

KAIST Summer Session 2018

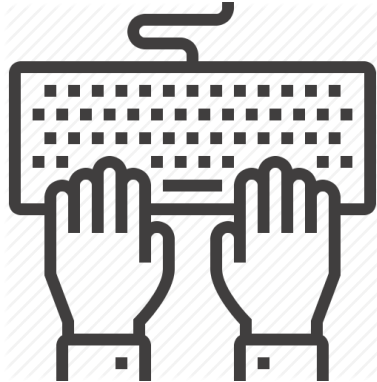
Module 3. Deep Learning with PyTorch

Web Data Extraction

KAIST College of Business

Jiyong Park

13 August, 2018



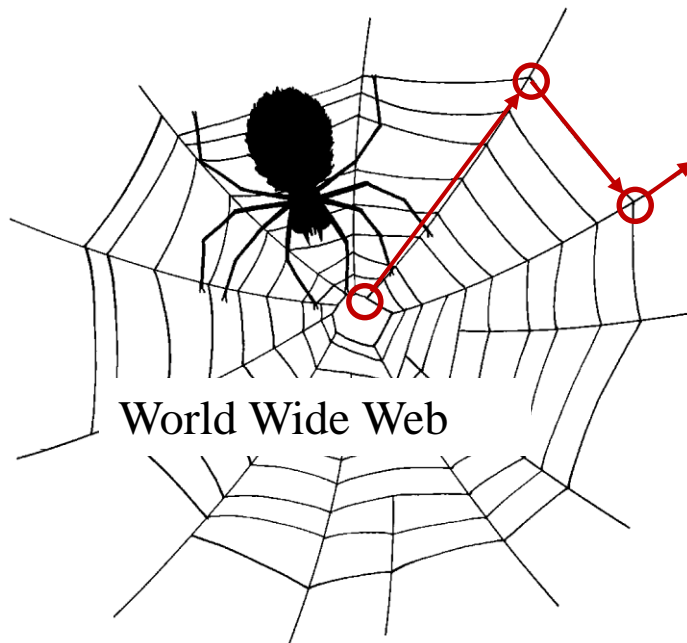
Python Tutorial in 10 minutes

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.



World Wide Web



○ webpage
→ hyperlink

Crawling versus Scraping

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

아일라 상영중

Ayla: The Daughter of War, 2017

관람객 ? ★★★★★ 9.08 기자·평론가 ★★★★★ 5.25

네티즌 ? ★★★★★ 9.55 내 평점 ★★★★★ 등록 >

개요 드라마, 전쟁 | 한국, 터키 | 123분 | 2018.06.21 개봉

감독 잔 울카이

출연 김설(아일라)

등급 [국내] 15세

당갈 상영중

Dangal, 2016

관람객 ? ★★★★★ 9.61 기자·평론가 ★★★★★ 7.00

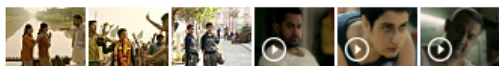
네티즌 ? ★★★★★ 9.55 내 평점 ★★★★★ 등록 >

개요 드라마, 액션 | 인도 | 161분 | 2018.04.25 개봉

감독 니테쉬 티와리

출연 아미르 칸(마하비르 싱 포갓), 파티마 사나 세이크(기타), 산야 말... [더보기](#) >

등급 [국내] 12세 관람가



덕구

Stand by me, 2017

관람객 ? ★★★★★ 9.29 기자·평론가 ★★★★★ 5.50

네티즌 ? ★★★★★ 9.47 내 평점 ★★★★★ 등록 >

개요 드라마 | 한국 | 91분 | 2018.04.05 개봉

감독 방수인

출연 이순재(덕구)

등급 [국내] 전체

원더 상영중

Wonder, 2017

관람객 ? ★★★★★ 9.43 기자·평론가 ★★★★★ 6.86

네티즌 ? ★★★★★ 9.42 내 평점 ★★★★★ 등록 >

개요 드라마 | 미국 | 113분 | 2017.12.27 개봉

감독 스티븐 크보스키

출연 제아콥 트럼블레이(어기 폴먼), 줄리아 로버츠(이자벨 폴먼), 오... [더보기](#) >

등급 [국내] 전체 관람가 [해외] PG ?



