

Term Project

Making Meaningful Dataset Using HTML Parsing

Parsing?

- 어떤 문장을 분석하거나 문법적 관계를 해석하는 행위를 말함.
- **BeautifulSoup** 같은 HTML 파싱 라이브러리를 활용하여 포털, 정부기관에서 제공하는 웹페이지에서 실시간 인기 검색어, 환율, 공공 데이터 등의 유용한 데이터를 HTML 소스로부터 추출할 수 있다.
- 1차적으로 추출한 데이터를 2차 가공 단계를 통하여 실생활에서 우리에게 유용한 정보 또는 기능을 제공하는 프로그램을 구현하시오.

Guide Line

- HTML 코드는 웹 페이지에서 마우스 오른쪽 클릭 -> 소스 보기를 누르면 볼 수 있음.

```
<html lang="ko" class="svgless">
<head>
<meta charset="utf-8">
<meta name="Referrer" content="origin">
<meta http-equiv="Content-Script-Type" content="text/javascript">
<meta http-equiv="Content-Style-Type" content="text/css">
<meta http-equiv="X-UA-Compatible" content="IE=edge">
<meta name="viewport" content="width=1100">
<meta name="apple-mobile-web-app-title" content="NAVER" />
<meta name="robots" content="index, nofollow"/>
<meta name="description" content="네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요"/>
<meta property="og:title" content="네이버">
<meta property="og:url" content="http://www.naver.com/">
<meta property="og:image" content="https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png">
<meta property="og:description" content="네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요"/>
<meta name="twitter:card" content="summary">
<meta name="twitter:title" content="">
<meta name="twitter:url" content="http://www.naver.com/">
<meta name="twitter:image" content="https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png">
<meta name="twitter:description" content="네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요"/>

<link rel="shortcut icon" type="image/x-icon" href="/favicon.ico" />

<link rel="stylesheet" type="text/css" href="https://s.pstatic.net/pm2/css/main_v171025.css"/>
<link rel="stylesheet" type="text/css" href="https://s.pstatic.net/pm2/css/webfont_v170623.css"/>
<link rel="stylesheet" type="text/css" href="https://ssl.pstatic.net/sstatic/search/pc/css/api_atcmp_170914.css"/>
<script type="text/javascript" src="https://s.pstatic.net/pm2/js/c/nlog_v170629.js"></script>
<script type="text/javascript">
var nsc = "navertop.v3";
document.domain = "naver.com";
var jindoAll = "https://s.pstatic.net/pm2/js/c/jindo_v171025.js";
if (!window.console) {window.console={};window.console["log"]=function(){};

var isLogin = "dam99001";
function refreshLcs(etc) {etc = etc ? etc : {};if(document.cookie.indexOf("nrefreshx=1") != -1) {etc["mrf"]="1";} else {etc["pan"]="sci";}return etc;}

lcs_do(refreshLcs());
```

Guide Line

- BeautifulSoup Module로 HTML을 파싱한다.
 - Python에 내장 되어 있는 Library가 아님
 - 직접 설치해야 함.
- Parsing한 데이터는 반드시 새로운 정보로 가공 되어야 함
- Parsing할 데이터의 종류 또는 사이트에 대한 제한은 없음
- 최종적으로 구현할 프로그램에 대한 주제 범위 역시 제한 없음.

