

파이썬 웹 크롤링

강사장철원

파이썬 웹 크롤링

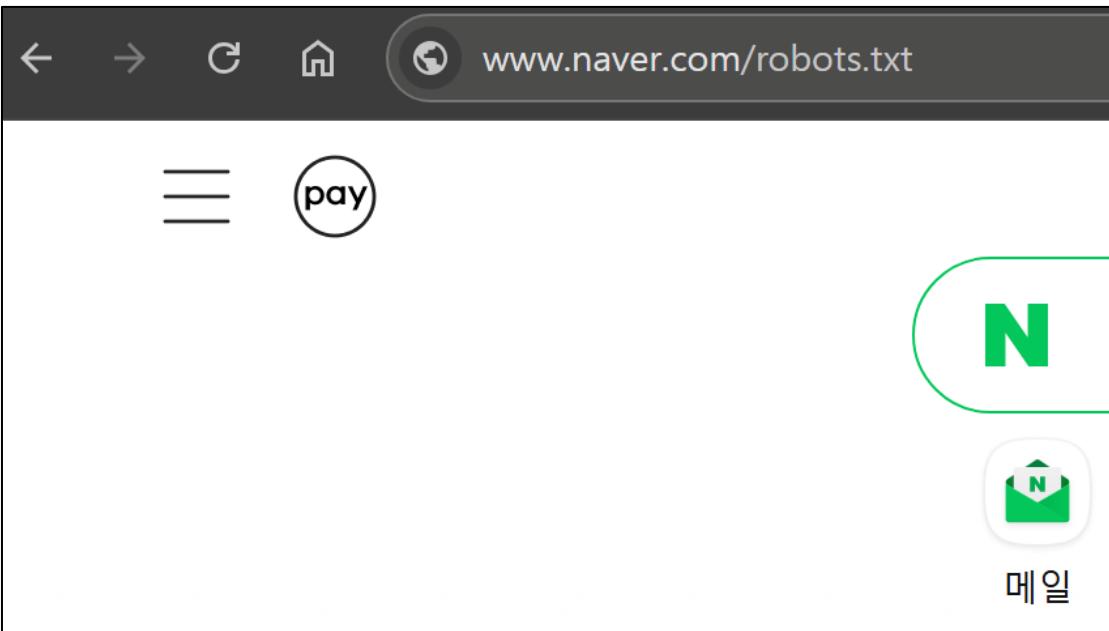
Section 3. 크롤링 기초

Section 3-2. 웹 크롤링 기초

크롤링 전에는 크롤링 허가된 사이트인지 확인(1)

<https://www.naver.com/robots.txt>

robots.txt 파일 확인



robots.txt

파일	편집	보기
User-agent: *	모든 크롤러 대상	
Disallow: /	루트의 하위 경로 모두 크롤링 금지	
Allow : /\$	루트 경로('/) 자체만 허용	
Allow : ./well-known/privacy-sandbox-attestations.json	해당 경로의 파일 접근 허용	

크롤링 전에는 크롤링 허가된 사이트인지 확인(2)

<https://www.danawa.com/robots.txt>



☰ 전체 카테고리

자동차 조립PC

danawa.com/robots.txt

User-agent: *	모든 크롤러 대상
Disallow: /user_report/	크롤링 금지 페이지
Disallow: /elec/Management	
User-agent: HMSE_Robot	HMSE_Robot은 크롤링 금지
Disallow: /	
User-agent: bingbot	bingbot은 한 시간에 한 번만 요청 가능
Crawl-delay: 3600	
User-agent: Mediapartners-Google	
User-agent: Googlebot	
User-agent: Googlebot-image	
User-agent: NaverBot	
User-agent: Yeti	
User-agent: Daumoa	
User-agent: Twitterbot	
Crawl-delay: 1	크롤링 간격 1초 이상 둘 것
Sitemap:	https://www.danawa.com/sitemap.xml

