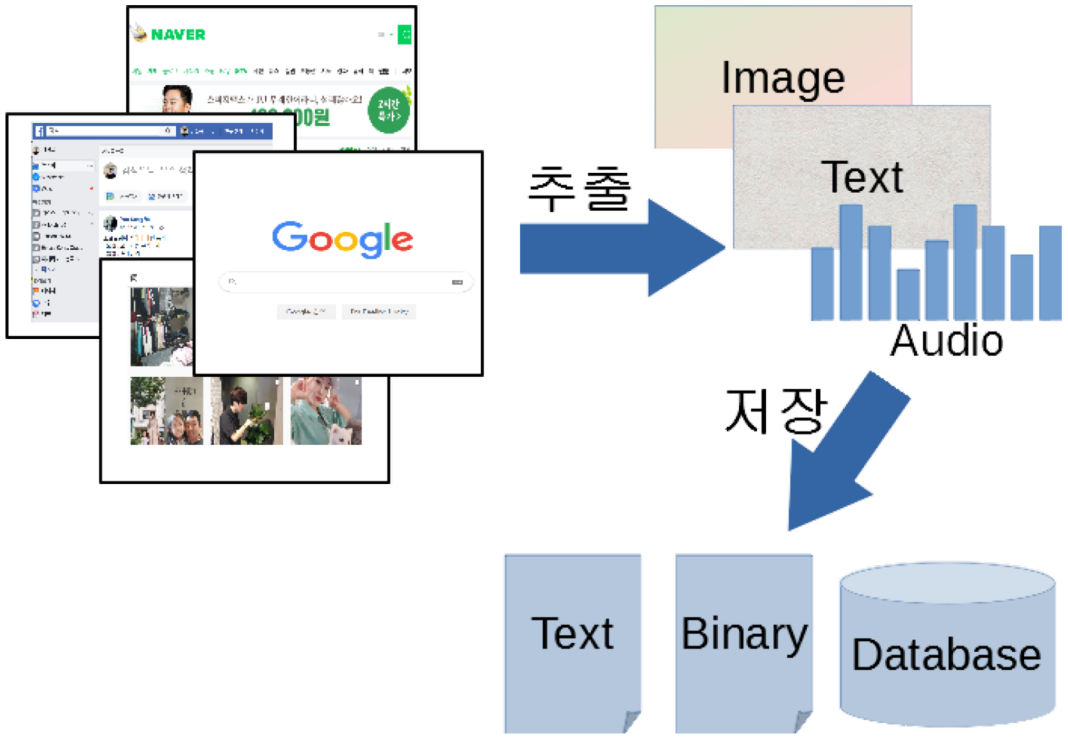
웹크롤링

구글 이미지 검색

AI

AI

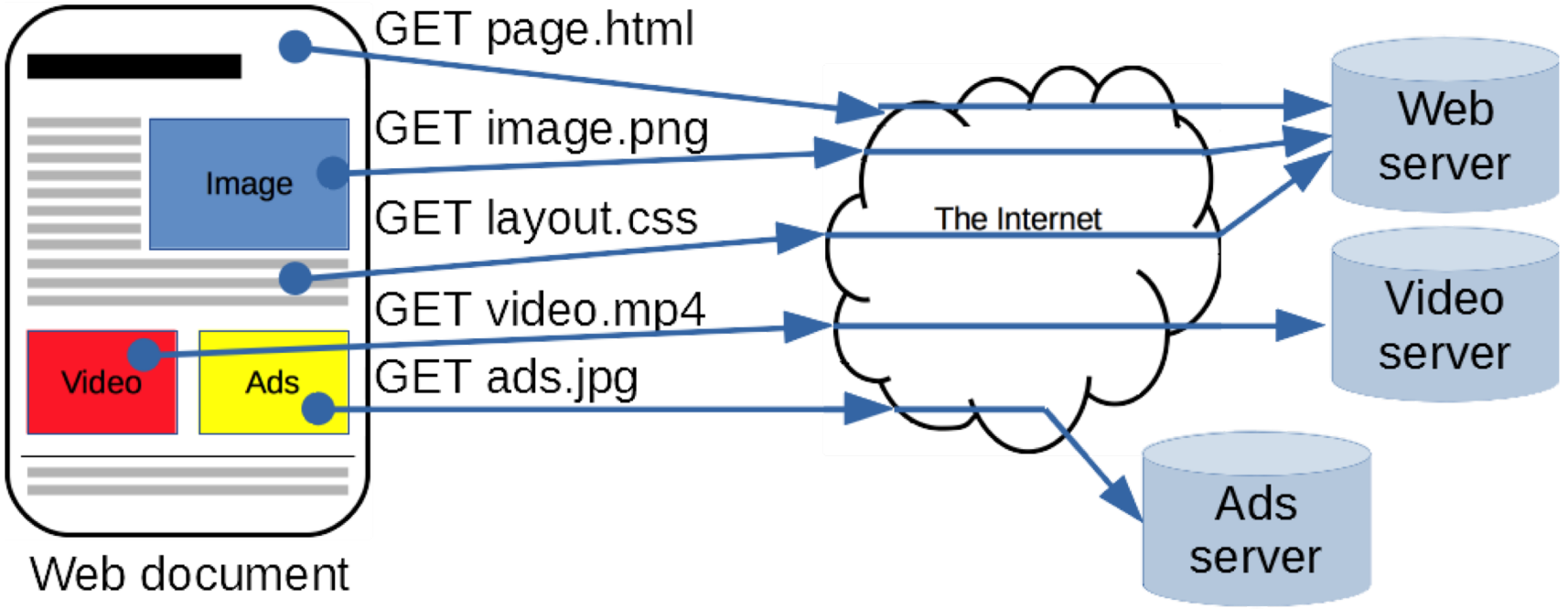
웹크롤링



- 웹 사이트에 있는 특정 정보를 추출하는 기술
  - 인스타그램, 페이스북, 네이버 카페, 블로그 등
  - Text, Image, Audio 등
- HTML 구조 분석 및 로그인 처리 필요