참고할 만한 문헌

- 파이썬으로 웹 크롤러 만들기
 - 라이언 미첼 지음 / 한빛 미디어

O'REILLY®

초간단 나만의 웹 크롤러로 원하는 데이터를 가져오는 방법

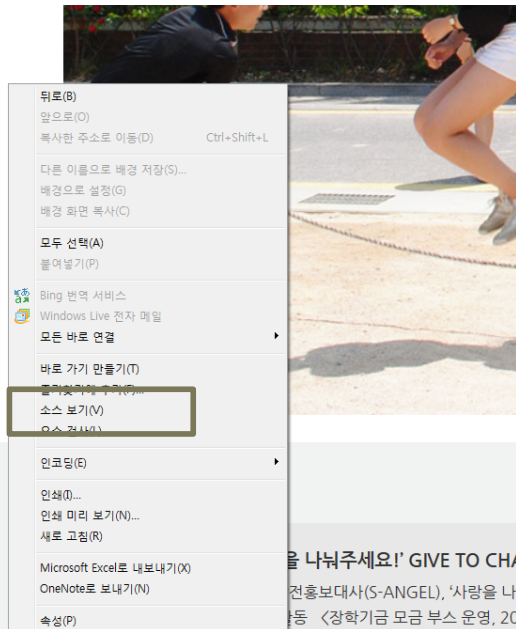


Web Scraping
with Python

파이썬으로 웹 크롤러 만들기

Appendix – HTML

- HyperText Markup Language
- 초 간단하게 말하자면 웹 페이지를 표현하기 위한 언어
- 모든 웹 페이지는 HTML를 통해 구성 됨
- “웹 페이지 -> 마우스 오른쪽 클릭 -> 소스보기 클릭” 하면 해당 페이지의 HTML 코드를 볼 수 있음.

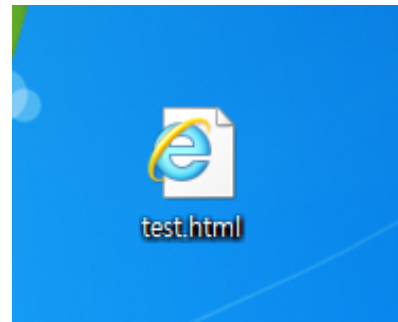


· 상트페테르부르크 국립대학 총장, 성균관대 방문
· 인권과 개발센터, 사회학과 BK21플러스사업단 공동 주최/ 인권

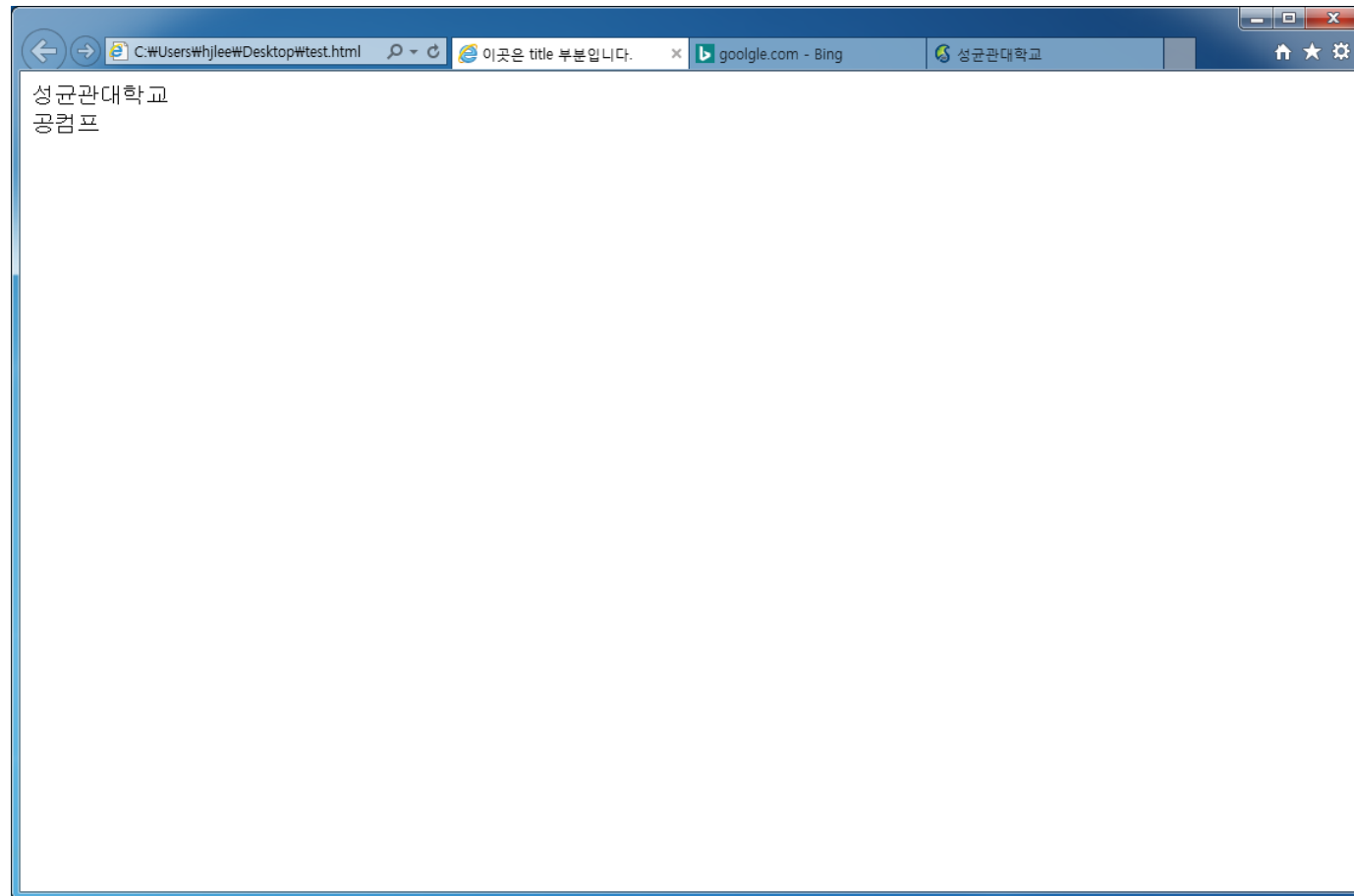
HTML 해보기

```
<html>
  <head>
    <title>이곳은 title 부분입니다. </title>
  </head>
  <body>
    성균관대학교 <br>
    공컴프
  </body>
</html>
```

