

웹 크롤링 좀 더 잘하기

College	University of seoil
Major	Computer electronic
Name	Wangwon Lee

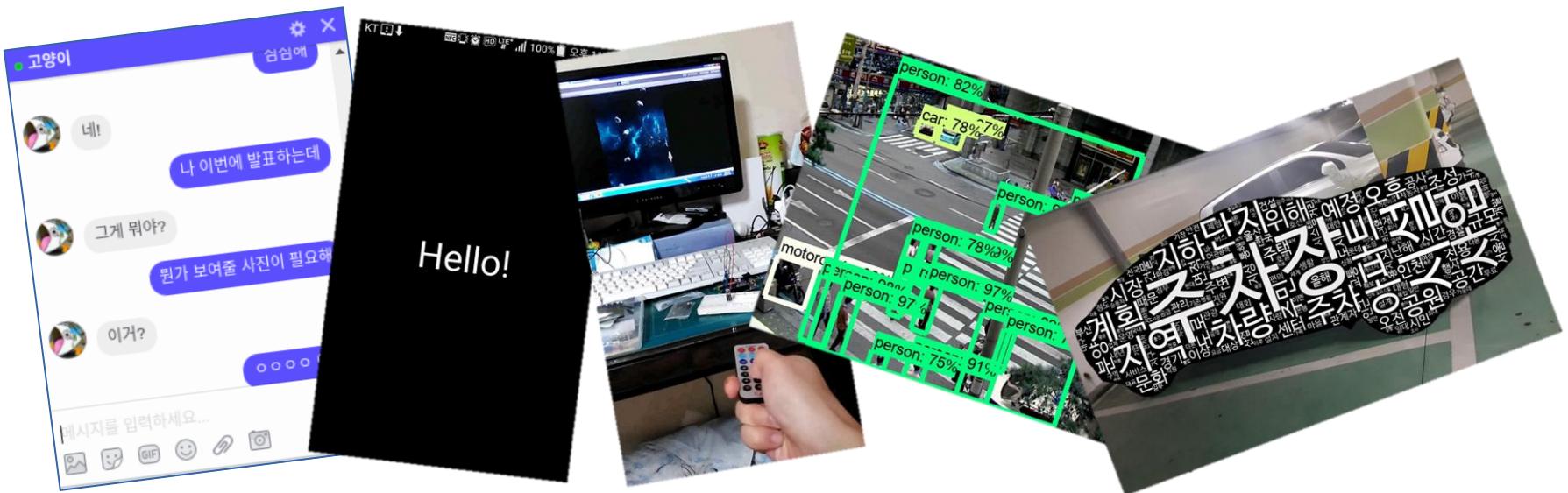
Contents

- 발표자 소개
- 이런 이야기를 하고 싶어요!
- 웹 크롤링 개요
- 기본적인 웹 크롤링
- 동적 페이지 개요
- 동적 페이지 크롤링
- 마무리

발표자 소개 - 이왕원



- 전자부품연구원 인공지능연구센터 위촉연구원
- 서일대학교 소프트웨어 공학과 재학 중
- 실시간 데이터 수집 프로젝트인 KoShort 컨트리뷰터
- 아직 많이 부족합니다. 여러분께 많이 배우고 싶어요!



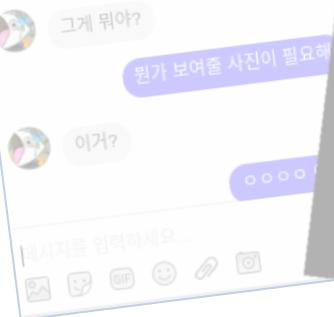
2018 데이터야 놀자

발표자 소개 - 이왕원

- 전자부품연구원 인공지능연구센터 위촉연구원
- 서일대학교 소프트웨어 공학과 재학 중
- 실시간 데이터 수집 프로젝트인 KoShort 컨트리뷰터

웹 크롤링은 자료 찾기가 너무 힘들어서
제가 먼저 공유하고자 이렇게 발표하게 되었어요.

이걸 계기로 웹크롤링도 많은 정보 공유가 있었으면 해요. 😊



Hello!

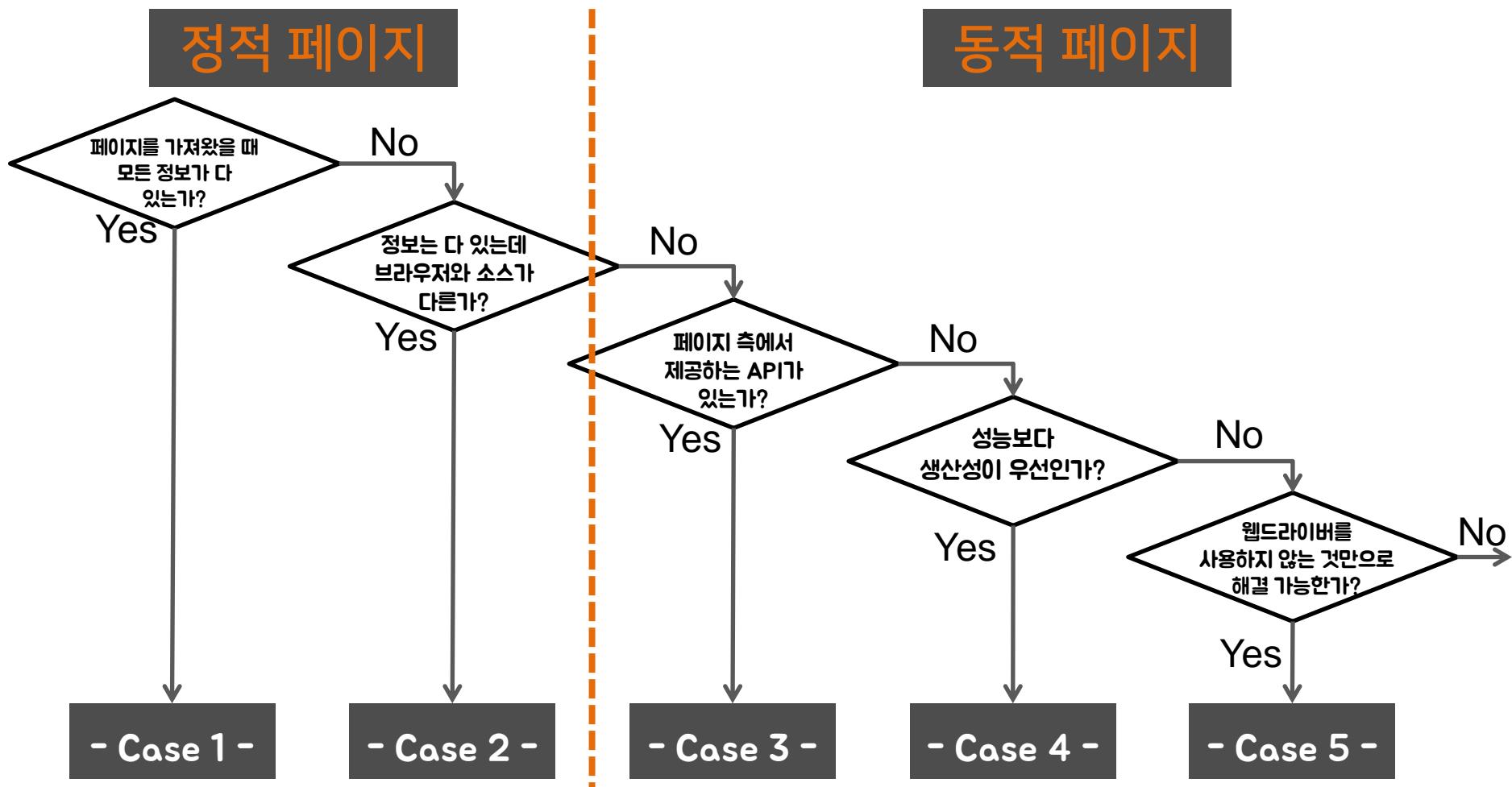


이런 이야기를 하고 싶어요!

- 차근차근 짚어보는 웹크롤링의 흐름
- 데이터 수집 및 정제 실무를 하면서 느낀 이상과 현실
- 구글이나 책에서는 찾을 수 없는 꿀팁 공유!
- ~~실시간 데이터 수집 오픈소스 Koshort 홍보~~

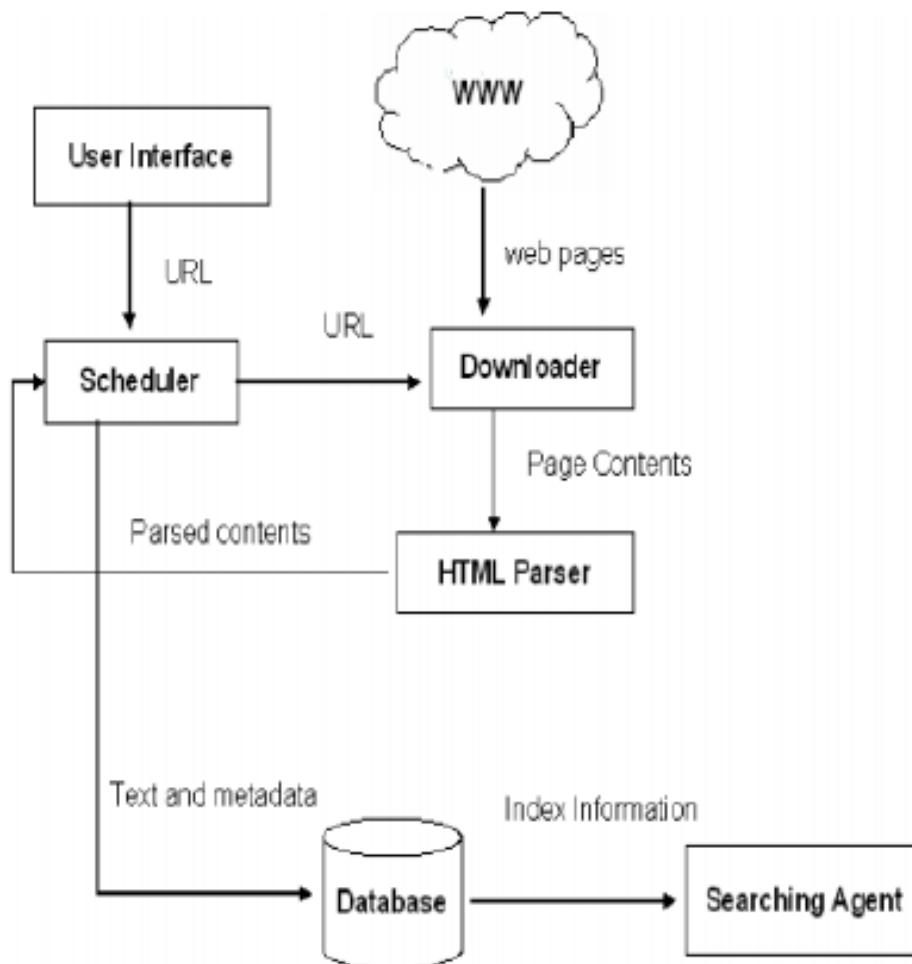
이런 이야기를 하고 싶어요!