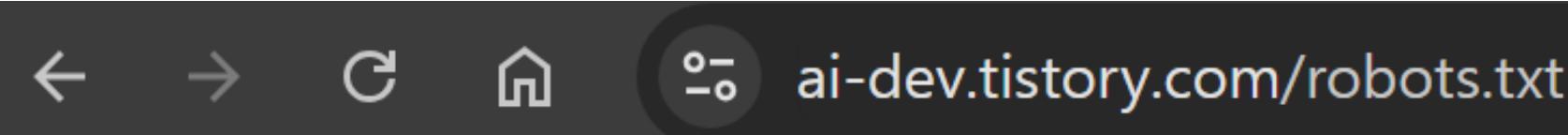
Section

웹크롤링기초

크롤링 사이트 확인 <https://ai-dev.tistory.com/1>



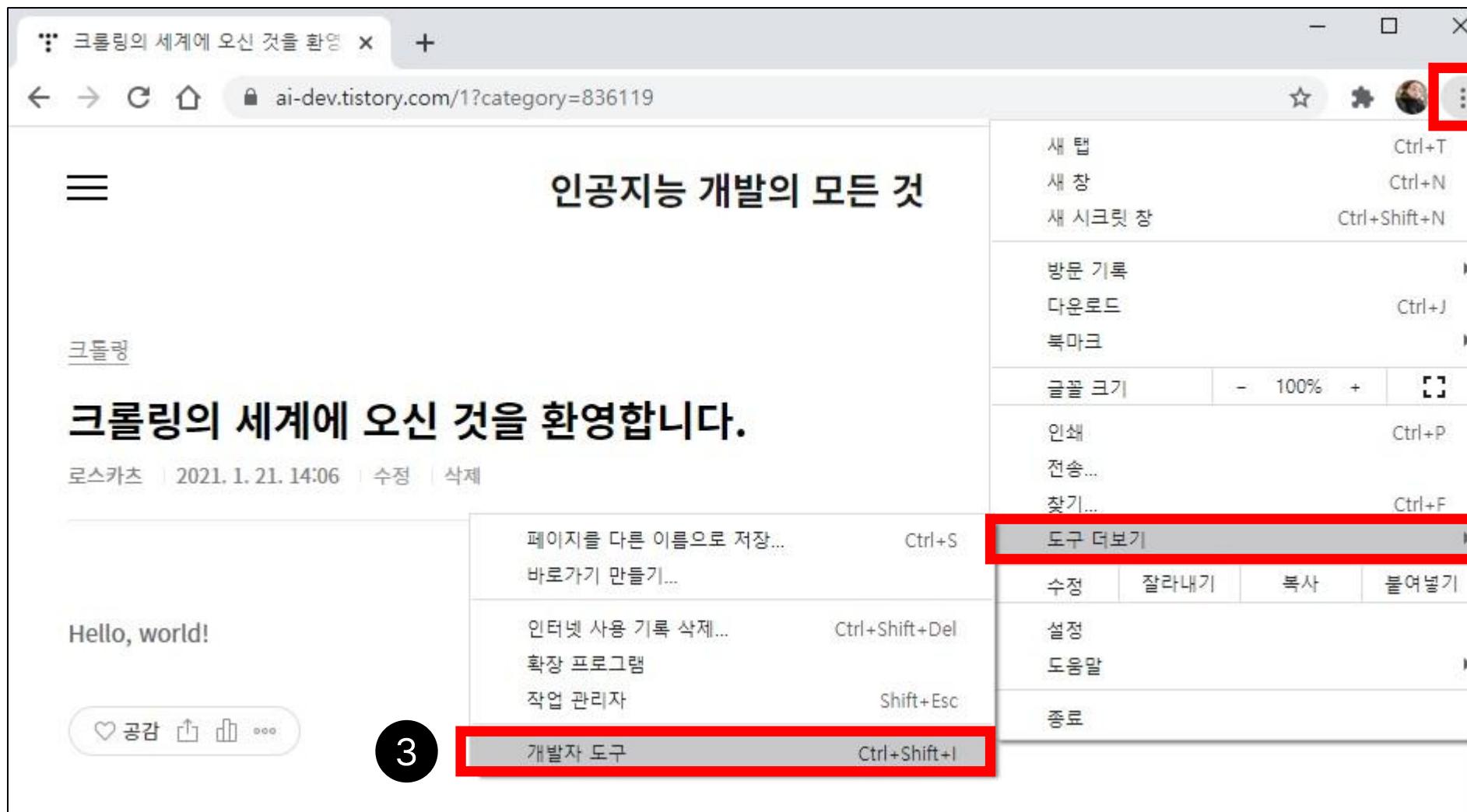
크롤링 허가 확인 <https://ai-dev.tistory.com/robots.txt>



```
User-agent: *
Disallow: /guestbook
Disallow: /m/guestbook
Disallow: /manage
Disallow: /owner
Disallow: /admin
Disallow: /search
Disallow: /m/search
```

```
User-agent: bingbot
Crawl-delay: 20
```