- 메모장으로도 웹 페이지를 만들고 편집 가능
- ←--- 왼쪽 코드를 메모장에 넣고 test.html 로 바탕화면에 저장하고 실행



HTML 해보기



HTML 구성

```
<html>
  <head>
    <title>이곳은 title 부분입니다. </title>
  </head>
  <body>
    성균관대학교 <br>
    공컴프
  </body>
</html>
```

- HTML은 크게 <head><body>로 구성되며 '태그'라고도 부름.
- 태그는 반드시 시작과 끝이 명시되어야 함
- <태그> </태그>
- 태그의 시작과 끝을 나타냄

태그	HTML	HEAD	BODY
설명	HTML의 문서 선언	HTML 문서에 필요한 정보나 요소를 선언하는 곳 (신경 안 써도 됨)	실제 웹 페이지 본문을 구성하는 곳 (제일 중요한 부분)

<title>안에 내용과 <body>안에 내용이 웹 페이지 어디에 출력됐나 웹페이지 출력화면을 보고 확인해보시오.

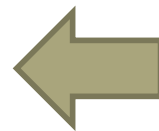
HTML 구성

- HTML에는 많은 태그들과 요소들이 있음
- Term Project를 하기 위해서는 기본적인 태그 몇 개만 알아도 복잡한 HTML 코드에서 원하는 정보만 추출 할 수 있음.
- 본 PPT에서는 그 몇 개만 소개해드리지만 HTML에 대한 충분한 내용을 전달하는 것은 아니니 참고도서 또는 인터넷을 통해 필요한 정보들을 검색하는 것을 권장

HTML 초간단 살펴보기

태그명	설명
 	줄바꿈
<p>...</p>	단락

```
<html>
<head>
  <title>이곳은 title 부분입니다. </title>
</head>
<body>
  성균관대학교 <br>
  공컴프 <br>
  성균관대학교
  <p>
    공컴프
  </p>
</body>
</html>
```



메모장에 넣고 html 확장자로 저장
한 뒤 실행해 결과를 보고

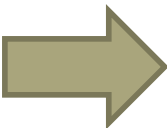
과 <p>의 차이를 이해하시오.

