

Week 10 • 데이터 저널리즘

# Social Data Mining 01

---

Joonhwan Lee

human-computer interaction + design lab.

## 오늘 다룰 내용

---

- Crawling from websites

# 1. Crawling from Websites

---

---

## 웹 데이터 수집

- ◆ RQ: 어떤 사람의 트위터 팔로워 구성을 통해 그 사람의 성향을 유추할 수 있을까?
  - ◆ 예1: A라는 사람의 트위터 팔로워는 모두 500명, 그 중에 30% 정치인, 60%는 연예인 → 연예 정보에 관심이 많은 사람.
  - ◆ 예2: A라는 사람이 팔로우하는 정치인 중, 보수성향 정치인 10%, 진보성향 정치인 90% → 진보적인 성향을 가진 사람.
  - ◆ Q1: 팔로우하는 사람의 속성 (연예인인지, 정치인인지, 보수성향의 정치인인지 등)은 어떻게 수집하나..?

# 웹 데이터 수집

[http://twtkr.com/fpl.php?d=3\\_1&n=](http://twtkr.com/fpl.php?d=3_1&n=)

The screenshot shows the twtkr directory website. The browser address bar displays [twtkr.olleh.com/fpl.php?d=3\\_1&n=](http://twtkr.olleh.com/fpl.php?d=3_1&n=). The page title is "twtkr 디렉토리". The main navigation menu includes: 홈 | 검색 | 디렉토리 | 이벤트 | 모임 | 장터 | 동네 | 도구 | 설정 | 도움말 | 로그인.

The left sidebar contains a "순위 디렉토리" (Ranking Directory) with categories such as: 전체, 전체(연예인 제외), 연예인(아이돌), 연예인, 스포츠, 정치인/공직자, 기업인/CEO, 전문가, 미디어, 기업, 기관/단체, 자치단체, 학교/대학교, 작가/출판사, 엔터테인먼트, 생활/문화서비스, 개인전문사업, 종교/종교인, 팬클럽/커뮤니티, 트위터 서비스, 인기/유명 트위터, and 활동 많은 트위터.

The main content area shows a search bar and a "사람찾기" (Find People) button. Below it, a filter for "정치인/공직자 : 정치인 순위 (1,076) > 포인트" is active. The list of users is as follows:

이름	아이디	포인트	순위
유시민	@u_simin	533,403	#1
정봉주	@BBK_Sniper	390,192	#2
김용민	@funronga	378,190	#3
문성근 (민주당,배우)	@actormoon	242,760	#4

Each user entry includes a profile picture, a "팔로우" (Follow) button, and statistics for 팔로워 (Followers), 팔로잉 (Following), 트윗 (Tweets), and 리스트됨 (Listed). For example, 유시민 has 518,664 followers and 40,966 following.

The right sidebar features a "twtkr스폰서" (twtkr Sponsors) section with an "이벤트" (Event) for "줄리엣성형외과 분당점" (@juleclinic). Below that is a "디렉토리 등록/추천" (Directory Registration/Recommendation) button and a "디렉토리 등록 사용자" (Directory Registration Users) count of 8,992. At the bottom, there is a "twtkr 프리미엄 서비스" (twtkr Premium Service) section.

# 웹 데이터 수집

[http://twtkr.com/fpl.php?d=3\\_1&n=](http://twtkr.com/fpl.php?d=3_1&n=)

The screenshot shows the twtkr directory website interface. The browser address bar displays `twtkr.olleh.com/fpl.php?d=3_1&n=`. The page title is "twtkr 디렉토리".

On the left side, there is a "순위 디렉토리" (Ranking Directory) menu with categories such as "전체", "연예인", "스포츠", "정치인/공직자", "기업인/CEO", "전문가", "미디어", "기업", "기관/단체", "자치단체", "학교/대학교", "작가/출판사", "엔터테인먼트", "생활/문화서비스", "개인전문사업", "종교/종교인", "팬클럽/커뮤니티", "트위터 서비스", "인기/유명 트위터", and "활동 많은 트위터".

The main content area shows a search bar and a "사람찾기" (Find People) button. Below this, there is a filter for "정치인/공직자 : 정치인 순위 (1,076) > 포인트". The current view is "포인트" with "20명씩 보기" (View 20 at a time).

Four red arrows point to specific user profiles in the list:

- Arrow 1 points to the profile of **유시민 @u\_simin** (533,403 followers, #1).
- Arrow 2 points to the profile of **정봉주 @BBK\_Sniper** (390,192 followers, #2).
- Arrow 3 points to the profile of **김용민 @funro** (378,190 followers, #3).
- Arrow 4 points to the profile of **문성근 (민주당,배우) @actormoon** (242,760 followers, #4).