AI

WEB



› HTTP는 HTML 문서와 같은 리소스들을 요청하는 프로토콜

› 하나의 완전한 문서는 다른 하위 문서, 텍스트,
레이아웃 설명, 이미지, 비디오, 스크립트 등으로 구성

AI

URL

- URL(Uniform Resource Locator)은 네트워크 상에서 자원이 어디 있는지를 알려주기 위한 규약
- 즉, 컴퓨터 네트워크와 검색 메커니즘에서의 위치를 지정하는, 웹 리소스에 대한 참조
- URL은 웹 사이트 주소뿐만 아니라 컴퓨터 네트워크상의 자원을 모두 나타낼 수 있음
- URL 형태
  scheme:[//[user:password@]host[:port]]
  [/]path [?query] [#fragment]

AI

URL

- http://www.example.com:80/path/to/myfile.html?key1=value1&key2=value2

파트	설명
http://	프로토콜(데이터를 교환/전송하기 위한 규약)
www.example.com	도메인 이름 또는 IP 주소
:80	포트(IP 내에서 프로세스 구분)
/path/to/myfile.html	웹서버에서 자원에 대한 경로
?key1=value1&key2=value2	웹서버에 제공하는 파라미터로 & 기호로 구분된 키/값쌍

AI

HTML

```
<!doctype html>
<html>
  <head>
    <title>제목</title>
  </head>
  <body>
    <img src="img.png”>
    <p id=“p1">내용</p>
  </body>
</html>
```

- HTML은 웹 문서를 기술하기 위한 마크업 언어
- 외관과 배치를 정의하는 CSS 같은 스크립트를 포함
- 꺽쇠 괄호로 둘러싸인 “태그”로 되어있는 HTML 요소 형태로 작성
- id 속성 값은 전 요소를 포함하여 유일하게 식별 가능한 값

AI

FORM

이름	값
이름	
성별	○ 남성 ◉ 여성
눈색	녹색 ▾
적합한 것을 골라주세요.	☐ 신장 180cm 이상 ☐ 체중 90kg 이상
자신의 신체 능력에 대해서 써주세요.	
보내기	

```
<from [속성="속성값"]>
    ...
</from>
```

- 사용자가 웹사이트로 어떤 정보를 보낼 수 있는 요소
- 웹사이트의 로그인 폼이나 회원가입 폼이 대표적
- 종류로는 텍스트 상자, 버튼, 레이블, 체크 상자 등이 있으며 type 속성으로 지정

AI

CSS(Cascading Style Sheets)

```
p{
    font-size: 110%;
    font-family: sans-serif;
}
.highlight{
    color: red;
    background: yellow;
    font-weight: bold;
}
#test_id {
    color: blue;
}
```

- CSS는 마크업 언어가 실제 표시되는 방법을 기술하는 언어
- HTML와 같은 마크업 언어가 내용과 논리적인 구조를 담당하고 CSS는 레이아웃과 스타일을 정의(내용와 디자인을 분리)
- 특정 Element, Class, Id 수준에서 적용

AI

XPath(XML Path Language)

› 문서 내의 위치 정보를 위한 언어

› / A / B / C
선두가 / 인 절대 경로이고 복수의 C요소를 선택, 선택되는 C 요소는 B 요소의 자식이며 B요소는 A 요소의 자식이며, A요소는 문서의 루트 요소

› A // B / * [1]
선두가 / 없는 상대 경로이며 A 요소는 현재 문맥 노드의 자식이고 B요소는 A요소의 직간접적인 자식 요소이며, B의 자식 요소들 중 첫 번째 요소를 선택

AI

XPath(XML Path Language)

- // a [@href='help.php']
  href 속성을 가지고 있고 그 속성 값이 'help.php'인 모든 a요소 노드를 지정
- // a [@href = 'help.php'] [name (..) = 'div'] [.. / @ class = 'header']
  부모 요소 이름이 div, 부모 요소 (div)의 class 속성의 값이 'header'이고 href 속성이 'help.php' 인 모든 a 요소 지정
- // item [@price> = 2 * @discount]
  price 속성의 수치가 discount 숫자의 2 배 이상인 item 요소를 지정

