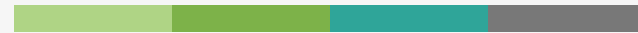


# CRAWLING

16기 DA팀 윤형준



# 크롤링이란?



Crawl : 1.(옆드려) 기다2.(곤충이) 기어가다3.기어가기, 서행



# 크롤링이란?



Crawl : 1.(옆드려) 기다2.(곤충이) 기어가다3.기어가기, 서행

Crawling: Web상에 존재하는 Contents를 수집하는 작업(프로그래밍으로 자동화 가능)

1. HTML 페이지를 가져와서, HTML/CSS등을 파싱하고, 필요한 데이터만 추출하는 방법
2. Open API를 제공하는 서비스에 Open API를 호출해서, 받은 데이터 중 필요한 데이터만 추출하는 기법
3. Selenium등 브라우저를 프로그래밍으로 조작해서, 필요한 데이터만 추출하는 기법



# 크롤링이란?



"웹 사이트에서 우리가 원하는 데이터를 수집하는 것"



# 크롤링이란?



"웹 사이트에서 우리가 원하는 데이터를 수집하는 것"

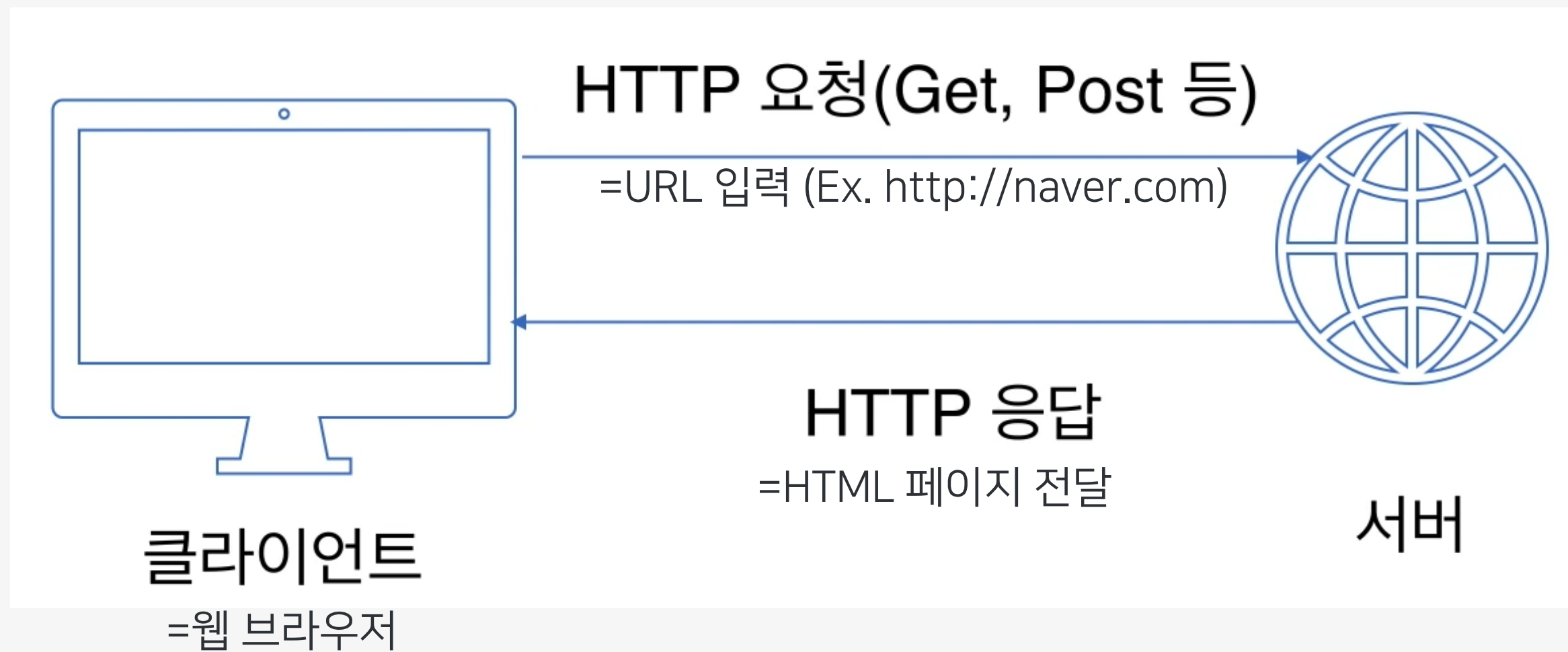
→ How?