Let's Scrap the Naver Movie Information

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

순위	영화명	평점	변동폭
1	아일라	★★★★★ 9.55 평점주기	- 0
2	당갈	★★★★★ 9.55 평점주기	- 0
3	덕구	★★★★★ 9.47 평점주기	- 0
4	원더	★★★★★ 9.42 평점주기	- 0
5	쇼생크 탈출	★★★★★ 9.41 평점주기	- 0
6	허스토리	★★★★★ 9.40 평점주기	- 0
7	프린스 앤 프린세스	★★★★★ 9.40 평점주기	- 0
8	터미네이터 2	★★★★★ 9.40 평점주기	- 0
9	인생은 아름다워	★★★★★ 9.39 평점주기	- 0
10	매트릭스	★★★★★ 9.39 평점주기	- 0

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

아일라 상영중

Ayla: The Daughter of War, 2017

관람객? ★★★★★ 9.08 기자·평론가 ★★★★★ 5.25

네타즌? ★★★★★ 9.55 내 평점 ★★★★★ 등록>

개요 드라마, 전쟁 | 한국, 터키 | 123분 | 2018.06.21 개봉

감독 장준하

출연 김설(아일라), 이스마일 하지오글루(슬레이만) 더보기>

등급 [국내] 15세 관람가

예매하기 다운로드 326



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

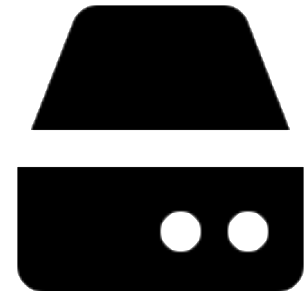
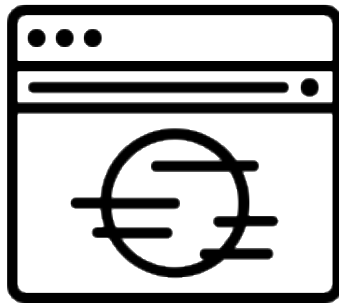
<https://movie.naver.com/movie/bi/mi/basic.nhn?code=169240>
<https://movie.naver.com/movie/bi/mi/basic.nhn?code=157243>
<https://movie.naver.com/movie/bi/mi/basic.nhn?code=154667>
<https://movie.naver.com/movie/bi/mi/basic.nhn?code=151196>

- 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)

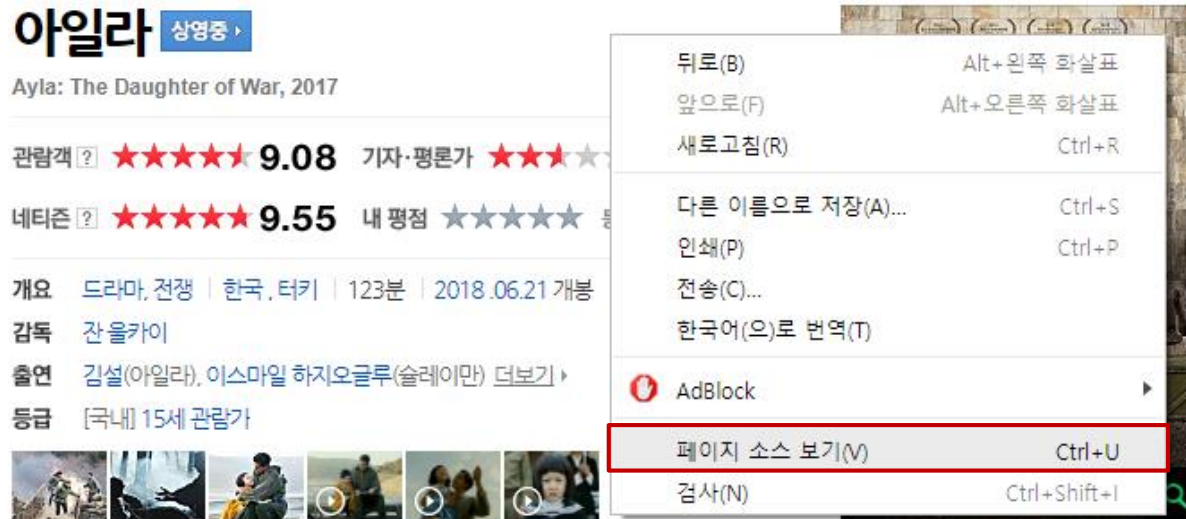


2. Response

- 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>7</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 인 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>