크롤링 예제 1-1 - urllib 사용



1

또는 F12

2

3

Section

웹크롤링기초

크롤링 예제 1-1

The screenshot shows a web browser window displaying a blog post from 'ai-dev.tistory.com'. The title of the post is '크롤링의 세계에 오신 것을 환영합니다.' (Welcome to the world of crawling). The post content is 'Hello, world!'. Below the content are social sharing icons (heart, comment, share) and a link to another post in the same category. The date of the post is 2021.01.21. On the right side of the browser, the developer tools are open, specifically the 'Elements' tab. The DOM tree shows the HTML structure of the page, including the header, body, and various script tags. Several script tags have their sources highlighted in blue, indicating they are being loaded from external CDNs.

```
...<!DOCTYPE html> == $0
<html lang="ko">
  <head>...
    <body id="tt-body-page" class="color-gray post-type-text paging-view-more">
      <script type="text/javascript">...</script>
      <script type="text/javascript" src="https://t1.daumcdn.net/tistory_admin/assets/blog/tistory-0a7992ffde7...cript-owner.js?version=tistory-0a7992f..."></script>
      <script type="text/javascript" src="https://t1.daumcdn.net/tistory_admin/assets/blog/tistory-0a7992ffde7...blog/common.js?version=tistory-0a7992f..."></script>
      <div style="margin:0; padding:0; border:none; background:none; float:none; clear:none; z-index:0"></div>
      <div id="acc-nav">...</div>
      <div id="wrap">...</div>
      <script src="https://tistory4.daumcdn.net/tistory/0/pg_Poster/images/script.js"></script>
      <script src="//search1.daumcdn.net/search/statics/common/g/search_dragselection.min.js"></script>
      <script>...</script>
      <script type="text/javascript" src="//t1.daumcdn.net/tiara/js/v1/tiara.min.js"></script>
      <script type="text/javascript">...</script>
      <script type="text/javascript" src="https://t1.daumcdn.net/tistory_admin/assets/blog/tistory-0a7992ffde7...tiara.min.js?version=tistory-0a7992f..." defer></script>
      <script type="text/javascript">...</script>
      <script type="text/javascript" src="//t1.daumcdn.net/midas/rt/dk_bt/roosevelt_dk_bt.js" async></script>
      <script type="text/javascript">...</script>
      <iframe style="position: absolute; width: 1px; height: 1px; left: -100px; top: -100px" src="//ai-dev.tistory.com/api" id="editEntry">...</iframe>
      <div id="tistoryEtcLayer" class="layer_post" style="display: none;"></div>
      <div id="tistorySnsLayer" class="layer_post" style="display: none;">...</div>
    ... <!doctype>
  ...
```

크롤링 예제 1-1

The screenshot illustrates a web development environment with three numbered callouts highlighting specific features:

- 1**: A red box highlights the "Elements" tab in the developer tools toolbar.
- 2**: A red box highlights the "크롤링" (Crawling) heading in the browser's sidebar.
- 3**: A red box highlights the "크롤링의 세계에 오신 것을 환영합니다." (Welcome to the world of crawling) text in the browser's content area, which is also selected in the Elements tab's DOM tree.

The browser window displays a blog post with the following details:

- Title:** 인공지능 개발의 모든 것
- Header:** 크롤링
- Text:** 크롤링의 세계에 오신 것을 환영합니다.
- Meta Information:** 2021. 1. 21. 14:06 | 스판 | 삭제
- Content Summary:** 크롤링 707 × 38
- Color:** #000000
- Font:** 27px "Noto Sans KR", sans-serif
- Margin:** 0px 0px 9px
- Accessibility:** Aa 21 (green checkmark)
- Name:** 크롤링의 세계에 오신 것을 환영합니다.
- Role:** heading
- Keyboard-focusable:** (checkbox)

The developer tools' Elements tab shows the following DOM structure for the selected element:

```
<!DOCTYPE html>
<html lang="ko">
  <head>...</head>
  <body id="tt-body-page" class="color-gray post-type-text paging-view-m...
    <script type="text/javascript">...</script>
    <script type="text/javascript" src="https://t1.daumcdn.net/tistory_a...
      s? version =tistory-aa685ca..."></script>
    <script type="text/javascript" src="https://t1.daumcdn.net/tistory_a...
      s? version =tistory-aa685ca..."></script>
    <div style="margin:0; padding:0; border:none; background:none; float...
      <div id="acc-nav">...</div>
    <div id="wrap">
      <header id="header">...</header>
      <hr>
      <section class="container">
        <article id="content">
          <div class="inner">
            <div class="hgroup">
              <div class="category">크롤링</div>
              <h1>크롤링의 세계에 오신 것을 환영합니다. </h1> == $0
            </div>
            <div class="post-meta">...</div>
          </div>
          <div class="entry-content">...</div>
          <div id="entry1Comment">...</div>
          <script type="text/javascript">loadedComments[1]=true;findFra...
            ::after
          </script>
        </article>
        <div class="pagination">...</div>
      </section>
    </div>
  </body>
</html>
```

크롤링 예제 1-1

```
[1]: from urllib.request import urlopen  
  
url = "https://ai-dev.tistory.com/1"  
html = urlopen(url)
```

```
[2]: type(html)
```

```
[2]: http.client.HTTPResponse
```

```
[3]: html.status
```

```
[3]: 200
```

웹페이지 url에 접속해서 HTML 데이터를 가져옴

즉, html 객체에 웹 서버로부터의 HTML 응답(response)을 받아 담음

- 1 urlopen(url) 실행
- 2 해당 URL로 HTTP 요청을 보냄(GET 방식)
- 3 웹 서버가 HTML 문서를 응답(response)
- 4 응답을 html 이라는 변수에 담음