오늘은 이 흐름을 토대로 웹 크롤링을 차근차근 짚어볼거에요



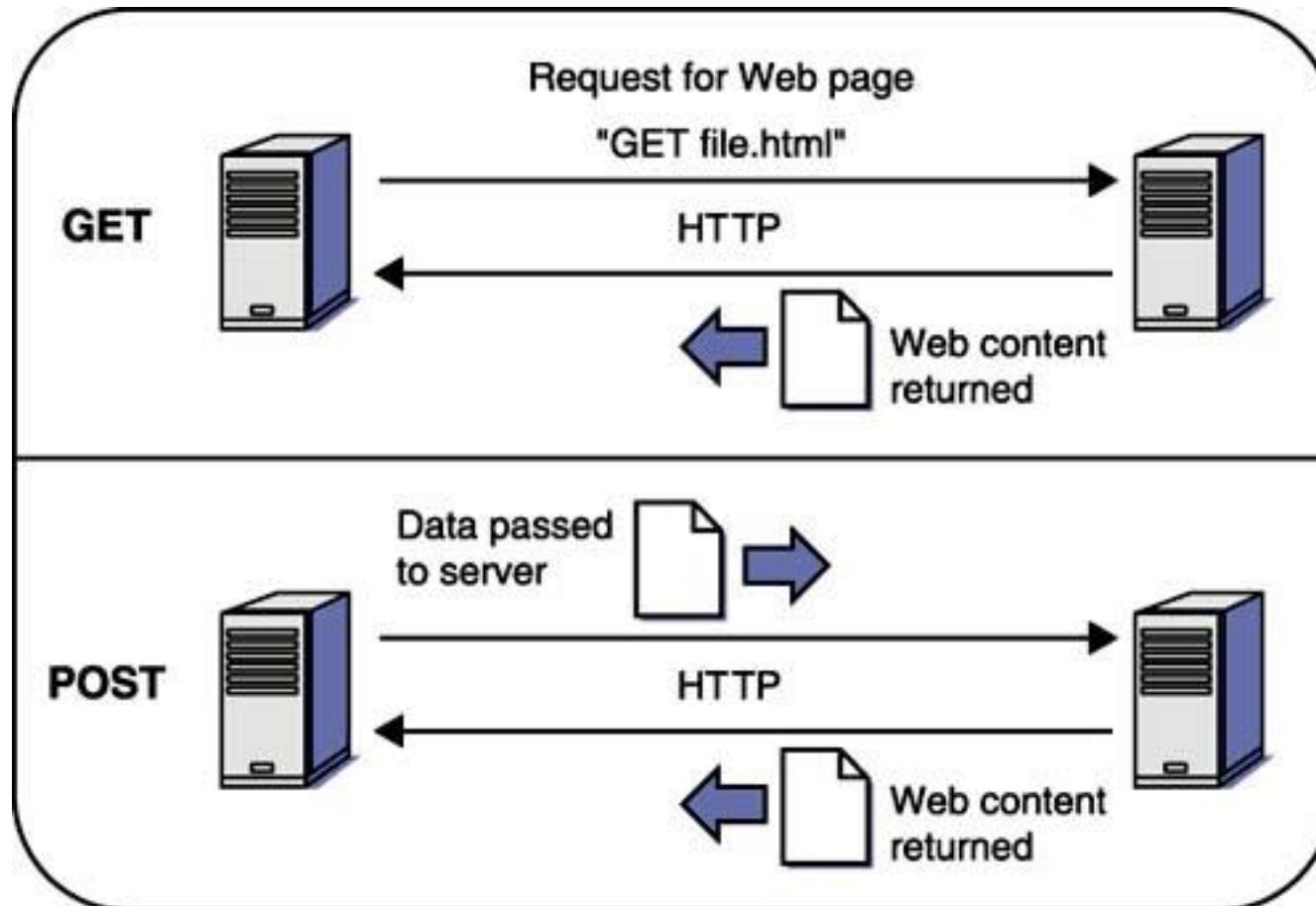
웹 크롤링 개요

흔한 웹 크롤러의 흐름.jpg



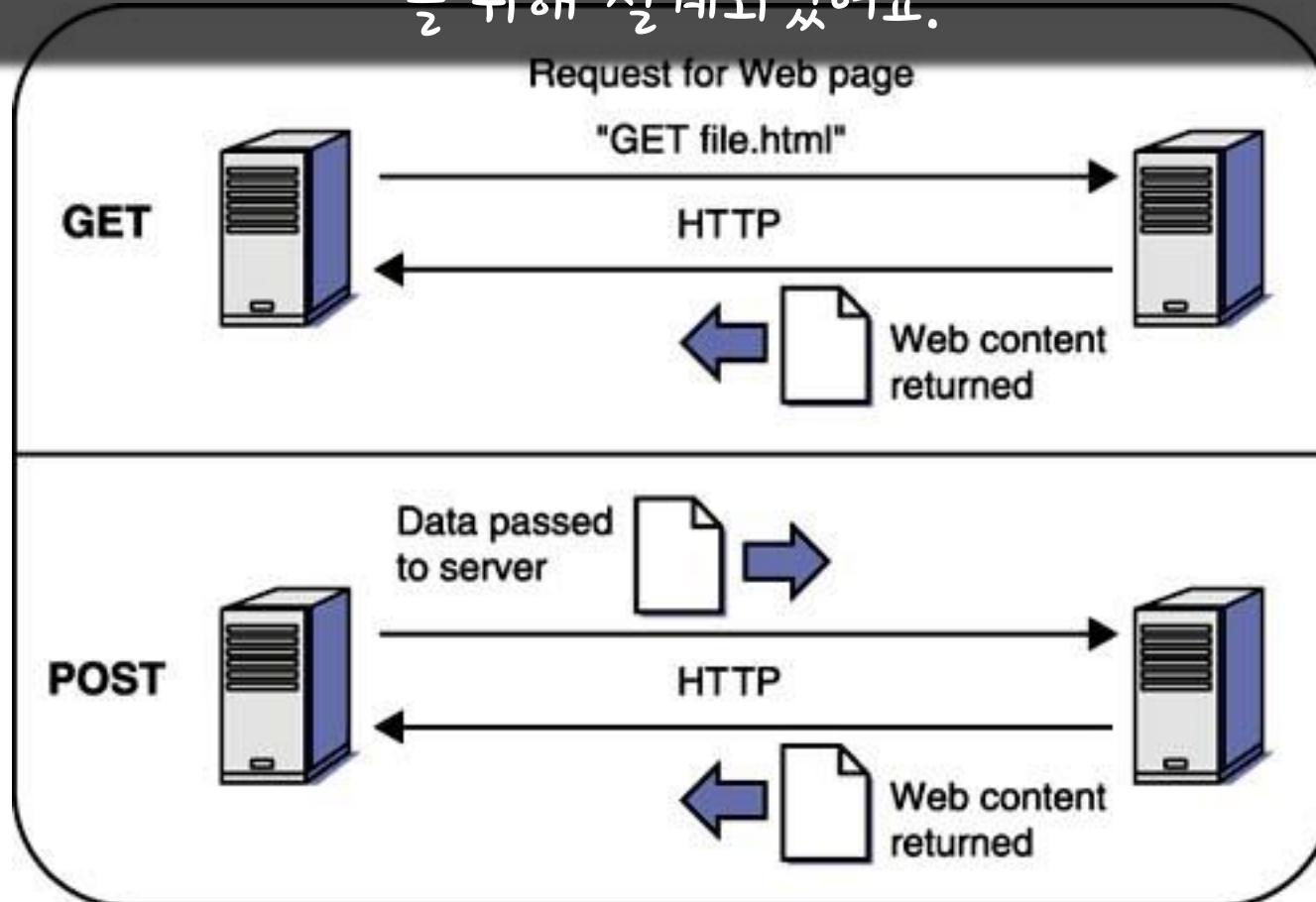
웹 크롤링 개요

HTTP에는, GET과 POST와 같은 메소드가 있는데



웹 크롤링 개요

GET은 가져오는 동작, POST는 수행하는 동작
을 위해 설계되었어요.



웹 크롤링 개요

우리가 쓰는 웹 브라우저라는 친구들은



웹 크롤링 개요

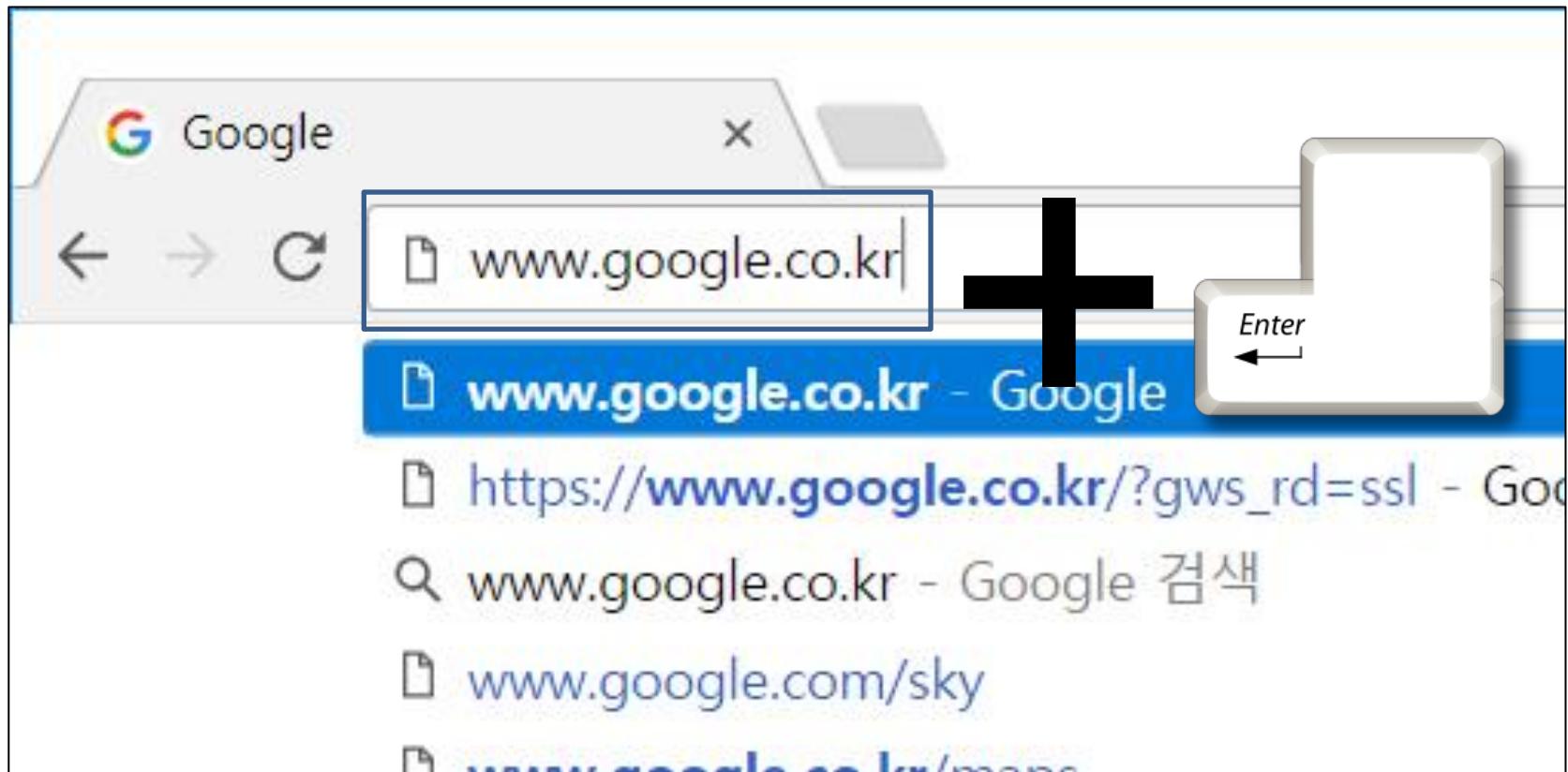
HTML 소스를 해석해서 웹 페이지를 그려주는 역할을 해요

The screenshot shows a browser window with the Google homepage loaded. A large black box labeled "표시된 화면" (Displayed Screen) covers the main content area. To the right, the browser's developer tools are open, specifically the "Elements" tab under the "Developer" menu. The "HTML" section of the elements panel is selected, displaying the full HTML source code of the Google page. The code includes meta tags, links, scripts, and various styles. At the bottom of the developer tools, there is a toolbar with buttons for "Styles", "Event Listeners", "DOM Breakpoints", and "Properties".

```
<html itemscope itemtype="http://schema.org/WebPage" lang="ko"> == $0
  <head>
    <meta content="/images/branding/googleg/1x/googleg_standard_color_128dp.png" itemprop="image">
    <link href="/images/branding/product/ico/googleg_lodp.ico" rel="shortcut icon">
    <meta content="origin" id="mref" name="referrer">
    <title>Google</title>
    <script src="https://apis.google.com/_scs/abc-static/_/js/k=gapi.gapi.en.OtWZA_d=1/ed=1/am=AAg/rse=AHpOoo-x_5rAkhg6nsZl4dxJHq9g08k6GA/cb=gapi.loaded_0" async></script>
  </head>
  <body>
    <div>
      <meta content="https://www.google.co.kr/?gws_rd=ssl" itemprop="og:url" rel="canonical">
      <link href="https://www.google.co.kr/" rel="canonical">
      <script>window.location.replace('https://www.google.co.kr/?gws_rd=ssl');</script>
      <div>
        <div>
          <div>
            <div>
              <div>
                <div>
                  <div>
                    <div>
                      <div>
                        <div>
                          <div>
                            <div>
                              <div>
                                <div>
                                  <div>
                                    <div>
                                      <div>
                                        <div>
                                          <div>
                                            <div>
                                              <div>
                                                <div>
                                                  <div>
                                                    <div>
                                                      <div>
                                                        <div>
                                                          <div>
                                                            <div>
                                                              <div>
                                                                <div>
                                                                  <div>
                                                                    <div>
                                                                      <div>
                                                                        <div>
                                                                          <div>
                                                                            <div>
                                                                              <div>
                                                                                <div>
                                                                                  <div>
                                                                                    <div>
                                                                                      <div>
                                                                                        <div>
                                                                                          <div>
                                                                                            <div>
                                                                                              <div>
                                                                                                <div>
                                                                                                  <div>
                                                                                                    <div>
                                                                                                      <div>
                                                                                                        <div>
                                                                                                          <div>
                                                                                                            <div>
                                                                                                              <div>
                                                                                                                <div>
                                                                                                                  <div>
                                                                                                                    <div>
                                                                                                                      <div>
                                                                                                                        <div>
                                                                                                                          <div>
                                                                                                                            <div>
                                                                                                                              <div>
                                                                                                                                <div>
                                                                                                                                  <div>
                                                                                                                                    <div>
                                                                                                                                      <div>
                                                                                                                                        <div>
                                                                                                                                          <div>
                                                                                                                                            <div>
                                                                                                                                              <div>
                                                                                                                                                <div>
                                                                                                  <div>
                                                                                                    <div>
                                                                                                      <div>
                                                                                                        <div>
                                                                                                          <div>
                                                                                                            <div>
                                                                                                              <div>
                                                                                                                <div>
                                                                                                                  <div>
                                                                                                                    <div>
                                                                                                                      <div>
                                                                                                                        <div>
                                                                                                                          <div>
                                                                                                                            <div>
                                                                                                                              <div>
                                                                                                                                <div>
                                                                                                                                  <div>
                                                                                                                                    <div>
                                                                                                                                      <div>
                                                                                                                                        <div>
                                                                                                                                          <div>
                                                                                                                                            <div>
................................................................
```

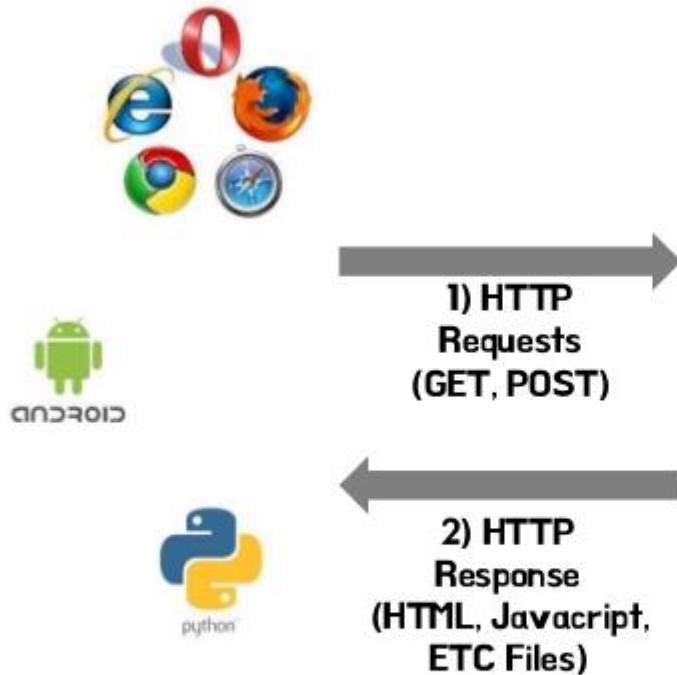
웹 크롤링 개요

따라서 브라우저 주소창에 주소를 입력하고 엔터를 치는 것은



웹 크롤링 개요

주소에 할당된 웹 서버로 GET 요청을 보내서
웹 페이지를 가져오는 동작이에요



```
1 <!DOCTYPE html PUBLIC "-//W3C//DTD HTML
2 <html>
3   <head>
4     <title>Example</title>
5     <link href="screen.css" rel="sty
6   </head>
7   <body>
8     <h1>
9       <a href="/">Header</a>
10    </h1>
11    <ul id="nav">
12      <li>
13        <a href="#">
14        </a>
15      <li>
16        <a href="#">
17        </a>

```

Web Server

```
1 function Greeter(greeting
2   this.greeting = greet
3 }
4 Greeter.prototype.greet =
5   return "Hello," + thi
6 }
7 var greeter = new Greeter({message: "world"});
8 var button = document.createElement("button");
9 button.innerText = "Say Hello"
10 button.onclick = function() {
11   alert(greeter.greet())
12 }
13 document.body.appendChild(button)
14

```

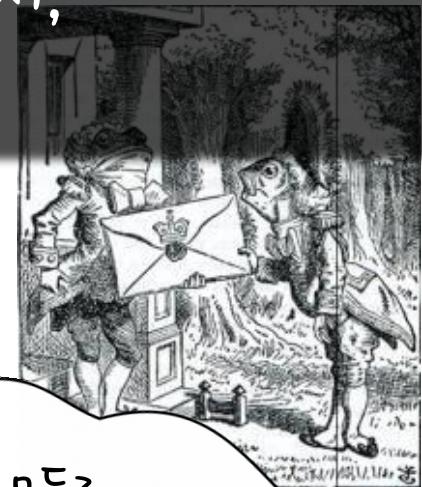
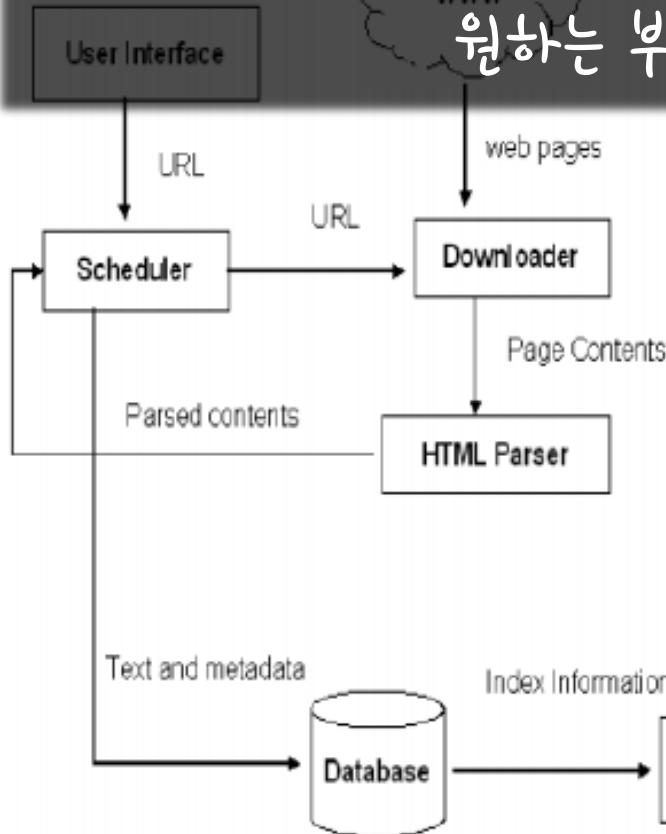
Web Crawler