# HOW?

HTTP (Hyper Text Transfer Protocol) : Hyper Text를 전송하기 위한 프로토콜(규약)

웹 문서를 구성하고 있는 언어인 HTML



# HOW?

HTML (Hyper Text Markup Language)

: 웹 사이트를 생성하기 위한 언어로 문서와 문서가 링크로 연결되어 있고, 태그를 사용하는 언어

HTML 문서의 기본 구조

<태그명 속성1 = '속성값1' 속성2='속성값2' />

<태그명 속성1 = '속성값1' 속성2='속성값2'> Value </태그명>

`<h3 id="articleTitle" class="tts_head">[날씨] 전국 대부분 지역에 폭염·열대야...내일 내륙 강한 소나기</h3>`

\* 우리가 원하는 값은 대부분 Value 위치에 있다!

# HOW?

```
<html>
<head>
  <title>BeautifulSoup test</title>
</head>
<body>
  <div id='upper' class='test'>
    <h3 title='Good Content Title'>Contents Title</h3>
    <p>Test contents</p>
    <img src='https://cdn.nba.net/nba-drupal-prod/styles/'
  </div>
  <div id='lower' class='test'>
    <p>Test Test Test 1</p>
    <p>Test Test Test 2</p>
    <p>Test Test Test 3</p>
  </div>
</body>
</html>
```

