

Week 08 • 소셜네트워크 데이터마이닝과 분석

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=

The screenshot shows the twtkr website interface. The browser address bar displays the URL http://twtkr.olleh.com/fpl.php?d=3_1&n=. The website header includes the twtkr logo and a navigation menu with links like 홈, 검색, 디렉토리, 이벤트, 모임, 장터, 동네, 도구, 설정, 도움말, and 로그인. The main content area is titled '순위 디렉토리' and lists various 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 right sidebar contains sections for twtkr 스폰서, 이벤트, twtkr 디렉토리 등록/추천, and twtkr 프리미엄 서비스. The main content area displays a list of users with their profiles, names, and follower counts. The first user listed is 유시민 (@u_simin) with 533,403 followers. Other users include 정봉주 (@BBK_Sniper) with 390,192 followers, 김용민 (@funronga) with 378,190 followers, and 문성근 (민주당, 배우) (@actormoon) with 242,760 followers. The bottom of the page features a 'powered by olleh' logo and a '포인트' section.

twtkr 디렉토리

twtkr.olleh.com/fpl.php?d=3_1&n=

twtkr 디렉토리

홈 | 검색 | 디렉토리 | 이벤트 | 모임 | 장터 | 동네 | 도구 | 설정 | 도움말 | 로그인

순위 디렉토리

- 전체
- 전체(연예인 제외)
- 연예인(아이돌)
- 연예인
- 스포츠
- 정치인/공직자
 - 정치인
 - 국무총리/장·차관
 - 광역단체장
 - 기초자치단체장
 - 지방자치단체의원
 - 교육감
 - 공공기관장
 - 기타
- 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
- 활동 많은 트위터

powered by olleh

포인트 팔로워 리스트됨

1 최시원

twtkr 디렉토리

사람찾기

> 정치인/공직자 : 정치인 순위 (1,076) > 포인트

포인트 팔로워 - 팔로잉 팔로워 리스트됨 팔로잉 트윗 20명씩 보기

Biz twtkr

경북 관광 알럼이 @GB_tour

twtkr 마케팅 @twtkr_mkt

유시민 @u_simin

533,403 #1

팔로워 : 518,664 | 팔로잉 : 40,966 | 트윗 : 1,052 | 리스트됨 : 17,381

정봉주 @BBK_Sniper

390,192 #2

팔로워 : 400,938 | 팔로잉 : 42,992 | 트윗 : 1,253 | 리스트됨 : 9,514

김용민 @funronga

378,190 #3

팔로워 : 364,652 | 팔로잉 : 20,471 | 트윗 : 9,172 | 리스트됨 : 10,385

문성근 (민주당, 배우) @actormoon

242,760 #4

팔로워 : 242,760 | 팔로잉 : 20,471 | 트윗 : 9,172 | 리스트됨 : 10,385

twtkr 스폰서 이벤트

줄리엣성형외과 분당점 @juleclinic

twtkr 디렉토리 등록/추천

8,992 명

twtkr 프리미엄 서비스

웹 데이터 수집

http://twtkr.com/fpl.php?d=3_1&n=

The screenshot shows the twtkr directory website. The left sidebar contains a '순위 디렉토리' (Rank Directory) with categories like '전체' (All), '연예인' (Celebrities), '스포츠' (Sports), '정치인/공직자' (Politicians/Officials), '기업인/CEO' (Businessmen/CEOs), '전문가' (Experts), '미디어' (Media), '기업' (Companies), '기관/단체' (Organizations/Groups), '자치단체' (Local Governments), '학교/대학교' (Schools/Universities), '작가/출판사' (Writers/Publishers), '엔터테인먼트' (Entertainment), '생활/문화서비스' (Lifestyle/Culture Services), '개인전문사업' (Personal Professional Services), '종교/종교인' (Religion/Religious Figures), '팬클럽/커뮤니티' (Fan Clubs/Communities), '트위터 서비스' (Twitter Services), '인기/유명 트위터' (Popular/Famous Twitter), and '활동 많은 트위터' (Active Twitter). The main content area displays a list of political figures under the heading '정치인/공직자 : 정치인 순위 (1,076) > 포인트'. The list includes profiles for 유시민 (@u_simin), 정봉주 (@BBK_Sniper), 김용민 (@funro), and 문성근 (민주당, 배우) (@actormoon). Each profile shows a photo, name, handle, and various statistics. Red arrows point to the profiles of @u_simin, @BBK_Sniper, @funro, and @actormoon. The right sidebar contains sections for 'twtkr스폰서' (twtkr Sponsors), '이벤트' (Events), '디렉토리 등록/추천' (Directory Registration/Recommendation), '디렉토리 등록 사용자' (Directory Registration Users), and 'twtkr 프리미엄 서비스' (twtkr Premium Services).