AI

탐색

A. 여러 웹 페이지의 링크를 무작위로 방문하여 웹문서의 내용을 탐색

B. 포털 사이트 등에서 제공하는 API를 이용하여 데이터를 수집

C. 특정 웹 페이지를 주기적으로 방문하여 데이터를 수집

› 수집 목적에 따라 웹 사이트를 탐색할 방법을 정함

AI

방문

A. 헤더 수정
B. 로그인 처리
C. 쿠키 제시
D. 타이밍
E. CAPTCHA 읽기

› 특정 사이트의 경우 해당 웹 페이지에 접근하기 위해서는 로그인과 같이 일정한 절차를 거쳐야 하는 경우가 존재

› 폼을 너무 빨리 완성하거나 단 시간내 너무 여러 페이지를 다니는 등 사람과 다른 이상한 행동을 보이면 차단 당할 수 있기에 각 경우에 따른 트릭이 필요

AI

추출

A. 정적 페이지 처리
B. 동적 페이지 처리

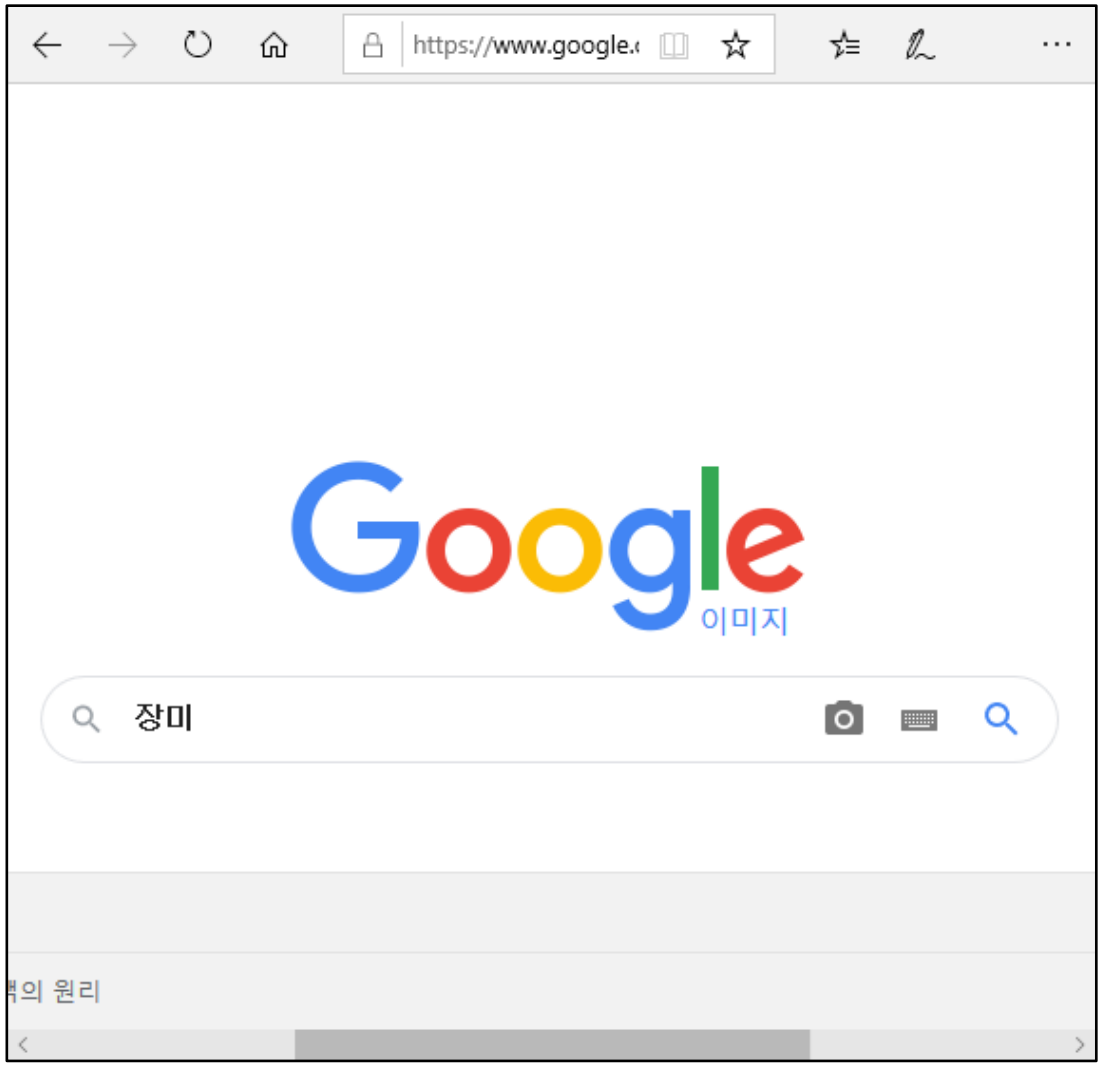
A. XPath
B. element id/class

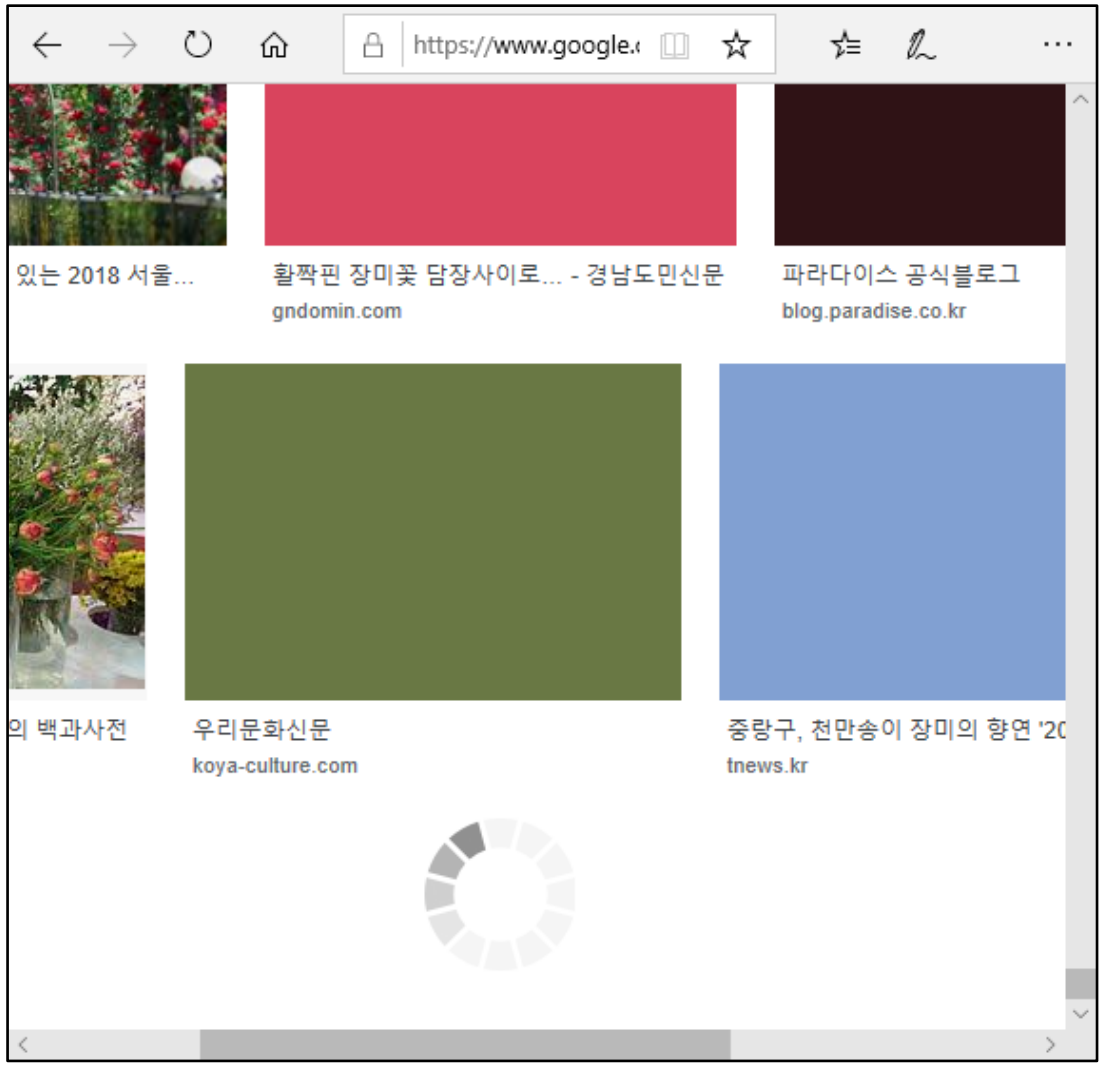
A. 정규표현식(텍스트)
B. URL(미디어)

› 웹 서버에서 받은 페이지는 스크립트에 의해 최초 내용이 변경될 수 있음(AJax 기술)
› HTML 문서로 부터 추출하고자 하는 정보를 포함하고 있는 특정 요소 지정
› 해당 요소가 포함하고 있는 정보를 추출(텍스트의 경우에는 정규표현식, 미디어의 경우에는 URL)