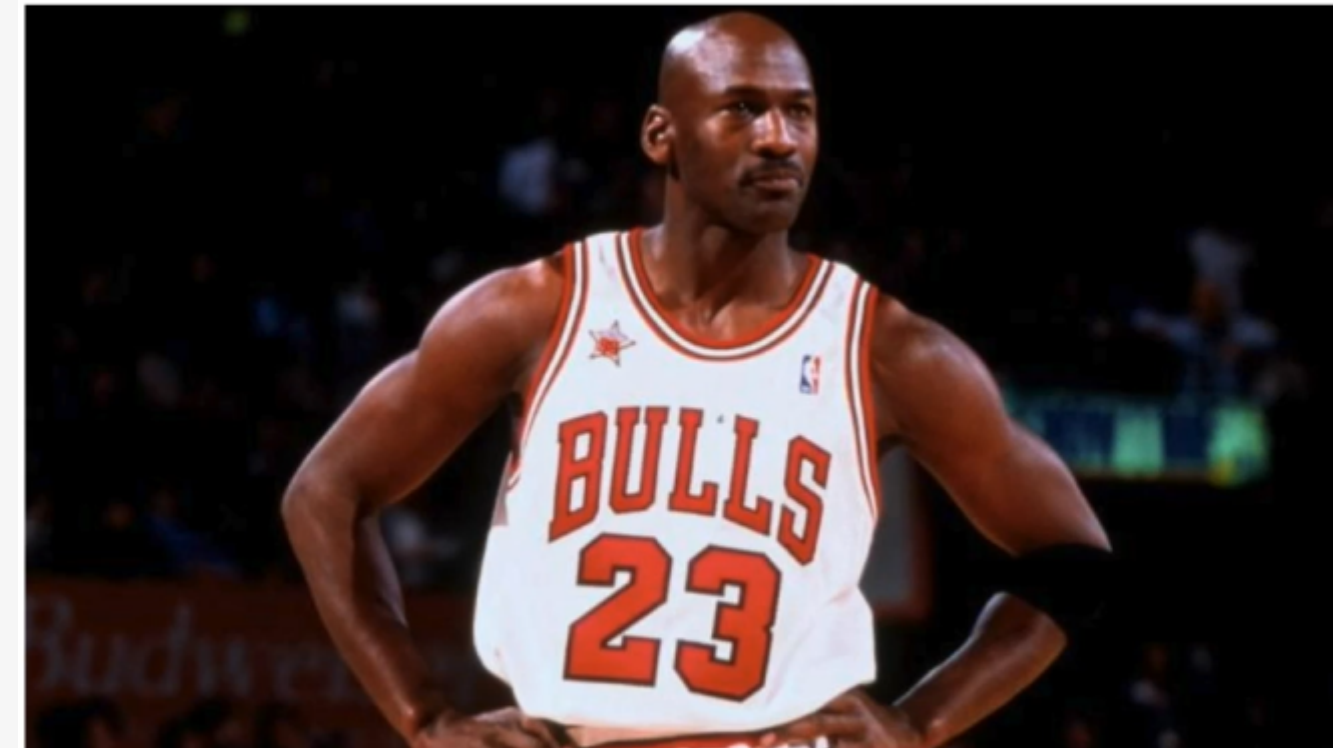
<html> : html 문서의 시작과 끝

<head> : 문서의 머리 (브라우저에 직접적으로 보이지 않음)

<body> : 문서의 콘텐츠 (브라우저에 직접적으로 보이는 부분)

Contents Title

Test contents



Test Test Test 1

Test Test Test 2

Test Test Test 3

<div> : 문서 구역(공간)을 나누는 기준

<h> : 머리글 (뒤의 숫자는 텍스트의 크기 결정)

<p> : 단락을 나누는 기준

\*크롤링 할때는 모든 태그의 역할을 알 필요는 없음



# HOW?



"웹 사이트에서 우리가 원하는 데이터를 수집하는 것"

→ How? HTML을 분석해서!



# HOW?



"웹 사이트에서 우리가 원하는 데이터를 수집하는 것"

→ How? HTML을 분석해서!

→ How?

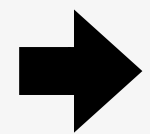


# HOW?

1



+ BeautifulSoup

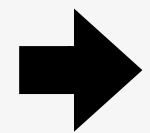


Requests:

HTTP Request를 웹 브라우저가 아닌 python에서 가능하게 해주는 모듈

ex) `resp = requests.get(https://www.naver.com/)`

`https://www.naver.com/` 라는 주소에 get 요청(request)를 하고 받은 응답(HTML 등)을 저장



BeautifulSoup:

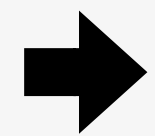
HTML과 같은 문서를 가져와 파싱(parsing)해주는 모듈

cf. parsing : <>와 같은 태그를 사용자가 입력하면 컴퓨터가 알아볼 수 있도록 바꿔주는 과정

Requests로 HTML 받아오고, BeautifulSoup로 파싱해서 HTML을 분석한다!

# HOW?

---



Selenium:

웹페이지 테스트 자동화용 모듈(실제 사용자가 사용하는 것 처럼 동작),  
크롤링을 할 때에도 사용이 가능하다

# HOW?



VS



+ BeautifulSoup

장점:

거의 모든 상황에서 정보를 받아올 수 있다.

단점:

속도가 느리다.

(직접 사용자가 사용하는 것 처럼 모든 정보를 다운로드 받으므로)

장점:

속도가 빠르다.

(HTML 태그를 통해 필요한 정보만 다운로드 받음)

단점:

정보를 받아오지 못하는 경우가 존재한다.

(외부에서 데이터를 받아오는 JS, iFrame 내의 정보를 가져올 수 없다.)

# HOW?



"웹 사이트에서 우리가 원하는 데이터를 수집하는 것"

→ How? HTML을 분석해서!