Table

- <table> 표를 생성하는 태그

```
<table border="1">
  <tr>
    <th>번호</th>
    <th>이름</th>
    <th>설명</th>
  </tr>
  <tr>
    <td>1 </td>
    <td>홍길동 </td>
    <td>호부호형</td>
  </tr>
  <tr>
    <td>2 </td>
    <td>성준향</td>
    <td>이몽룡</td>
  </tr>
  <tr>
    <td>3 </td>
    <td>심청이</td>
    <td>공양미</td>
  </tr>
</table>
```

<th>내용</th>: table head 약자로, 표의 제목을 쓰는 역할 (기본값은 굵은 글씨체에, 중앙 정렬)
<tr> 내용 </tr>: table row 약자로, 가로줄을 만드는 역할 (기본값은 보통 글씨체에, 왼쪽 정렬)
<td>내용 </td>: table data 약자로, 셀을 만드는 역할 (기본값은 보통 글씨체에, 왼쪽 정렬)



번호	이름	설명
1	홍길동	호부호형
2	성준향	이몽룡
3	심청이	공양미

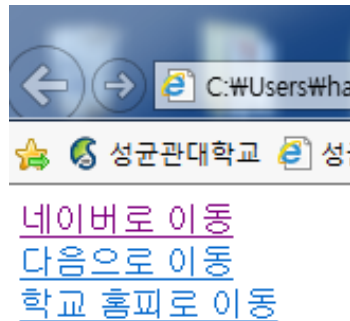
하이퍼링크

-<a> 태그는 Text에 링크를 걸어줄 때 사용함 : text

네이버로 이동

다음으로 이동

학교 홈페이지로 이동



이미지

- : 이미지를 삽입할 때 사용.

Ex)

img1.jpg 이미지를 가로 100 세로 200 px 크기로 두께 1의 테두리를 씌어 삽입한다.

목록

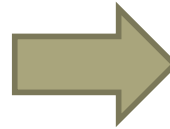
- 와 태그는 순서 목록을 만듦
- 는 각 항목을 나열하는 태그

순서 있는 1

순서 있는 2

순서 있는 3

순서 있는 4

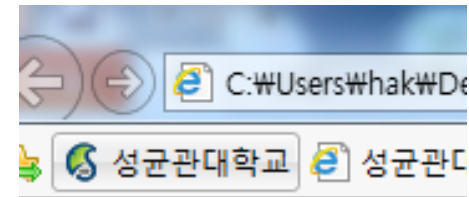


순서 없는 1

순서 없는 2

순서 없는 3

순서 없는 4



1. 순서 있는 1
2. 순서 있는 2
3. 순서 있는 3
4. 순서 있는 4

- 순서 없는 1
- 순서 없는 2
- 순서 없는 3
- 순서 없는 4

작년 Term Project 주제 <학부생 대상>

- Naver News 파싱하여 핵심 키워드 추출
 - 추출된 핵심 키워드를 입력하면 기사 헤드라인을 볼 수 있음
- Tripadvisor 에서 한국의 맛집을 추출
 - 여행 지역을 입력 -> 한, 중, 일식 입력
- 영화 관객수를 추출 현재 가장 인기 있는 영화, 장르별로 분류하여 추천하는 프로그램

[공지 - 숙제]

- 다음주 까지 **반드시** 자신의 컴퓨터나 노트북에 beautifulsoup을 설치해볼 것.
- 독창성 있는 주제일 수록 좋음.
- 틀에 박힌 주제들은 비교 대상이 많아지기 때문에 상대적으로 점수 받기 어려움

참고 사이트

- BeautifulSoup 설치
 - <http://shaeod.tistory.com/900>
- BeautifulSoup Document
 - <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- 간단하게 따라하기
 - <http://www.dreamy.pe.kr/zbxe/CodeClip/163260>