

Web Scraper

in 30 Minutes

강철 <kang@cheol.net>

발표자 소개



KAIST 전산학과

2015년부터 G사에서 일합니다.



teampopong에서 **대한민국 정치의 모든 것**을 개발하고 있습니다.

목표

웹 스크래퍼를 프레임웍 없이 처음부터 작성해 본다.

목표

웹 스크래퍼를 프레임웍 없이 처음부터 작성해 본다.



스크래퍼/크롤러의 작동 원리를 이해한다.

목표

웹 스크래퍼를 프레임웍 없이 처음부터 작성해 본다.



스크래퍼/크롤러의 작동 원리를 이해한다.



언제든 필요할 때 동시성 크롤러를 직접 구현

목차

개론

- 스크래퍼(Scraper), 크롤러(Crawler)
- 웹(WWW)의 동작 방식

실습

1. Chrome Developer Tools로 웹사이트 구조 파악하기
2. Requests로 웹페이지 요청하기
3. scrapy.selector를 이용해 데이터 추출하기
4. gevent로 여러 페이지 동시에 요청하기

Scraper, Crawler

Web crawling:

웹에서 링크를 타고다니며 웹페이지들을 수집하는 행위

예) Google bot이 전 세계의 웹사이트를 수집하여 검색 서비스 제공

Web scraping:

웹사이트에서 정보를 추출하는 행위

예) 쇼핑몰 가격비교 - 각 쇼핑몰 상품 페이지에서 상품 이름, 가격 등을 추출

Scraper, Crawler

Google 검색 결과
(structured)

명량

영화 2014

★★★★☆ 3.9/5 - 씨네21
★★★★☆ 4.2/5 - 왓차

■ 박스오피스 2위 - 8. 24.

〈명량〉은 2014년 7월 30일에 개봉된 대한민국의 영화이다. 정유재란 중 명량 해전을 소재로 하고 있다. 위키백과

개봉일: 2014년 7월 30일
감독: 김한민
상영등급: 15세 이상 관람가
상영 시간: 128분
관련사이트: 12vs330.co.kr
언어: 한국어
각본: 전철홍, 김한민
장르: 드라마 영화, 전쟁 영화, 액션 영화

출연진

5개 이상 항목 더보기



최민식



노민우



류승룡



이정현



조진웅

출처: 씨네21, 왓차

피드백

왜 필요한가

자세히보기

국가안전보장회의법 일부개정법률안

심사진행단계

접수 → **위원회 심사** → 체계자구 심사 → 본회의 심의 → 정부 이송 → 공포

의안접수정보

의안번호	제안일자	제안자	문서	제안이유 및 주요내용	제안회기
1911465	2014-08-25	유대운의원 등 10인	▶ 목록 의안원문		제19대 (2012~2016) 제328 회

소관위 심사정보

소관위원회	회부일	상정일	처리일	처리결과	문서
국회운영위원회	2014-08-26				

부가정보

국회 의안정보시스템.

왜 필요한가

1. 의원 이름을 클릭하면 그 의원의 다른 법안들도 볼 수 있음 좋겠는데...
2. 이 의원의 출석률은 얼마나 될까?
3. 이 법안을 발의한 의원들의 정당 분포가 어떻게 될까?
전부 새누리당? 아니면 반반?

제공되는 기능 외에 하고 싶은 게 너무 많아!

왜 필요한가



국회, 선거관리위원회 등 국가 정보시스템 데이터를 수집해 만든

대한민국 정치의 모든 것 <http://pokr.kr/>

웹(WWW)의 동작 방식



server



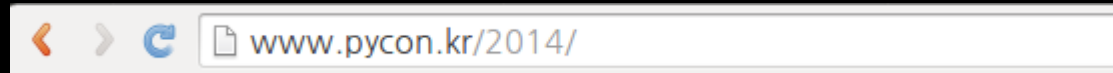
client

웹(WWW)의 동작 방식

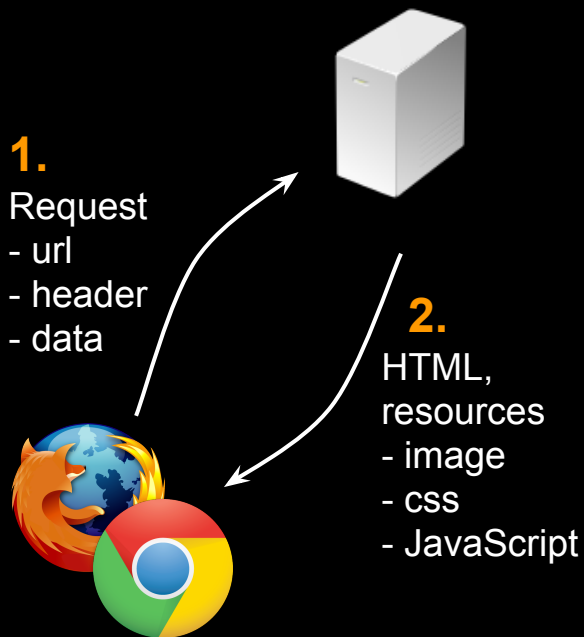
1.

