



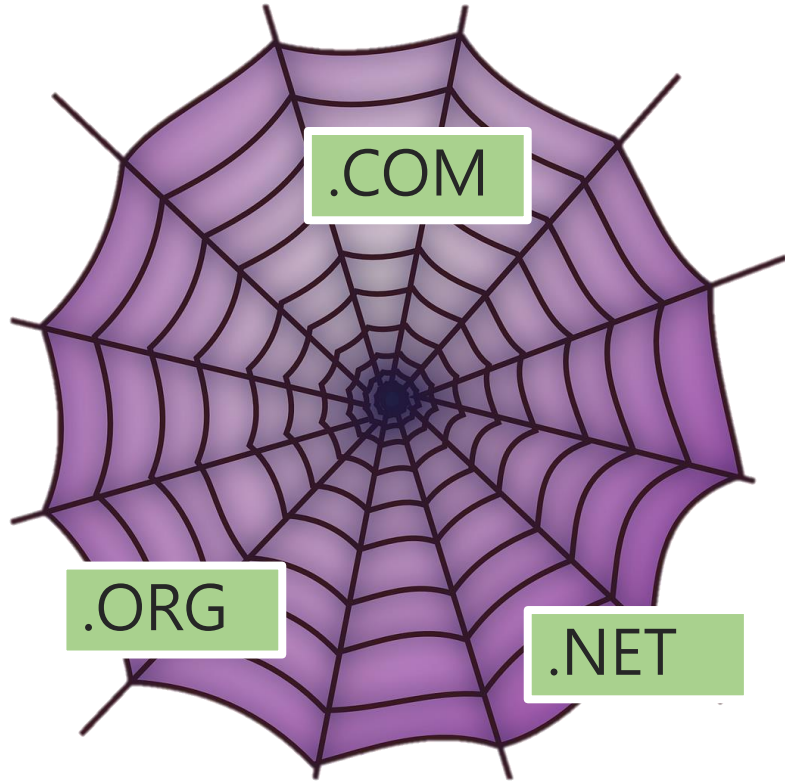
크롤링 기초 & HTML

웹 크롤러(Web crawler)

1. 웹 크롤러(web crawler)는 조직적, 자동화된 방법으로 월드 와이드 웹을 탐색하는 컴퓨터 프로그램이다.
2. 웹 크롤러가 하는 작업을 '웹 크롤링'(web crawling) 혹은 '스파이더링'(spidering)이라 부른다. 검색 엔진과 같은 여러 사이트에서는 데이터의 최신 상태 유지를 위해 웹 크롤링한다. 웹 크롤러는 대체로 방문한 사이트의 모든 페이지의 복사본을 생성하는 데 사용되며, 검색 엔진은 이렇게 생성된 페이지를 보다 빠른 검색을 위해 인덱싱한다.

[출처:위키백과]

웹 크롤러(Web crawler)



웹 크롤러

웹페이지를 방문하며 자동적으로 자료 수집

'크롤러'('로봇' 또는 '스파이더'라고도 함)는 한 웹페이지에서 다른 웹페이지로 연결되는 링크를 따라가며 웹사이트를 자동으로 검색하는 데 사용되는 프로그램을 가리키는 일반적인 용어입니다. Google의 기본 크롤러를 Googlebot이라고 합니다.

웹 스크래핑(Web Scraping)

웹 크롤링과 웹 스크래핑의 장점

- 심층 분석과 실시간 정보 제공에 유용한 “웹 크롤링”

웹 크롤링은 웹상을 돌아다니며 방대한 양의 정보를 수집하기 때문에, 특정 키워드에 대한 심층 분석이 필요할 때 유용합니다.

- 정확한 정보를 요구할 때 쓰이는 “웹 스크래핑”

웹 스크래핑은 특정 사이트나 페이지에 대한 정보를 찾는데 집중하므로 데이터 포인트를 정확히 잡고 확실한 정보만을 수집할 수 있다는 점에서 유용합니다.