Section

웹크롤링기초

크롤링 예제 1-1

```
[4]: response = html.read()  
print(response)
```

```
b'<!doctype html>\n<html lang="ko">\n\n<head>\n    <script>\n        type="text/javascript">if (!window.T) { window.T = {} }</script>\n        window.T.config = {"TOP_SSL_URL": "https://www.tistory.com", "PREVIEW": false, "ROLE": "guest", "PREV  
_PAGE": "", "NEXT_PAGE": "", "BLOG": {"id": 4442027, "name": "ai-dev", "title": "\xec\x9d\xb8\xea\xb3\xb5\xec\x a7\x80\xeb\x8a\x a5 \xea\xb0\x9c\xeb\xb0\x9c\xec\x  
9d\x98 \xeb\xaa\x a8\xeb\x93\x a0 \xea\xb2\x83", "isDormancy": false, "nickName": "\xeb\x a1\x9c\xec\x a8\x a4\xec\xb9\xb4\xec\xb8\x a0", "status": "open", "profile  
Status": "normal"}, "NEED_COMMENT_LOGIN": false, "COMMENT_LOGIN_CONFIRM_MESSAGE": "", "LOGIN_URL": "https://www.tistory.com/auth/login/?redirectUrl=https://ai  
-dev.tistory.com/1", "DEFAULT_URL": "https://ai-dev.tistory.com", "USER": {"name": null, "homepage": null, "id": 0, "profileImage": null}, "SUBSCRIPTION": {"statu
```

```
[15]: type(response)
```

```
[15]: bytes
```

```
[5]: response.decode('utf-8')
```

```
'<!doctype html>\n<html lang="ko">\n\n<head>\n    <script t  
ype="text/javascript">if (!window.T) { window.T = {} }</script>\n    window.T.config = {"TOP_SSL_URL": "https://www.tistory.com", "PREVIEW": false, "ROLE": "guest", "PREV  
_PAGE": "", "NEXT_PAGE": "", "BLOG": {"id": 4442027, "name": "ai-dev", "title": "인공지능 개발의 모든 것", "isDormancy": false, "nickName": "로스카츠", "status": "ope  
n", "profileStatus": "normal"}, "NEED_COMMENT_LOGIN": false, "COMMENT_LOGIN_CONFIRM_MESSAGE": "", "LOGIN_URL": "https://www.tistory.com/auth/login/?redirectUrl  
=https://ai-dev.tistory.com/1", "DEFAULT_URL": "https://ai-dev.tistory.com", "USER": {"name": null, "homepage": null, "id": 0, "profileImage": null}, "SUBSCRIPTIO  
N": {"status": "none", "isConnected": false, "isPending": false, "isWait": false, "isProcessing": false, "isNone": true}, "IS_LOGIN": false, "HAS_BLOG": false, "IS_SUPP
```

```
[14]: type(response.decode('utf-8'))
```

```
[14]: str
```

Section

웹크롤링기초

크롤링 예제 1-1

```
[4]: from urllib.request import urlopen  
from bs4 import BeautifulSoup
```

```
url = "https://ai-dev.tistory.com/1"  
html = urlopen(url)  
bs_obj = BeautifulSoup(html, "html.parser")
```

```
[5]: print(bs_obj)
```

```
<!DOCTYPE html>  
  
<html lang="ko">  
<head>  
<script type="text/javascript">if (!window.T)  
window.T.config = {"TOP_SSL_URL":"https://www.  
i-dev","title":"인공지능 개발의 모든 것","isDoI  
COMMENT_LOGIN_CONFIRM_MESSAGE":"","LOGIN_URL":  
"dev.tistory.com","USER":{"name":null,"homepa  
ge,"isWait":false,"isProcessing":false,"isNone  
tistory.com","JOIN_URL":"https://www.tistory.c  
om"},  
window.T.entryInfo = {"entryId":1,"isAuthor":  
true},  
window.appInfo = {"domain":"tistory.com","top  
tistory.com/auth/logout"}:
```

Section

웹크롤링기초

크롤링 예제 1-1

```
[6]: title = bs_obj.find_all("h1")      HTML 문서 내에서 <h1> 태그를 모두 찾아서 리스트 형태로 리턴
print(title)

[<h1><a href="https://ai-dev.tistory.com/">인공지능 개발의 모든 것</a></h1>, <h1>크롤링의 세계에 오신 것을 환영합니다. </h1>]
```

```
[7]: print(title[1])

<h1>크롤링의 세계에 오신 것을 환영합니다. </h1>
```

```
[8]: print(title[1].text)

크롤링의 세계에 오신 것을 환영합니다.
```

```
[9]: contents = bs_obj.find_all("p")
print(contents)

[<p>POWERED BY TISTORY</p>, <p>Hello, world!</p>, <p class="copyright">DESIGN BY <a href="#">TISTORY</a> <a class="admin manage">관리자</a></p>, <p class="desc_g"></p>]
```

```
[10]: print(contents[1])

<p>Hello, world!</p>
```

```
[11]: print(contents[1].text)

Hello, world!
```