AI

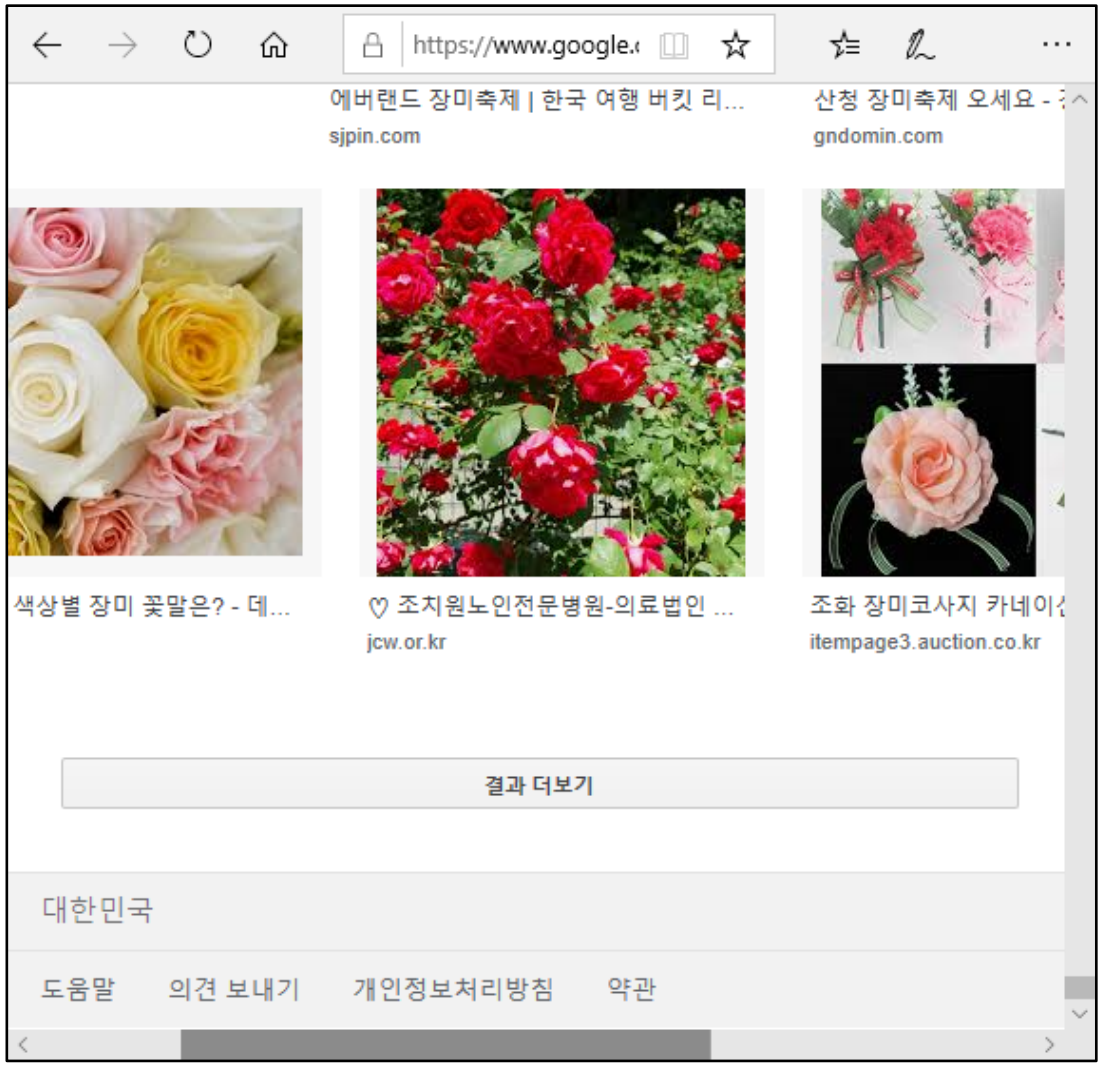
구글 이미지 검색 예

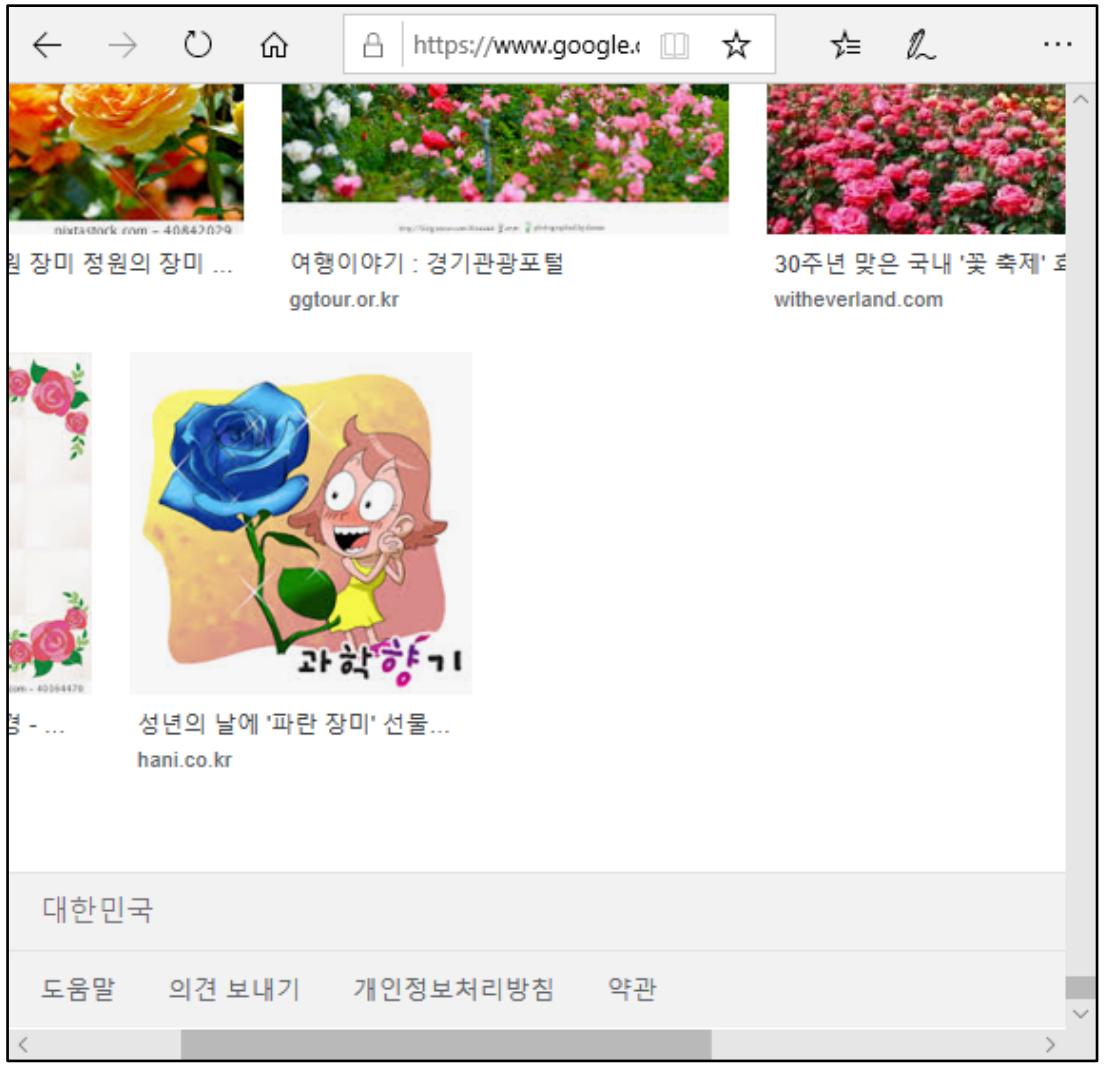

https://www.google.
Google
이미지
장미
의 원리


https://www.google.
있는 2018 서울...
활짝핀 장미꽃 담장사이로... - 경남도민신문
gndomin.com
파라다이스 공식블로그
blog.paradise.co.kr
의 백과사전
우리문화신문
koya-culture.com
중랑구, 천만송이 장미의 향연 '20
tnews.kr

AI

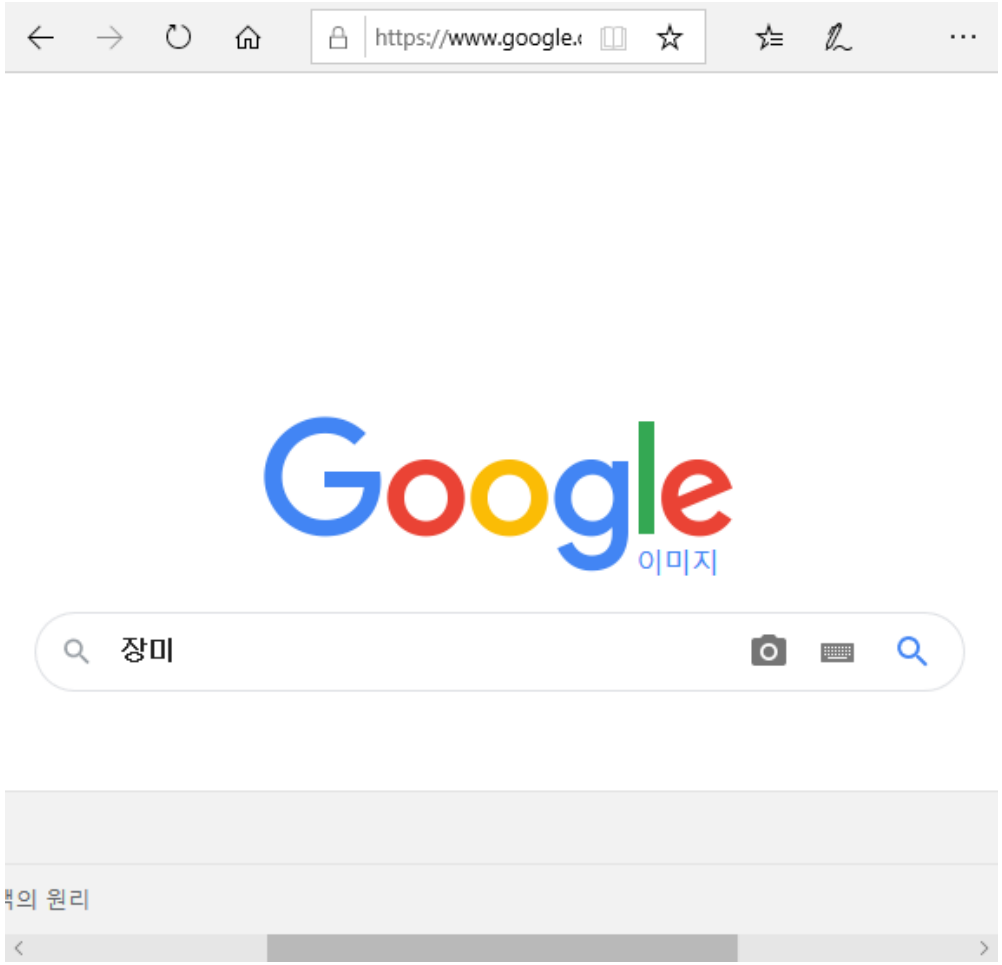
구글 이미지 검색 예


https://www.google.
에버랜드 장미축제 | 한국 여행 버킷 리...
sjpin.com
산청 장미축제 오세요 -
gndomin.com
색상별 장미 꽃말은? - 데...
♡ 조치원노인전문병원-의료법인 ...
jcw.or.kr
조화 장미코사지 카네이
itempage3.auction.co.kr
결과 더보기
대한민국
도움말
의견 보내기
개인정보처리방침
약관


https://www.google.
장미 정원의 장미 ...
여행이야기 : 경기관광포털
ggtour.or.kr
30주년 맞은 국내 '꽃 축제'
witheverland.com
과학향기
성년의 날에 '파란 장미' 선물...
hani.co.kr
대한민국
도움말
의견 보내기
개인정보처리방침
약관