웹 스크래핑(Web Scraping)

웹 스크래핑은 특정 웹 사이트나 페이지에서 필요한 데이터를 자동으로 추출해 내는 것을 의미합니다 [출처:위키백과]

크롤링하는 방법

1. 웹 페이지 정보를 가져온다. -> 파이썬에서는 **requests** 라이브러리를 사용.
2. **html** 소스를 파싱(분석한다)하여 원하는 정보를 얻는다. -> 파이썬에서는 **BeautifulSoup** 라이브러리를 사용.

사전 준비하기

```
!pip install bs4
```

```
!pip install lxml
```

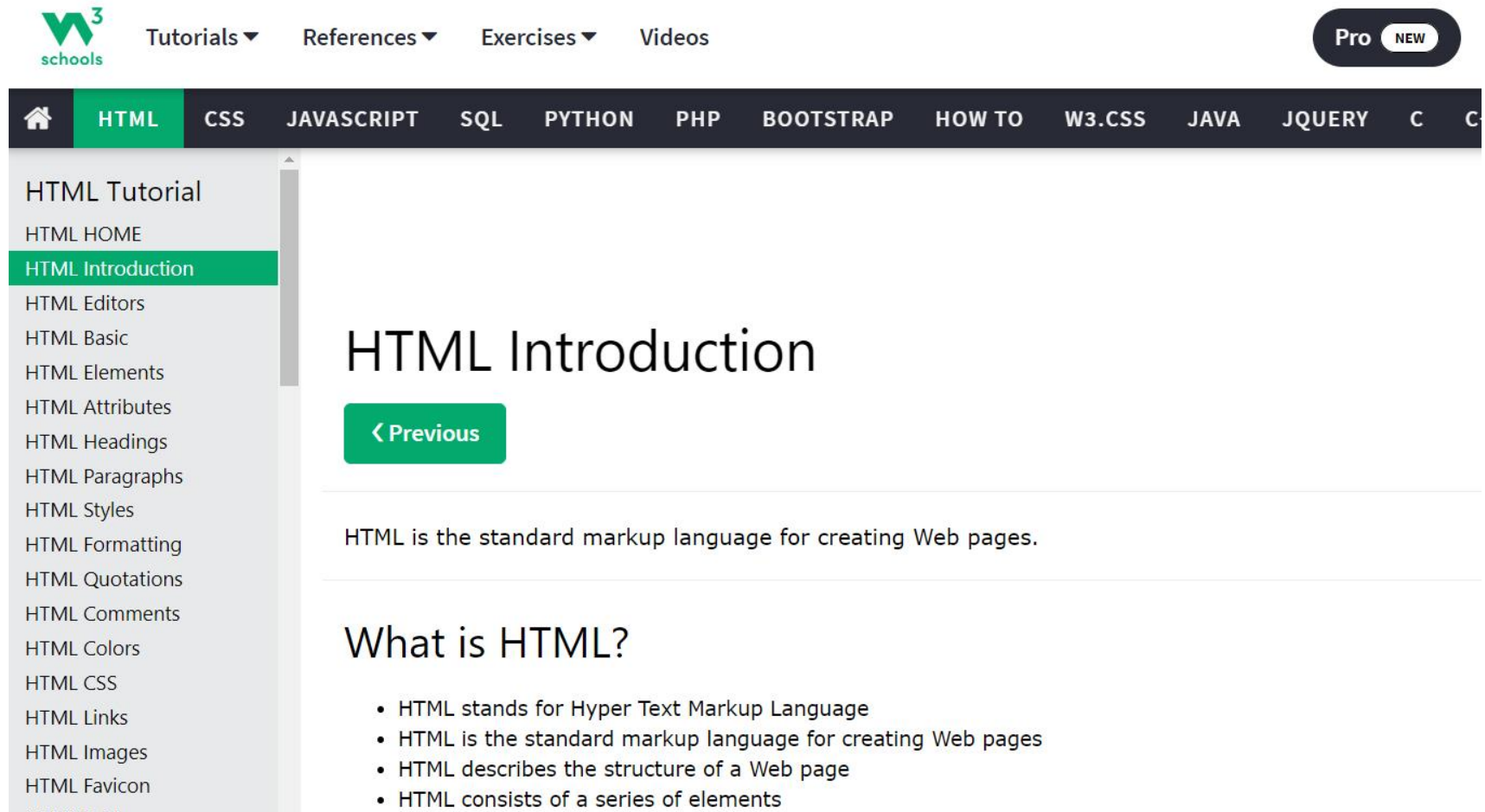
- BeautifulSoup 라이브러리

Beautiful Soup의 자세한 정보는 아래 사이트를 참고하세요.

<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

- HTML구조 학습

HTML: w3schools.com



The screenshot shows the w3schools.com website. At the top, there is a navigation bar with the w3schools logo on the left and links for Tutorials, References, Exercises, and Videos. On the right of the navigation bar, there is a dark button labeled 'Pro' and a white button labeled 'NEW'. Below the navigation bar is a dark horizontal menu with various technology categories: HTML (highlighted in green), CSS, JAVASCRIPT, SQL, PYTHON, PHP, BOOTSTRAP, HOW TO, W3.CSS, JAVA, JQUERY, C, and