# 웹 크롤링

## 개요

- 웹 크롤링 (Web Crawling)
  - 웹 크롤링의 정식명칭은 "Web Scraping"이다. (외국자료는 Web Scraping으로 찾으면 더 많음..)
  - 웹 크롤링이란 컴퓨터 소프트웨어 기술로 웹 사이트들에서 원하는 정보를 추출하는 것을 의미한다.

- 웹 크롤러 (Web Crawler)
  - 인터넷에 있는 웹 페이지를 방문해서 자료를 수집하는 일을 하는 프로그램
  - 웹은 기본적으로 HTML 태그 형태로 이루어져 있다.
    HTML 태그는 개발자가 직접 코딩한다.
    그러면 특정 웹 페이지의 HTML 태그의 내용은 정형화된 형태로 존재할 것이고 규칙이 생길 것이다.
    그러한 규칙을 분석하여 원하는 정보만 추출하는 것이 웹 크롤러이다.

## 예시) 웹 크롤러의 작업방식

※ http://www.example-domain.com (예시 도메인임으로 인터넷상엔 존재하지 않음)

➢ 도메인에 해당되는 서버에 저장된 index.html의 내용

```
<html>
        <head>
                <title> 예시 HTML 문서 </title>
        </head>
        <body>
                <div id="important">
                        <p>웹 크롤러가 필요로하는 정보1</p>
                        <p>웹 크롤러가 필요로하는 정보2</p>
                </div>

                <div id="garbage">
                        <p>웹 크롤러가 필요로하지 않는 정보1</p>
                        <p>웹 크롤러가 필요로하지 않는 정보2</p>
                </dev>
        </body>
</html>
```

1. 위 도메인 주소(http://example-domain.com)를 통해 index.html 웹 문서 요청 (HTTP Request)

2. index.html의 모든 내용을 저장 (HTTP Response)

3. HTML 태그 중 태그를 찾기
   : 여기서는 id가 garbage와 important인 모든 을 찾아낸다.

4. 찾아낸 태그 중 id가 important인 태그 찾기
   : 여기서 id가 important인 만을 찾아낸다.

5. 의 모든 내용을 크롤링하기

6. 웹 크롤러가 얻어낸 정보 (HTML 태그)

```
<div id="important">
        <p>웹 크롤러가 필요로하는 정보1</p>
        <p>웹 크롤러가 필요로하는 정보2</p>
</div>
```

7. 얻어낸 정보를 활용하여 다음 작업 수행

## 예제 코드 저장소

- Python code : https://github.com/HoDoLi123/web-crawling-python

## 참고문헌


- Junee01 (2016). https://m.blog.naver.com/potter777777/220605598446