Each profile includes a profile picture, name, handle, follower count, and a brief bio. The "팔로워" (Followers) count is highlighted in a box for each profile.

On the right side, there are sections for "twtkr스폰서" (twtkr Sponsors) and "이벤트" (Events), including "줄리엣성형외과 분당점" and "@twtkr\_dir". There is also a "디렉토리 등록/추천" (Directory Registration/Recommendation) section and a "디렉토리 등록 사용자" (Directory Registration Users) section showing 8,992 users.

At the bottom left, there is a "powered by olleh" logo and a "포인트 팔로워 리스트됨" (Points Follow List) section.



---

# 웹 데이터 수집

- ◆ 실습: 소스코드 분석

- ◆ 수집하려는 웹 페이지의 소스를 분석하여, 필요한 데이터가 담긴 반복되는 패턴블럭을 찾아낸다.
- ◆ 반복되는 패턴블럭의 계층 구조를 찾아내 각각의 요소를 정리한다.
- ◆ 계층 구조 내에서 필요한 요소를 따로 찾아 정리한다.
- ◆ `twtkr_example.html`을 열고 주요한 데이터의 반복되는 패턴블럭을 찾고, 내부 데이터를 구조화 하시오.

# 웹 데이터 수집

## ◆ 실습

```
<div class="total_ranking">
```

```
<div class="stream">
```

```
<div class="avatar">
```

```
<div class="article">
```

```
<div class="header">
```

```
<cite>
```

```
...
```

```
<div class="stream">
```

```
<div class="stream">
```