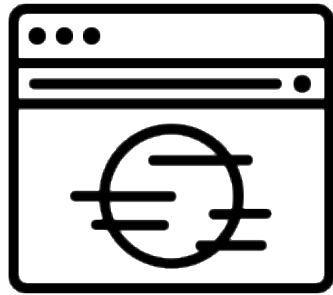
Let them (or us) be free!

Option 2:

Let computers do

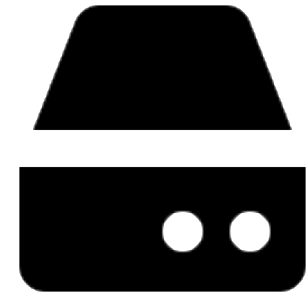


Web Browsers are not Necessarily Needed for Requests



1. Request

- URL
- header
- data



Server

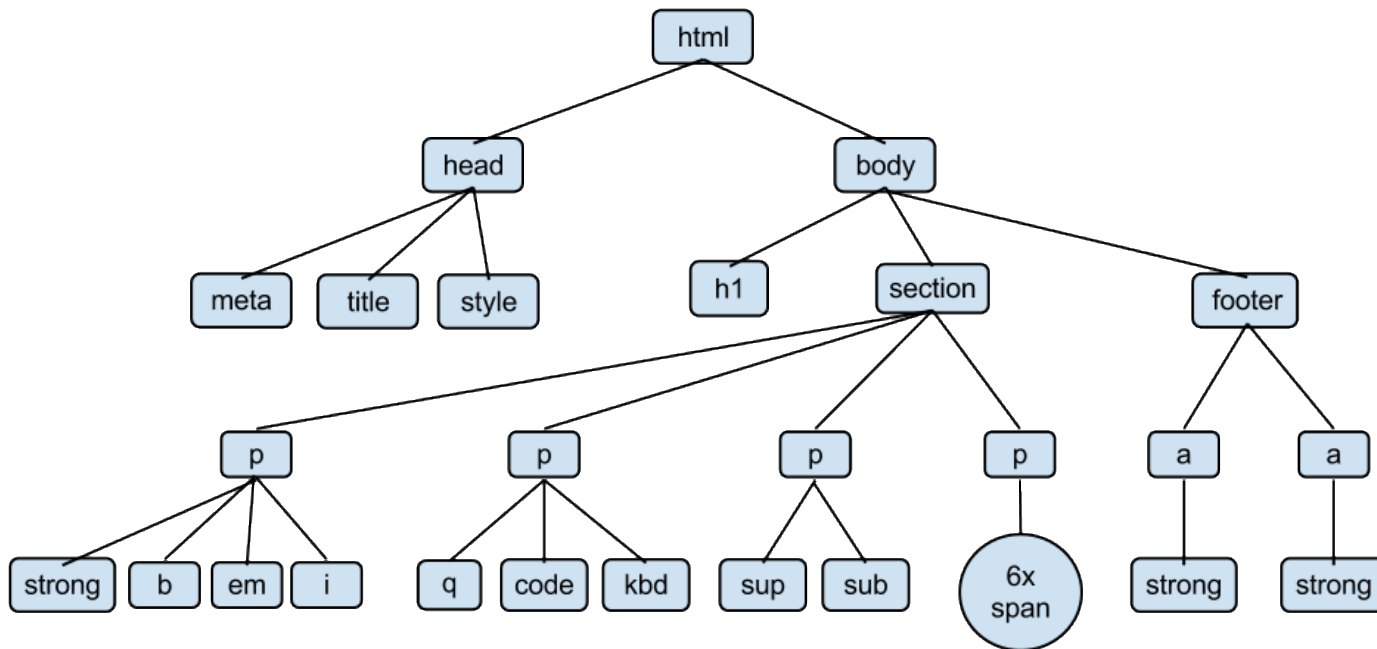


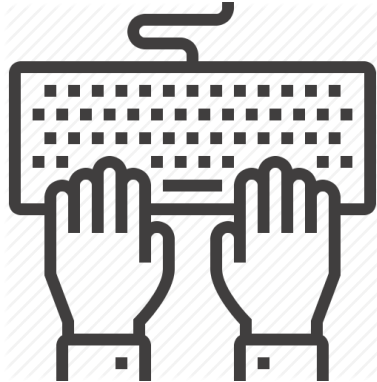
2. Response

- HTML
- image
- JavaScript

Parsing HTML

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





Naver Movie Scraper

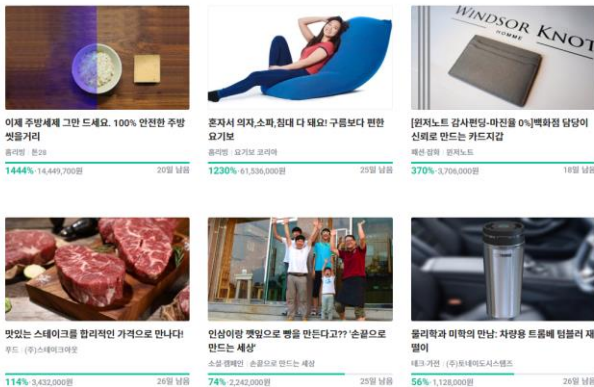
M3.3 Web Scraping_Naver.ipynb

Scraping with Python + Web Module

Sometimes, Information is not Visible at a Glance

Wadiz 투자 리워드 캐스트 ...

전체보기



이제 주방세제 그만 드세요. 100% 안전한 주방
