KAIST Summer Session 2018

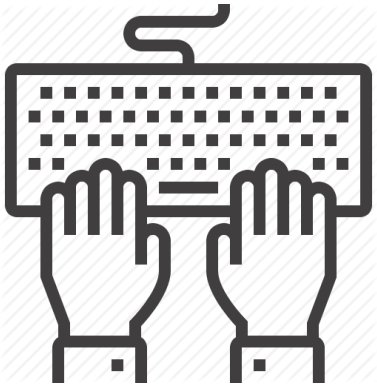
Module 3. Deep Learning with PyTorch

Web Data Extraction

KAIST College of Business

Jiyong Park

13 August, 2018



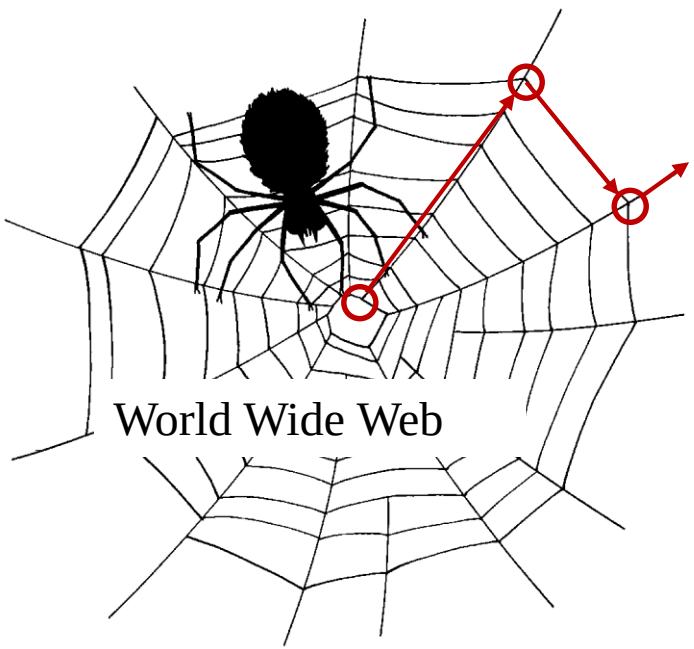
Python Tutorial in 10 minutes

M3.3 Python Tutorial.ipynb

Scraping with Python

Crawling versus Scraping

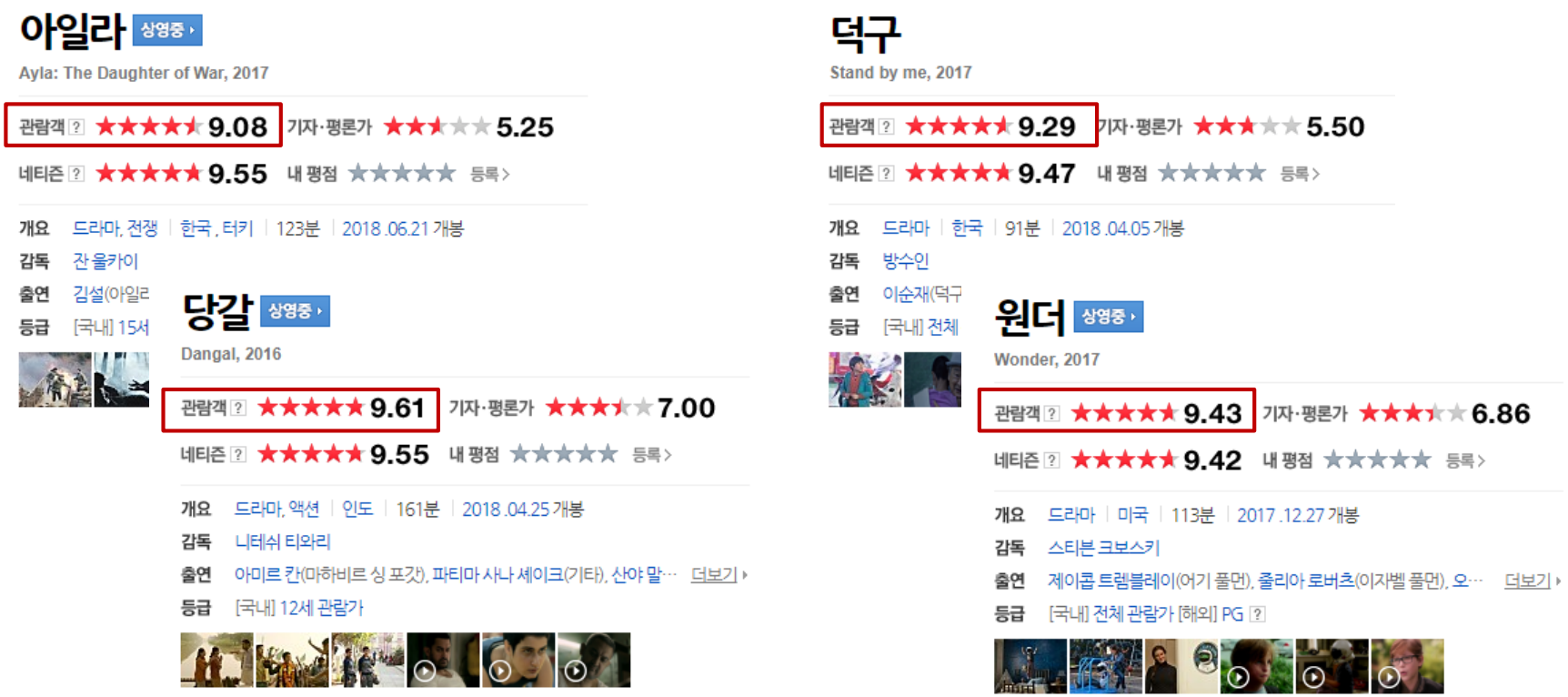
- A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner.