C++. A sidebar on the left contains a list of HTML topics, with 'HTML Introduction' highlighted in green. The main content area features the title 'HTML Introduction' in a large font, a green button with a left arrow and the text '< Previous', and a paragraph stating 'HTML is the standard markup language for creating Web pages.' Below this is a section titled 'What is HTML?' followed by a bulleted list of five points.

w3schools

Tutorials ▼ References ▼ Exercises ▼ Videos

Pro NEW

HTML CSS JAVASCRIPT SQL PYTHON PHP BOOTSTRAP HOW TO W3.CSS JAVA JQUERY C C++

HTML Tutorial

HTML HOME

HTML Introduction

HTML Editors

HTML Basic

HTML Elements

HTML Attributes

HTML Headings

HTML Paragraphs

HTML Styles

HTML Formatting

HTML Quotations

HTML Comments

HTML Colors

HTML CSS

HTML Links

HTML Images

HTML Favicon

HTML Introduction

< Previous

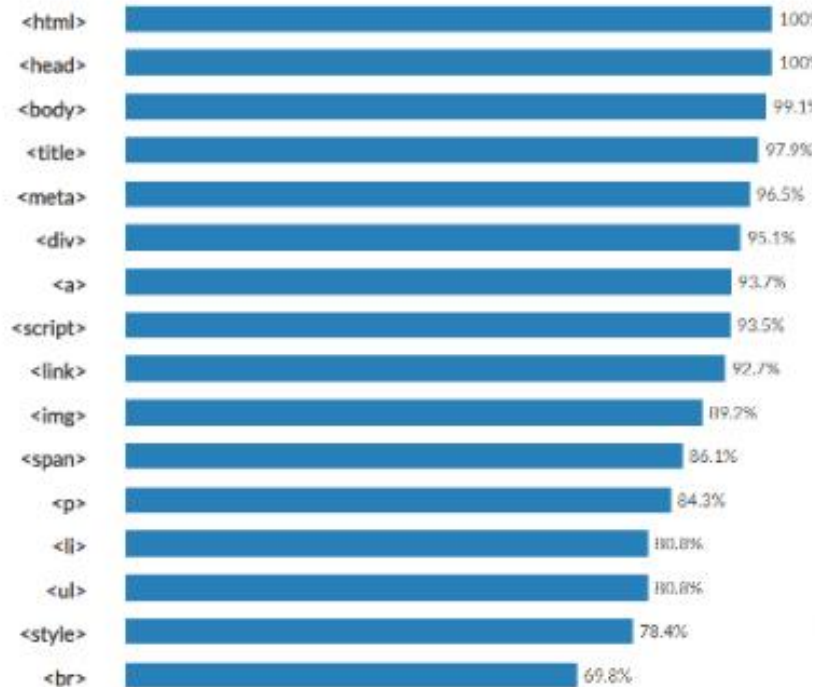
HTML is the standard markup language for creating Web pages.