크롤링 예제 1-2 - requests 사용

```
[1]: import requests
```

```
url = "https://ai-dev.tistory.com/1"
html = requests.get(url)
```

```
[2]: type(html)
```

```
[2]: requests.models.Response
```

```
[3]: html.text
```

```
[3]: '<!doctype html>\n<html lang="ko">\n\n
  <head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0, maximum-scale=1.0, user-scalable=0" />
    <title>Tistory</title>
  </head>
  <body>
    <div id="wrap">
      <div id="header">
        <div class="inner">
          <div class="left">
            <a href="#">로그인</a>
            <a href="#">회원가입</a>
          </div>
          <div class="right">
            <span>로그인</span>
            <span>회원가입</span>
          </div>
        </div>
        <div id="main">
          <div class="inner">
            <div class="content">
              <div class="list">
                <div class="item">
                  <div class="info">
                    <div class="title">
                      <h2>인공지능 개발의 모든 것</h2>
                    </div>
                    <div class="desc">
                      <p>인공지능 개발의 모든 것</p>
                    </div>
                    <div class="meta">
                      <span>작성일: 2023-07-10</span>
                      <span>조회수: 100</span>
                      <span>댓글: 0</span>
                    </div>
                  </div>
                  <div class="img">
                    
                  </div>
                </div>
              </div>
              <div class="comment">
                <div class="list">
                  <div class="item">
                    <div class="info">
                      <div class="meta">
                        <span>작성일: 2023-07-10</span>
                        <span>조회수: 100</span>
                        <span>댓글: 0</span>
                      </div>
                    </div>
                    <div class="text">
                      <p>인공지능 개발의 모든 것</p>
                    </div>
                  </div>
                </div>
              </div>
            </div>
            <div class="bottom">
              <div class="left">
                <div class="button"><span>목록</span></div>
                <div class="button"><span>상세보기</span></div>
                <div class="button"><span>수정</span></div>
                <div class="button"><span>삭제</span></div>
              </div>
              <div class="right">
                <div class="button"><span>목록</span></div>
                <div class="button"><span>상세보기</span></div>
                <div class="button"><span>수정</span></div>
                <div class="button"><span>삭제</span></div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </body>
</html>'
```

디코딩을 자동으로 해줌

Section

웹크롤링기초

크롤링 예제 1-2

```
[4]: import requests
from bs4 import BeautifulSoup

url = "https://ai-dev.tistory.com/1"
html = requests.get(url).text
bs_obj = BeautifulSoup(html, "html.parser")
```

```
[5]: print(bs_obj)
```

```
<!DOCTYPE html>

<html lang="ko">
<head>
<script type="text/javascript">if (!wi
window.T.config = {"TOP_SSL_URL":"http
-dev","title":"인공지능 개발의 모든 것",
MENT_LOGIN_CONFIRM_MESSAGE":"","LOGIN
v tistory com" "IISFR"··"name"·null "ho
```

```
[6]: title = bs_obj.find_all("h1")
print(title)
```

```
[<h1><a href="https://ai-dev.tistory.com/">인공지능 개발의 모든 것</a></h1>, <h1>크롤링의 세계에 오신 것을 환영합니다. </h1>]
```

```
[7]: print(title[1])
```

```
[<h1>크롤링의 세계에 오신 것을 환영합니다. </h1>]
```

```
[8]: print(title[1].text)
```

```
크롤링의 세계에 오신 것을 환영합니다.
```

```
[9]: contents = bs_obj.find_all("p")
print(contents)
```

```
[<p>POWERED BY TISTORY</p>, <p>Hello, world!</p>, <p class="copyright">DESIGN BY <a href="#">TISTORY</a> <a class="admin" href="m/manage">관리자</a></p>, <p class="desc_g"></p>]
```

```
[10]: print(contents[1])
```

```
<p>Hello, world!</p>
```

```
[11]: print(contents[1].text)
```

```
Hello, world!
```

urllib vs requests

	urllib.request.urlopen()	requests.get()
라이브러리 형태	내장 라이브러리	외부 라이브러리
설치 방법	설치 필요 없음	pip install requests
수준	저수준	고수준
인코딩/디코딩 처리	수동	자동

Section

웹크롤링기초

크롤링 예제 2

<https://ai-dev.tistory.com/2>

크롤링

크롤링 예제 페이지 - 02

로스카츠 | 2021. 8. 8. 00:12 | 수정 | 삭제

본 페이지는 크롤링을 테스트 하기 위한 페이지입니다.

다음 표에 속하는 텍스트를 크롤링 해봅시다.

상품	색상	가격
셔츠1	빨강	20000
셔츠2	파랑	19000
셔츠3	초록	18000
바지1	검정	50000
바지2	파랑	51000

이번에는 리스트에 속하는 텍스트를 크롤링 해봅시다.

다음은 컴퓨터 구성 요소를 나타내는 리스트입니다.

- 모니터
- CPU
- 메모리
- 그래픽카드
- 하드디스크
- 키보드
- 마우스

크롤링 예제 2 <https://ai-dev.tistory.com/2>

크롤링

크롤링 예제 페이지 - 02

로스카츠 | 2021. 8. 8. 00:12 | 수정 | 삭제

본 페이지는 크롤링을 테스트하기 위한 페이지입니다.

table 805 x 229 뷰스터를 크롤링 해봅시다.

