

Data Crawling 1



| 목차

01 Crawling?

- 크롤링이란? / Robots.txt / HTML, CSS, Javascript

02 HTML / CSS

- HTML / HTML 문서의 기본 구조 / 태그 별 주요 요소 / class와 id / CSS

03 정규표현식

- 정규표현식



Crawling?



Crawling

자동으로 웹페이지 데이터를 수집하는 행위

NAVER 증권

종목명 지수명 입력

Q

통합검색

로그인

증권 홈

국내증시

해외증시

시장지표

리서치

뉴스

MY

최근조회종목

MY STOCK

최근조회종목이 없습니다.

실시간속보

더보기

지앤비에스 예코, 한화큐셀과 45억원 태양광 장...
우리자산운용, 2년 연속 ESG펀드 인증 따내
한국투자증권, 바이샬스탠다드와 토근증권 상품 ...
테스트테크, 우수 인재 지속적 채용 확대 ...“반도...
KB증권, ‘계좌개설한 김에 KB증권 중개형ISA’ 이...
‘홍콩오피스 손실’ 해외부동산 우려 고조...금감원...

TOP종목

더보기

거래상위	상승	하락	시가총액 상위
자연과환경	1,678	▲ 325	+24.02%
KODEX 200선물인버스2X	2,430	▲ 15	+0.62%
삼부토건	5,010	▲ 215	+4.48%
KODEX 코스닥150선물...	3,800	-	0.00%
헬바이도텍	4,030	▲ 670	+19.94%
네오셀	6,480	▲ 1,140	+21.35%
에스유흐olding스	1,020	▼ 219	-17.68%
폴라리스오피스	4,175	▲ 960	+29.86%
대유펙러스	1,059	▲ 72	+7.29%
웨이버스	1,845	▲ 108	+6.22%

오늘의 증시

실시간 2023.07.17 13:57 장중

코스피

2,616.71 ▼11.59 -0.44%

개인 +3,877 외국인 -2,756 기관 -1,100 (억원)

↑ 1 ▲ 341 - 42 ▼ 550 ↓ 0

코스닥

898.13 ▲1.85 +0.21%

코스피 200

345.66 ▼1.27 -0.37%

업종상위

더보기

1	화학	POSCO홀딩스	신스틸
	+4.57%	+6.82%	+5.45%
2	소프트웨어	폴라리스오...	오브젠
	+3.43%	+29.86%	+28.86%
3	IT서비스	삼성에스디..	토마토시스..
	+3.08%	+9.55%	+6.85%

해외 증시

더보기

다우산업(07.14)

34,509.03 ▲ 113.89

나스닥(07.14)

14,113.70 ▼ 24.87

홍콩H(07.14)

6,558.88 ▲ 14.97

상해종합(07.17)

3,199.17 ▼ 38.53

니케이225(07.14)

32,391.26 ▼ 28.07

한화투자증권

미리 챙기는 해외주식 혜택

(~8/31)

최대 \$30 (미리만) 이상 (가점)

수수료 0원 (완전무대)

※ 이벤트 신청 완료 고객 대상 조건 충족 시, 투자 전 실적 향상 및 상품에서 이익률 특별 예금자보유금보상금 이상 보 지급 가능
※ 선물, 환율변동 등에 따른 환금손실 0-100% 발생 가능 및 투자지 위수
※ 온라인 표준 수수료는 0.25% (미국 기준, 인도시 0.0008% 제외)을 부과하며, 홈페이지 참조 ※ 한국금융투자협회 심사일 제23-0234호
2023-07-01~2023-08-31

인기 검색 종목

더보기

1.	POSCO홀딩스	477,500	▲ 30,500
2.	금양	110,100	▲ 8,800
3.	에코프로	1,001,000	▲ 13,000
4.	삼부토건	5,010	▲ 215
5.	삼성전자	73,200	▼ 200

danawa

비교하고 잘 사는, 다나와

보수 EVENT! 골드바 당첨의 주인공은?