웹 크롤링 개요

통상적인 웹 크롤러 또한 같은 원리로,

GET 메소드를 통해 웹 페이지를 가져와서,

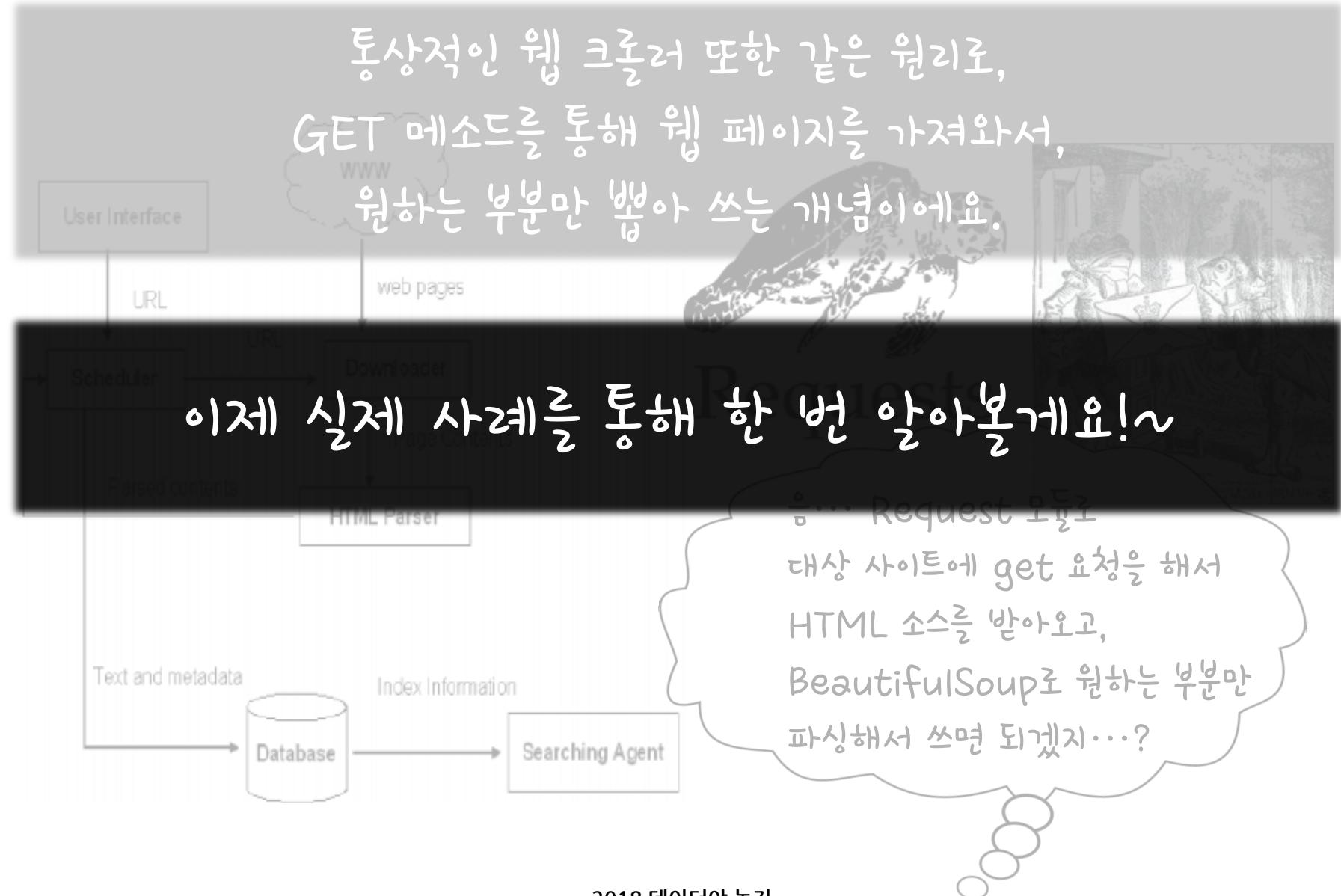
원하는 부분만 뽑아 쓰는 개념이에요.



Requests

음… Request 모듈로
대상 사이트에 get 요청을 해서
HTML 소스를 받아오고,
BeautifulSoup로 원하는 부분만
파싱해서 쓰면 되겠지…?

웹 크롤링 개요



기본적인 웹 크롤링

- Case 1 -

페이지를 받아오면 정보가 그대로 있는 경우

The screenshot shows a Microsoft Edge browser window displaying the Naver homepage (<https://www.naver.com>). The developer tools sidebar is open, showing the Element tab with the following code snippet:

```
<div class="area_hotkeyword" role="listbox">
  <div class="ah_roll" PM_CL_realtimeKeyword_rolling_base="C1539612535076591488" aria-hidden="true">
    <h3 class="blind">급상승 검색어</h3>
    <div class="ah_roll_area PM_CL_realtimeKeyword_rolling">
      queryId="C1539612535076591488"
      <ul class="ah_l" queryId="C1539612885115404932">
        <li class="ah_item">박지원</li>
        <li class="ah_item">멸달</li>
        <li class="ah_item">돌아와요 부산항해</li>
        <li class="ah_item">가수김세환나이</li>
        <li class="ah_item">문숙</li>
        <li class="ah_item">동덕여대 알몸남</li>
        <li class="ah_item">탁석산</li>
        <li class="ah_item">제이쓴</li>
        <li class="ah_item">현철</li>
        <li class="ah_item">사마천</li>
      </ul>
    </div>
  </div>
</div>
```

The main content area of the browser shows the Naver homepage with various news sections, advertisements, and a sidebar for "급상승 검색어" (Rising Search Terms) which lists the top 20 keywords from the provided code.

네이버 인기검색어

기본적인 웹 크롤링

- Case 1 -

페이지를 받아오면 정보가 그대로 있는 경우

```
soup.find_all("a", attrs={"class": "sister"})  
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,  
# <a class="sister" href="http://example.com/latie" id="link2">Lacie</a>,  
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]
```

HTML 태그와 그에 따른 속성(class, id 등)을 토대로
필요한 정보만 파싱해서 추출.

```
soup.select("#link1")  
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>]  
  
soup.select("a#link2")  
# [<a class="sister" href="http://example.com/latie" id="link2">Lacie</a>]
```

네이버 인기검색어

기본적인 웹 크롤링

- Case 2 -

정보는 다 있는데, 브라우저 상에서 내가 받은 소스가 다른 경우

The screenshot shows a browser window displaying the YouTube trending page at <https://www.youtube.com/feed/trending>. The page lists several video thumbnails and titles. To the right, the browser's developer tools are open, specifically the Elements tab, which shows the HTML structure of one of the video cards. A yellow box highlights the title '아는 형님(Knowing bros) 150회 예고편' and its details. Another yellow box highlights the title '친친모 시즌2] 교실을 벗어나라! 단체 게임 방 탈출 놀이' and its details. A third yellow box highlights the title '백종원에 혼쭐난 자한당 의원님! 의원님! 그건 너무 하신 거 아닙니까!' and its details. The developer tools' element inspector shows the DOM structure for these titles, including class names like 'yt-dash-card', 'yt-dash-card__content', and 'yt-dash-card__text'. The properties panel on the right shows styles for the highlighted elements, such as 'color: #e64a19;', 'font-size: 1.8rem;', and 'font-weight: 400;'. The bottom of the image features a large blue rounded rectangle containing the text '유튜브 실시간 인기 동영상'.

유튜브 실시간 인기 동영상

2018 데이터야 놀자

기본적인 웹 크롤링

- Case 2 -

정보는 다 있는데, 브라우저 상에서와 내가 받은 소스가 다른 경우

The screenshot shows a browser window with the URL <https://www.youtube.com/feed/trending>. The developer tools are open, specifically the Elements tab, which displays the HTML structure of a video thumbnail. The thumbnail for '아는 형님(Knowing bros) 150회 예고편' is selected, and its style properties are visible in the right panel. The main content area shows several video thumbnails, including one for '방탄소년단 네덜란드 콘서트 밖 충격적인 현장'.

당황하지 말고, user-agent를 변경해보자!
(보통 'i-explorer'가 디폴트인 경우가 많음)

유튜브 실시간 인기 동영상

2018 데이터야 놀자

기본적인 웹 크롤링

- Case ? -

내가 받은 페이지에서는 정보가 일부만 있거나 아예 안보이는 경우

The image shows a web browser window with four tabs open, each displaying a different website:

- Vogue Korea**: A fashion news website. A specific section is highlighted with a blue box and labeled "- Vogue Korea -".
- Google Trends**: A search results page showing trending topics. A specific topic is highlighted with a blue box and labeled "- Google Trend -".
- Reddit**: A social news and discussion platform. A specific post is highlighted with a blue box and labeled "- Reddit -".
- Facebook**: A social media platform. A specific post is highlighted with a blue box and labeled "- Facebook -".

기본적인 웹 크롤링

- Case ? -

내가 받은 페이지에서는 정보가 일부만 있거나 아예 안보이는 경우

The screenshot shows a split-screen view of a web browser. The left half displays the Vogue Korea website, which has a large 'VOGUE' logo at the top and a 'Fashion' section below it. The right half shows a Google Trends search results page for the query 'Flowers'. The results show a single trend entry: '1. 오스카르 로메로·성인·바티칸 시국·교황 요한 바오로 2세·시성...' (Oscar Romero, Pope John Paul II, etc.). The overall context suggests a comparison between what is visible on the page and what might be missing or obscured.

일단 Javascript를 비활성화 시켜보자!

- Vogue Korea -

The screenshot shows a split-screen view of a web browser with three tabs open. The left tab is Reddit, showing a post from r/popular. The middle tab is DataTaya, a Korean news website, displaying a list of news articles. The right tab is Facebook, showing a news feed with various posts and interactions. A large black rectangular box covers the central portion of the screen, obscuring the content of the middle and right tabs. This visual effect serves as a metaphor for how JavaScript can hide or manipulate content on a webpage.

you just have to get an A on.

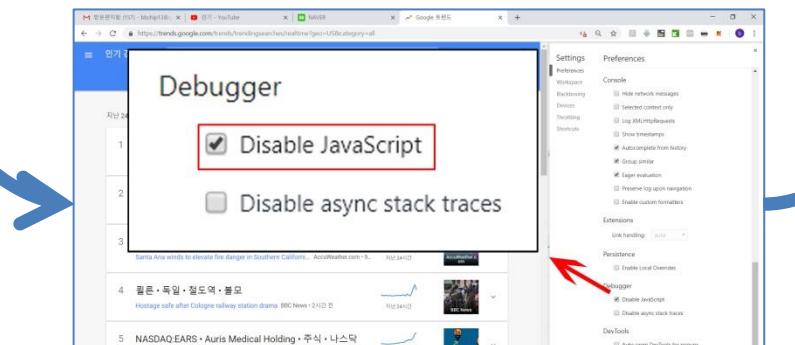
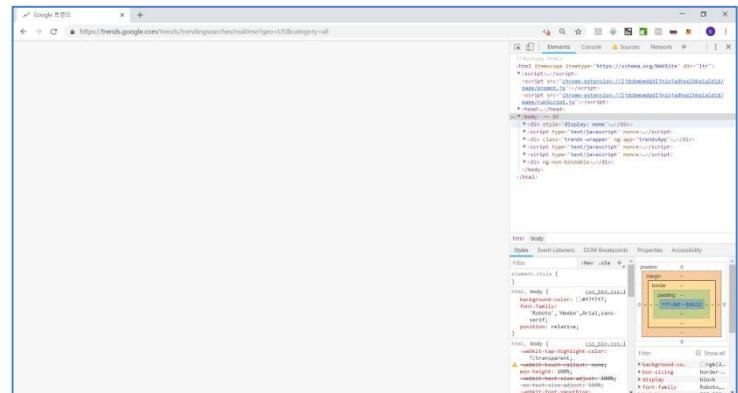
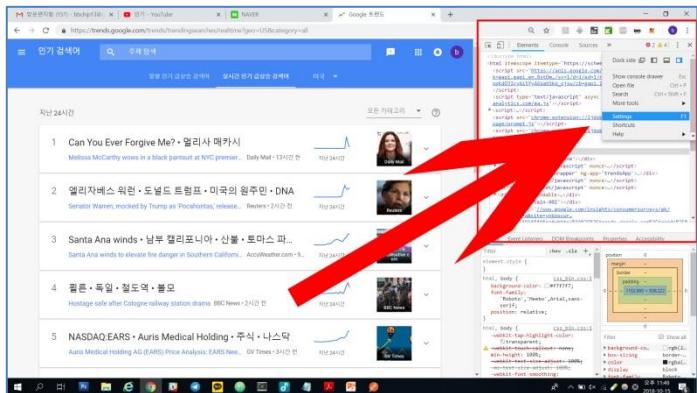
- Reddit -

2018 데이터야 놀자

- Facebook -

기본적인 웹 크롤링

만약 크롬 개발자 도구에서 자바스크립트를 비활성화 했을 때
웹브라우저 상에서도 동일한 증상이 보인다면?

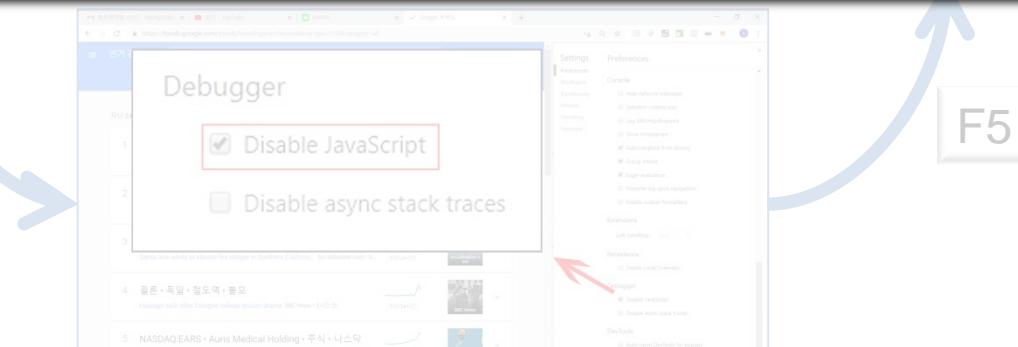


구글 트렌드

기본적인 웹 크롤링

만약 크롬 개발자 도구에서 자바스크립트를 비활성화 했을 때
웹브라우저 상에서도 동일한 증상이 보인다면?

이건 100% 동작 페이지!

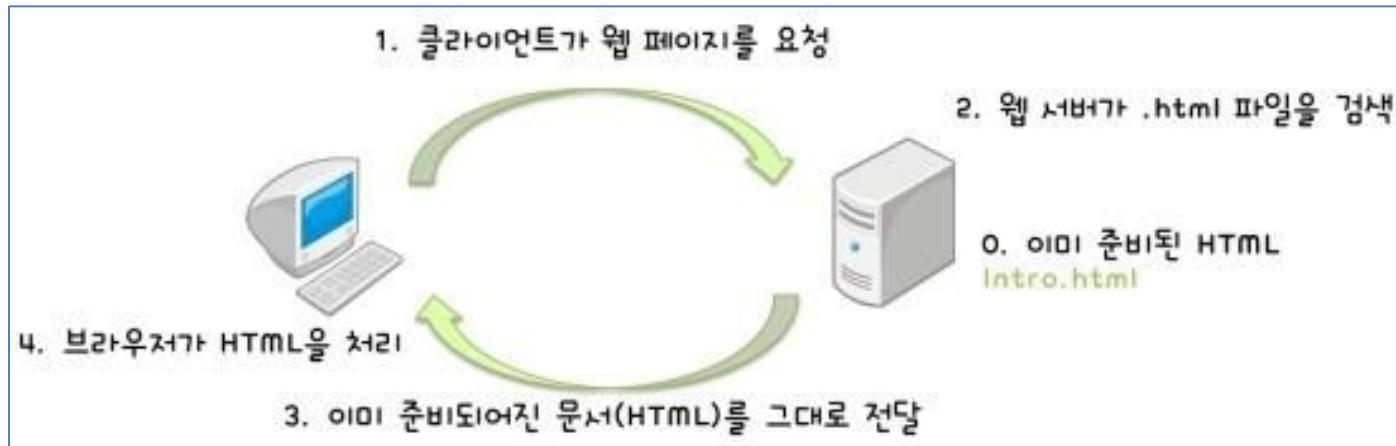


구글 트렌드

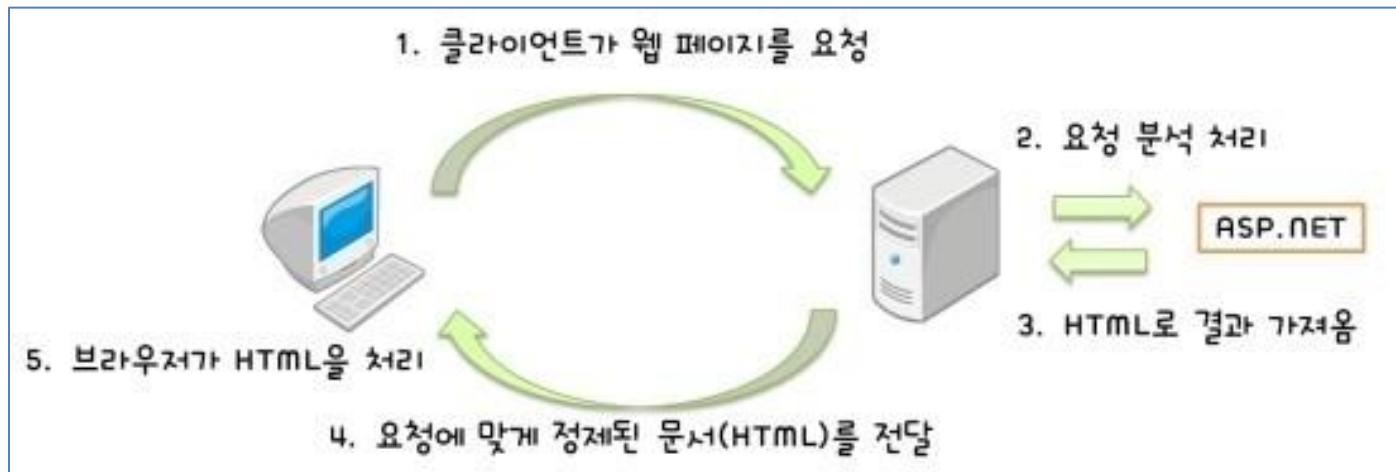
동적 페이지 개요

페이지가 Javascript를 통해 동적으로 렌더링되는 경우에는

정적:

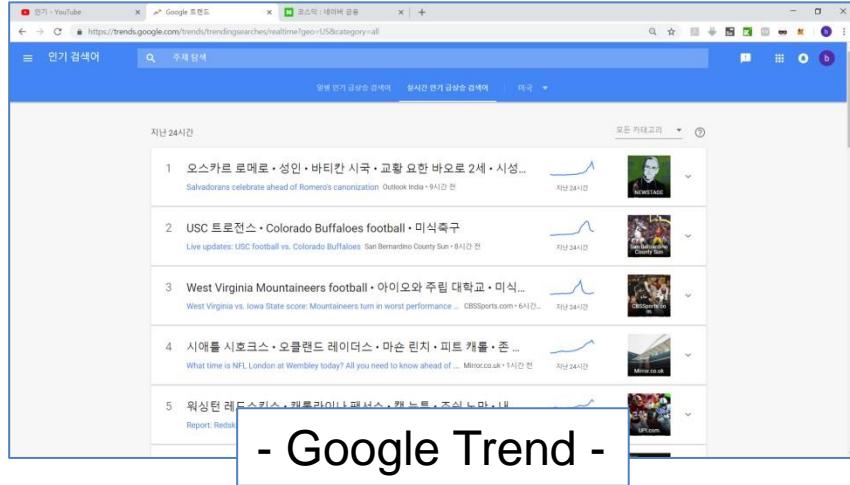


동적:



동적 페이지 개요

아까 말한 단순한 방식으로는 크롤링할 수 없어요 π τ



- 페이지를 받아오면 아무 정보도 없는 경우 -
Google trend, 과거의 youtube 등이
대표적.

HTML로 페이지의 틀을 잡고, 페이지가 로딩
되면서 동적으로 정보를 받아서 표시하는 방식.



- 페이지를 받아오면 일부의 정보만 있는 경우 -
Facebook, naver datalab, vogue korea
등 대부분의 사이트가 이런 식.

처음에는 일부분의 정보만 표시하고 이후에
사용자가 정보를 더 요청하는 동작을 한다면
추가적으로 정보를 받아서 표시하는 방식.

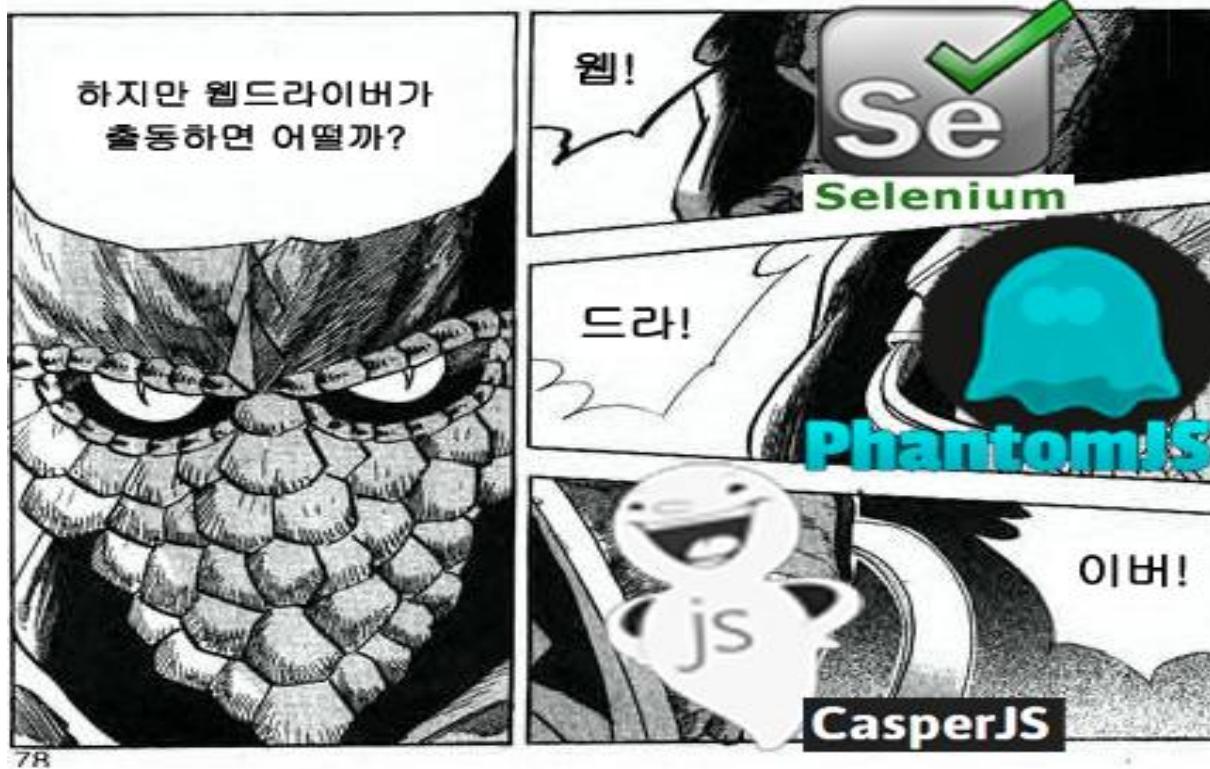
동적 페이지 개요

물론 포기하라는 법은 없기에 다 방법이 있는데요.



동적 페이지 개요

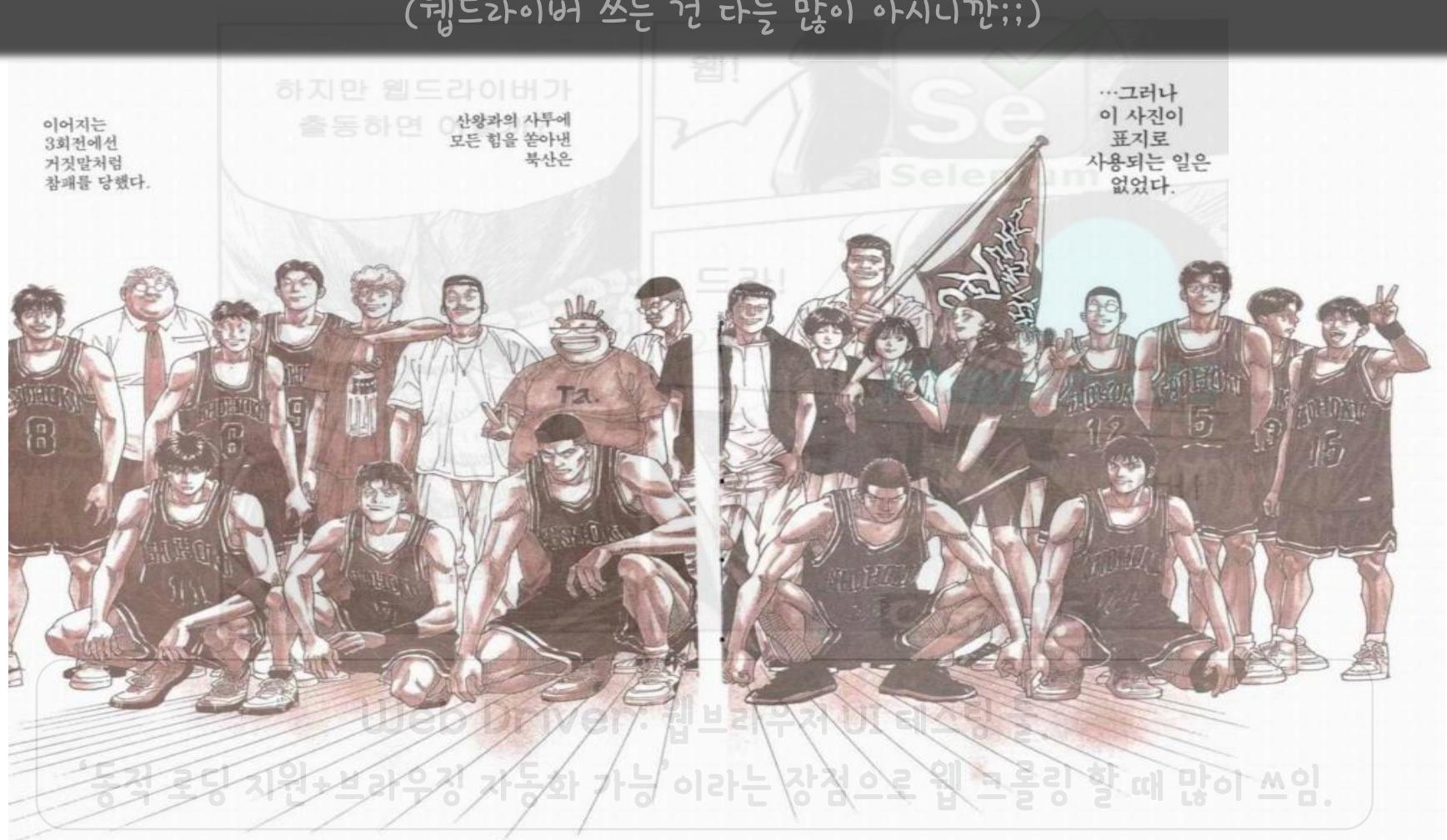
좀 해보신 고수님들은 흐름상 이 생각을 하고 계시겠지만



웹 드라이버 (Web Driver): 웹브라우저 UI 테스팅 툴.
‘동적 로딩 지원+브라우징 자동화’라는 장점으로 크롤링 할 때 많이 쓰임.

동적 페이지 개요

오늘은 웹드라이버를 쓰지 않고 이 문제를 해결하는 것도 다룰거에요.
(웹드라이버 쓰는 건 다들 많이 아시니깐;;)



동적 페이지 개요

오늘은 웹드라이버를 쓰지 않고 이 문제를 해결하는 것도 다룰거에요.
(웹드라이버 쓰는 건 구글링하면 금방 찾으니깐;;)

이어지는
3회전에선
거짓말처럼
챔피언 당했다.

하지만 웹드라이버가
쓸 줄 아는데
산왕과의 사투에
모든 힘을 쏟아낸
복산은

…그러나
이 사진이
표지로
사용되는 일은
없었다.

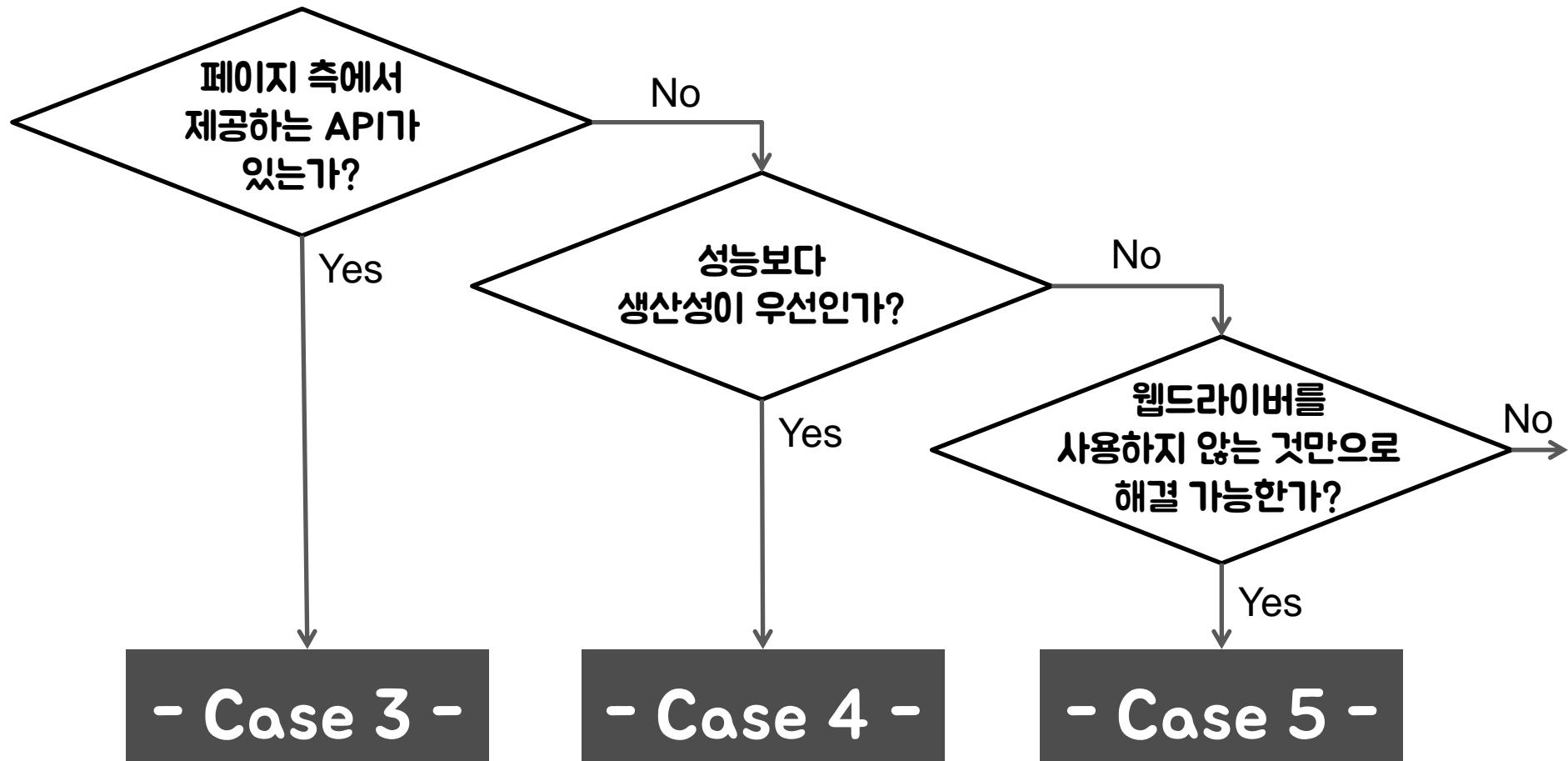
그러면 이제부터 동적 페이지 수집 방법과,
그에 따른 장·단점을 같이 알아봐요!



동적 로딩 지원+브라우징 가능이라는 장점으로 웹 크롤링 할 때 많이 쓰임.

동적 페이지 크롤링

동적 페이지를 크롤링하기 전에, 저 3가지를 꼭 체크해보는 것이 중요해요.



동적 페이지 크롤링

- Case 3 -

페이지에서 제공하는 API가 있는 경우 => API 활용

The screenshot shows a browser window displaying the YouTube trending page at <https://www.youtube.com/feed/trending>. The page lists several trending videos with their titles, descriptions, and thumbnails. On the right side of the browser, the developer tools are open, specifically the Elements tab. A video card is selected, and the DOM structure is visible. The title of the video, "아는 형님(Knowing bros) 150회 예고편", is highlighted, and its corresponding element in the DOM is shown. The developer tools also show the styles applied to this element, including colors, fonts, and sizes.