AI

구글 이미지 검색 예



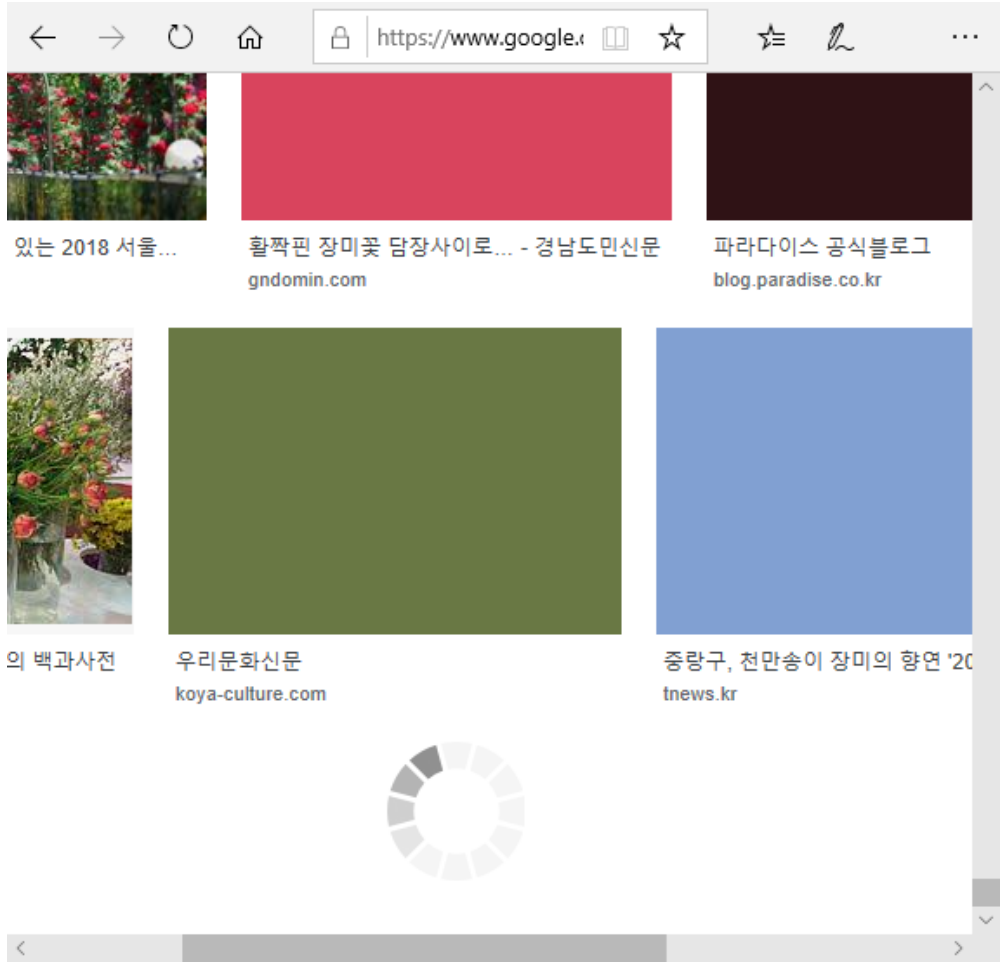
① 이미지 검색 URL로 이동 (www.google.co.kr/imghp)