→ How? 분석 라이브러리를 이용해서



# HOW?



HTML 분석 모듈을 이용해 웹 사이트의 HTML을  
분석해서 원하는 정보를 찾아내는 것



# 실습



Selenium의 Webdriver를 사용하기 위해 'Chromedriver'설치

<https://sites.google.com/a/chromium.org/chromedriver/downloads>

(크롬 버전 확인 : Chrome://version)

```
chrome://version
```

```
Chrome: 84.0.4147.125 (공식 빌드) (64비트) (cohort: 84_Win_125)  
개정: d0784639447f2e10d32ebaf9861092b20cfde286-refs/branch-  
heads/4147@{#1059}
```










# 실습

## Current Releases

- If you are using Chrome version 85, please download [ChromeDriver 85.0.4183.38](#)
- If you are using Chrome version 84, please download [ChromeDriver 84.0.4147.30](#)
- If you are using Chrome version 83, please download [ChromeDriver 83.0.4103.39](#)
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

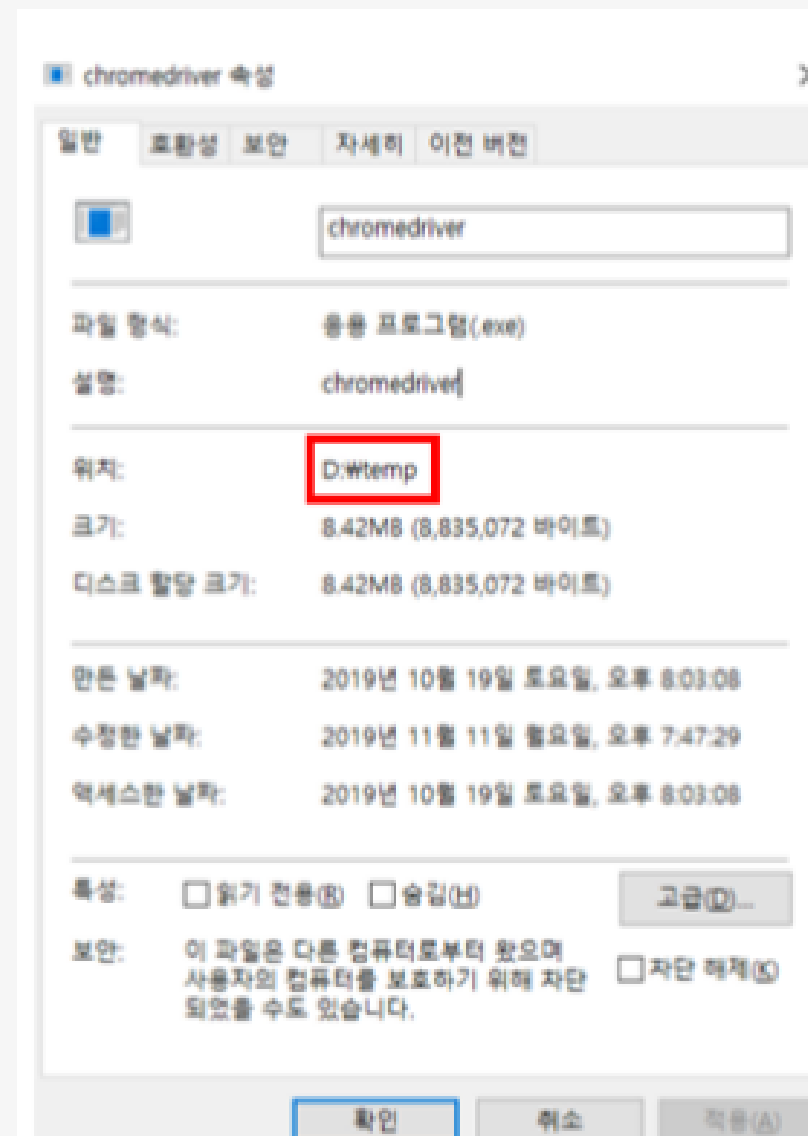
## Index of /84.0.4147.30/

	<u>Name</u>	Last modified	Size	ETag
	<a href="#">Parent Directory</a>		-	
	<a href="#">chromedriver_linux64.zip</a>	2020-05-28 21:05:07	5.06MB	beffb1bca07d8f4fd23213b292ef963b
	<a href="#">chromedriver_mac64.zip</a>	2020-05-28 21:05:09	6.99MB	b2ff30e148ae11a78e0f13e93b29f271
	<a href="#">chromedriver_win32.zip</a>	2020-05-28 21:05:11	4.63MB	3bf0e106a93382efd7a5bb3b55b182a6
	<a href="#">notes.txt</a>	2020-05-28 21:05:15	0.00MB	a505de7f878e415f1b06a44935f109bf

\*32/64 비트 상관 없음

# 실습

압축파일 해제 후 .exe 파일을 본인이 원하는 경로에 위치시키기



\*.exe 파일의 로컬 주소 필요

실습



실습시작!