유튜브 실시간 인기 동영상

2018 데이터야 놀자

동적 페이지 크롤링

유튜브, 페이스북 같은 경우는

자체적으로 제공하는 ‘Data API’가 존재해요

The screenshot shows a web browser displaying the [YouTube Data API \(v3\)](https://developers.google.com/youtube/v3/docs/) documentation. The page has a red header bar with navigation links for Home, 안내 (Documentation), 참조 (Reference), 샘플 (Samples), and 지원 (Support). On the left, there's a sidebar with a tree view of API resources: Overview, Activities, Captions (English), ChannelBanners, Channels, ChannelSections, Comments, CommentThreads, GuideCategories, PlaylistItems, Playlists, Search, Subscriptions, Thumbnails, VideoAbuseReportReasons, VideoCategories, Videos, Watermarks, 표준 매개변수 (English), and 오류 (Errors). The main content area features a title 'API Reference' with a star rating of 5 stars. Below it is a detailed description of the API's purpose and usage. A sidebar on the right lists various resource names: 목차 (Table of Contents), API 호출 (API Calls), 리소스 유형 (Resource Types), activities, channelBanners, channels, guideCategories, playlistItems, playlists, search, subscriptions, thumbnails, videoCategories, and videos. At the bottom, there's a footer with a link to the API Reference Terms of Service.

동적 페이지 크롤링

유튜브, 페이스북 같은 경우는

자체적으로 제공하는 ‘Data API’가 존재해요

The screenshot shows a browser window displaying the YouTube Data API (v3) documentation at <https://developers.google.com/youtube/v3/docs/>. The page has a red header bar with the title 'YouTube > YouTube Data API (v3)'. Below the header, there are tabs for 'Home', 'Introduction', 'Samples', and 'Tutorials'. The main content area is titled 'API Reference' and contains a sidebar with various API endpoints like 'Activities', 'Captions', 'Comments', etc. The main content area features two large, bold, red text blocks: '장점 : 신뢰도 보장, 빠름' (Pros: Trustworthy, Fast) and '단점 : API 키 필요, 쓰는법 배워야 함' (Cons: API key required, need to learn how to use it). Below these, there is a section titled 'API 호출' (API Calls) with instructions for making requests, mentioning the need for an API key and OAuth 2.0 tokens.

장점 : 신뢰도 보장, 빠름

단점 : API 키 필요, 쓰는법 배워야 함

여기 있는 줄 모르고 삽질하다가 나중에 발견하면 복장터짐

API Reference

API 호출

다음은 YouTube Data API 요청에 적용되는 요구 사항입니다.

- 모든 요청은 API 키(`key` 매개변수 포함)를 지정하거나 OAuth 2.0 토큰을 제공해야 합니다. API 키는 APIs Console에서 프로젝트의 API Access 항에 있습니다.
- 모든 삽입, 업데이트 및 삭제 요청 시에는 반드시 인증 토큰을 전송해야 합니다. 또한 인증된 사용자의 비공개 데이터를 검색하는 모든 요청 시에도 인증 토큰을 전송해야 합니다.

리소스 검색을 위한 일부 API 메소드의 경우 인증이 필요한 매개변수를 지원하거나 요청이 인증될 때 추가 메타데이터를 포함할 수 있습니다. 예를 들어 사용자가 업로드한 동영상을 검색하는 요청에는 특정 사용자가 요청을 인증할 경우 비공개 동영상도 포함될 수 있습니다.

동적 페이지 크롤링

- Case 4 -

성능보다 생산성이 우선인 경우 => 웹 드라이버(Web Driver)



웹 드라이버 (Web Driver): 웹브라우저 UI 테스팅 툴.

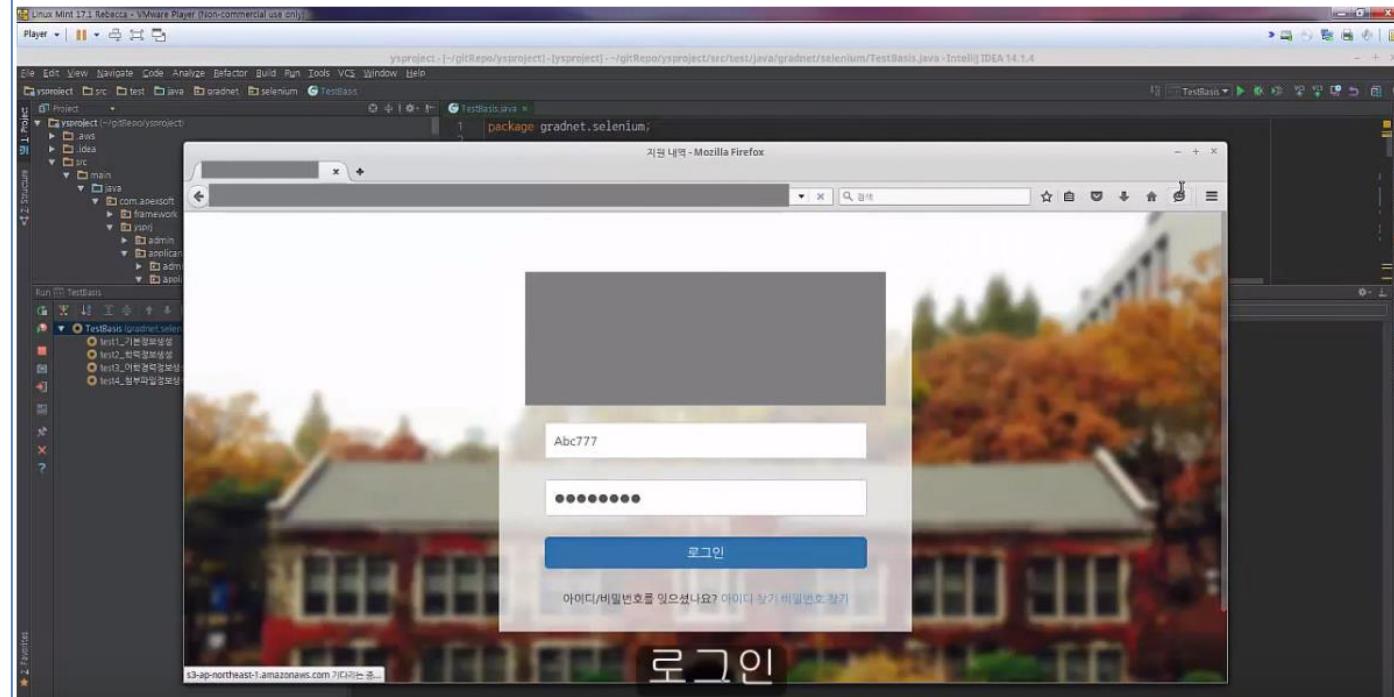
‘동적 로딩 지원+브라우징 자동화’라는 장점으로 크롤링 할 때 많이 쓰임.

동적 페이지 크롤링

가져오려는 데이터가 그리 크지 않으면서 한 번 쓰고 버릴 용도와 같이
성능보다는 일단 당장 되는 것을 만들어야 할 경우 뿐만 아니라



검색

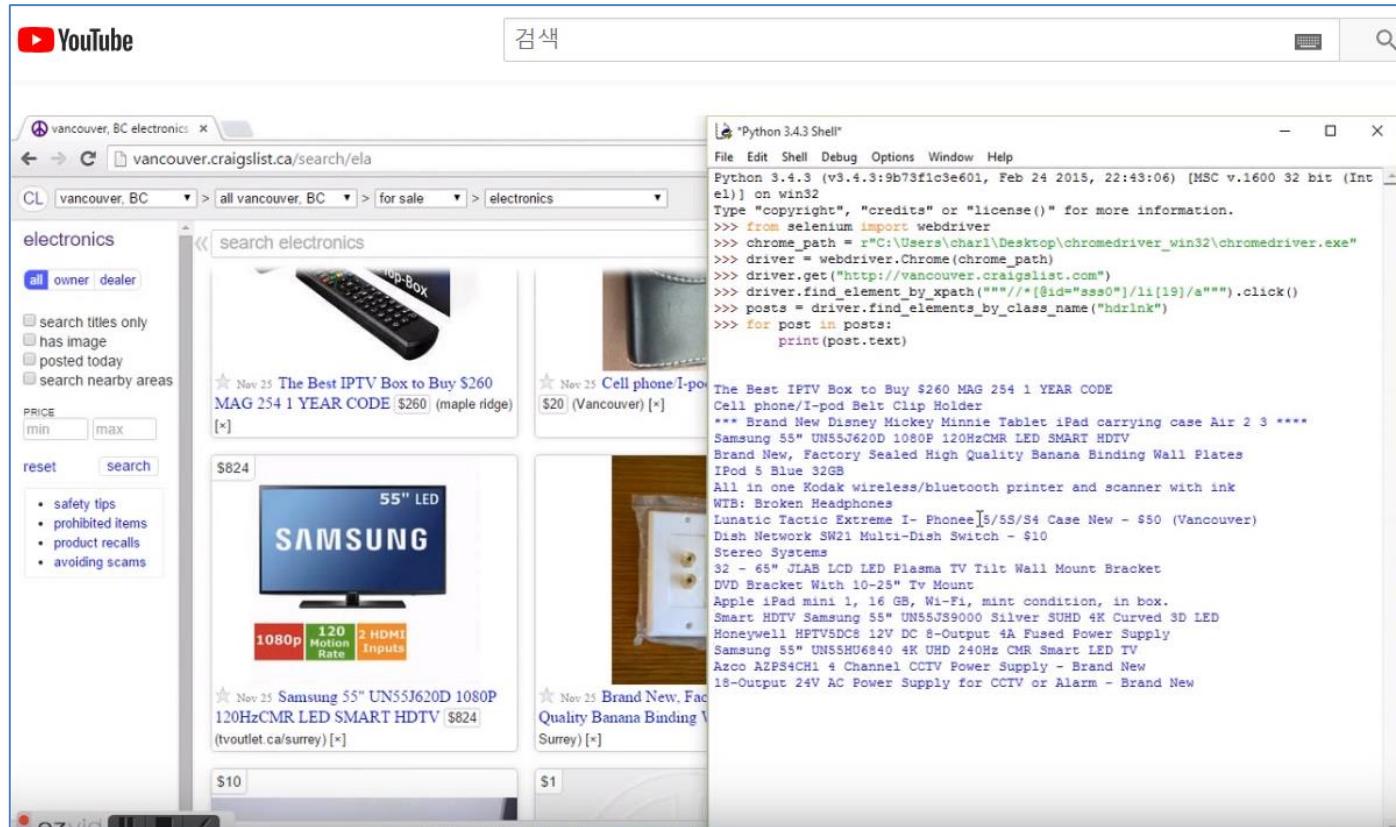


참고 영상

오명운님 - Selenium 테스트 영상

동적 페이지 크롤링

그냥 귀찮을 때 보통 동적 페이지 수집할 때는 이걸 많이 써요.

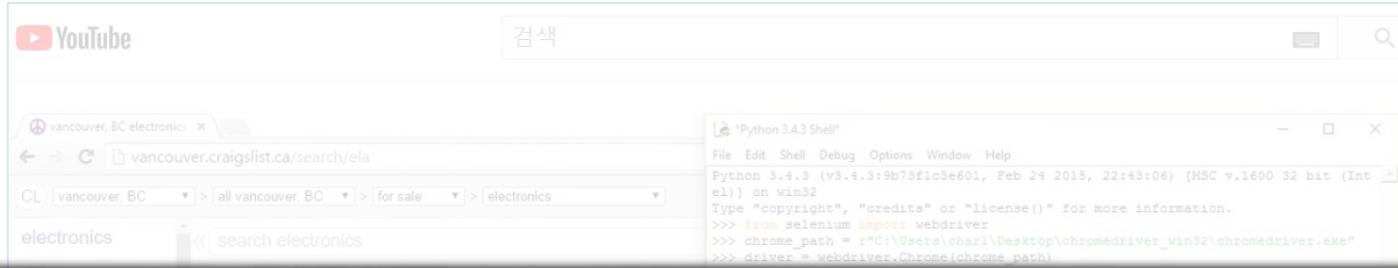


참고 영상

[How to Web Scrape with Python \(Selenium/ChromeDriver\)](#)

동적 페이지 크롤링

그냥 귀찮을 때 보통 동적 페이지 수집할 때는 이걸 많이 써요.



장점 : 생산성, 익혀놓으면 웹페이지 반복작업에도 쓸 수 있음

**단점 : 느림, 무거움, Python과 독립된 도구 필요
, 네트워크 상태의 영향을 크게 받음**



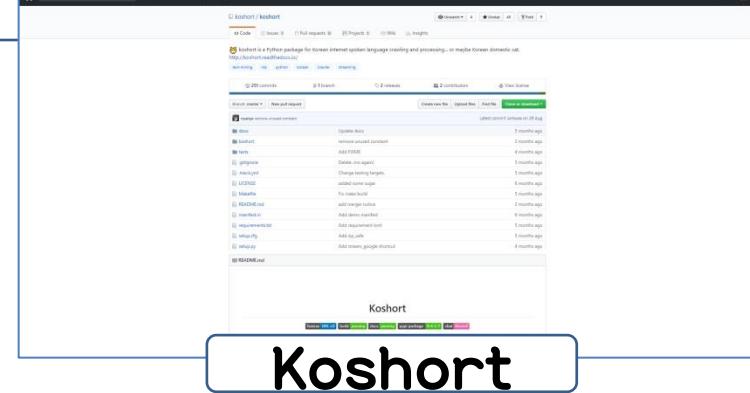
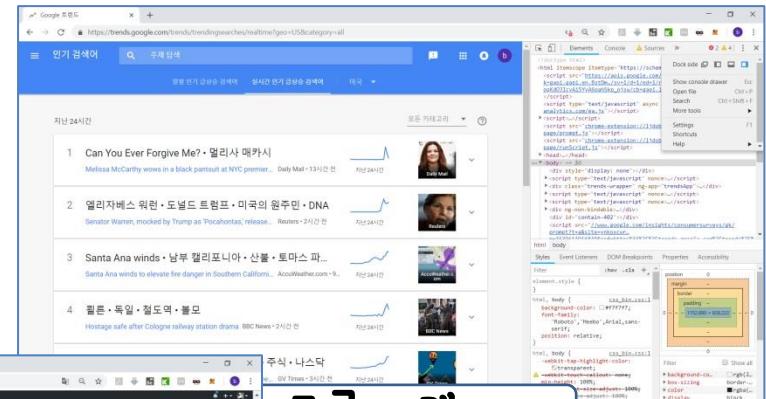
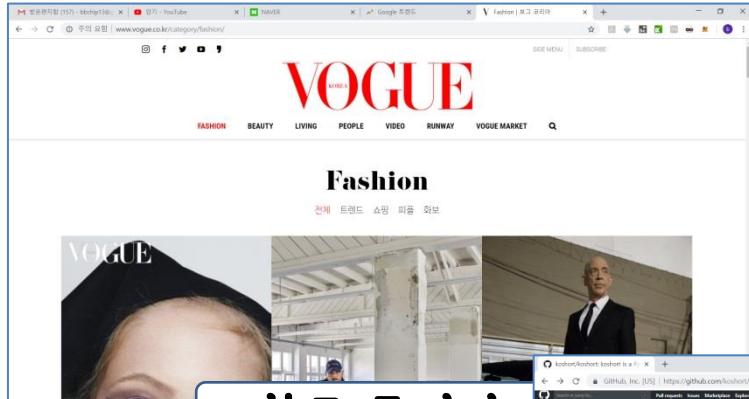
참고 영상

[How to Web Scrape with Python \(Selenium/ChromeDriver\)](#)

동적 페이지 크롤링

- Case 5 -

최적화가 필요한 경우 => 네트워크 리버싱을 통한 데이터 송·수신 구조 추정



Koshort

동적 페이지 크롤링

보그 코리아, 구글 트렌드와 같이 점진적으로 조딩되는 페이지에서
실시간·지속적인 데이터 수집이 필요한 경우나

Vogue | 보그 코리아 | www.vogue.co.kr/category/fashion/
FOOTWEAR FASHION BEAUTY LIVING PEOPLE VIDEO RUNWAY VOGUE MARKET