② 검색 폼에 키워드 입력

③ 검색 버튼 클릭

AI

구글 이미지 검색 예



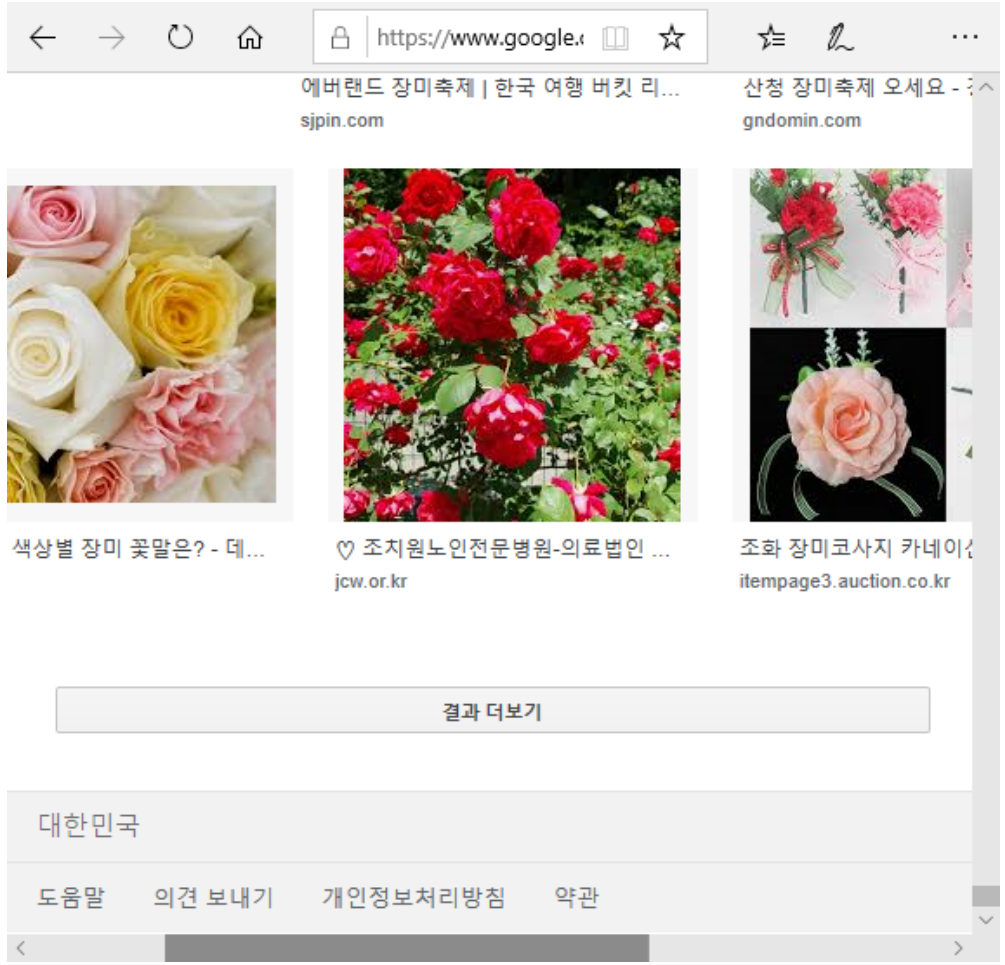
④ 최초 검색 결과 100개 표시

⑤ 스크롤링을 반복하여 나머지 검색 결과가 동적으로 페이지에 추가 표시(대기 시간 필요)

⑥ 페이지 하단에 footer등이 존재하나 숨김 상태

AI

구글 이미지 검색 예

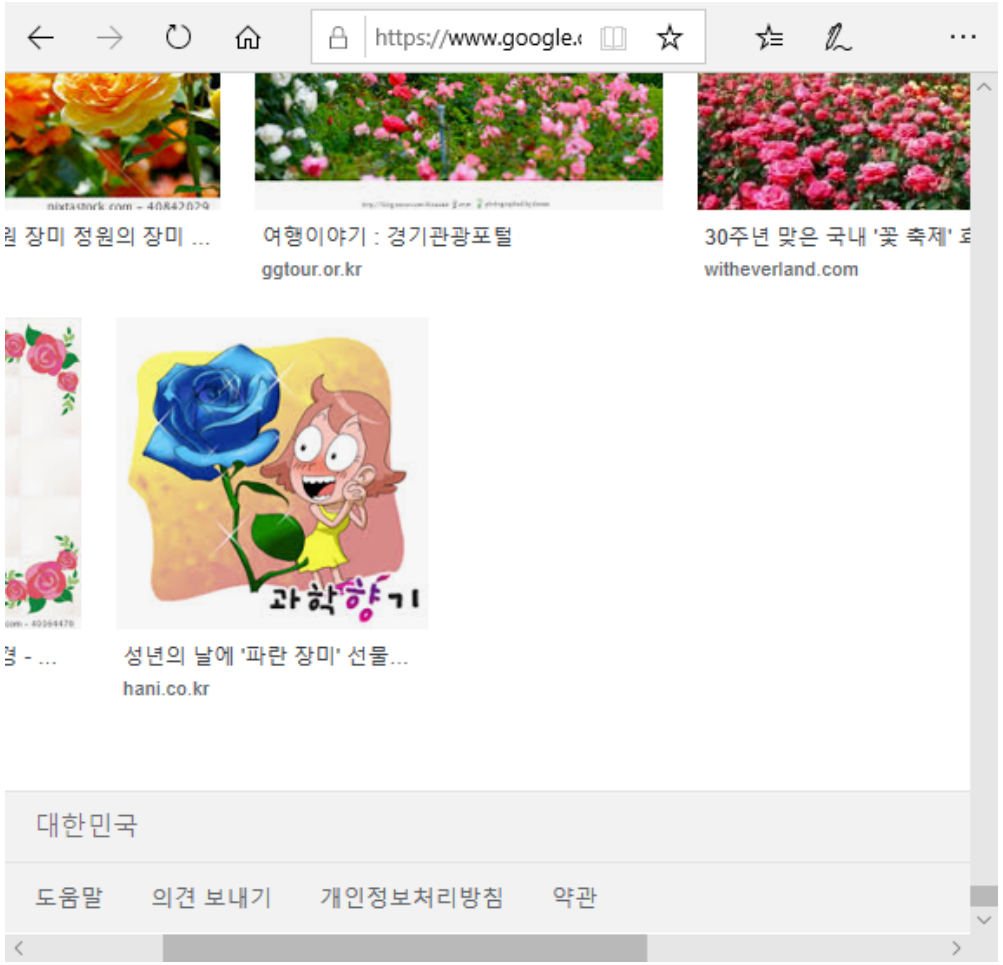


⑦ 일정 검색 결과가 표시된 후에는 페이지 하단에 숨김 상태로 있던 Footer와 '결과 더보기' 버튼이 표시

⑧ '결과 더보기' 버튼 클릭

AI

구글 이미지 검색 예



⑨ 동적으로 추가되는 결과를 모두 표시하고 나면 페이지 끝에 숨김 상태로 있던 footer 표시

AI

› 실습 해야 됨....
› 중간평가!