What is HTML?

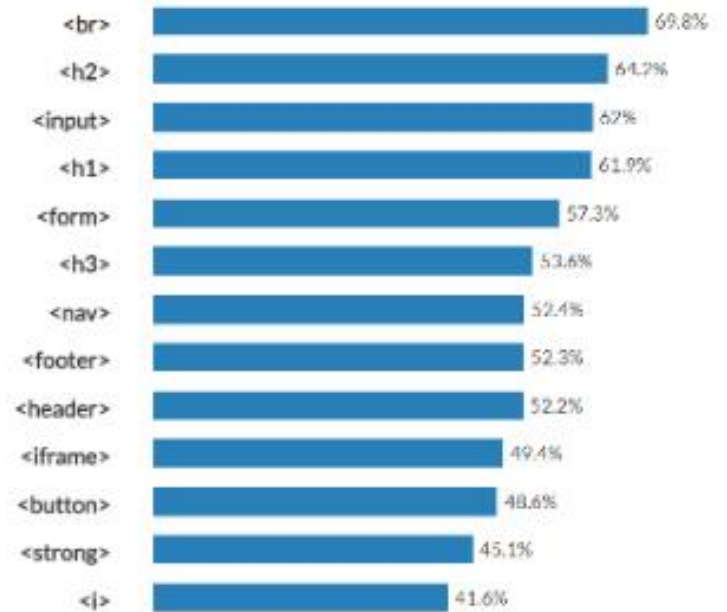
- HTML stands for Hyper Text Markup Language
- HTML is the standard markup language for creating Web pages
- HTML describes the structure of a Web page
- HTML consists of a series of elements

HTML: w3schools.com

The thirty-two elements used on most pages, ordered by appearance frequency:



Total 11,264,652 pages



출처 : <https://www.advancedwebranking.com/html/>

BeautifulSoup

Beautiful Soup 4.9.0 documentation » Beautiful Soup Documentation

Table of Contents

- Beautiful Soup Documentation
 - Getting help
- Quick Start
- Installing Beautiful Soup
 - Installing a parser
- Making the soup
- Kinds of objects
 - Tag
 - Name
 - Attributes
 - Multi-valued attributes
 - NavigableString
 - BeautifulSoup
 - Comments and other special strings
- Navigating the tree

Beautiful Soup Documentation

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates your expectations.


This document covers Beautiful Soup version 4.11.0. The examples in this documentation were written for Python 3.8.


You might be looking for the documentation for Beautiful Soup 3. If so, you should know that Beautiful Soup 3 is no longer being developed and that all support for it was dropped on December 31, 2020. If you want to learn about the differences between Beautiful Soup 3 and Beautiful Soup 4, see [Porting code to BS4](#).



<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

웹 스토어 > Web Scraper

 chrome 웹 스토어

 chrome 웹 스토어

chrome.google.com/webstore/search/scraper?hl=ko

chrome 웹 스토어

scraper

« 홈

☐ 확장 프로그램

☐ 테마

평점

☐ ★★★★★

☐ ★★★★★ 이상

☐ ★★★★★ 이상


☐ ★★★★★ 이상

개인정보처리방침

서비스 약관 Updated

Chrome 웹 스토어 정보


확장 프로그램 ⓘ




Web Scraper - Free Web Scraping

Web data extraction tool with an easy point-and-click interface for modern web

★★★★★ 761 생산성



BigSeller - Product Scraper

 bigseller.com

BigSeller Product Scraper, scrape products from marketplaces to your multiple stores.

★★★★★ 42 생산성

확장 프로그램 더보기