씻을거리
올리빙 톤28
1444% 14,449,700원 20일 남음

혼자서 의자, 소파, 침대 다 돼요! 구름보다 편한
요기보
올리빙 요기보 크리미
1230% 61,536,000원 25일 남음

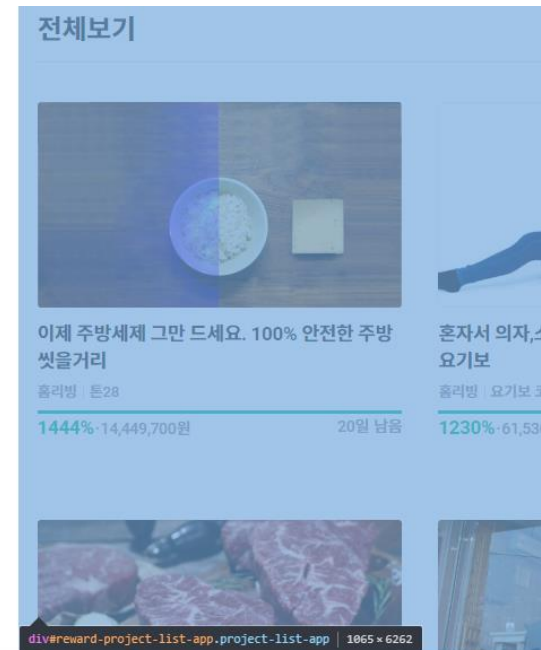
[원저노트 감사편당·마진율 0%]백화점 담당이
신뢰로 만드는 카드지갑
웨산 영화 원저노트
370% 3,706,000원 16일 남음

맛있는 스테이크를 합리적인 가격으로 만나자!
후드 (후)스테이크야로
114% 3,432,000원 29일 남음

인생이란 맛있는 밥을 만든다고?? 손끝으로
만드는 세상
소문 캠페인, 손끝으로 만드는 세상
74% 2,242,000원 25일 남음

올리학과 미학의 만남: 차량용 트롤러 텀블러 재
벌어
네크-거친 (후)올리학과미학트롤러
56% 1,128,000원 26일 남음

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



전체보기

이제 주방세제 그만 드세요. 100% 안전한 주방
씻을거리
올리빙 톤28
1444% 14,449,700원 20일 남음

```
Elements Console Sources Network Performance Memory Application Security Audits Web Scraper AdBlock