상품	색상	가격
셔츠1	빨강	20000
셔츠2	파랑	19000
셔츠3	초록	18000
바지1	검정	50000
바지2	파랑	51000

```
<script type="text/javascript" src="https://t1.daumcdn.net/tistory_admin/assets/blog/tistory-bd96dd17334...c1...istory-bd96dd1..."></script>
<script type="text/javascript" src="https://t1.daumcdn.net/tistory_admin/assets/blog/tistory-bd96dd17334...b1...istory-bd96dd1..."></script>
<div style="margin:0; padding:0; border:none; background:none; float:none; clear:none; z-index:0"></div>
▶<div id="acc-nav">...</div>
▼<div id="wrap">
  ▶<header id="header">...</header>
  <hr>
  ▼<section class="container">
    ▼<article id="content">
      ▼<div class="inner">
        ▶<div class="hgroup">...</div>
        ▼<div class="entry-content">
          ▼<div class="tt_article_useless_p_margin contents_style">
            <p data-ke-size="size16">본 페이지는 크롤링을 테스트하기 위한 페이지입니다.&nbsp;</p>
            <p data-ke-size="size16">&nbsp;</p>
            <p data-ke-size="size16">&nbsp;</p>
            ...
            ▼<table style="border-collapse: collapse; width: 100%; border="1" data-ke-align="alignLeft" data-ke-cellspacing="0" data-ke-cellpadding="5" data-ke-rows="2" data-ke-cells="3" data-ke-rowspacing="10" data-ke-cellsizing="auto">
              ▼<tbody>
                ▼<tr>
                  <td style="width: 33.3333%; text-align: center;">상품</td>
                  <td style="width: 33.3333%; text-align: center;">색상</td>
                  <td style="width: 33.3333%; text-align: center;">가격</td>
                </tr>
                ▼<tr>
                  <td style="width: 33.3333%; text-align: center;">셔츠1</td>
                  <td style="width: 33.3333%; text-align: center;">빨강</td>
                  <td style="width: 33.3333%; text-align: center;">20000</td>
                </tr>
              </tbody>
            </table>
          </div>
        </div>
      </div>
    </article>
  </section>
</div>
```

Section

웹크롤링기초

크롤링

크롤링 예제 페이지 - 02

로스카츠 2021.8.8.00:12 | 수정 | 삭제

본 페이지는 크롤링을 테스트 하기 위한 페이지입니다.

다음 표에 속하는 텍스트를 크롤링 해봅시다.

상품	색상	가격
셔츠1	빨강	20000
셔츠2	파랑	19000
셔츠3	초록	18000
바지1	검정	50000
바지2	파랑	51000

이번에는 리스트에 속하는 텍스트를 크롤링 해봅시다.

PC100 컴퓨터 구성 요소를 나타내는 리스트입니다.

ul 805 × 168

- 모니터
- CPU
- 메모리
- 그래픽카드
- 하드디스크
- 키보드
- 마우스

```
... <div class="entry-content">
  <p data-ke-size="size16">본 페이지는 크롤링을 테스트 하기 위한 페이지입니다.&ampnbsp</p>
  <p data-ke-size="size16">&nbsp;</p>
  <p data-ke-size="size16">&nbsp;</p>
  <p data-ke-size="size16">다음 표에 속하는 텍스트를 크롤링 해봅시다.</p>
  <table style="border-collapse: collapse; width: 100%;" border="1" data-ke-align="alignLeft" data-ke-size="size16">
    <tbody>
      <tr>
        <td style="width: 33.3333%; text-align: center;">상품</td>
        <td style="width: 33.3333%; text-align: center;">색상</td>
        <td style="width: 33.3333%; text-align: center;">가격</td>
      </tr>
      <tr>
        <td style="width: 33.3333%; text-align: center;">셔츠1</td>
        <td style="width: 33.3333%; text-align: center;">빨강</td>
        <td style="width: 33.3333%; text-align: center;">20000</td>
      </tr>
      <tr>...</tr>
      <tr>...</tr>
      <tr>...</tr>
      <tr>...</tr>
    </tbody>
  </table>
  <p data-ke-size="size16">&nbsp;</p>
  <p data-ke-size="size16">이번에는 리스트에 속하는 텍스트를 크롤링 해봅시다.</p>
  <p data-ke-size="size16">다음은 컴퓨터 구성 요소를 나타내는 리스트입니다.</p>
  <p data-ke-size="size16">&nbsp;</p>
  ... <ul style="list-style-type: disc;" data-ke-list-type="disc"> == $0
    <li>...</li>
    <li>...</li>
    <li>...</li>
    <li>...</li>
    <li>...</li>
    <li>...</li>
    <li>...</li>
    <li>...</li>
  </ul>
  <p data-ke-size="size16">&nbsp;</p>
  <p data-ke-size="size16">&nbsp;</p>
  <p data-ke-size="size16">&nbsp;</p>
  <div class="container_postbtn #post_button_group">...</div>
  <div class="another_category another_category_color_gray">...</div>
  </div>
  </div>
  <div class="related-articles">...</div>
  <div id="entry2Comment">...</div>
... ::text-paging-view-more div#wrap section.container article#content div.inner div.entry-content div.tt_article_useless_p
:: Console What's New X
```

