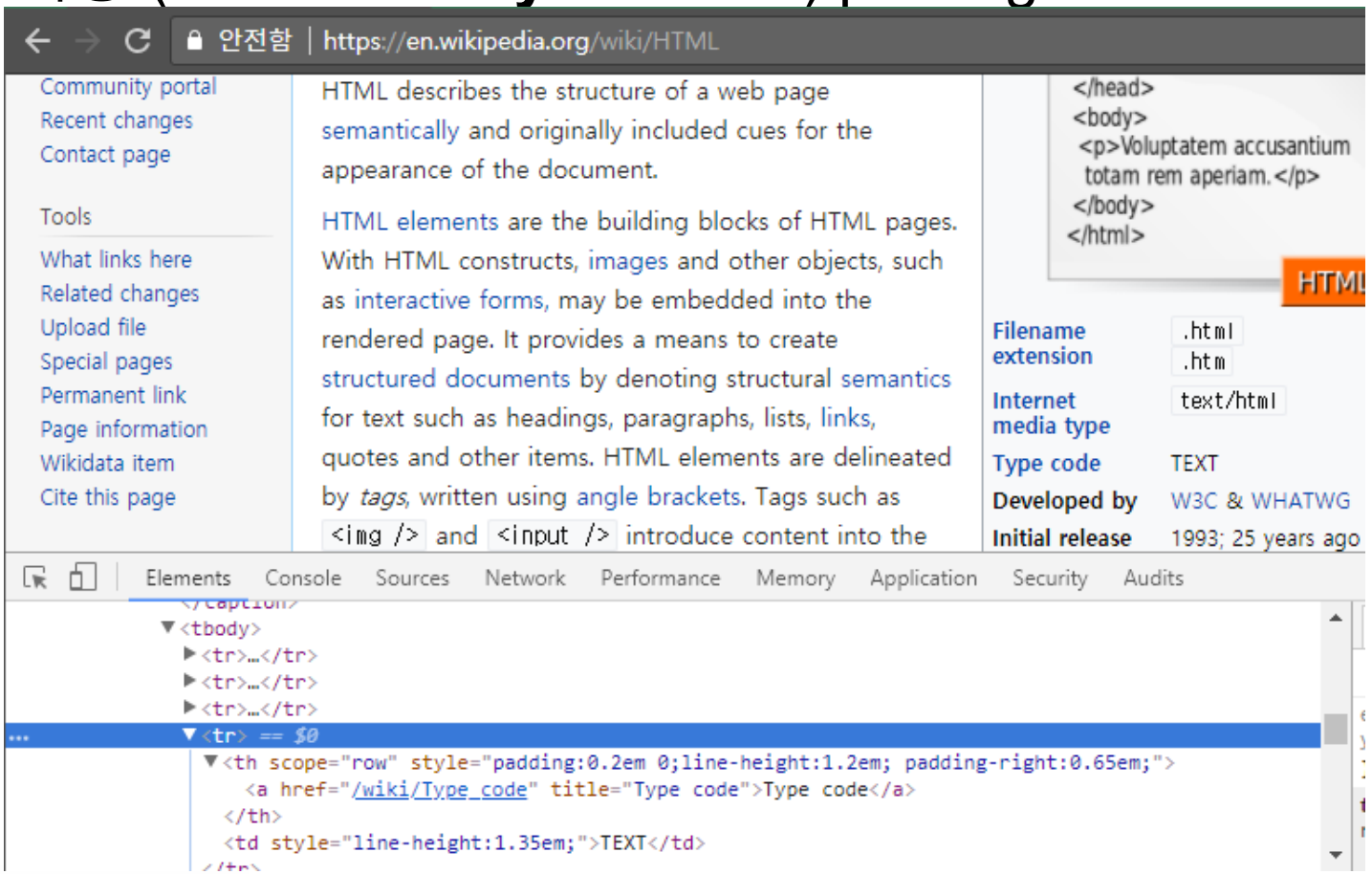
```
from bs4 import BeautifulSoup
from urllib.request import urlopen

url = 'https://en.wikipedia.org/wiki/HTML'
data = urlopen(url).read()
doc = BeautifulSoup(data, 'html.parser')
title = doc.find("h1")
print(title)
```

```
<h1 class="firstHeading" id="firstHeading" lang="en">HTML</h1>
```

BeautifulSoup 예시 코드

- Wikipedia ‘HTML’ 페이지 parsing
  - https://en.wikipedia.org/wiki/HTML
  - 표 내용 (**table > tbody > tr** 요소) parsing



BeautifulSoup 예시 코드

- Wikipedia ‘HTML’ 페이지 parsing
  - https://en.wikipedia.org/wiki/HTML
  - 표 내용 (**table > tbody > tr** 요소) parsing
    - 태그가 여러 개일 경우 속성과 그 이름으로 원하는 요소 선택
    - .find()를 연속적으로 나열하여 특정 요소 안에 포함되는 다른 자식 요소(child element) 접근 가능
    - .find_all()를 사용하여 해당하는 요소를 모두 가져올 수 있음

```
rows = doc.find("table", class_="infobox").find_all('tr')
row = rows[3]
th = row.find('th').string
td = row.find('td').string
print(th, td)
```

```
Type code TEXT
```

BeautifulSoup 참조 링크

- https://www.crummy.com/software/BeautifulSoup/bs4/doc/

프로젝트 팀 구성 공지

- 팀 구성 방법: 자유
- 팀 구성 인원: 3~5인
- 팀 구성에 어려움이 있을 경우 조교에게 메일
- **4/3(화) 낮 12:00까지** 팀 구성 완료 후 조교에게 메일
  - 메일 없을 경우 랜덤 배정