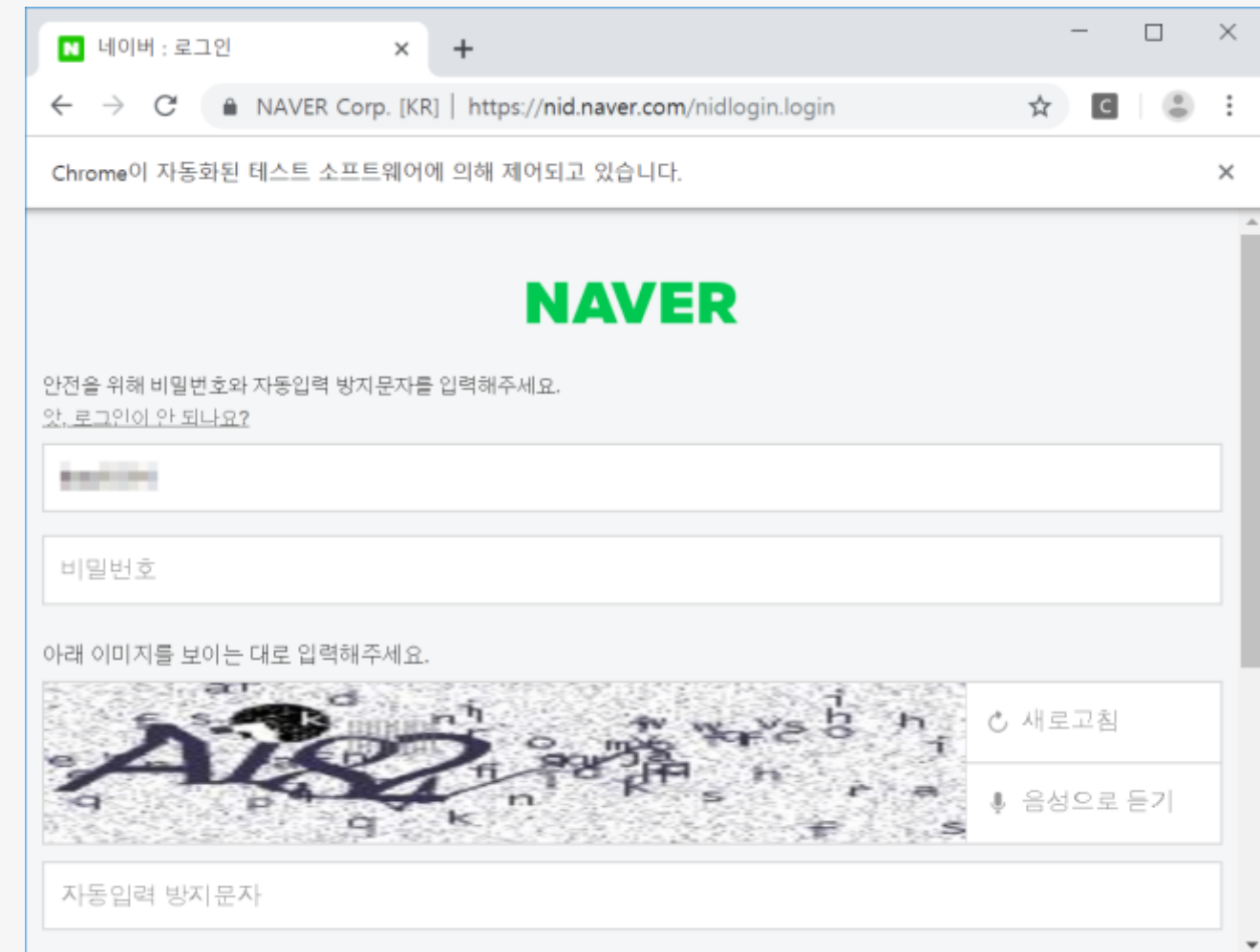
# ETC

네이버 검색 결과는 각 탭마다 제공되는 페이지수의 한계 존재



\*검색 날짜 설정 등을 통해 나누어서 검색할 수 있음

네이버는 봇을 이용한 로그인이 막혀있습니다  
(selenium으로도 로그인 불가능)



\*뚫는 방법도 존재

<https://neung0.tistory.com/34>

# ETC

## 크롤링 해도 되는 정보인가?

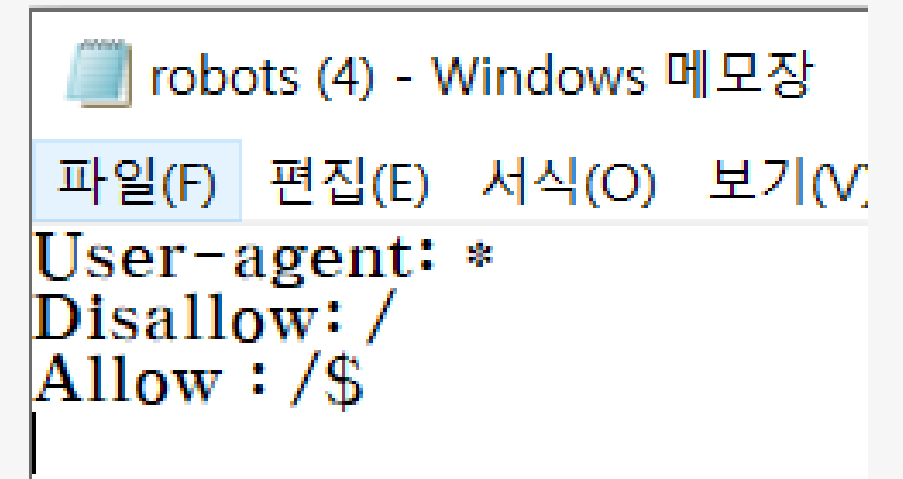
잡코리아, 무단 크롤링한 사람인HR에 승소...손해배상금 등 4억5000만원 받는다

메인 도메인 + /robots.txt를 통해 확인 가능

ex) <http://www.naver.com/robots.txt>

```
User-agent: Yeti
Disallow: /

User-agent: *
Disallow: /PostList.nhn
Disallow: /PostPrint.nhn
Disallow: /NBlogPostPreview.nhn
Disallow: /NBlogHidden.nhn
Disallow: /BlogInfo.nhn
Disallow: /PostExportDoc.nhn
Disallow: /PostPreview.nhn
Disallow: /NVisitor4Ajax.nhn
Disallow: /NVisitor4Ajax.nhn
Disallow: /NBuddyList.nhn
Disallow: /WidgetListAsync.nhn
Disallow: /socialapp/SocialAppAppBoxMyAppListAsync.nhn
Disallow: /buddy/
Disallow: /export/
Disallow: /common/
Disallow: /post/
Disallow: /npost/
Disallow: /main/
Disallow: /guestbook/
Disallow: /intro.nhn
Disallow: /history.nhn
Disallow: /comment.nhn
Disallow: /socialapp/
Disallow: /upload/
Disallow: /connect/
```