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

웹 데이터 수집

http://twtkr.com/fpl.php?d=3_1&n=

The screenshot displays a web browser window showing a Twitter-like interface. The main content area lists several user profiles with their names, avatars, and follower counts. The profiles are: 유시민 (@u_simin) with 533,403 followers, 김용민 (@funronga) with 378,190 followers, and 문성근 (민주당, 배우) (@actormoon) with 242,760 followers. The interface includes a sidebar with navigation links and a right-hand panel with additional user information and a 'twtkr' logo.

Overlaid on the bottom half of the browser window is a developer tool. The left pane shows the file explorer with a tree view containing 'Frames', 'Images', 'Scripts', 'Stylesheets', and 'Extension Scripts'. The middle pane displays the HTML source code, highlighting the 'div.tatal_ranking' element. The right pane shows the 'Type' and 'Location' tabs, with the 'Location' tab selected, displaying the full URL: `http://twtkr.olleh.com/fpl.php?d=3_1&s=&p=1`.

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.html을 열고 주요한 데이터의 반복되는 패턴블럭을 찾고, 내부 데이터를 구조화 하시오.

웹 데이터 수집

◆ 실습

```
<div class="tatal_ranking">
```

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

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

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

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

```
<cite>
```

```
...
```

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

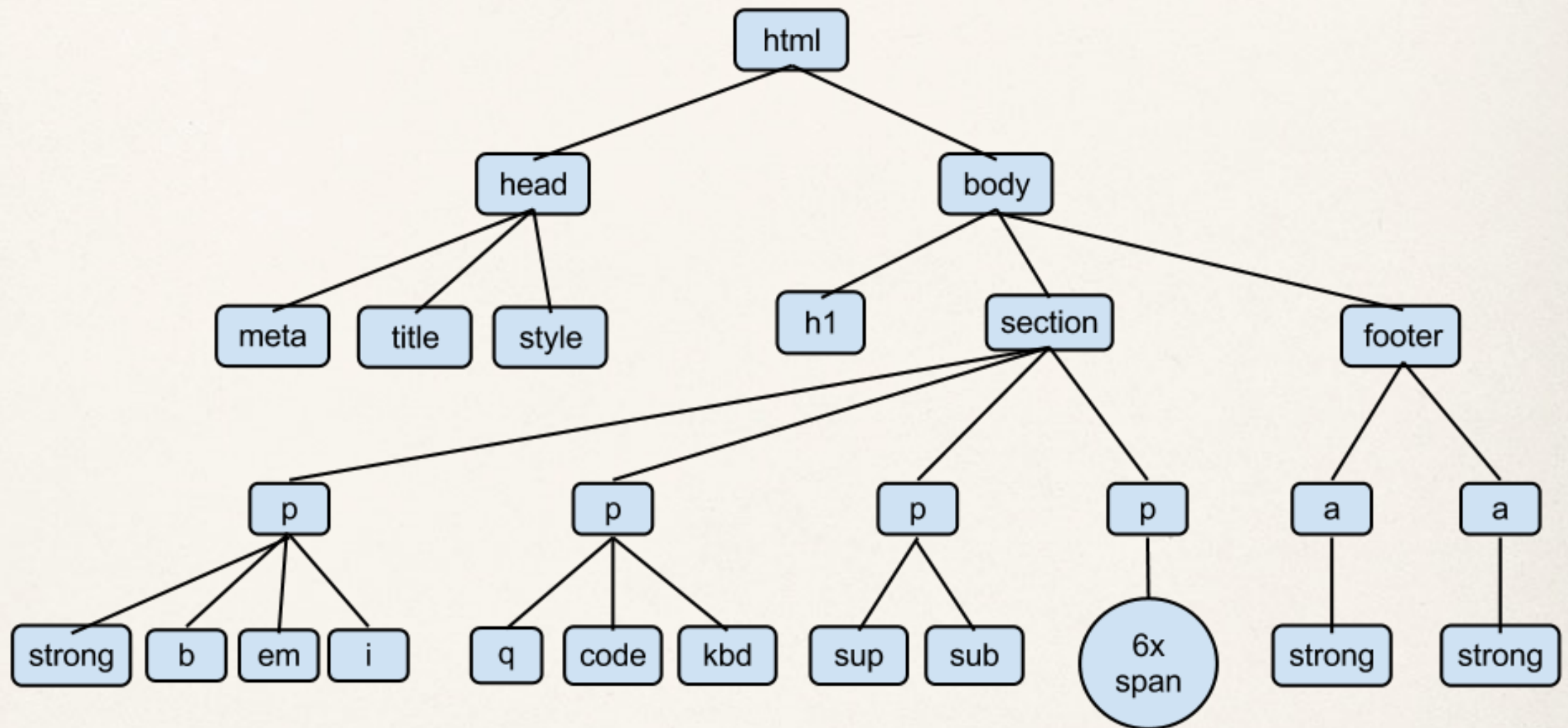
```
<div class="stream">
```

```
...
```

BeautifulSoup을 이용한 웹페이지 수집 및 분석

- ✦ 웹 문서로 부터 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">~</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...?