webpage
hyperlink

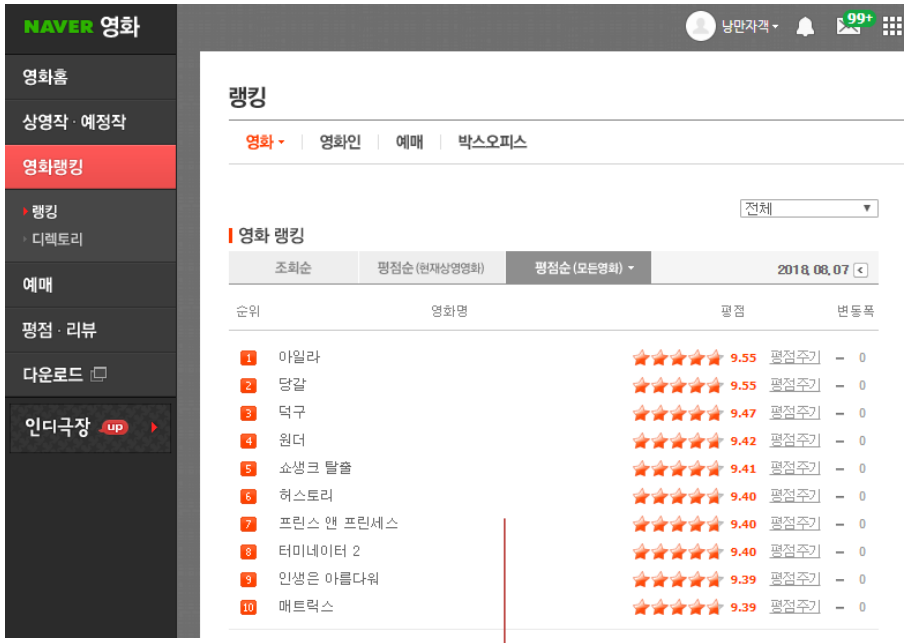
Crawling versus Scraping

- Web scraping (also known as web data extraction) is an automated technique of extracting information from web.



Let’s Scrap the Naver Movie Information

- Predicting movie success is a popular task in the literature (e.g., Lee et al. 2018)



https://movie.naver.com/movie/sdb/rank/rmovie.nhn?sel=pnt

Lee, K., Park, J., Kim, I. and Choi, Y., 2018. Predicting Movie Success with Machine Learning Techniques: Ways to Improve Accuracy. *Information Systems Frontiers*, 20(3), pp.577-588.