Request

- url
- header
- data

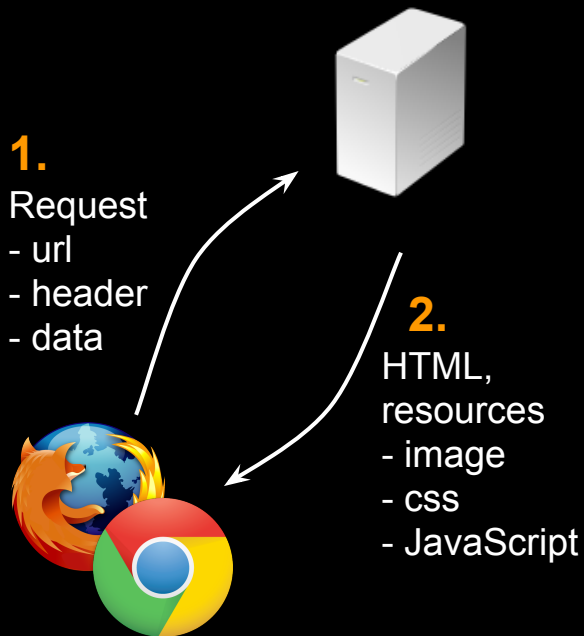


웹(WWW)의 동작 방식



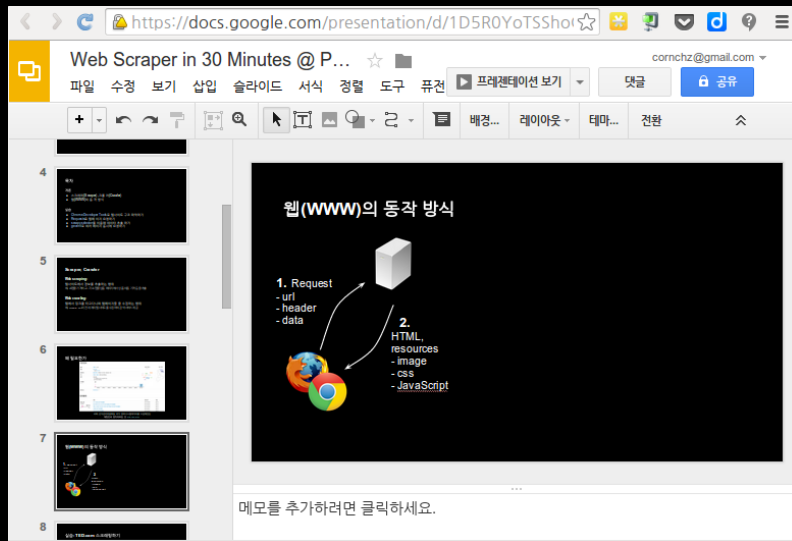
```
84 <li class="dropdown">
85   <a href="#" class="dropdown-toggle" data-toggle="dropdown">
86     <span class="glyphicon glyphicon-python"></span> 파이콘 한국
87   <ul class="dropdown-menu" role="menu">
88
89     <li class="">
90       <a href="/about/pyconkr">파이콘 한국 2014</a>
91     </li>
92
93     <li class="">
94       <a href="/about/coc">파이콘 성명서</a>
95     </li>
96
97     <li class="">
98       <a href="/about/detail">행사 상세 정보</a>
99     </li>
100
101     <li class="">
102       <a href="/about/announcements">알림</a>
103     </li>
104
```

웹(WWW)의 동작 방식




3.

브라우저는 그걸 받아 렌더링








실습: TED.com 스크래핑하기


TED Watch Read Attend Participate About

Search...  Log in Sign up


1800+ talks to stir your curiosity

Find just the right talk: Topics  Events  Languages  Rating (funny, etc.) 


Sort by: Newest 




Ziyah Gafic
Everyday objects, tragic histories
112K views • Aug 2014
Informative, Courageous




Laurel Braitman
Depressed dogs, cats with OCD — what animal madness means for us humans
188K views • Aug 2014
Informative, Inspiring




Jarrett J. Krosoczka
Why lunch ladies are heroes
184K views • Aug 2014
Inspiring, Beautiful



Aziza Chaouni
How I brought a river, and my city, back to life
212K views • Aug 2014
Inspiring, Informative



Tim Berners-Lee
A Magna Carta for the web
225K views • Aug 2014
Informative, Inspiring



Clint Smith
The danger of silence
936K views • Aug 2014
Inspiring, Beautiful

실습: TED.com 스크래핑하기

목표:

최신 비디오 목록을 긁어모아 json으로 저장한다.

