웹 파싱(Parsing)과 크롤링(Crawling)의 차이점 및 개념

**파싱 (Parsing)**

**파싱**이란 ?

​

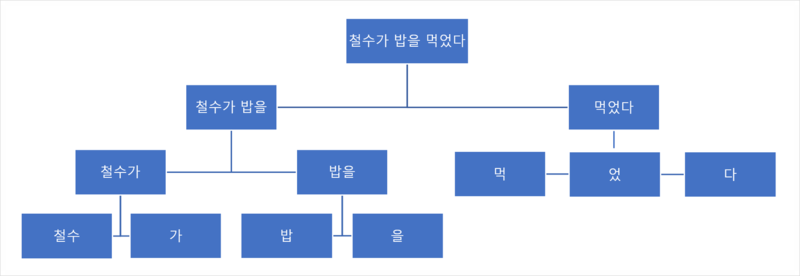
**파싱(Parsing)** 은 어떠한 웹 페이지에서 내가 원하는 데이터를 특정 패턴이나 순서로 추출하여 정보로 가공하는것을 말합니다.

인터프린터나 컴파일러의 구성 요소 중 하나로, 입력 토큰에 내제되어있는 자료구조를 빌드하고 문법을 검사하는 역할을 합니다.

​

쉽게말해서 파싱은 일련의 정보를 분석하고 가공하는 것이라고 할 수 있습니다.

​

[](https://m.blog.naver.com/yoodh0713/221546039399)

**크롤링 (Crawling)**

**웹 크롤링**이란?

​

**크롤링(Crawling)** 은 웹 상의 페이지를 수집해서 분류하고 저장한 후에 나중에 쉽게 찾아볼 수 있도록 하는 역할을 하는 일종의 로봇입니다.

데이터를 저장한 후 쉽게 찾을 수 있도록 인덱싱 합니다.

​

쉽게말해서 수많은 데이터를 수집해서 나중에 꺼내보기 쉽도록 정리 해두는 것이라고 생각하면 됩니다.

​

추가적으로 **스크래핑(Scraping)** 이라는 것은 데이터(자료)를 수집하는 모든 과정을 말합니다.

​

즉, 크롤링은 스크래핑의 방법 중 하나가 되는 것입니다.

흔히 알고있는 웹 퍼블리셔나 개발자가 개발을 하는 과정을 위 내용을 토대로 해석하자면

여러곳에 흩어져 있는 수많은 자료를 스크래핑(크롤링)하여 수집된 데이터를 토대로 파싱(가공)하여 원하는 데이터를 빼내어 활용하는 것.

이것이 웹 크롤링한다, 파싱한다 라는 표현으로 사용되고 있습니다.

**크롤링(crawling)**

크롤링이란 단어는 웹 크롤러(crawler)라는 단어에서 시작한 말이다.크롤러란 조직적, 자동화된 방법으로 월드와이드 웹을 탐색하는 컴퓨터 프로그램이다.

크롤링은 크롤러가 하는 작업을 부르는 말로, 여러 인터넷 사이트의 페이지(문서, html 등)를 수집해서 분류하는 것이다.

대체로 찾아낸 데이터를 저장한 후 쉽게 찾을 수 있게 인덱싱한다.

**파싱(parsing)**

파싱이란 어떤 페이지(문서, html 등)에서 내가 원하는 데이터를 특정 패턴이나 순서로 추출하여 정보를 가공하는 것이다.

위 문장만 보면 굉장히 간단해 보이지만 컴퓨터 과학적 정의를 보면 파싱이란 일련의 문자열을 의미있는 토큰(token)으로 분해하고 이들로 이루어진 파스 트리(parse tree)를 만드는 과정을 말한다.(출처: 위키백과)

인터프리터나 컴파일러의 구성 요소 가운데 하나로, 입력 토큰에 내제된 자료 구조를 빌드하고 문법을 검사하는 역할을 한다.

**스크래핑(scraping)**

스크래핑이란 HTTP를 통해 웹 사이트의 내용을 긁어다 원하는 형태로 가공하는 것이다. 쉽게 말해 웹 사이트의 데이터를 수집하는 모든 작업을 뜻한다.

크롤링도 일종의 스크래핑 기술이라고 할 수 있다.

[**[Python][Crawling][Scraping]크롤링과 스크래핑, 그리고 원리**](https://kamang-it.tistory.com/entry/PythonCrawlingScraping%ED%81%AC%EB%A1%A4%EB%A7%81%EA%B3%BC-%EC%8A%A4%ED%81%AC%EB%9E%98%ED%95%91-%EA%B7%B8%EB%A6%AC%EA%B3%A0-%EC%9B%90%EB%A6%AC1)

현재 웹 크롤링과 웹 스크래핑은 매우 중요하다고 할 수 있다. 그 이유는 데이터 사이언스와 머신러닝이 도래하면서 빅 데이터가 중요해졌다. 하지만 빅 데이터를 그냥 얻기는 좀 힘드므로... 어쩔 수 없이 인터넷에서 긁어오는 경우가 많다. 그리고 굳이 빅 데이터가 아니더라도 특정 서비스를 위해서 혹은 개인의 편리를 위해서 사용하는 경우가 많다.

***웹 크롤링 - 웹 크롤러(자동화 봇)가 일정 규칙으로 웹페이지를 브라우징 하는 것***

이를 보통 우리는 웹 크롤링이라고 부른다.

웹 스크래핑이라는 말은 쓰는 사람은 쓰고 안 쓰는 사람은 안쓰는 묘한 단어이다.

***웹 스크래핑 - 웹 사이트 상에서 원하는 정보를 추출하는 기술***

엄밀히 말하면 둘은 다르고 구분되어야 한다.

쉽게 말해서 웹 크롤링은 그냥 돌아만 다니는 거고 스크래핑은 긁어 오는 것이다.

그래서 우리가 보통 크롤링을 한다고 하는것은 사실은 스크래핑이라고 할 수 있다.

하지만 의미 없는 단어 논쟁은 안하는게 좋으므로 여기까지만 말하겠다.

어쨋든 우리는 웹 크롤링과 스크래핑에서 배울 것이며 특별한 일이 없다면 필자는 이 둘을 퉁쳐서 크롤링이라고 부르겠다.

크롤링이 파이썬만 가능한 것도 아니고 여러 언어에서 크롤링이 가능하다.

그럼 간단한 예제를 한번 보도록하자.

from urllib.request import urlopen  
  
html = urlopen('http://pythonscraping.com/pages/page1.html')  
print(html.read())

정말 간단한 예제다.

위의 사이트는 파이썬 크롤링을 위한 예제 사이트인데 필자가 만든건 아니고 파이썬 크롤링을 가르치는 곳에서 만든 토이 사이트이다.

위의 예제의 결과를 한번 보도록 하자.

https://t1.daumcdn.net/cfile/tistory/99BE663E5CBC5F2023

해당 페이지의 html을 통째로 긁어온 것을 확인할 수 있다.

이런식으로 가져와서 페이지를 긁어서 작업할 수 있다는 걸 여러분은 알 수 있다.