크롤링 예제 2

```
[1]: from urllib.request import urlopen

url = "https://ai-dev.tistory.com/
html = urlopen(url)
```

```
[2]: print(html.read())
```

크롤링 예제 2

```
[3]: from urllib.request import urlopen  
      from bs4 import BeautifulSoup
```

```
url = "https://ai-dev.tistory.com/2"
html = urlopen(url)
bs_obj = BeautifulSoup(html, "html.parser")
```

```
[4]: print(bs_obj)
```

```
<!DOCTYPE html>

<html lang="ko">
<head>
<script type="text/javascript">if (!window.T) { window.T = {}; }
window.T.config = {"TOP_SSL_URL":"https://www.tistory.com","TOP_DOMAIN_NAME": "tistory.com", "TOP_TITLE": "인공지능 개발의 모든 것", "isDormancy": false, "TOP_LOGIN_CONFIRM_MESSAGE": "", "LOGIN_URL": "https://www.tistory.com", "USER": {"name": null, "homepage": null, "id": 0, "isProcessing": false, "isNone": true}, "IS_LOGIN": false, "JOIN_URL": "https://www.tistory.com/member/join", "PHASE": 1, "entryInfo": {"entryId": 2, "isAuthor": false, "category": null}, "appInfo": {"domain": "tistory.com", "topUrl": "https://www.tistory.com/auth/logout"}, "initData": {}};</script>
</head>
<body>
<div id="main"></div>
</body>
</html>
```

Section

웹크롤링기초

크롤링 예제 2

```
[5]: table_tag = bs_obj.find_all("table")  
  
[6]: print(table_tag)  
  
[<table border="1" data-ke-align="alignLeft" data-ke-style="si  
<tbody>  
<tr>  
<td style="width: 33.3333%; text-align: center;">상품 색상 | 가격 |  
<tr>  
<td style="width: 33.3333%; text-align: center;">셔츠1 빨강 | 20000 |  
<tr>  
<td style="width: 33.3333%; text-align: center;">셔츠2 파랑 | 19000 |  
<tr>  
<td style="width: 33.3333%; text-align: center;">셔츠3 초록 | 18000 |  
<tr>  
<td style="width: 33.3333%; text-align: center;">바지1 검정 | 50000 |  
<tr>  
<td style="width: 33.3333%; text-align: center;">바지2 파랑 | 51000 |  
</tbody>  
</table>. <table>
```

크롤링 예제 2

```
[7]: table_tag[0]
```

```
[7]: <table border="1" data-ke-align="alignLeft" data-ke-style="style1" st
<tbody>
<tr>
<td style="width: 33.3333%; text-align: center;">상품</td>
<td style="width: 33.3333%; text-align: center;">색상</td>
<td style="width: 33.3333%; text-align: center;">가격</td>
</tr>
<tr>
<td style="width: 33.3333%; text-align: center;">셔츠1</td>
<td style="width: 33.3333%; text-align: center;">빨강</td>
<td style="width: 33.3333%; text-align: center;">20000</td>
</tr>
<tr>
<td style="width: 33.3333%; text-align: center;">셔츠2</td>
<td style="width: 33.3333%; text-align: center;">파랑</td>
<td style="width: 33.3333%; text-align: center;">19000</td>
</tr>
<tr>
<td style="width: 33.3333%; text-align: center;">셔츠3</td>
<td style="width: 33.3333%; text-align: center;">초록</td>
<td style="width: 33.3333%; text-align: center;">18000</td>
</tr>
<tr>
<td style="width: 33.3333%; text-align: center;">바지1</td>
<td style="width: 33.3333%; text-align: center;">검정</td>
<td style="width: 33.3333%; text-align: center;">50000</td>
</tr>
<tr>
<td style="width: 33.3333%; text-align: center;">바지2</td>
<td style="width: 33.3333%; text-align: center;">파랑</td>
<td style="width: 33.3333%; text-align: center;">51000</td>
</tr>
</tbody>
```

크롤링 예제 2

```
[8]: table_tag01 = table_tag[0].find_all("td")
print(table_tag01)
```

```
<td style="width: 33.3333%; text-align: center; vertical-align: center;">가격</td>, <td style="width: 33.3333%; text-align: center;">20gn: center;">파랑</td>, <td style="width: 33.3333%; text-align: center;">초록</td>, <td style="width: 33.3333%; text-align: center;">바지1</td>, <td style="width: 33.3333%; text-align: center;">바지2</td>, <td style="width: 33.3333%; text-align: center;">51000</td>]
```

```
[9]: for idx, element in enumerate(table_tag01):
    print(idx, element.text)
```

```
0 상품
1 색상
2 가격
3 셔츠1
4 빨강
5 20000
6 셔츠2
7 파랑
8 19000
9 셔츠3
10 초록
11 18000
12 바지1
13 검정
14 50000
15 바지2
16 파랑
17 51000
```