[https://ko.wikipedia.org/wiki/%EB%A1%9C%EB%B4%87\\_%EB%B0%B0%EC%A0%9C\\_%ED%91%9C%EC%A4%80](https://ko.wikipedia.org/wiki/%EB%A1%9C%EB%B4%87_%EB%B0%B0%EC%A0%9C_%ED%91%9C%EC%A4%80)

위키피디아 - 로봇 배제 표준

# ETC

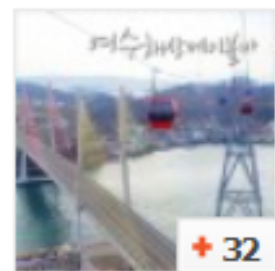
## 특정 검색어에 대한 블로그, 뉴스 등 본문을 보고 싶을때는?

[여수 여행 -- 오동도 가는 길](#) 어제

해결하고</p><p>다시 여수로 달려 갑니다. 여수로 가는 길이... 갑니다</p><p>&nbsp;</p><p>여수 여행

- 오동도 앞 바다 보트...

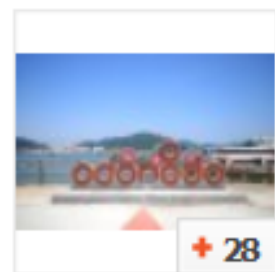
하늘하늘의 공간사랑 [blog.daum.net/csmsjy/7092677](http://blog.daum.net/csmsjy/7092677) | 블로그 내 검색



[전남여행 여수 해상케이블카 자산공원~돌산탐승장 왕복..](#) 2020.04.14.

전남여행 갈 곳도 많고 좋은 곳도 많지만 여수도 지나치면 섭섭할 곳이죠. 여수여행을 하면서 숙박여행을 한다면 고민거리 하나가 숙소를 선택하는 것일텐데요....

♥마리안의 여행이야기-마음... [anndam.blog.me/22190...](http://anndam.blog.me/22190...) | 블로그 내 검색 | 약도 ▾



[여수 오동도 산책하기 좋은 여행 코스](#) 어제

아니더라도 여수 오동도는 다양한 볼거리가 있고 푸른 자연 속에서 산책하기에 좋은 곳이라 한번쯤 다녀오면 참 좋은것 같더라고요. 물론 저는 다음 여행에서는 꼭 봄에...

둔켈의 여행 그리고 이야기 [blog.naver.com/uni77...](http://blog.naver.com/uni77...) | 블로그 내 검색 | 약도 ▾

```
<a class="sh_blog_title _sp_each_url _sp_each_title"
href="http://blog.daum.net/csmsjy/7092677" target=
_blank" onclick="return goOtherCR(this,
'a=blg*d.tit&r=1&i=a00000fa_0d16dc3cf7aa8954c087bb92&u='
+urlencode(this.href))" title="여수 여행 -- 오동도 가는
길">...</a> == $0
```

사실 이렇게 본문에 접근해도 블로그 html 구조가 조금씩 다른경우가 있습니다.. 따로 설정해 줘야됨.....

# ETC

---

iframe 등으로 막혀있어서 정보가 안보일 때는  
모바일 사이트로 접근해보기!  
(실습파일 ㄱㄱ)

# ETC

---

사실 얘기하자면 끝도 없는 예외상황들...



# ETC

---

사실 얘기하자면 끝도 없는 예외상황들...

\*DA팀 세션때 더 자세히 다룰꺼예요~

DA팀으로 오세요~

# 과제

---



---

네이버 검색창에 '여수' 검색해서 나오는 블로그 제목  
20개(1,2페이지) 크롤링한 화면 캡처해서 제출  
(두가지 모듈 각각 따로)