Chapter. 01

데이터 크롤링 개요

웹 사이트 크롤링을 위한 기본 지식

FASTCAMPUS
ONLINE
금융공학/퀀트

강사. 서찬웅

이번시간에 배울 내용

1. 웹과 인터넷에 대한 기본 지식
2. http, get, post 방식
3. html 구성 및 기본 태그

I html에 대해서 알아보겠습니다.

- HTML의 역사
  - HTML은 1990년에 CERN(유럽원자핵연구기구)에서 일하던 팀 버너스 리에 의해 개발
  - 거의 동시에 HTTP와 URI, WWW와 같은 기본적인 구조와 세계 최초의 웹 브라우저도 개발
  - 그 당시에는 HTML은 CERN 내의 연구 공유와 교환을 위해 사용
  - 1993년 CERN이 WWW를 공개한 것과 인터넷 접속 서비스가 시작된 것을 계기로 표준 규격으로 자리 잡게 됨.

- HTML의 버전

1993년	HTML 1.0	1997년	HTML 4.0
1995년	HTML 2.0	1999년	HTML 4.01
1997년	HTML 3.0	2014년	HTML 5

I 인터넷의 기본 용어에 대해서 알아보겠습니다.

- URI(Uniform Resource Identifier)
  - '통합 자원 식별자'라고 하며, 인터넷상에서 존재하는 자원을 나타내기 위한 주소

- HTTP
  - WWW에서 정보를 주고 받을 때 사용하는 프로토콜, 클라이언트와 서버 간의 요청과 응답으로 정보를 전달

- HTML(Hyper Text Markup Language)
  - 웹 문서를 만들기 위해서 사용하는 프로그래밍 언어의 한 종류
  - HTML 태그라는 특수 기호 사용하여 페이지를 작성

- Web Crawler
  - 조직적, 자동화된 방법으로 WWW을 탐색하는 컴퓨터 프로그램
  - Crawler가 하는 작업을 web crawling 혹은 spidering이라 부른다.

I HTTP 요청 방식에 대해서 알아보겠습니다.

- HTTP 요청 방식에는 2가지 방식이 있습니다.
  - GET 방식
    - URL에 변수를 포함시켜 요청
    - 데이터가 헤더에 포함되어 전달
    - 길이 제한이 있고, URL의 데이터가 노출
  - Post 방식
    - 데이터가 본문에 포함되어 전달
    - URL에 데이터가 노출되지 않음
    - 길이 제한이 없음
- 네이버 검색엔진에서 검색할 때는 GET방식으로 동작합니다.

```python
import requests
import json
url = 'http://dart.fss.or.kr/corp/searchCorpL.ax'

data = {'currentPage':10,
        'maxResults' :None,
        'maxLinks' :None,
        'searchIndex':0,
        'textCrpNmAddPer':None,
        'textCrpNm':None,
        'corporationType':None
        }

res = requests.post(url, data=data)
```

I html5의 대해서 공부하겠습니다.

- HTML 파일은 텍스트 파일로, 메모장 같은 텍스트 에디터로도 생성할 수 있습니다.
- 텍스트 에디터로 'test.html'이라는 파일을 아래와 같이 생성해 보겠습니다.

```
1  <!DOCTYPE html>
2  <html>
3      <head>
4          <title>
5              퀀트
6          </title>
7      </head>
8      <body>
9          퀀트를 위한 파이썬
10     </body>
11 </html>
```

'<'와 '>'의 기호로 둘러싼 기호를 '태그(tag)'라고 합니다.
태그 안에는 속성을 지정하여 기능과 정보를 부여할수 있습니다.

Data를 수집할 때 해당 페이지에서 원하는 정보의 태그 위치를 찾아서 가져오는 방식이 scrapping하는 방법입니다.

1 Line : Html5라는 것을 선언합니다.
4~6 Line : html문서의 제목을 설정합니다.
8~10 Line : body 부분에 해당 내용을 적습니다.

태그는 <>로 시작해서 </>로 끝나야 합니다. 파이썬의 문자열을 작성할 때 ""안에 내용을 작성하는 것처럼 html의 태그도 <>..... </> 로 끝납니다.

I 자주 사용하는 html 태그에 대해서 정리하겠습니다.

- 아래는 자주 사용하는 html의 태그 목록입니다.

태그	기능
<a>	하이퍼링크를 추가. href 속성을 사용하여 링크할 url을 지정. 메일 주소나 페이지 안의 다른 부분도 링크 가능
<h1>~ <h6>	문장의 제목을 생성. 숫자가 낮을수록 상위 레벨을 의미.
<p>	단락을 나타낸다. 달리 표시할 요소가 없는 경우에 사용하기를 권장.
 	줄바꿈을 나타낸다. 줄을 바꿀 문자의 끝에 를 붙인다. 이 태그는 </br>은 필요없다.
	번호가 없는 항목 쓰기의 범위를 나타낸다. 항목은 로 지정
	번호가 붙은 항목 쓰기의 범위를 나타낸다. 항목은 으로 지정

I a 태그를 이용하면 링크를 설정할 수 있습니다.

- <a>.. 태그
- a href="이동할 주소"를 설정합니다.

```
1  <!DOCTYPE html>
2  <html>
3      <head>
4          <title>
5              퀀트
6          </title>
7      </head>
8      <body>
9          퀀트를 위한 파이썬 <br>
10         <a href='http://www.naver.com'> 여기를 클릭하면 이동합니다. </a>
11     </body>
12 </html>
```

퀀트를 위한 파이썬
여기를 클릭하면 이동합니다.

-
은 줄바꿈을 나타냅니다.

I <ul> 태그에 대해서 알아보겠습니다.