Three Steps for Manual Scraping

- First, generate the list of URLs to download
- Second, download the webpage HTML from the URLs
- Third, parse the relevant information from the downloaded HTML

Three Steps for Manual Scraping

- First, generate the list of URLs to download

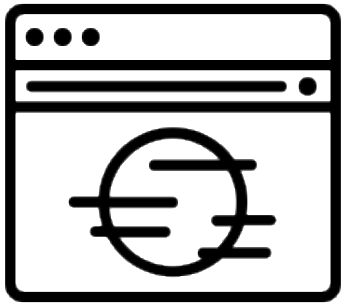
https://movie.naver.com/movie/bi/mi/basic.nhn?code=169240
https://movie.naver.com/movie/bi/mi/basic.nhn?code=157243
https://movie.naver.com/movie/bi/mi/basic.nhn?code=154667
https://movie.naver.com/movie/bi/mi/basic.nhn?code=151196

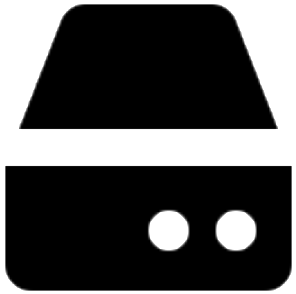
- Second, download the webpage HTML from the URLs

- Third, parse the relevant information from the downloaded HTML

Three Steps for Manual Scraping

1. Request
- URL
- header
- data





Server

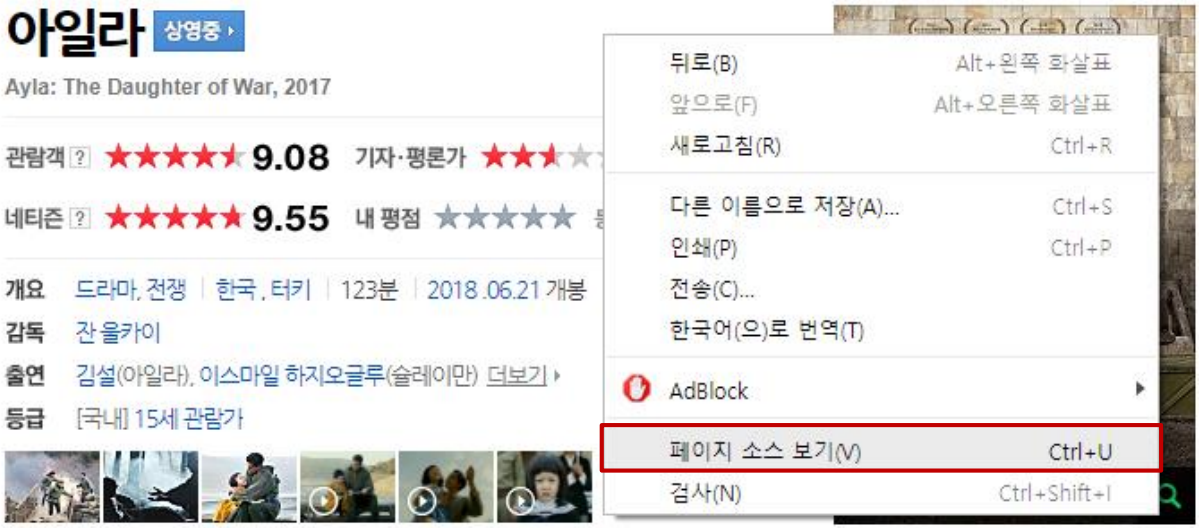
3. Rendering (through Web browser)

2. Response
- HTML
- image
- JavaScript

Three Steps for Manual Scraping

- First, generate the list of URLs to download

- Second, download the webpage HTML from the URLs



- Third, parse the relevant information from the downloaded HTML

Three Steps for Manual Scraping

- First, generate the list of URLs to download
- Second, download the webpage HTML from the URLs
- Third, parse the relevant information from the downloaded HTML

```
<div class="score">
    <div class="uio_ntz_btn see">
        <span class="ntz _actualPointHelpWide">
            <em class="blind">관람객 평점</em>
            <a href="#" id="actualPointHelpButtonWide" class="help _actualPointHelpWide">관람객 평점 도움말</a>
        </span>
        <div class="ly_ntz _actualPointHelpWide" id="actualPointHelpWide" style="display:none">
            <span></span>
            관람객 평점은 네이버영화에서<br>예매하고 실제 관람 후 이용자들이<br>작성한 평점입니다.
            <button type="button" class="btn_close _actualPointHelpWide" id="actualPointHelpCloseButtonWide"><em>닫기</em>

        </div>
        <div class="ly_count" id="actualPointCountWide" style="display:none">
            <span></span>
            참여 <em>71</em>명
        </div>
    </div>
    <a id="actualPointPersentWide" href="./point.nhn?code=169240&onlyActualPointYn=Y#pointAfterTab" class="ntz_score">
        <div class="star_score">


                    <span class="st_off"><span class="st_on" style="width:90.8%">관람객 평점 9.08점</span></span><em
K</em><em class="dot">.</em><em class="num0">0</em><em class="num8">8</em>
```

How about scraping more than 1,000 movies?