Q

최근

관심

로그인

전체 카테고리

자동차

조립PC

PC견적

기업구매상담

여행

쇼핑기획전

DPG

이벤트/체험단

더보기

한 눈에 보는 가구 쇼핑 가이드

가전-TV

컴퓨터-노트북-조립PC

태블릿-모바일-디카

아웃도어스포츠-골프

자동차-용품-공구

가구-조명

식품-유아완구

생활-주방-건강

패션-잡화-뷰티

사무-취미-반려동물

여행-항공-호텔

danawa

중고 노트북도 이젠, 안심하고 다나와!

합리적인 가격으로 구성된

다나와 인증 중고 노트북 한정 판매!

WIN10 정품 설치로 수령 후 바로 사용 가능!

보러가기

이네오스, 더블 캡 콕업 트럭
'올-뉴 그랜저' 디어 퀵터마...
론진, 새로운 '플래그십' 엘리...
터치 컬렉션' 공개...시대들...
퍼실, 탑재된 4in1 디스크 클...
러 캡슐제 파우치 26개입...

쇼핑정보

주요이슈

컴퓨터

테크

자동차

사용기

HIT브랜드

4/6

드디어 나왔다. 아이오닉 5 N

BMW 미니병 완치할 수 있을까?

아우디 역대급 엔진이 들어갔다?

레이 EV 4년 만에 돌아왔다

실내외 부분변경? 이견 못 함지!

이제 바로 말트만 뜯

주형 별란스 잡은 입문용 바이크!

뜨겁다! 데이브더다이버

에어컨 빠르게 사라 가기

구매 인증하고 만 원 받기

쾌적한 여름을 위한 리빙템!

보수 골드바 당첨 기회!

가성비 워치를 찾아라!

삼성 스마트 모니터

M7 S27CM701 체험단

체험단

삼성전자

포터블 SSD T7 Shield (1TB)

체험단

질만 Z10 PLUS

체험단 모집

OUTTA

야놀자, 크롤링 논란 여기어때 상대로 민사 이겼다

보도1팀 | 승인 2022.08.26 16:22 | 댓글 0



국내를 대표하는 숙박 플랫폼의 양대산맥 격인 야놀자와 여기어때가 크롤링을 이용한 정보수집 일*, 크롤링 논란으로 장기적인 법적 공방이 이어지고 있는 가운데 야놀자가 민사소송에서 승기를 잡으면서 귀추가 주목되고 있다.

이코노믹 리뷰 매체에 따르면 야놀자와 여기어때가 크롤링 논란으로 수년간 공방을 펼치는 가운데, * 일 민사소송에서 야놀자가 여기어때를 대상으로 민사소송 2심 항소심에서 원고일부승 판결을 끌어냈다고 전했다.

야놀자가 여기어때를 대상으로 2016년 6월부터 10월까지 크롤링 프로그램을 이용해 야놀자 콘텐츠의 정보를 수집한 것을 문제 삼으면서 갈등이 시작됐다. 크롤링이란 자동으로 웹페이지 데이터를 수집하는 행위를 뜻하며 여기어때가 크롤링 프로그램을 통해 야놀자의 가격 정보와 숙박 업소 목록 및 주소 등의 정보를 수집한 것으로 드러나면서 법적 공방이 이어져 왔다.



IP 차단



사이트에 연결할 수 없음

stopit에 오타가 있는지 확인하세요.

DNS_PROBE_FINISHED_NXDOMAIN

새로고침

coupang

您没有权限访问此页面。如需帮助，请发送电子邮件至helpseller_global@coupang.com。

You don't have permission to access this page. Please contact us for assistance at helpseller_global@coupang.com.

Back to previous page

Go to homepage

Reference : 18.8e973b17.1636340099.178ec268

Client IP : 203.128.163.99

Path : <https://www.coupang.com/>

Time : 2021. 11. 8. 11:54:59 GMT+0900 (한국 표준시)

UserAgent : Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/95.0.4638.69 Safari/537.36

© Forward Ventures Co.,Ltd. All rights reserved.