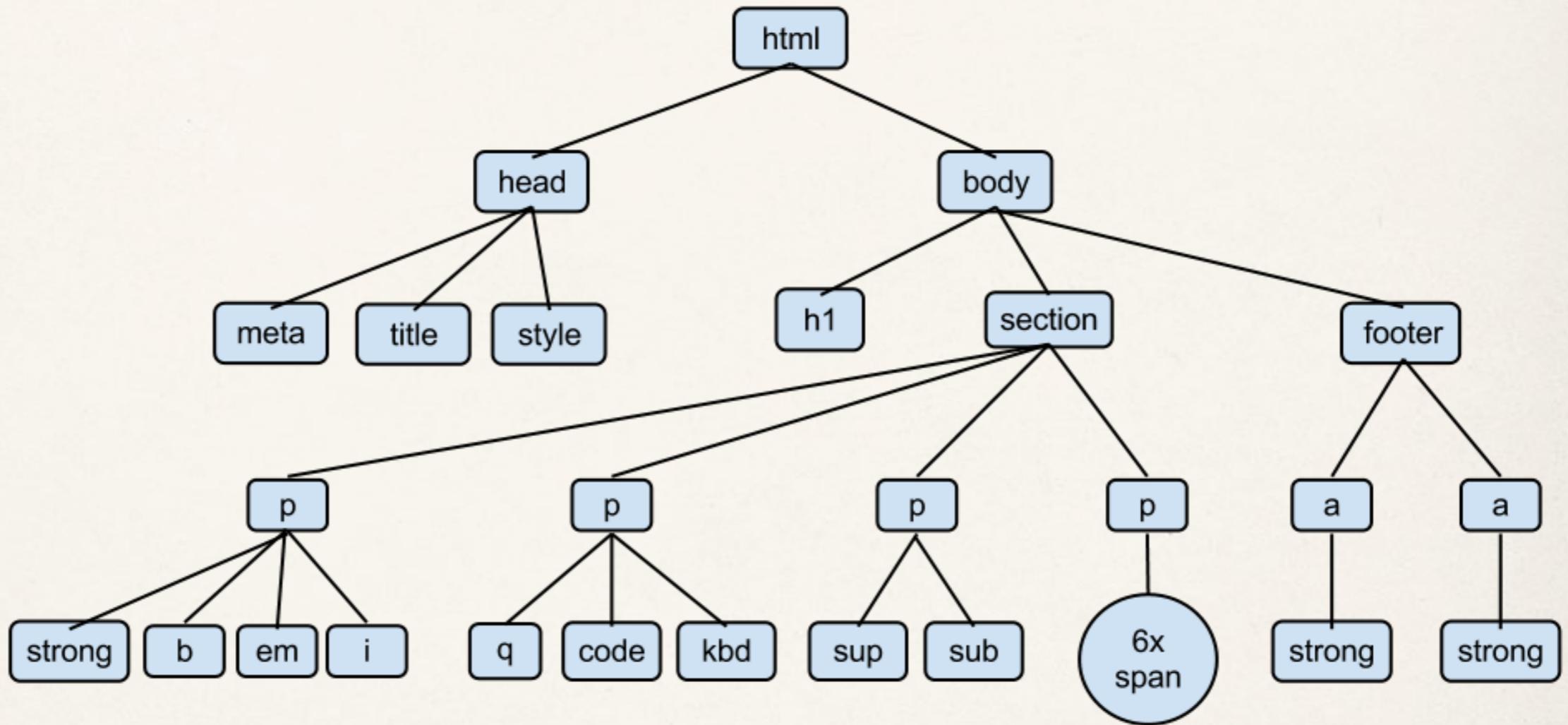
```
...
```

---

# BeautifulSoup을 이용한 웹페이지 수집 및 분석

- ◆ 웹 문서로부터 특정한 데이터를 추출하기 위해서는 HTML 문서를 읽고 구조를 해석할 수 있는 소프트웨어가 필요.
- ◆ BeautifulSoup은 HTML, XML 등을 읽고 해석할 수 있는 소프트웨어 (parser)
  - ◆ BS4는 문서를 파싱한 후 DOM Tree 를 만든다.
- ◆ BeautifulSoup 설치
  - ◆ `pip install beautifulsoup4`

# HTML Document 와 DOM Tree



# HTML Document 와 DOM Tree

The Document

```
<html>
<body>
<h1>Title</h1>
<p>A <em>word</em></p>
</body>
</html>
```

The DOM Tree

```
DOCUMENT
├── ELEMENT: html
│   ├── TEXT: '\n'
│   ├── ELEMENT: body
│   │   ├── TEXT: '\n'
│   │   ├── ELEMENT: h1
│   │   │   └── TEXT: 'Title'
│   │   ├── TEXT: '\n'
│   │   ├── ELEMENT: p
│   │   │   ├── TEXT: 'A'
│   │   │   └── ELEMENT: em
│   │   │       └── TEXT: word
│   └── TEXT: '\n'
└── TEXT: '\n'
```

---

# BS4를 이용한 HTML Parsing

- ◆ BeautifulSoup의 사용

```
> from bs4 import BeautifulSoup
> html_doc = "<html><body><h1>Mr. Belvedere Fan
Club</h1></body></html>"

> soup = BeautifulSoup(html_doc, "html.parser")
> soup
=> <html><body><h1>Mr. Belvedere Fan Club</h1></
body></html>

> print(soup.prettify())

> heading = soup.find_all("h1")
=> [<h1>Mr. Belvedere Fan Club</h1>]

> heading[0].get_text()
=> 'Mr. Belvedere Fan Club'
```

---

# BS4를 이용한 HTML Parsing

## ◆ find\_all 의 사용법

- ◆ `find_all("h1")`

- ◆ `<h1>~</h1>` 태그 안의 내용

- ◆ `find_all("div")`

- ◆ `<div>~</div>` 태그 안의 내용

- ◆ `find_all("div", class_="footer")`

- ◆ `<div class="footer">~</div>` 태그 안의 내용

- ◆ `find_all("div", id="footer")`

- ◆ `<div id="nav">~</div>` 태그 안의 내용

- ◆ `divs = soup.find_all("div", class_="header")`

- `for div in divs:`

- `if div.a["href"] == "twitter_anywhere":`

- ◆ `<div class="header"><a href="" id="twitter_anywhere">~</a></div>` 태그 안의 내용

---

# BS4를 이용한 HTML Parsing

## ◆ find\_all의 사용법

- ◆ find\_all이 반환하는 값은 array (한 페이지에 같은 요소가 여럿 있을 것을 가정하므로...)
- ◆ 따라서 find\_all이 수집한 데이터를 처리하기 위해서는 for-loop 등의 iterator 를 사용한다.

```
id_list = []
divs = soup.find_all("div", class_="header")
for div in divs:
    if div.a["href"] == "twitter_anywhere":
        id_list.append(div.a.text)
```

---

## twitter 아이디와 사용자 이름 수집

- ✦ twtkr\_example.html 파일을 읽어 트위터 아이디와 사용자 이름을 수집해 보자. 수집된 id 에서 @ 기호를 삭제하여 출력한다.

- ✦ 예: u\_simin, 유시민

- ✦ (참고) HTML 파일 불러오는 방법

```
with open("data/twtkr_example.html") as  
file:
```

```
    html_doc = file.read()
```

---

## 웹에서 직접 데이터 수집

- ◆ 항상 저장된 페이지에서 파일을 수집할 수 없음.
- ◆ 실시간으로 웹페이지에 접속해서 저장된 페이지를 수집해야 함.
- ◆ 인터넷에 접속하여 페이지의 소스코드를 받아 처리하기 위해서는 다음과 같은 명령어를 사용.

```
◆ import urllib.request
  with urllib.request.urlopen("http://
    twtkr.com/fpl.php?d=3&n=20") as url:
      doc = url.read()
```

**Questions?**

---