Option 1:

Let graduate students do

교육을 빙자한 대학원생 노동 착취

2017년 08월 10일(목) 제516호

홍덕구 (인문학협동조합 조합원) webmaster@sisain.co.kr

가- 가+

- 지금 이 순간에도 많은 대학원생이 '교육'을 빙자한 '도제식 노동'에 시달리고 있다. 대학원생의 노동 착취와 인권 문제에 대해 대학과 정부는 방관한다.

'Y대학교에서 테러로 추정되는 폭발.' 스마트폰으로 기사들을 검색해본 동료가 말했다. "교수 연구실이라는데?" 순간 그 자리에 있던 다섯 명의 시선이 '오복성 패스'처럼 교차했다. 감히 누구도 입 밖으로 꺼내지 못했지만, 모두가 같은 생각을 하고 있음이 분명했다. '대학원생이네.' '대학원생이군.' '대학원생이야.' '대학원생일걸.' '대학원생이다.' 말하지 않아도 알 수 있었다.

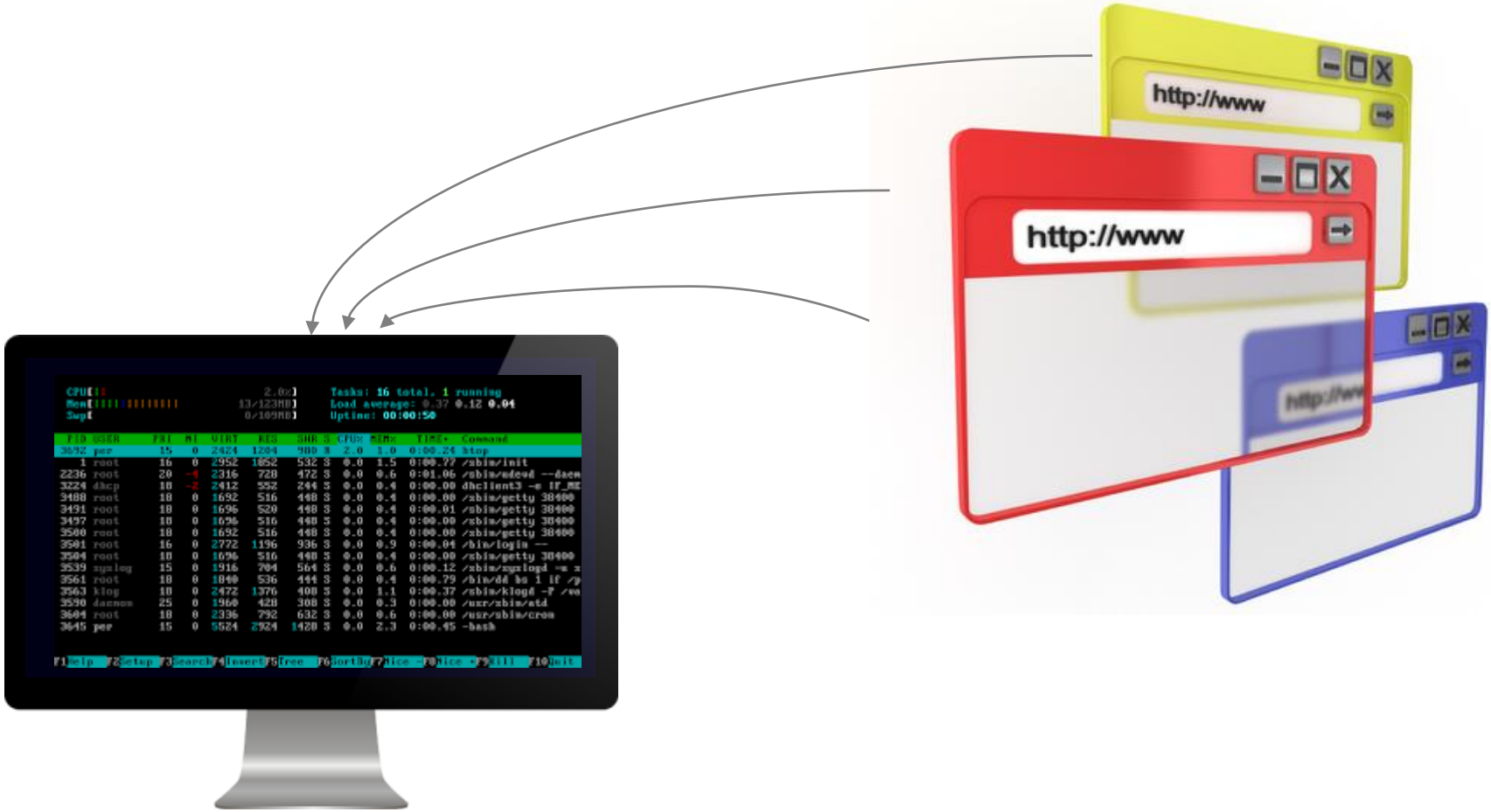
Source: http://www.sisain.co.kr/?mod=news&act=articleView&idxno=29777