Givenchy 2018 가을
겨울 컬렉션 'Night
Noir'
겐틀몬스터의 런던 플레이
그십 스토어
이성경이 클로에의 꽃이
되었다

V
VOGUE
보그 코리아

Google 트렌드 | https://trends.google.com/trends/trendingsearches/realtimetime?geo=US&category=all
인기 검색어 주제 탐색

15 세크라멘토 킹스·유타 재즈·코스타 루포스
Regular season starts tonight as Jazz travel to face Sacramento Kings SLC Dunk•9시간 전
지난 24시간

16 밀워키 벅스·밀워키·아니스 안데로쿤보·마이크 버든홀저·201...
What to expect from Coach Budenholzer in Milwaukee thesportsdaily.com•3시간 전
지난 24시간

17 보스턴 브루인스·캘거리 플레이스·파트리스 베르제풀·2017~...
Five things to watch in the Bruins' game against the Flames tonight The Boston Globe•8시간 전
지난 24시간

18 올랜도 매직·マイアミ ヒート·올랜도·모하메드 봄바
Richardson late missstep trips up Heat in season-opening loss in Orlando Sun Sentinel•1시간 전
지난 24시간

다운로드하기

Google Trends 개인정보 보호 약관 고객센터 의견 보내기 정보

구글 트렌드

동적 페이지 크롤링

저처럼 라이브러리로서 사람들에게 제공해야 할 때와 같이
특수한 케이스가 왕왕 있어요

Koshort

koshort / koshort

Code Issues Pull requests Projects Wiki Insights

251 commits 1 branch 2 releases 2 contributors View license

Branch: master New pull request

nyanye remove unused constant

File	Commit Message	Time Ago
docs	Update docs	5 months ago
koshort	remove unused constant	2 months ago
tests	Add FIXME	4 months ago
.gitignore	Delete .mo again!	5 months ago
.travis.yml	Change testing targets.	5 months ago
LICENSE	added some sugar	6 months ago
Makefile	Fix make build	5 months ago
README.md	add merger notice	2 months ago
manifest.in	Add demo manifest	6 months ago
requirements.txt	Add requirement lxml	5 months ago
setup.cfg	Add zip_safe	5 months ago
setup.py	Add stream_google shortcut	4 months ago
README.md		

라이브러리라는 특성상 ‘API 키’, ‘웹드라이버’와 같이
파이썬에 독립적인 요소를 사용하는 것은 최대한 피해야 하며,
생산 시간이 더 들더라도 성능을 최대화해야 함.

동적 페이지 크롤링

저처럼 라이브러리로서 사람들에게 제공해야 할 때와 같이
특수한 케이스가 왕왕 있어요

The screenshot shows a GitHub repository page for 'koshort/koshort'. The page includes a navigation bar with links for 'Code', 'Issues', 'Pull requests', 'Projects', 'Wiki', and 'Insights'. Below the navigation bar, there is a brief description of the project: 'koshort is a Python package for Korean internet spoken language crawling and processing... or maybe Korean domestic cat.' A link to 'http://koshort.readthedocs.io/' is also provided. The main content area displays code snippets and a commit history table.

File	Commit Message	Date
Makefile	Fix make build	5 months ago
README.md	add merger notice	2 months ago
manifest.in	Add demo manifest	6 months ago
requirements.txt	Add requirement.lxml	5 months ago
setup.cfg	Add zip_safe	5 months ago
setup.py	Add stream_google shortcut	4 months ago
README.md		

Koshort

라이브러리라는 특성상 ‘API 키’, ‘웹드라이버’와 같이
파이썬에 독립적인 요소를 사용하는 것은 최대한 피해야 하며,
생산 시간이 더 들더라도 성능을 최대화해야 함.

동적 페이지 크롤링

- Case 5-1 -

페이지를 받아왔는데, 정보가 일부 밖에 없는 경우

The screenshot shows a browser window with multiple tabs open. The active tab is for the Vogue Korea website, specifically the 'Fashion' category page. The page header includes the Vogue logo, social media links, and navigation menus for Fashion, Beauty, Living, People, Video, Runway, and Vogue Market. Below the header, a large 'Fashion' section title is visible, along with a sub-menu for 전체 (All), 트렌드 (Trends), 쇼핑 (Shopping), 피플 (People), and 화보 (Photo). A large image of a woman's face is partially visible on the left. In the center, there are three smaller images: a man in a baseball cap, a concrete pillar in an industrial setting, and a man in a suit. A white callout box with a blue border and black text '보그 코리아' (Vogue Korea) is overlaid at the bottom. The browser interface shows other tabs for 'Google Trends' and 'Fashion | 보그 코리아'.

보그 코리아

동적 페이지 크롤링

- Case 5-1 -

페이지를 받아왔는데, 정보가 일부 밖에 없는 경우

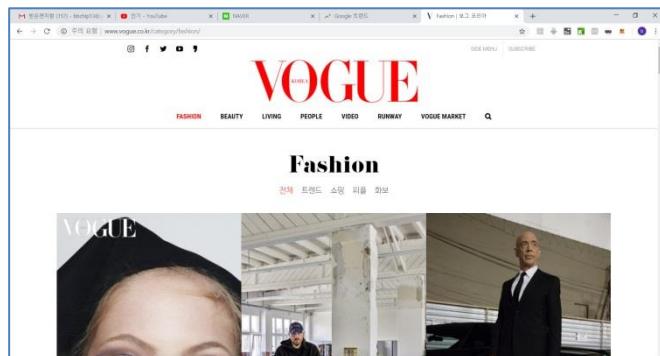


데이터가 로딩되는 시점의 Ajax 요청을 분석한다!

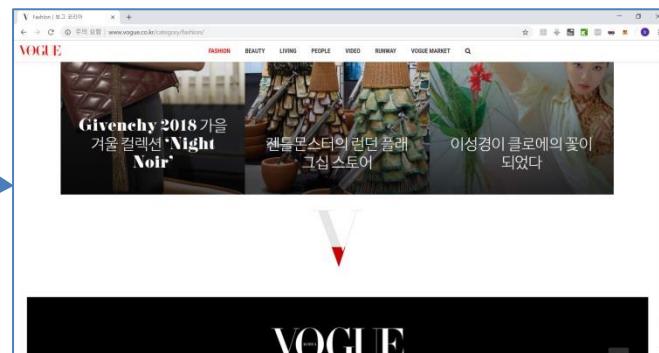
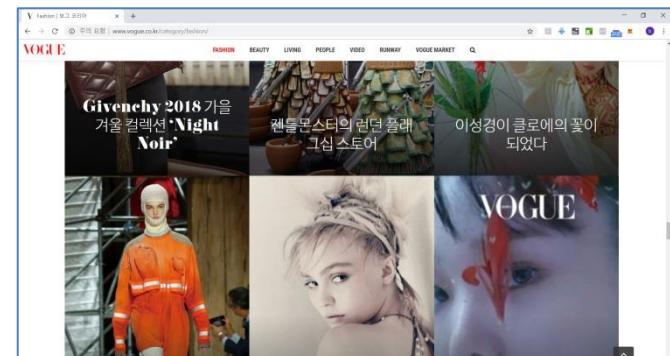


동적 페이지 크롤링

많은 사이트에서 처음에는 일부 데이터만 보여주다가
‘스크롤 다운’ 동작이나 ‘더 로드하기’ 버튼 등을 통해서
동적으로 사이트를 보여주는 방식을 사용하는데요.

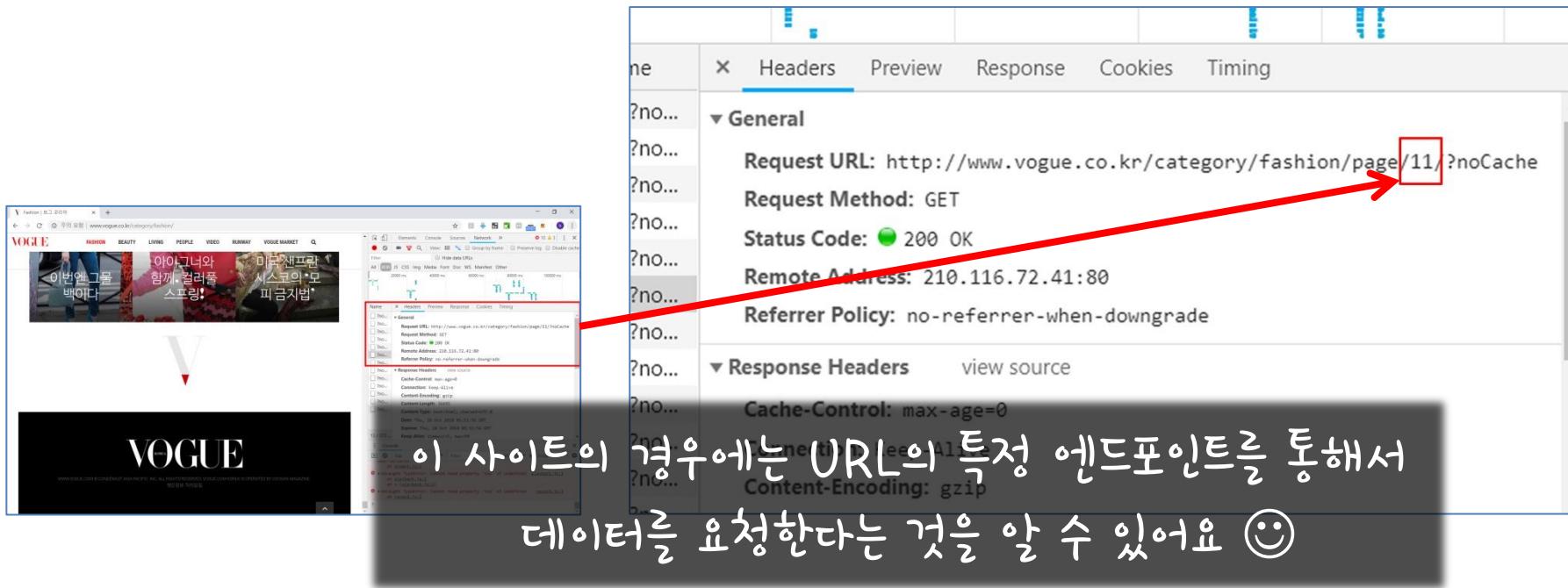


Scroll down!



동적 페이지 크롤링

이런 경우에는 데이터가 로딩되는 동작을 한 시점의 Ajax 요청을 보면서 그 데이터가 어디서 오는지 분석해요.



The screenshot shows a browser window for Vogue Korea's website. The Network tab of the developer tools is open, displaying a list of requests. One specific request is highlighted with a red box and a red arrow pointing to it from the text below. The request details are as follows:

- Request URL: <http://www.vogue.co.kr/category/fashion/page/11/?noCache>
- Request Method: GET
- Status Code: 200 OK
- Remote Address: 210.116.72.41:80
- Referrer Policy: no-referrer-when-downgrade

Below the request details, there are sections for Response Headers and Response Body. The Response Headers include Cache-Control: max-age=0 and Content-Encoding: gzip.

On the left side of the screenshot, the actual content of the Vogue website is visible, showing a fashion category page with various articles and advertisements.

이 사이트의 경우에는 URL의 특정 엔드포인트를 통해서 데이터를 요청한다는 것을 알 수 있어요 😊

동적 페이지 크롤링

시험삼아 10페이지 정도 크롤링 해 봤을 때,
무려 3배 가까운 성능 향상을 보였다는 사실!

```
[vir_webcrawling) /cygdrive/c/Users/BbChip/Desktop/2018_datayanolja_webcrawling
bbchip@DESKTOP-KNOT35K /cygdrive/c/Users/BbChip/Desktop/2018_datayanolja_webcrawling
$ python time_test.py
--vogue_korea_title_webdriver--
['지지 하디드의 블명은 지지 하디드가 아니다!', '육망립스틱을 든 설리', 'Lost In The Country', '헤어 아트의 신세계', '모델 칼리 클로스', 이방카 트럼프와 동서지간되다']
00:00:53
['지지 하디드의 블명은 지지 하디드가 아니다!', '육망립스틱을 든 설리', 'Lost In The Country', '헤어 아트의 신세계', '모델 칼리 클로스', 이방카 트럼프와 동서지간되다']
00:00:53
['지지 하디드의 블명은 지지 하디드가 아니다!', '육망립스틱을 든 설리', 'Lost In The Country', '헤어 아트의 신세계', '모델 칼리 클로스', 이방카 트럼프와 동서지간되다']
00:00:52
['지지 하디드의 블명은 지지 하디드가 아니다!', '육망립스틱을 든 설리', 'Lost In The Country', '헤어 아트의 신세계', '모델 칼리 클로스', 이방카 트럼프와 동서지간되다']
00:00:52
['지지 하디드의 블명은 지지 하디드가 아니다!', '육망립스틱을 든 설리', 'Lost In The Country', '헤어 아트의 신세계', '모델 칼리 클로스', 이방카 트럼프와 동서지간되다']
00:00:52
['지지 하디드의 블명은 지지 하디드가 아니다!', '육망립스틱을 든 설리', 'Lost In The Country', '헤어 아트의 신세계', '모델 칼리 클로스', 이방카 트럼프와 동서지간되다']
00:00:52
['지지 하디드의 블명은 지지 하디드가 아니다!', '육망립스틱을 든 설리', 'Lost In The Country', '헤어 아트의 신세계', '모델 칼리 클로스', 이방카 트럼프와 동서지간되다']
00:00:50
['지지 하디드의 블명은 지지 하디드가 아니다!', '육망립스틱을 든 설리', 'Lost In The Country', '헤어 아트의 신세계', '모델 칼리 클로스', 이방카 트럼프와 동서지간되다']
00:01:01
['지지 하디드의 블명은 지지 하디드가 아니다!', '육망립스틱을 든 설리', 'Lost In The Country', '헤어 아트의 신세계', '모델 칼리 클로스', 이방카 트럼프와 동서지간되다']
00:00:51
average_time: 00:00:52 , 52 [sec]
```

웹드라이버 사용 : 52[s]



--vogue_korea_title_no_webdriver--
['지지 하디드의 본명은 지지 하디드가 아니다!', '육망린스틱을 든 셜리', 'Lost in the country', '헤어 아트의 신세계', '모델 칼리 플로스, 이방카 트럼프와 둘서지간되다']
00:00:28
['지지 하디드의 본명은 지지 하디드가 아니다!', '육망린스틱을 든 셜리', 'Lost in the country', '헤어 아트의 신세계', '모델 칼리 플로스, 이방카 트럼프와 둘서지간되다']
00:00:16
['지지 하디드의 본명은 지지 하디드가 아니다!', '육망린스틱을 든 셜리', 'Lost in the country', '헤어 아트의 신세계', '모델 칼리 플로스, 이방카 트럼프와 둘서지간되다']
00:00:16
['지지 하디드의 본명은 지지 하디드가 아니다!', '육망린스틱을 든 셜리', 'Lost in the country', '헤어 아트의 신세계', '모델 칼리 플로스, 이방카 트럼프와 둘서지간되다']
00:00:16
['지지 하디드의 본명은 지지 하디드가 아니다!', '육망린스틱을 든 셜리', 'Lost in the country', '헤어 아트의 신세계', '모델 칼리 플로스, 이방카 트럼프와 둘서지간되다']
00:00:17
['지지 하디드의 본명은 지지 하디드가 아니다!', '육망린스틱을 든 셜리', 'Lost in the country', '헤어 아트의 신세계', '모델 칼리 플로스, 이방카 트럼프와 둘서지간되다']
00:00:17
['지지 하디드의 본명은 지지 하디드가 아니다!', '육망린스틱을 든 셜리', 'Lost in the country', '헤어 아트의 신세계', '모델 칼리 플로스, 이방카 트럼프와 둘서지간되다']
00:00:15
['지지 하디드의 본명은 지지 하디드가 아니다!', '육망린스틱을 든 셜리', 'Lost in the country', '헤어 아트의 신세계', '모델 칼리 플로스, 이방카 트럼프와 둘서지간되다']
00:00:16
['지지 하디드의 본명은 지지 하디드가 아니다!', '육망린스틱을 든 셜리', 'Lost in the country', '헤어 아트의 신세계', '모델 칼리 플로스, 이방카 트럼프와 둘서지간되다']

웹드라이버 미사용 : 17[s]

00:00:16
average_time: 00:00:17 , 17 [sec]

웹드라이버 미사용 : 17[s]

동적 페이지 크롤링

시험삼아 10페이지 정도 크롤링 해 봤을 때,

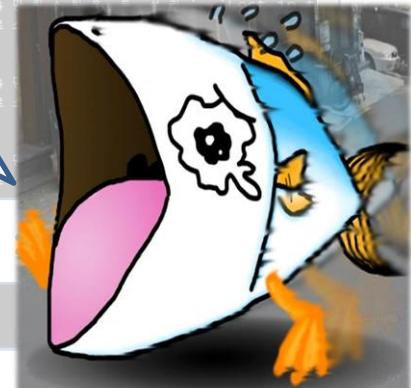
테스트는 데스크탑 유선랜 환경에서 실행하였고,
네트워크 환경에 따라서 결과가 달라질 수 있어요!