Robots.txt와 사용자 에이전트(user agent)

Robots.txt

- 웹 사이트 및 웹 페이지를 수집하는 로봇들의 무단 접근을 방지하기 위해 만들어진 로봇 배제 표준/국제 권고안
 - 일부 스팸 봇이나 악성 목적을 지닌 가짜 클라이언트 로봇은 웹 사이트에 진짜 클라이언트처럼 접근
 - 무단으로 웹 사이트 정보를 훑어가거나, 웹 서버에 부하를 줌
- > 이런 로봇들의 무분별한 접근을 통제하기 위해 마련

User agent

- 웹 서버에 요청을 보내도 요청을 거부 당하는 경우 발생 -> 무단 봇으로 짐작하고 웹 서버에서 접근을 막는 것
 - 우리가 스팸 봇이 아니라 사람이라는 것을 브라우저에게 알려줘야 함
- > 이때 브라우저에게 전달하는 것이 사용자 에이전트 정보

서버에 과도한 부하를 주지 않는다.

가져온 정보를 사용할 때(특히 상업적으로) 저작권과 데이터베이스권에 위배되지 않는지 주의한다.



Robots.txt와 사용자 에이전트(user agent)

```
# robots.txt for http://www.danawa.com/  
  
User-agent: HMSE_Robot  
Disallow: /  
  
User-agent: bingbot  
Crawl-delay: 3600  
  
User-agent: *  
Disallow: /user_report/  
Disallow: /elec/Management  
  
Sitemap: https://www.danawa.com/seo_data/www/WWW_main.xml
```

[사이트 URL/robots.txt](#)



HTML



Markup Language
Content

CSS



Style sheet Language
Presentation

JS



Programming Language
Behavior



What's the Difference?



HTML

Hypertext Markup Language

Create the structure

- Controls the layout of the content
- Provides structure for the web page design
- The fundamental building block of any web page



CSS

Cascading Style Sheet

Stylize the website

- Applies style to the web page elements
- Targets various screen sizes to make web pages responsive
- Primarily handles the "look and feel" of a web page



Javascript

Increase interactivity

- Adds interactivity to a web page
- Handles complex functions and features
- Programmatic code which enhances functionality





HTML/CSS



HTML

문서나 데이터의 구조를 표현하는 웹 페이지를 위한 마크업 언어

HTML 문서의 기본 구성 : 태그와 속성

```

```

①

②

- ① 태그 : HTML에서 콘텐츠를 표현하거나 처리하기 위해 사용하는 명령어 / <tag>
- ② 속성 : 태그보다 구체화된 명령어 체계 / 속성 = 속성값의 형태로 표현



HTML

오픈 태그

```
<div id="u_skip">
```

태그 안의 태그

```
  <a href="#menu" tabindex="1">  
    <span>메인 메뉴로 바로가기</span>  
  </a>  
  <a href="#start" tabindex="2">  
    <span>본문으로 바로가기</span>  
  </a>
```

클로징 태그

```
</div>
```



HTML 문서의 기본 구조

```
① <html>  
  <head>  
    <title>HTML 문서</title>  
    <meta charset = 'utf-8'>  
  </head>  
  <body>  
    <b>Hello World</b>  
  </body>  
② </html>
```

- ① 문서에서 가장 먼저 사용, 해당 문서가 HTML 언어 사용했음을 나타냄
- ② <html> 으로 문서 시작, </html>로 끝을 알림
- ③ <head> 웹 페이지를 전체적으로 아우르는 기본 내용 (화면에 표시되지 않는 내용)
- ④ <body> 실제로 화면에 표현되는 내용



<head> 태그와 주요 요소

```
<html>
  <head>
    <title>HTML 문서</title>
    <meta charset = 'utf-8'>
  </head>
```

① <title>

- <title> 문서 제목 </title> 형태로 표현
- 문서의 제목을 나타냄



<head> 태그와 주요 요소

```
<html>
  <head>
    <title>HTML 문서</title>
    <meta charset = 'utf-8'>
  </head>
```

② <meta> : 문서에 대한 정보 포함

<meta charset = 인코딩 방식> : 문서에서 사용된 인코딩 방식 표시

- 어떤 문자들을 깨지지 않게 출력하고 취급할 수 있는지 나타냄
- Utf-8과 같이 여러 종류의 문자 집합을 취급할 수 있는 코드가 아니라면 문자가 깨져서 표현되기 때문

<body> 태그와 주요 요소

(1)