KAIST COLLEGE OF BUSINESS

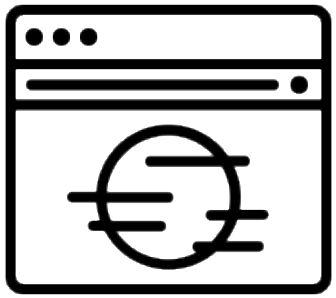
Let them (or us) be free!

Option 2:

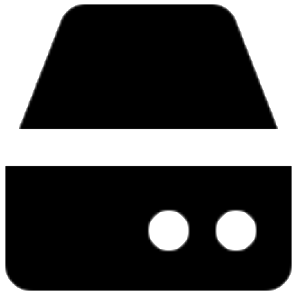
Let computers do



Web Browsers are not Necessarily Needed for Requests



1. Request
- URL
- header
- data

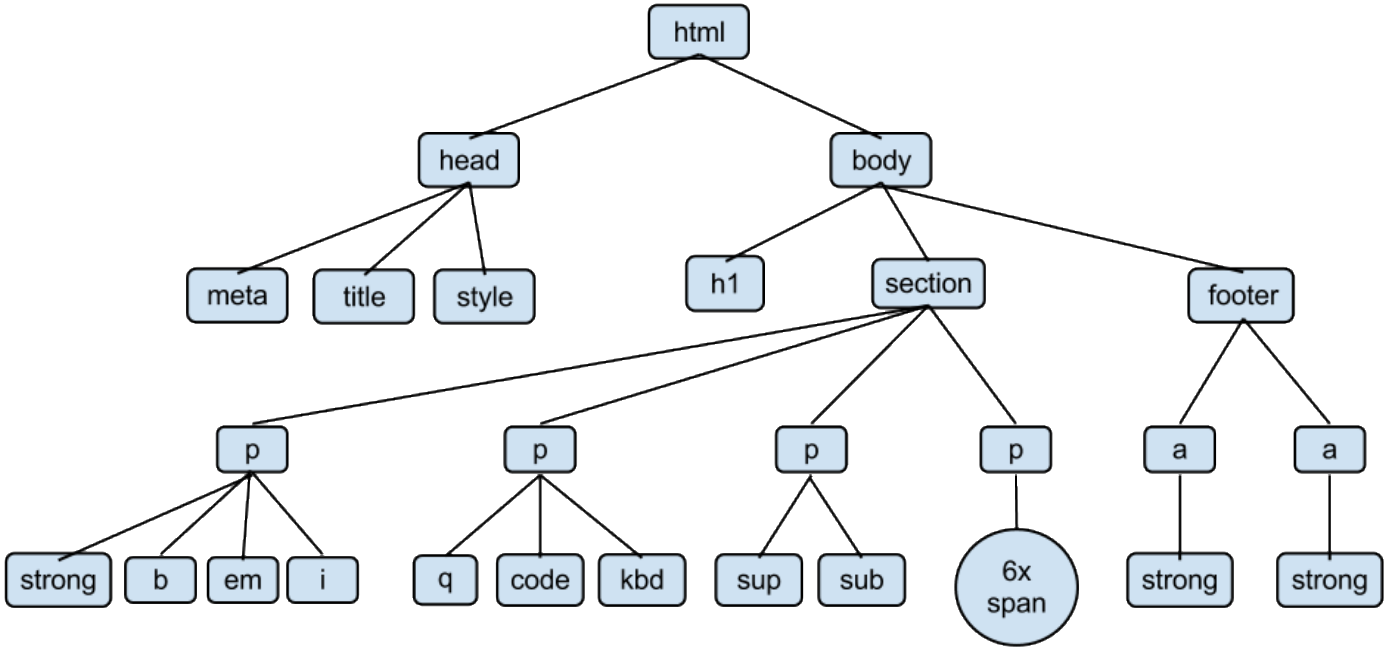


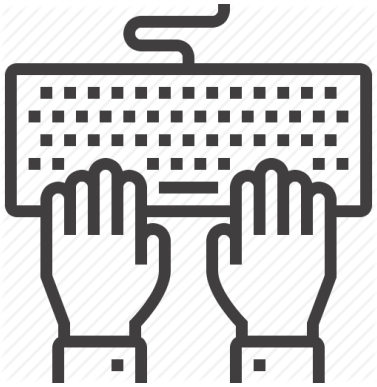
Server

2. Response
- HTML
- image
- JavaScript

Parsing HTML

- The process of analyzing a string of symbols, either in natural language, computer languages or data structures
- Beautiful Soup is a Python package for parsing HTML and XML documents.