멀티스레딩, 멀티프로세싱 등으로 분산처리를 구현한다면,
더 높은 퍼포먼스를 기대할 수 있어요 😊

웹드라이버 사용 : 52[s]

2018 데이터야 놀자

웹드라이버 미사용



동적 페이지 크롤링

- Case 5-2 -

페이지를 받아왔는데, 정보가 하나도 없는 경우

The screenshot shows the Google Trends interface with the URL <https://trends.google.com/trendingsearches/realtimedatageo=US&category=all>. The search bar has '인기 검색어' (Popular Searches) and '주제 탐색' (Topic Search). The main area displays the top 5 trending searches for the last 24 hours:

순위	검색어	설명	출처	최신 업데이트	미디어
1	알렉스 코라 • 보스턴 레드삭스 • 푸에르토리코 • 휴스턴 애스트로...	Cora guides Red Sox, inspires team back home in Puerto Rico	Yahoo Sports	4시간 전	
2	보스턴 레드삭스 • 휴스턴 애스트로스 • 앤드루 베닌텐디 • 조쉬 레...	Red Sox 8, Astros 6	Chron.com	6시간 전	
3	빅 버드 • 세서미 스트리트 • 캐롤 스피니 • 꼭두각시 공연자 • 오스...	LeBron debuts with Lakers, Big Bird milestone, Saudi case: 5 things you need ...	USA TODAY	3시간 전	
4	Chris Odoi-Atsem • D.C. 유나이티드 • 호지킨 림프종 • 워싱턴 D.C.	DC United's Chris Odoi-Atsem diagnosed with Hodgkin's lymphoma	WJLA	20시간 전	
5	홍역 • 돌발 • 학교 • 예방 접종 • 빈니차주	In several schools in Vinnytsia region have suspended training due to a ...	The Siver Telegram	17시간 전	

A large blue callout bubble at the bottom left contains the text "구글 트렌드".

동적 페이지 크롤링

- Case 5-2 -

페이지를 받아왔는데, 정보가 하나도 없는 경우



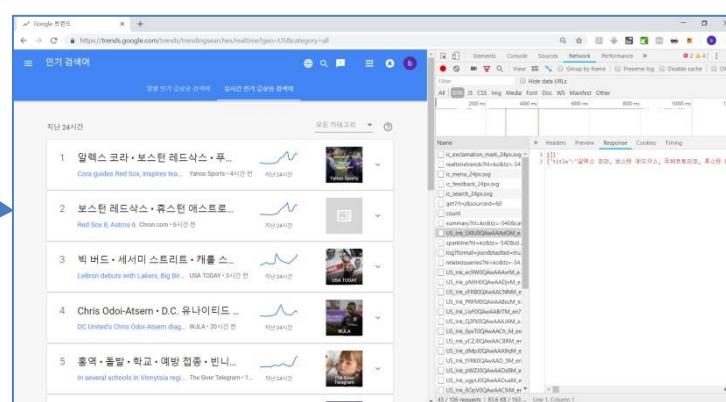
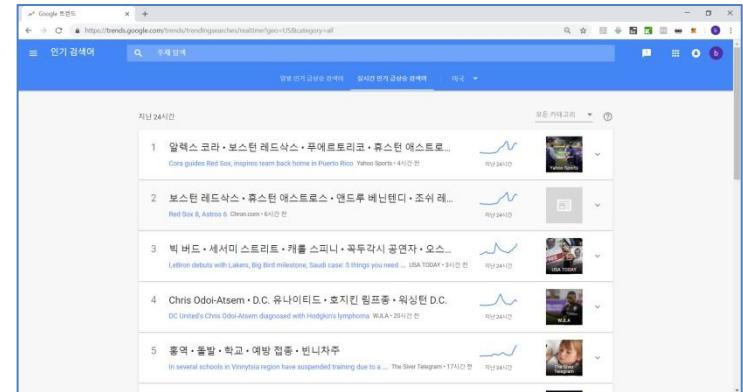
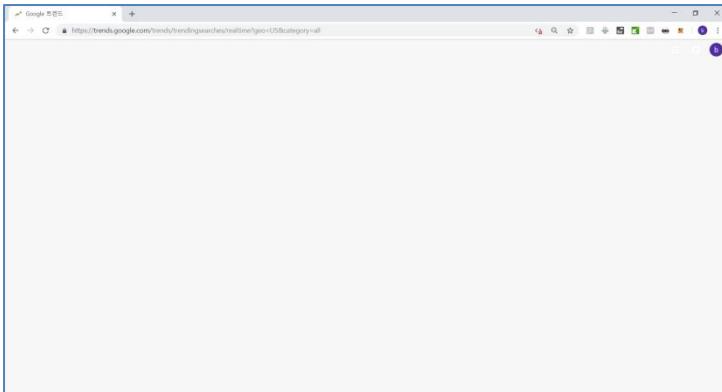
처음 페이지가 로딩되는 시점부터

Ajax 요청 흐름을 따라가며 파악한다!



동적 페이지 크롤링

정적인 HTML은 틀을 만드는 용도로만 쓰고,
데이터는 모조리 JS로 처리하는 방식인데요.



동적 페이지 크롤링

구글 트렌드의 실시간 급상승 검색어의 경우,
페이지를 로드하면 일단 키워드에 해당하는 전체 리스트를 가져오고

The screenshot shows a browser window for Google Trends with the URL <https://trends.google.com/trends/trendingsearches/realtime?geo=US&category=m>. The left side displays the JSON response from the API, which includes a list of trending story IDs. A red box highlights this list, and a red arrow points from it to the Network tab in the developer tools on the right. The Network tab shows the XHR request for the same URL, with the response body identical to the one on the left.

trendingStoryIds: ["US_1nk_tENW0QAwAA",
 [0 ... 99]
 0: "US_1nk_tENW0QAwAADiIM_en"
 1: "US_1nk_PRFM0QAwAABxcM_en"
 2: "US_1nk_B71J0QAwAABO2M_en"
 3: "US_1nk_CxRW0QAwAABddM_en"
 4: "US_1nk_84hG0QAwAAC16M_en"
 5: "US_1nk_FzRD0QAwAABUVM_en"
 6: "US_1nk_v5JI0QAwAAD38M_en"
 7: "US_1nk_8XdL0QAwAAC6FM_en"
 8: "US_1nk_KgdW0QAwAAB8ZM_en"
 9: "US_1nk_fD9K0QAwAAA2XM_en"
 10: "US_1nk_GVpJ00AwAABOOM_en"

Google Trends API Response (Network Tab):

```
        "date": "2018. 10. 18.",  
        "featuredStoryIds": [],  
        "hideAllImages: false",  
        "storySummaries": { "featuredStories": [] },  
        "trendingStoryIds": ["US_1nk_tENW0QAwAADiIM_en", "US_1nk_PRFM0QAwAABxcM_en",  
                      "US_1nk_B71J0QAwAABO2M_en", "US_1nk_CxRW0QAwAABddM_en", "US_1nk_84hG0QAwAAC16M_en",  
                      "US_1nk_FzRD0QAwAABUVM_en", "US_1nk_v5JI0QAwAAD38M_en", "US_1nk_8XdL0QAwAAC6FM_en",  
                      "US_1nk_KgdW0QAwAAB8ZM_en", "US_1nk_fD9K0QAwAAA2XM_en", "US_1nk_GVpJ00AwAABOOM_en",  
                      "US_1nk_SJFQ0QAwAAAY8M_en", "US_1nk_as9P0QAwAAA1M_en", "US_1nk_Nvp0Q0QAwAAB4mM_en",  
                      "US_1nk_R1lG0QAwAAC6GM_en", "US_1nk_mINM0QAwAAU4M_en", "US_1nk_2qS0QAwAAC1zM_en",  
                      "US_1nk_DXP0QAwAACBCFM_en", "US_1nk_3bhN0QAwAACQ2M_en", "US_1nk_BmpT0QAwABCVM_en",  
                      "US_1nk_2dxK0QAwAACTvM_en", "US_1nk_HVRV0QAwABIM_en", "US_1nk_wy9W0QAwAACVTM_en",  
                      "US_1nk_Zm2N0QAwAAarbM_en", "US_1nk_14k50QAwADDu6M_en", "US_1nk_FwpS0QAwAABFaM_en",  
                      "US_1nk_17J0QAwAAC2MM_en"]]
```

구글 트렌드

동적 페이지 크롤링

그 리스트의 요소들을 하나씩 요청해서 가져오는 방식이에요.

The screenshot shows the browser's developer tools Network tab. On the left, a list of API endpoint URLs is shown, each ending with a question mark and parameters like 'hl=ko&tz=-540'. On the right, the corresponding JSON response is displayed. A large red arrow points from the list of URLs to the JSON object, indicating that each request corresponds to one item in the JSON array.

date: "2018. 10. 18."
featuredStoryIds: []
hideAllImages: false
► storySummaries: {featuredStories: [...]}
▼ trendingStoryIds: ["US_Lnk_tENW0QAwAADiIM_en", "US_..."]
▼ [0 ... 99]
0: "US_Lnk_tENW0QAwAADiIM_en"
1: "US_Lnk_PRFM0QAwAABxcM_en"
2: "US_Lnk_B71J0QAwAABO2M_en"
3: "US_Lnk_CxRW0QAwAABddM_en"
4: "US_Lnk_84hG0QAwAAC16M_en"
5: "US_Lnk_FzRD0QAwAABUVM_en"
6: "US_Lnk_v5JI0QAwAAD38M_en"
7: "US_Lnk_8XdL0QAwAAC6FM_en"
8: "US_Lnk_KgdW0QAwAAB8ZM_en"
9: "US_Lnk_fD9K0QAwAAA2XM_en"
10: "US_Lnk_GVpJ0QAwAABQOM_en"
11: "US_Lnk_SJFQ0QAwAAAY8M_en"
12: "US_Lnk_as9P0QAwAAA1rM_en"
13: "US_Lnk_NvpO0QAwAAB4mM_en"
14: "US_Lnk_RI1G0QAwAAC6M_en"
15: "US_Lnk_mINM0QAwAADU4M_en"
... "US_Lnk_..._0QAwAACIzM_en"
... "US_Lnk_..._0QAwABCfM_en"

구글 트렌드

동적 페이지 크롤링

역시나 이 경우에도 9배 정도의 향상을 확인할 수 있었어요!

이 경우도 멀티스레딩 등을 통해서 더 최적화가 가능하답니다 😊

```
E:/cygdrive/c/Users/BbChip/Desktop/2018_datayanolja_webcrawling
$ clear
(vir_webcrawling)
BbChip@DESKTOP-KNOT35K /cygdrive/c/Users/BbChip/Desktop/2018_datayanolja_webcrawling
$ python time_test.py
--google_trend_title_webdriver--
['Tennessee Volunteers men's basketball ? 사우스 이스턴 컨퍼런스', 'Australian Energy Regulator ? 오스트레일리아', '열대 저기압 ? 맥시코 ? 폭풍', '피의자 ? 그위닛 카운티 ? 경찰관', '샌프란시스코 ? 오락 ? 오픈월드 ? 오락 ? ?']
00:00:56
['Kansas Jayhawks men's basketball ? AP Poll', 'Australian Energy Regulator ? 오스트레일리아', '열대 저기압 ? 맥시코 ? 폭풍', '오락 ? 클라우드 컴퓨팅 ? 오락 ? 오픈월드', '텍사스주 ? 푸포 ? 후보 ?']
00:00:58
['Kansas Jayhawks men's basketball ? AP Poll', 'Australian Energy Regulator ? 오스트레일리아', '열대 저기압 ? 맥시코 ? 폭풍', '오락 ? 클라우드 컴퓨팅 ? 오락 ? 오픈월드', '텍사스주 ? 푸포 ? 후보 ?']
00:00:49
['Kansas Jayhawks men's basketball ? AP Poll', 'Australian Energy Regulator ? 오스트레일리아', '열대 저기압 ? 맥시코 ? 폭풍', '오락 ? 클라우드 컴퓨팅 ? 오락 ? 오픈월드', '텍사스주 ? 푸포 ? 후보 ?']
00:00:53
['Kansas Jayhawks men's basketball ? AP Poll', 'Australian Energy Regulator ? 오스트레일리아', '열대 저기압 ? 맥시코 ? 폭풍', '오락 ? 클라우드 컴퓨팅 ? 오락 ? 오픈월드', '텍사스주 ? 푸포 ? 후보 ?']
00:00:57
['Kansas Jayhawks men's basketball ? AP Poll', 'Australian Energy Regulator ? 오스트레일리아', '열대 저기압 ? 맥시코 ? 폭풍', '오락 ? 클라우드 컴퓨팅 ? 오락 ? 오픈월드', '텍사스주 ? 푸포 ? 후보 ?']
00:00:49
['Kansas Jayhawks men's basketball ? AP Poll', 'Australian Energy Regulator ? 오스트레일리아', '열대 저기압 ? 맥시코 ? 폭풍', '오락 ? 클라우드 컴퓨팅 ? 오락 ? 오픈월드', '텍사스주 ? 푸포 ? 후보 ?']
00:00:49
['Tennessee Volunteers men's basketball ? 2018 ? 웹캠 헌봉 ?', '피의자 ? 그위닛 카운티 ? 경찰관', '오락 ? 클라우드 컴퓨팅 ? 오락 ? 오픈월드 ?']
00:01:00
['Tennessee Volunteers men's basketball ? 사우스 이스턴 컨퍼런스', '열대 저기압 ? 맥시코 ? 폭풍', 'Wentworth by-election, 2018 ? 웹캠 헌봉 ?', '피의자 ? 그위닛 카운티 ? 경찰관', '오락 ? 클라우드 컴퓨팅 ? 오락 ? 오픈월드 ?']
```