- 은 unordered list의 약자로 순서가 필요 없는 목록을 만듭니다.
- 은 ordered list의 약자로 순서가 있는 목록을 만듭니다.

```
<!DOCTYPE html>
<html>
    <head>
        <title>
            퀀트
        </title>
    </head>
    <body>
        퀀트를 위한 파이썬 <br>
        <a href='http://www.naver.com'> 여기를 클릭하면 이동합니다. </a>
        <ul>
            <li>test 1</li>
            <li>test 2</li>
            <li>test 3</li>
            <li>test 4</li>
            <li>test 5</li>
        </ul>
    </body>
</html>
```

퀀트를 위한 파이썬
여기를 클릭하면 이동합니다.

- test 1
- test 2
- test 3
- test 4
- test 5

Q1. 위에 는 무엇인가?
A. list item의 약자로 의 항목들을 나열할 때 사용합니다.

Q2. 순서가 필요한 목록도 만들 수 있나요?
A. 대신 를 사용하면 됩니다.

I 태그에 대해서 알아보겠습니다.

- .. 는 p는 paragraph의 약자로 문단을 의미합니다.
- .. 사이에 있는 내용이 하나의 단락을 구성합니다.

```
<!DOCTYPE html>
<html>
    <head>
        <title>
            퀀트
        </title>
    </head>
    <body>
        퀀트를 위한 파이썬 <br>
        <a href='http://www.naver.com'> 여기를 클릭하면 이동합니다. </a>
        <p> This is some text </p>
        <p> This is some text </p>
        <p> This is some text </p>
    </body>
</html>
```

퀀트를 위한 파이썬
여기를 클릭하면 이동합니다.

This is some text

This is some text

This is some text

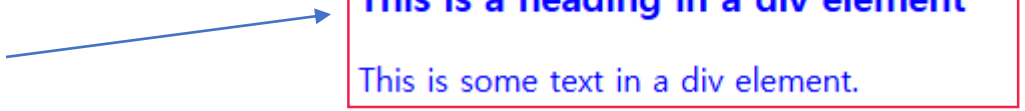
- 과 은 여러 개의 문장을 하나의 단위로 묶는 태그입니다.
  - 콘텐츠를 묶으면 작성하는 css(Cascading Style Sheets)를 적용할 수 있습니다.

태그	기능
	콘텐츠의 단위를 나타낸다. 요소의 전후에는 줄바꿈이 들어간다. (이러한 요소를 '블록요소'라고 한다.)
	콘텐츠의 단위를 나타낸다. 요소의 전후에 줄바꿈이 들어가지 않는다. (이러한 요소를 '인라인 요소'라고 한다.)

I 태그에 대해서 알아보겠습니다.

- 태그를 사용하여 div 태그로 묶여 있는 부분만 스타일을 적용 받습니다.

```
<!DOCTYPE html>
<html>
    <head>
        <title>
            퀀트
        </title>
    </head>
    <body>
        퀀트를 위한 파이썬 <br>
        <a href='http://www.naver.com'> 여기를 클릭하면 이동합니다. </a>
        <p> This is some text </p>
        <div style="color:#0000FF">
            <h3>This is a heading in a div element</h3>
            <p>This is some text in a div element.</p>
        </div>
    </body>
</html>
```



퀀트를 위한 파이썬
여기를 클릭하면 이동합니다.

This is some text

This is a heading in a div element

This is some text in a div element.

- 를 사용하여 해당 부분만 적용하였습니다.

```
<!DOCTYPE html>
<html>
    <head>
        <title>
            퀀트
        </title>
    </head>
    <body>
        퀀트를 위한 파이썬 <br>
        <a href='http://www.naver.com'> 여기를 클릭하면 이동합니다. </a>
        <p> This is some text </p>
        <p>My mother has <span style="color:blue;font-weight:bold">blue</span>
        eyes and my father has <span style="color:darkolivegreen;font-weight:bold">dark green</span> eyes.</p>
    </body>
</html>
```

퀀트를 위한 파이썬
여기를 클릭하면 이동합니다.

This is some text

My mother has **blue** eyes and my father has **dark green** eyes.

I <table> 태그에 대해서 알아보겠습니다.

- Table 태그는 데이터를 모아서 표로 나타낼 때 사용하는 태그입니다.
- <table> 태그와 <tr> 태그, <td>로 작성하면 됩니다.

태그	기능
<table>	표를 작성한다. <table>요소로 둘러싼 부분이 표가 된다.
<tr>	표에 행을 추가. <tr> 요소로 둘러싼 부분이 행이 된다.
<th>	표에 헤더가 되는 셀을 추가한다. <th> 요소로 둘러싼 부분이 셀의 제목.
<td>	표에 셀을 추가. <td> 요소로 둘러싼 부분이 셀의 내용.
<caption>	표의 타이틀을 나타낸다.

퀀트를 위한 파이썬

Firstname	Lastname	gender
seo	chan	male
smith	Jackson	male

```
<!DOCTYPE html>
<html>
    <head>
        <title>
            퀀트
        </title>
    </head>
    <body>
        퀀트를 위한 파이썬 <br>
        <table>
          <tr>
            <th>Firstname</th>
            <th>Lastname</th>
            <th>gender</th>
          </tr>
          <tr>
            <td>seo</td>
            <td>chan</td>
            <td>male</td>
          </tr>
          <tr>
            <td>smith</td>
            <td>Jackson</td>
            <td>male</td>
          </tr>
        </table>
    </body>
</html>
```

I 정리

- html의 개요
- post, get 방식
- html의 자주 사용하는 태그
  - <a>
  - <p>
  -
  - <table>

FAST CAMPUS
ONLINE
서찬웅 강사.

감사합니다

FAST CAMPUS
ONLINE
서찬웅 강사.