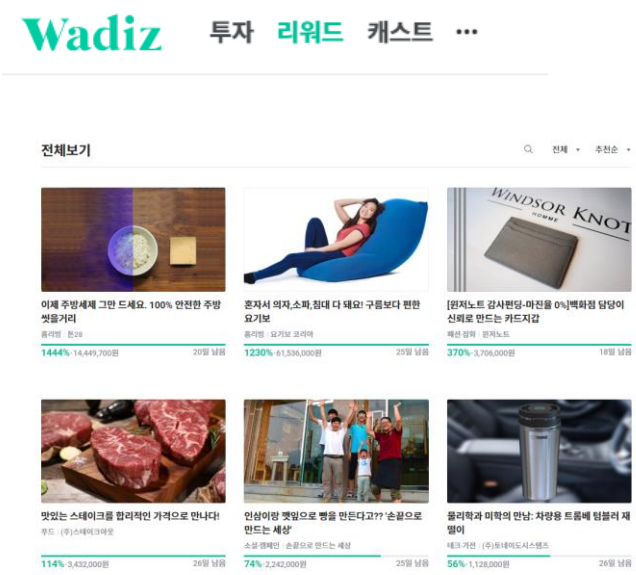


Naver Movie Scraper

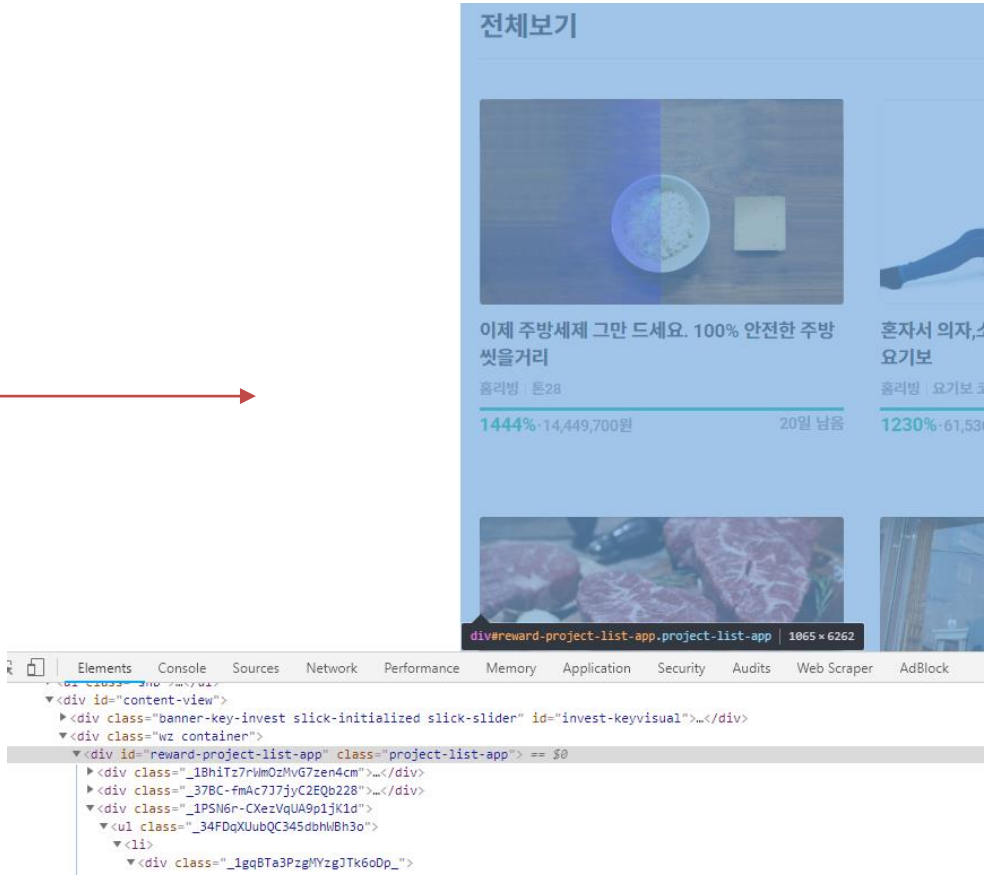
M3.3 Web Scraping_Naver.ipynb

Scraping with Python + Web Module

Sometimes, Information is not Visible at a Glance