웹드라이버 사용 : 54[s]



```
E:/cygdrive/c/Users/BbChip/Desktop/2018_datayanolja_webcrawling
[', 'Wentworth by-election, 2018, 웹캠 헌봉, 스콧 모리슨, 케린 월프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부', '피의자, 그위닛 카운티, 경찰관', '오락', 클라우드 컴퓨팅, 오락 ? 오픈월드, 오락 ? 클라우드 ?]
00:00:06
['Tennessee Volunteers men's basketball', 사우스 이스턴 컨퍼런스, 농구, Charlotte 49ers football, 그랜트 월리엄스, AP Poll, 스코필드 제독', '열대 저기압, 맥시코, 폭풍', 'Wentworth by-election, 2018, 웹캠 헌봉, 스콧 모리슨, 케린 월프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부', '피의자, 그위닛 카운티, 경찰관', '오락', 클라우드 컴퓨팅, 오락 ? 오픈월드, 오락 ? 클라우드 ?]
00:00:06
['Tennessee Volunteers men's basketball', 사우스 이스턴 컨퍼런스, 농구, Charlotte 49ers football, 그랜트 월리엄스, AP Poll, 스코필드 제독', '열대 저기압, 맥시코, 폭풍', 'Wentworth by-election, 2018, 웹캠 헌봉, 스콧 모리슨, 케린 월프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부', '피의자, 그위닛 카운티, 경찰관', '오락', 클라우드 컴퓨팅, 오락 ? 오픈월드, 오락 ? 클라우드 ?]
00:00:06
['Tennessee Volunteers men's basketball', 사우스 이스턴 컨퍼런스, 농구, Charlotte 49ers football, 그랜트 월리엄스, AP Poll, 스코필드 제독', '열대 저기압, 맥시코, 폭풍', 'Wentworth by-election, 2018, 웹캠 헌봉, 스콧 모리슨, 케린 월프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부', '피의자, 그위닛 카운티, 경찰관', '오락', 클라우드 컴퓨팅, 오락 ? 오픈월드, 오락 ? 클라우드 ?]
00:00:06
['Tennessee Volunteers men's basketball', 사우스 이스턴 컨퍼런스, 농구, Charlotte 49ers football, 그랜트 월리엄스, AP Poll, 스코필드 제독', '열대 저기압, 맥시코, 폭풍', 'Wentworth by-election, 2018, 웹캠 헌봉, 스콧 모리슨, 케린 월프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부', '피의자, 그위닛 카운티, 경찰관', '오락', 클라우드 컴퓨팅, 오락 ? 오픈월드, 오락 ? 클라우드 ?]
00:00:06
['Tennessee Volunteers men's basketball', 사우스 이스턴 컨퍼런스, 농구, Charlotte 49ers football, 그랜트 월리엄스, AP Poll, 스코필드 제독', '열대 저기압, 맥시코, 폭풍', 'Wentworth by-election, 2018, 웹캠 헌봉, 스콧 모리슨, 케린 월프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부', '피의자, 그위닛 카운티, 경찰관', '오락', 클라우드 컴퓨팅, 오락 ? 오픈월드, 오락 ? 클라우드 ?]
00:00:06
['Tennessee Volunteers men's basketball', 사우스 이스턴 컨퍼런스, 농구, Charlotte 49ers football, 그랜트 월리엄스, AP Poll, 스코필드 제독', '열대 저기압, 맥시코, 폭풍', 'Wentworth by-election, 2018, 웹캠 헌봉, 스콧 모리슨, 케린 월프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부', '피의자, 그위닛 카운티, 경찰관', '오락', 클라우드 컴퓨팅, 오락 ? 오픈월드, 오락 ? 클라우드 ?]
00:00:05
['Tennessee Volunteers men's basketball', 사우스 이스턴 컨퍼런스, 농구, Charlotte 49ers football, 그랜트 월리엄스, AP Poll, 스코필드 제독', '열대 저기압, 맥시코, 폭풍', 'Wentworth by-election, 2018, 웹캠 헌봉, 스콧 모리슨, 케린 월프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부', '피의자, 그위닛 카운티, 경찰관', '오락', 클라우드 컴퓨팅, 오락 ? 오픈월드, 오락 ? 클라우드 ?]
00:00:05
['Tennessee Volunteers men's basketball', 사우스 이스턴 컨퍼런스, 농구, Charlotte 49ers football, 그랜트 월리엄스, AP Poll, 스코필드 제독', '열대 저기압, 맥시코, 폭풍', 'Wentworth by-election, 2018, 웹캠 헌봉, 스콧 모리슨, 케린 월프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부', '피의자, 그위닛 카운티, 경찰관', '오락', 클라우드 컴퓨팅, 오락 ? 오픈월드, 오락 ? 클라우드 ?']
```

웹드라이버 미사용 : 6[s]

average_time: 00:00:06 , 6 [sec]

동적 페이지 크롤링

역시나 이 경우에도 9배 정도의 향상을 확인할 수 있었어요!
이 경우도 멀티스레딩 등을 통해서 더 최적화가 가능하답니다 😊

```
E:/cygdrive/c/Users/BbChip/Desktop/2018_datayanolja_webcrawling
$ clear
(vir_webcrawling)
BbChip@DESKTOP-KNOT35K /cygdrive/c/Users/BbChip/Desktop/2018_datayanolja_webcrawling
$ python time_test.py
--google_trend_title_webdriver--
["Tennessee Volunteers men's basketball ? 사우스 이스랜드 퍼포먼스 ", 'Australian Energy Regulator ? 오스트레일리아 ', '열대 저기압 ? 멕시코 ? 푸른 ', '오락을 ? 클라우드 컴퓨팅 ? 오락을 오픈월드 ', '텍사스 주 ? 후보 ? 후보 ']
```

```
E:/cygdrive/c/Users/BbChip/Desktop/2018_datayanolja_webcrawling
[0:00:06
["Wentworth by-election, 2018, 열정 힘풀, 스콧 모리슨, 케린 웰프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부 ', '피의자, 그위닛 카운티, 경찰관 ', '오락을, 클라우드 컴퓨팅, 오락을 오픈월드, 오락을 클라우드 ']
00:00:06
["Tennessee Volunteers men's basketball, 사우스 이스랜드 퍼포먼스, 뉴구, Charlotte 49ers football, 그린트 월리엄스, AP Poll, 스코폴드 제독 ", '열대 저기압, 멕시코, 푸른 ', 'Wentworth by-election, 2018, 열정 힘풀, 스콧 모리슨, 케린 웰프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부 ', '피의자, 그위닛 카운티, 경찰관 ', '오락을, 클라우드 컴퓨팅, 오락을 오픈월드, 오락을 클라우드 ']
```

장점 : 빠름, 가벼움

단점 : 번거로움 :, 웹개발에 대한 지식 필요

```
[0:00:57
["Kansas Jayhawks men's basketball ? AP Poll", 'Australian Energy Regulator ? 오스트레일리아 ', '열대 저기압 ? 멕시코 ? 푸른 ', '오락을 ? 클라우드 컴퓨팅 ? 오락을 오픈월드 ', '텍사스 주 ? 후보 ? 후보 ']
00:00:49
["Kansas Jayhawks men's basketball ? AP Poll", 'Australian Energy Regulator ? 오스트레일리아 ', '열대 저기압 ? 멕시코 ? 푸른 ', '오락을 ? 클라우드 컴퓨팅 ? 오락을 오픈월드 ', '텍사스 주 ? 후보 ? 후보 ']
00:00:49
["Tennessee Volunteers men's basketball ? 사우스 이스랜드 퍼포먼스 ", '열대 저기압 ? 멕시코 ? 푸른 ', 'Wentworth by-election, 2018 ? 열정 힘풀 ?, '피의자 ? 그위닛 카운티 ? 경찰관 ', '오락을 ? 클라우드 컴퓨팅 ? 오락을 오픈월드 ']
00:01:00
["Tennessee Volunteers men's basketball ? 사우스 이스랜드 퍼포먼스 ", '열대 저기압 ? 멕시코 ? 푸른 ', 'Wentworth by-election, 2018 ? 열정 힘풀 ?, '피의자 ? 그위닛 카운티 ? 경찰관 ', '오락을 ? 클라우드 컴퓨팅 ? 오락을 오픈월드 ']
00:00:06
["Tennessee Volunteers men's basketball, 사우스 이스랜드 퍼포먼스, 뉴구, Charlotte 49ers football, 그린트 월리엄스, AP Poll, 스코폴드 제독 ", '열대 저기압, 멕시코, 푸른 ', 'Wentworth by-election, 2018, 열정 힘풀, 스콧 모리슨, 케린 웰프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부 ', '피의자, 그위닛 카운티, 경찰관 ', '오락을, 클라우드 컴퓨팅, 오락을 오픈월드, 오락을 클라우드 ']
00:00:05
["Tennessee Volunteers men's basketball, 사우스 이스랜드 퍼포먼스, 뉴구, Charlotte 49ers football, 그린트 월리엄스, AP Poll, 스코폴드 제독 ", '열대 저기압, 멕시코, 푸른 ', 'Wentworth by-election, 2018, 열정 힘풀, 스콧 모리슨, 케린 웰프스, 오스트레일리아, Division of Wentworth, 오스트레일리아 자유당, 정부 ', '피의자, 그위닛 카운티, 경찰관 ', '오락을, 클라우드 컴퓨팅, 오락을 오픈월드, 오락을 클라우드 ']
```

웹드라이버 사용 : 54[s]

average_time: 00:00:54 , 54 [sec]



웹드라이버 미사용 : 6[s]

average_time: 00:00:06 , 6 [sec]

마무리



마무리



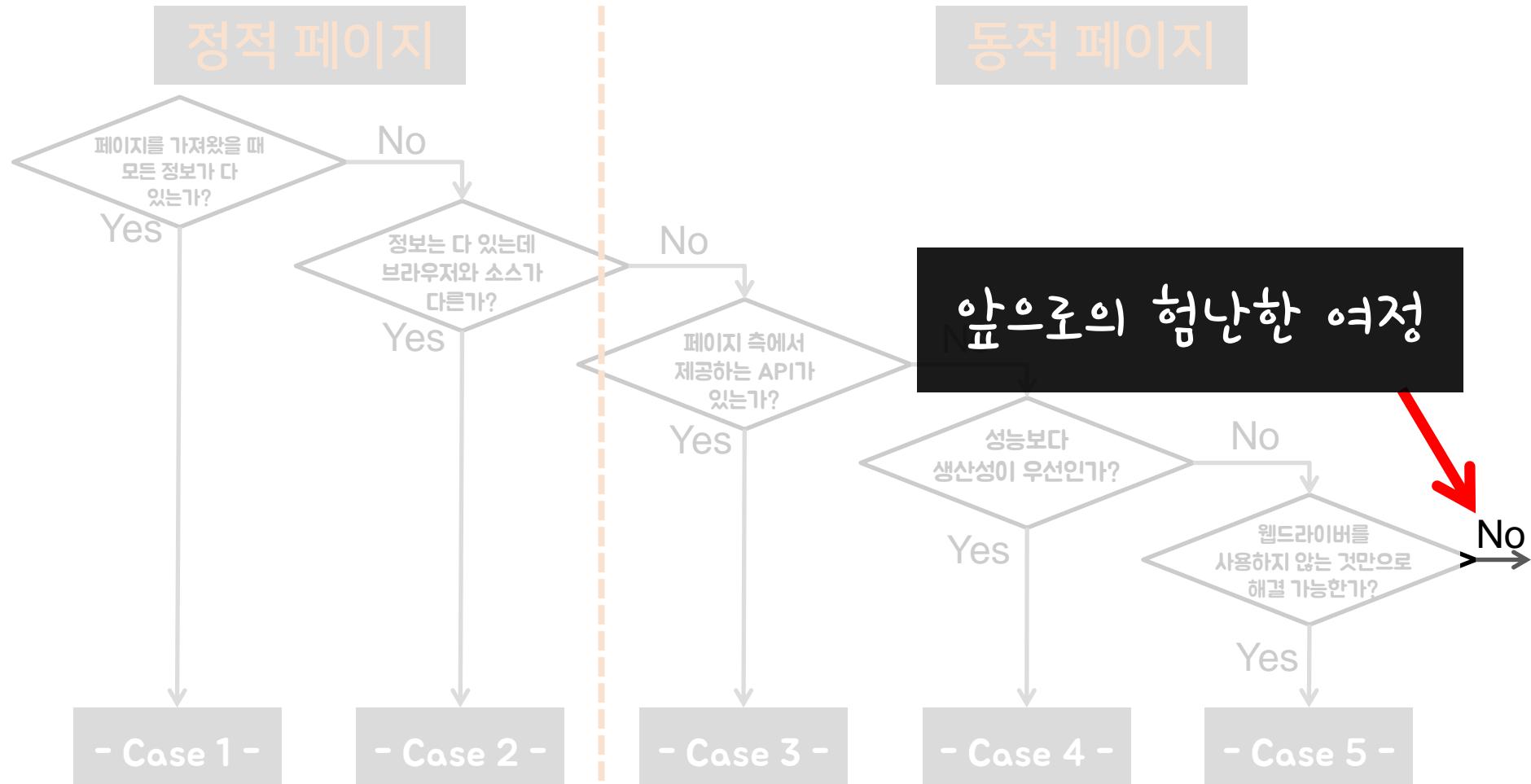
해냈다!
크롤링 끝!

이면 참 좋으련만.....



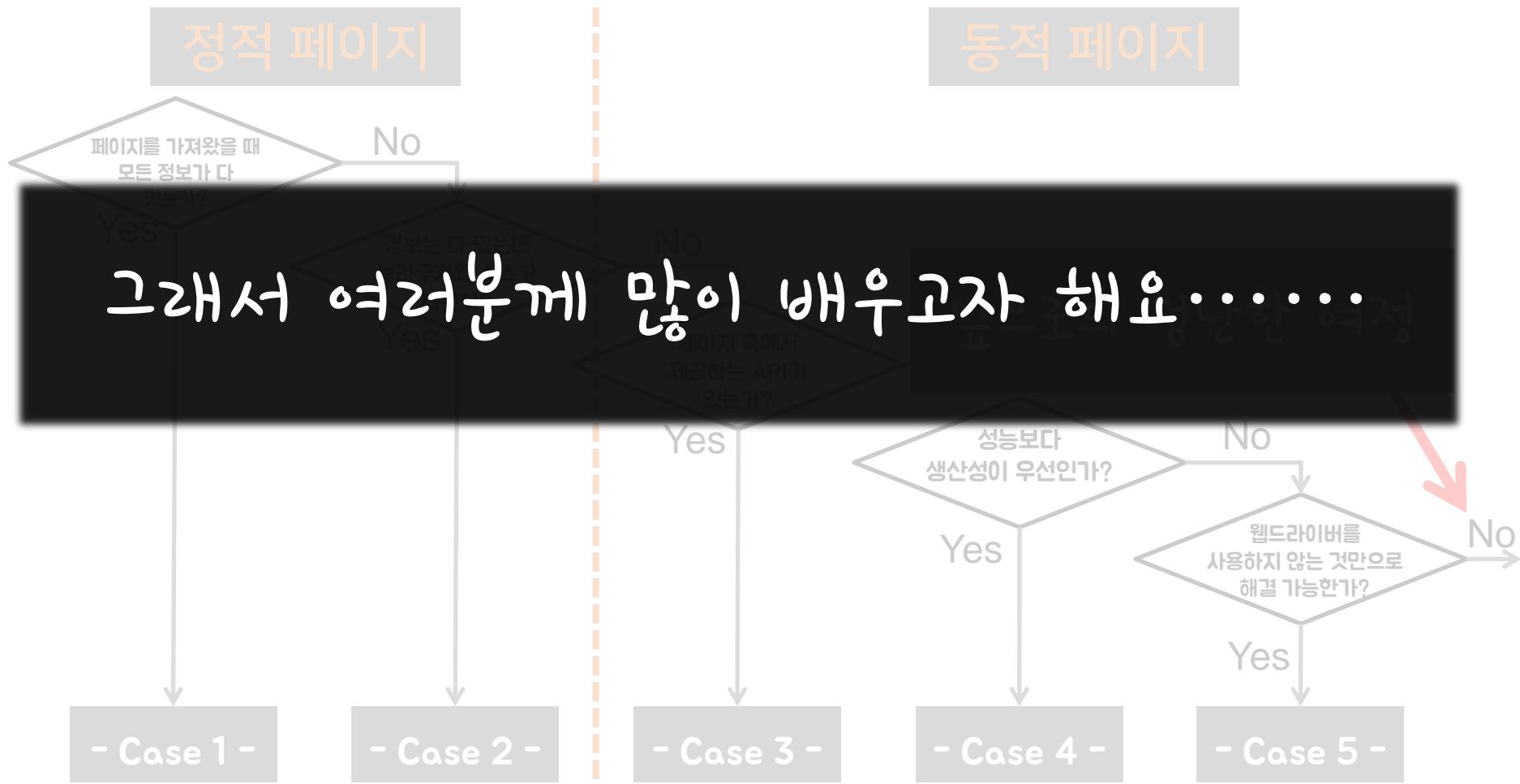
마무리

오늘 얘기하지 못한 것도, 앞으로 헤쳐나가야 할 것도 너무너무 많아요 ㅠㅠ



마무리

오늘 얘기하지 못한 것도, 앞으로 헤쳐나가야 할 것도 너무너무 많아요ㅠㅠ



마무리

잘 부탁 드립니다 π π

마무리

https://github.com/BbChip0103/2018_dataqanolja_webcrawling.git

Case 5의 소스코드는 여기서 확인하실 수 있어요 😊