크롤링 예제 2

```
[10]: table01 = bs_obj.find_all("td")
      print(table01)
```

```
[<td style="width: 33.3333%; text-align: center;">상품</td>, <td style="width: 3
-align: center;">가격</td>, <td style="width: 33.3333%; text-align: center;">셔츠
tyle="width: 33.3333%; text-align: center;">20000</td>, <td style="width: 33.333
gn: center;">파랑</td>, <td style="width: 33.3333%; text-align: center;">19000</
le="width: 33.3333%; text-align: center;">초록</td>, <td style="width: 33.3333%; 
center;">바지1</td>, <td style="width: 33.3333%; text-align: center;">검정</td>,
="width: 33.3333%; text-align: center;">바지2</td>, <td style="width: 33.3333%; 
enter;">51000</td>, <td>2021.01.21</td>]
```

```
[11]: table = bs_obj.find_all("td", {"style":"width: 33.3333%; text-align: center;"})
      print(table)
```

```
[<td style="width: 33.3333%; text-align: center;">상품</td>, <td style="width: 3
-align: center;">가격</td>, <td style="width: 33.3333%; text-align: center;">셔츠
tyle="width: 33.3333%; text-align: center;">20000</td>, <td style="width: 33.333
gn: center;">파랑</td>, <td style="width: 33.3333%; text-align: center;">19000</
le="width: 33.3333%; text-align: center;">초록</td>, <td style="width: 33.3333%; 
center;">바지1</td>, <td style="width: 33.3333%; text-align: center;">검정</td>,
="width: 33.3333%; text-align: center;">바지2</td>, <td style="width: 33.3333%; 
enter;">51000</td>]
```

크롤링 예제 2

enumerate: 반복문에서 인덱스도 함께 제공

```
[12]: for idx, element in enumerate(table):
    print(idx, element)
```

```
0 <td style="width: 33.3333%; text-align: center;">상품</td>
1 <td style="width: 33.3333%; text-align: center;">색상</td>
2 <td style="width: 33.3333%; text-align: center;">가격</td>
3 <td style="width: 33.3333%; text-align: center;">셔츠1</td>
4 <td style="width: 33.3333%; text-align: center;">빨강</td>
5 <td style="width: 33.3333%; text-align: center;">20000</td>
6 <td style="width: 33.3333%; text-align: center;">셔츠2</td>
7 <td style="width: 33.3333%; text-align: center;">파랑</td>
8 <td style="width: 33.3333%; text-align: center;">19000</td>
9 <td style="width: 33.3333%; text-align: center;">셔츠3</td>
10 <td style="width: 33.3333%; text-align: center;">초록</td>
11 <td style="width: 33.3333%; text-align: center;">18000</td>
12 <td style="width: 33.3333%; text-align: center;">바지1</td>
13 <td style="width: 33.3333%; text-align: center;">검정</td>
14 <td style="width: 33.3333%; text-align: center;">50000</td>
15 <td style="width: 33.3333%; text-align: center;">바지2</td>
16 <td style="width: 33.3333%; text-align: center;">파랑</td>
17 <td style="width: 33.3333%; text-align: center;">51000</td>
```

```
[13]: for idx, element in enumerate(table):
    print(idx, element.text)
```

```
0 상품
1 색상
2 가격
3 셔츠1
4 빨강
5 20000
6 셔츠2
7 파랑
8 19000
9 셔츠3
10 초록
11 18000
12 바지1
13 검정
14 50000
15 바지2
16 파랑
17 51000
```

Section

웹크롤링기초

크롤링 예제 2

```
[14]: com_list = bs_obj.find_all("li")
print(com_list)
```

```
[<li class=""><a class="link_tit" href="/category"> 분류 전체보기 <span class="c_cnt">(2)</span> </a>
<ul class="category_list"><li class=""><a class="link_item" href="/category/%ED%81%AC%EB%A1%A4%EB%A7%81"> 크롤링 <span class="c_cnt">(2)</span> </a></li>
</ul>
</li>, <li class=""><a class="link_item" href="/category/%ED%81%AC%EB%A1%A4%EB%A7%81"> 크롤링 <span class="c_cnt">(2)</span> </a></li>, <li>모니터</li>, <li>CPU</li>, <li>메모리</li>, <li>그래픽카드</li>, <li>하드디스크</li>, <li>키보드</li>, <li>마우스</li>, <li>
<a href="/1?category=836119">
<span class="thum">
</span>
<span class="title">크롤링의 세계에 오신 것을 환영합니다.</span>
</a>
</li>]
```

```
[15]: com_list01 = bs_obj.find_all("ul", {"style": "list-style-type: disc;"})
print(com_list01)
```

```
[<ul data-ke-list-type="disc" style="list-style-type: disc;">
<li>모니터</li>
<li>CPU</li>
<li>메모리</li>
<li>그래픽카드</li>
<li>하드디스크</li>
<li>키보드</li>
<li>마우스</li>
</ul>]
```

크롤링 예제 2

```
[16]: com_list02 = com_list01[0].find_all("li")
```

```
[17]: for idx, element in enumerate(com_list02):
        print(idx, element.text)
```

0 모니터
1 CPU
2 메모리
3 그래픽카드
4 하드디스크
5 키보드
6 마우스

감사합니다.

Q & A