https://www.wadiz.kr/web/wreward/main



Combining Python Scraper with Web Modules

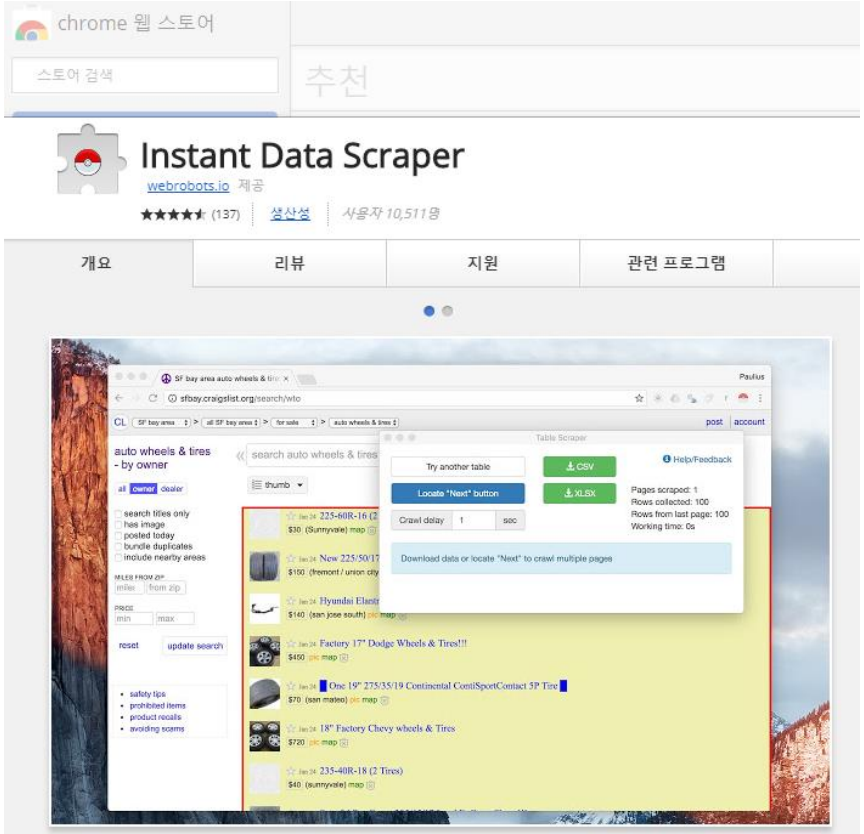
- First, generate the list of URLs to download