- 텍스트 내의 줄바꿈을 나타낼 때 사용
- HTML 파일에서는 엔터를 사용해 줄바꿈 하더라도 화면에 적용 X
- 단독 태그로 따로 클로징 태그가 필요 X

(2)

- 텍스트를 굵게 나타내기 위해 사용

(3) <p>

- 문단을 정의할 때 사용하는 태그
- 브라우저는 자동으로 p 태그 안 콘텐츠의 위쪽과 아래쪽에 약간의 여백을 추가함



<body> 태그와 주요 요소

(4) <h1> ~ <h6>

- 제목을 정의할 때 사용
- H 뒤쪽의 숫자가 커질수록 중요도는 작아짐

(5) <a>

- 하나의 페이지에서 다른 페이지나 문서를 연결할 때 사용

(6)

- 이미지 삽입 시 사용되는 태그

(7) <div>

- CSS와 함께 활용되며 웹 사이트의 전체적인 레이아웃을 만드는 데 사용



<body> 태그와 주요 요소

(8) <table>

- 행과 열로 구성된 표 테이블을 정의할 때 사용

① <thead> : 각 열의 타이틀과 관련된 부분

- <th> : 각 열의 타이틀 입력

② <tbody> : 타이틀 제외 본문 내용

- <td> : 각 행의 열, 셀 속에 들어가는 데이터

③ <tr> : 테이블의 행을 생성하는 태그

```
<table>
  <thead>
    <tr>
      <th>이름</th>
      <th>성적</th>
    </tr>
  </thead>
  <tbody>
    <tr>
      <td>김김김</td>
      <td>90</td>
    </tr>
    <tr>
      <td>이이이</td>
      <td>92</td>
    </tr>
  </tbody>
</table>
```

이름	성적
김김김	90
이이이	92



class와 id

(1) class 속성

- 반복되는 태그들을 유형별로 분류하고 싶을 때 사용
- 크롤링을 하는 데에 키워드가 되는 경우가 많음
- .으로 선택

(2) id 속성

- 특정 요소에 이름을 붙이는 데에 사용, 중복 불가
- 크롤링 하는 데에 종종 쓰일 수 있음
- #으로 선택

```
<div id = "wrap">  
  <div class = "home_spot view_off">  
    <div class = "bx_spot">  
    </div>  
  </div>  
</div>
```



CSS

style 태그 확인

- 태그 안에 적용된 property 확인하기
- property: value;
property: value; 형태로 입력
- 데이터를 가져오는 데에는 그렇게 중요하지는 않음

```
<div id = "wrap">
  <div class = "home_spot view_off">
    <div class = "bx_spot">
      </div>
    </div>
  </div>

<style scope = "iron-ally-announcer">
  iron-ally-announcer {
    display: inline-block;
    position: fixed;
    clip: rect(0px, 0px, 0px, 0px);
  }
</style>
```