-


```

Downloaded HTML

```
<div class="wz container">
  <div id="reward-project-list-app" class="project-list-app"></div>
</div>
```

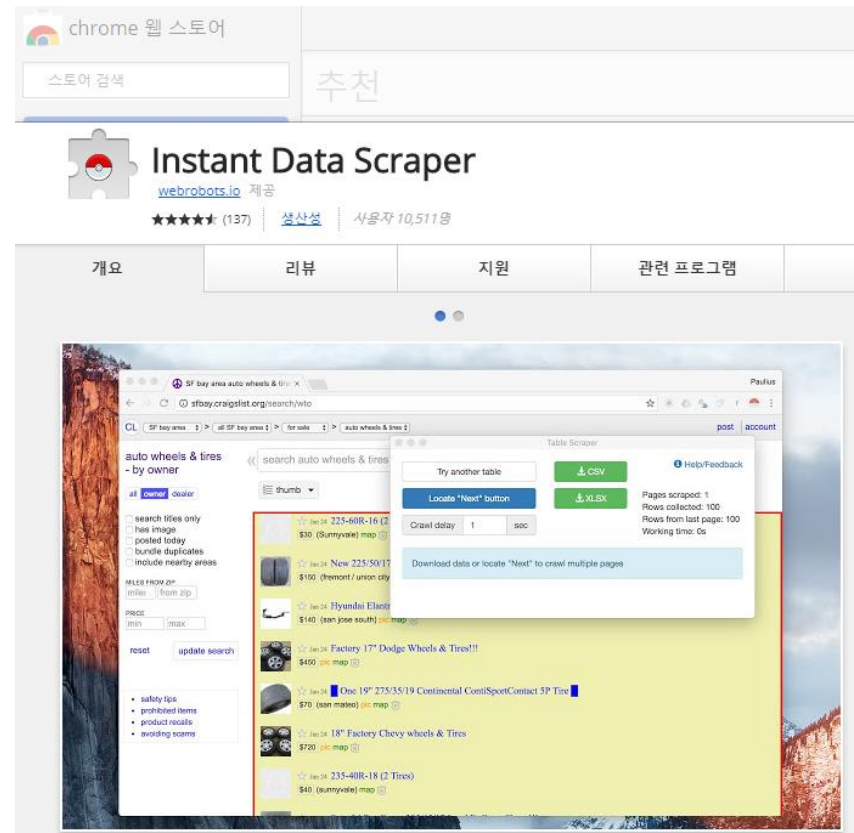
Combining Python Scraper with Web Modules

- 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

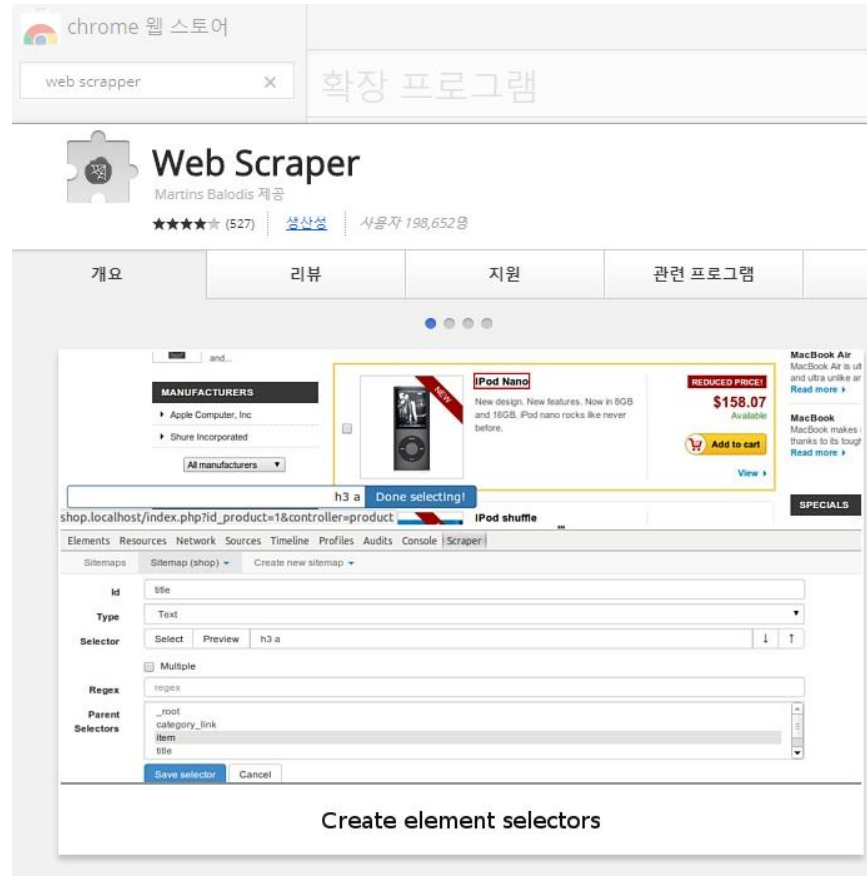
- Third, parse the relevant information from the downloaded HTML