- Second, download the webpage HTML from the URLs

These steps can be easily executed using Web modules

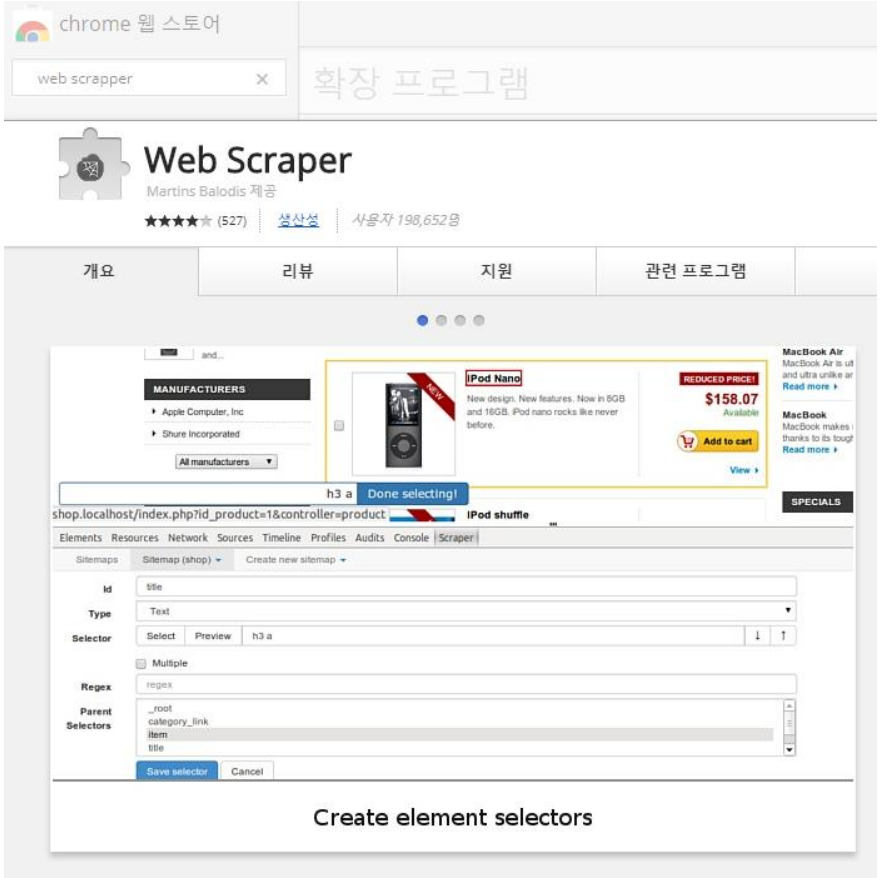
- Third, parse the relevant information from the downloaded HTML

1) Instant Data Scraper (Chrome Extension)



https://chrome.google.com/webstore/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah

2) Web Scraper (Chrome Extension)

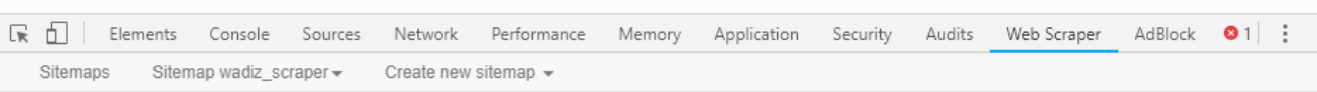


https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn

2) Web Scraper (Chrome Extension)

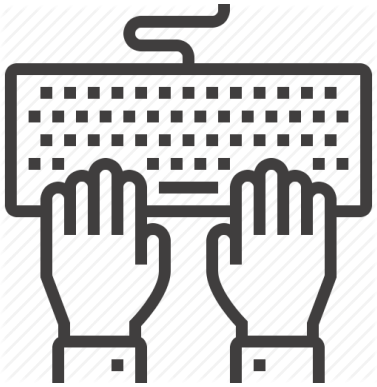
- Press F12 to execute the Web Scrapper

- Design your selector graph



➢ (Example) Import the sitemap JSON

{"_id":"wadiz_scraper","startUrl":["https://www.wadiz.kr/web/wreward/main"],"selectors":[{"id":"project_element","type":"SelectorElementClick","selector":"div._1gqBTa3PzgMYzgJTk6oDp_","parentSelectors":["_root"],"multiple":true,"delay":0,"clickElementSelector":"button._3U3i7gSPry8Nq5-cHL_8Gk","clickType":"clickMore","discardInitialElements":false,"clickElementUniquenessType":"uniqueText"},{"id":"project_link","type":"SelectorLink","selector":"div._1ExmN6wQskSC6MqnxiF7BD > a","parentSelectors":["project_element"],"multiple":false,"delay":0},{"id":"performance_html","type":"SelectorHTML","selector":"div.wd-ui-sub-opener-info div.state-box","parentSelectors":["project_link"],"multiple":false,"regex":"","delay":0},{"id":"profile_html","type":"SelectorHTML","selector":"div.maker-info","parentSelectors":["project_link"],"multiple":false,"regex":"","delay":0}]}



Wadiz Scraper

M3.3 Web Scraping_Wadiz.ipynb

End of Document