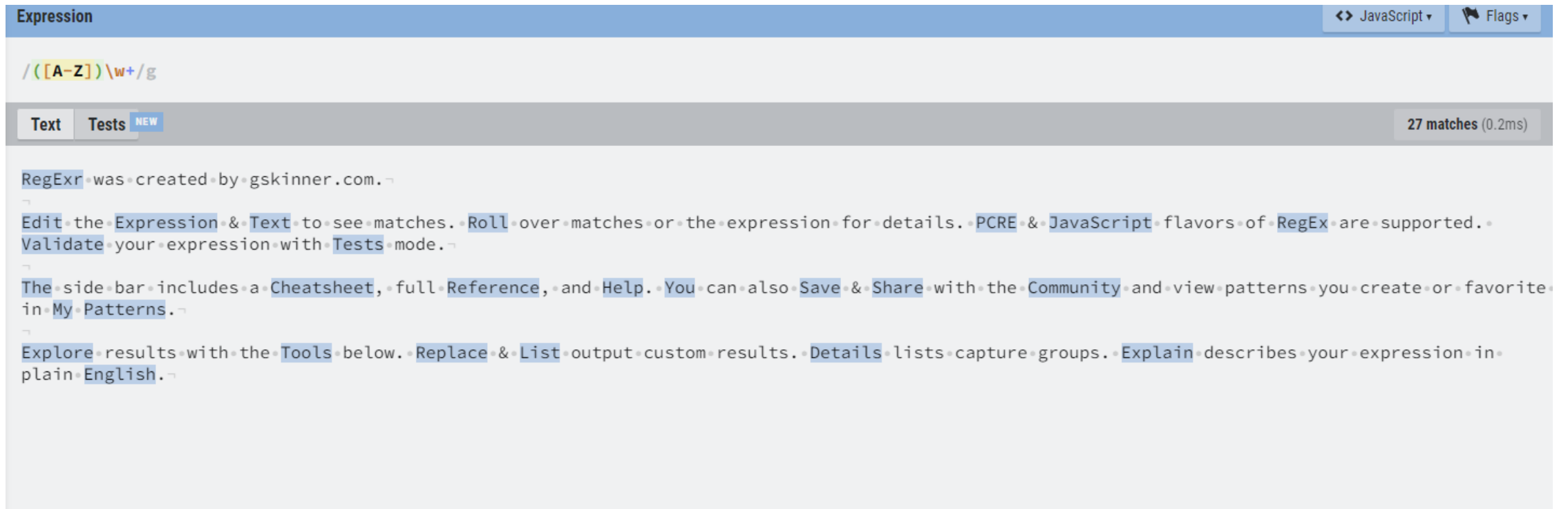


정규표현식



정규표현식

정규표현식을 이용해 정교하게 원하는 데이터를 추출 / 삭제 / 변환 가능



The screenshot shows the regexr.com interface. At the top, there's a header with 'Expression' and a dropdown menu set to 'JavaScript'. Below this, the regular expression `/([A-Z])\w+/g` is entered. The 'Text' tab is selected, and the test text is: `RegExp was created by gskinner.com. Edit the Expression & Text to see matches. Roll over matches or the expression for details. PCRE & JavaScript flavors of RegEx are supported. Validate your expression with Tests mode. The side bar includes a Cheatsheet, full Reference, and Help. You can also Save & Share with the Community and view patterns you create or favorite in My Patterns. Explore results with the Tools below. Replace & List output custom results. Details lists capture groups. Explain describes your expression in plain English.` The results show 27 matches in 0.2ms.

<https://regexr.com/>



Dot, 반복

Dot

- `.`(Dot) = 문자 하나 (숫자, 특수문자 포함)

? 는 앞 문자가 0번 또는 1번 표시되는 패턴

* 는 앞 문자가 0번 또는 그 이상 반복되는 패턴

+ 는 앞 문자가 1번 또는 그 이상 반복되는 패턴

{n} 는 앞 문자가 n번 반복되는 패턴

{m, n} 는 앞 문자가 m번 반복되는 패턴부터 n번 반복되는 패턴까지



괄호와 하이픈

[] 괄호 안에 들어가는 문자가 들어 있는 패턴

Ex) [abc] 는 a, b, c 중 하나가 들어 있는 패턴을 의미

하이픈(-)을 이용하면 알파벳 전체를 나타낼 수 있음

Ex) [a-c] 는 a, b, c 중 하나가 들어 있는 패턴을 의미

() 괄호는 괄호 안에 있는 단어 자체를 반환함

Ex) (abc) 는 abc가 들어 있는 패턴을 의미

Expression	
/[a-c]+/g	
Text	Tests
ab acb accccc afdcadsabc	

Expression	
/[abc]+/g	
Text	Tests
ab acb accccc afdcadsabc	

Expression	
/(abc)/g	
Text	Tests
ab acb accccc afdcadsabc	



정규표현식 라이브러리 함수 사용법

Match : 문자열 처음부터 정규식과 매칭되는 패턴을 찾아서 리턴

Search : 문자열 전체를 검색해서 정규식과 매칭되는 패턴을 찾아서 리턴

Findall : 정규표현식과 매칭되는 모든 문자열을 리스트 객체로 리턴

Split : 찾은 정규표현식 패턴 문자열을 기준으로 문자열을 분리

Sub : 찾은 정규표현식 패턴 문자열을 다른 문자열로 변경





감사합니다