1) Instant Data Scraper (Chrome Extension)



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

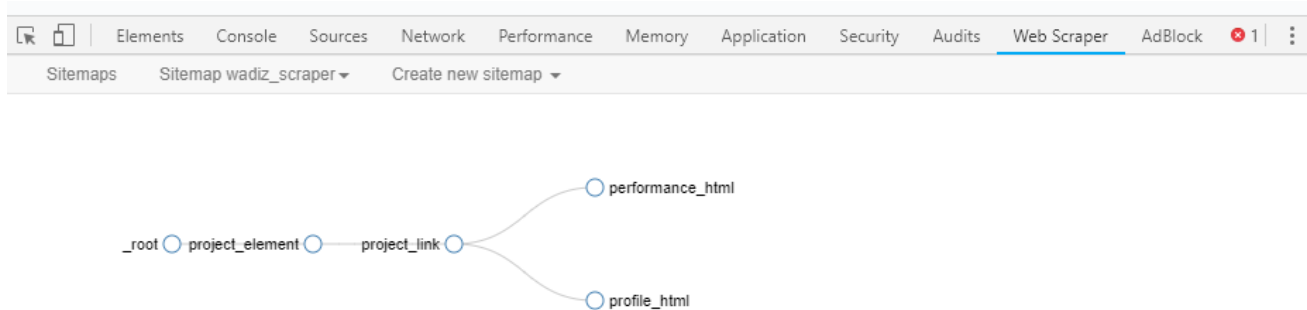
2) Web Scraper (Chrome Extension)



<https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehlklipmbmhn>

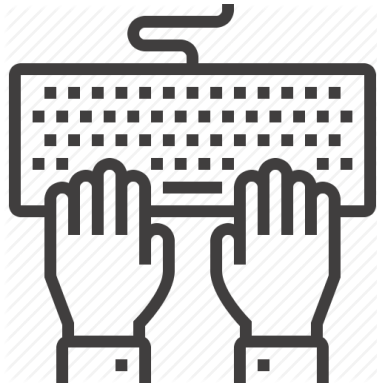
2) Web Scraper (Chrome Extension)

- Press F12 to execute the Web Scraper
- 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._1ExmN6wQskSC6MqnxiF7BDa", "parentSelectors": ["project_element"], "multiple": false, "delay": 0 }, { "id": "performance_html", "type": "SelectorHTML", "selector": "div.wd-ui-sub-opener-info", "parentSelectors": ["project_link"], "multiple": false, "delay": 0 }, { "id": "profile_html", "type": "SelectorHTML", "selector": "div.state-box", "parentSelectors": ["project_link"], "multiple": false, "delay": 0 }, { "id": "div_maker_info", "type": "SelectorHTML", "selector": "div.state-box", "parentSelectors": ["project_link"], "multiple": false, "delay": 0 } ] }
```

Wadiz Scraper

M3.3 Web Scraping_Wadiz.ipynb

End of Document