1. Chrome Developer Tools로 웹사이트 구조 파악하기

The screenshot displays the Chrome Developer Tools interface. The top navigation bar includes tabs for Elements, Network, Sources, Timeline, Profiles, Resources, Audits, and Console. Below this, the Network tab is active, showing a list of requests. The first request, 'browse?page=2 /talks', is selected and highlighted in blue. To the right of the request list, the 'Headers' sub-tab is open, displaying the details of the selected request. The 'Request Headers' section is expanded, showing various headers such as 'Accept', 'Accept-Encoding', 'Accept-Language', 'Cache-Control', 'Connection', 'Cookie', 'Host', and 'User-Agent'. The 'Query String Parameters' section is also expanded, showing 'page: 2'.

Elements | Network | Sources | Timeline | Profiles | Resources | Audits | Console

Preserve log | Disable cache

Name
Path

browse?page=2
/talks

global.css?1408478750
assets2.tedcdn.com/stylesheets

talks.css?1408478750
assets2.tedcdn.com/stylesheets

core.js?1408478750
assets2.tedcdn.com/javascripts

analytics.js
www.google-analytics.com

gpt.js
www.googletagmanager.com/tag/js

7cf9611626ed32b38dfde06ee5fd53cdddf13cad_2400x1800.jpg?quality...

Headers | Preview | Response | Cookies | Timing

Remote Address: 192.237.224.18:80
Request URL: http://www.ted.com/talks/browse?page=2
Request Method: GET
Status Code: 200 OK

Request Headers view source

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Accept-Encoding: gzip, deflate, sdch
Accept-Language: ko-KR,ko;q=0.8,en-US;q=0.6,en;q=0.4
Cache-Control: max-age=0
Connection: keep-alive
Cookie: __gads=ID=fbd7d152cdf49f39:T=1408851332:S=ALNI_MZbvBbq0FYS64R6v30IgGDive3hgQ; _ga=GA1.2.12075530
Host: www.ted.com
User-Agent: Mozilla/5.0 (X11; Linux i686) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/36.0.1985.143 Safari/537.36

Query String Parameters view source view URL encoded

page: 2

Request URL, method, parameter 등을 확인한다.

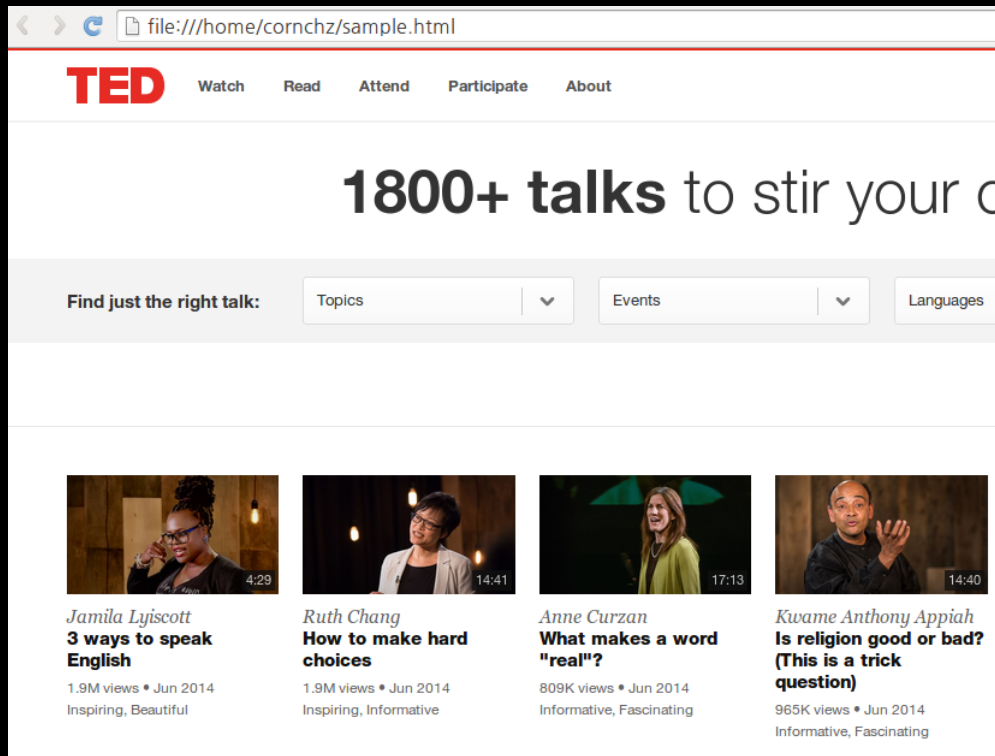
2. Requests 이용해 HTML 다운로드

```
import requests
```

```
def fetch_page(url):  
    r = requests.get(url)  
    return r.text
```


```
with open("sample.html", 'w') as f:  
    html = fetch_page("http://www.ted.com/talks/browse?page=2")  
    f.write(html.encode('utf-8'))
```

2. Requests 이용해 HTML 다운로드




sample.html


3. DOM 구조 파악하기




Jamila Lyiscott
3 ways to speak English
1.9M views • Jun 2014
Inspiring, Beautiful



Ruth Chang
How to make hard choices
1.9M views • Jun 2014
Inspiring, Informative



Anne Curzan
What makes a word "real"?
809K views • Jun 2014
Informative, Fascinating



Kwame Anthony Appiah
Is religion good (This is a trick question)
965K views • Jun 2014
Informative, Fascinating

www.ted.com/talks/yoruba_richen_what_the_gay_rights_movement_learned_from_the_civil_rights_movement

Elements | Network | Sources | Timeline | Profiles | Resources | Audits | Console

```
<div class="filters m1" id="filters">...</div>
<div class="container results" id="browse-results">
  <div class="row row-sm-4up row-lg-6up row-skinny">
    <div class="col">
      <div class="m3">
        <div class="talk-link">
          <div class="media media--sm-v">
            <div class="media_image media_image--thumb talk-link_image">
              <a href="/talks/jamila_lyiscott_3_ways_to_speak_english">
                <span class="thumb thumb--video thumb--crop-top">
                  <span class="thumb_size">
                    <span class="thumb_tugger">
                      
                      <span class="thumb_aligner"></span>
                    </span>
                  </span>
                <span class="thumb_duration">4:29</span>
              </span>
            </div>
          </div>
          <div class="media_message">
            <h4 class="h12 talk-link speaker">Jamila Lyiscott</h4>
            <h4 class="h9 m5">
              <a href="/talks/jamila_lyiscott_3_ways_to_speak_english">
                3 ways to speak English
              </a>
            </h4>
          </div>
        </div>
      </div>
    </div>
  </div>
</div>
```

4. 목록에서 게시물 링크들 추출하기

```
from urlparse import urljoin
from scrapy.selector import Selector
```

```
def talk_links_from_listpage(url):
    html = fetch_page(url)
    sel = Selector(text=html)
    talk_links = sel.css('.talk-link .media__message a::attr(href)').extract()
    talk_links = [urljoin(url, talk_link) for talk_link in talk_links]
    return talk_links
```

```
from pprint import pprint
pprint(talk_links_from_listpage('http://www.ted.com/talks/browse?page=2'))
```

실행결과

```
[u'http://www.ted.com/talks/jamila_lyiscott_3_ways_to_speak_english',
 u'http://www.ted.com/talks/ruth_chang_how_to_make_hard_choices',
 u'http://www.ted.com/talks/anne_curzan_what_makes_a_word_real',
 ...
]
```

5. 각 게시물 세부정보 요청하기

```
import re
```

```
download_re = re.compile(r'http://download.ted.com/talks/[^"]+')
```

```
def talk_from_page(url):
```

```
    html = fetch_page(url)
```

```
    sel = Selector(text=html)
```

```
    download_m = download_re.search(html)
```

```
    return {
```

```
        'title': sel.css('.talk-hero__title::text').extract(),
```

```
        'description': sel.css('.talk-description::text').extract(),
```

```
        'download': download_m.group(0) if download_m else None,
```

```
    }
```

```
def latest_talks(page=1):
```

```
    list_url = 'http://www.ted.com/talks/browse?page={0}'.format(page)
```

```
    talk_links = talk_links_from_listpage(list_url)
```

```
    talks = [talk_from_page(url) for url in talk_links]
```

```
    return talks
```

```
pprint(latest_talks())
```

6. 동시에 여러 페이지 요청하기

```
from gevent import monkey; monkey.patch_all()
from gevent.pool import Pool
```

```
def latest_talks(page=1):
    list_url = 'http://www.ted.com/talks/browse?page={0}'.format(page)
    talk_links = talk_links_from_listpage(list_url)
    # talks = [talk_from_page(url) for url in talk_links]
    pool = Pool(20) # XXX: constant
    talks = pool.map(talk_from_page, talk_links)
    return talks
```

단 5줄 변경만으로: 34초 → 8초

```
pprint(latest_talks())
```


정리

Python의 쉽고 간결한 문법과 강력한 라이브러리를 이용하면
간단한 스크래퍼 정도는 30분이면 충분히 짜고 남는다.

Q&A

Scrapy 좋아요

- command line tools
- referer & user-agent
- item pipeline
- export to file/DB
- crawling: link extraction, queueing, scheduling
- retry & failover
- logging
- daemonize
- distributed execution

감사합니다