응용: BeautifulSoup4를 활용한 데이터 크롤링

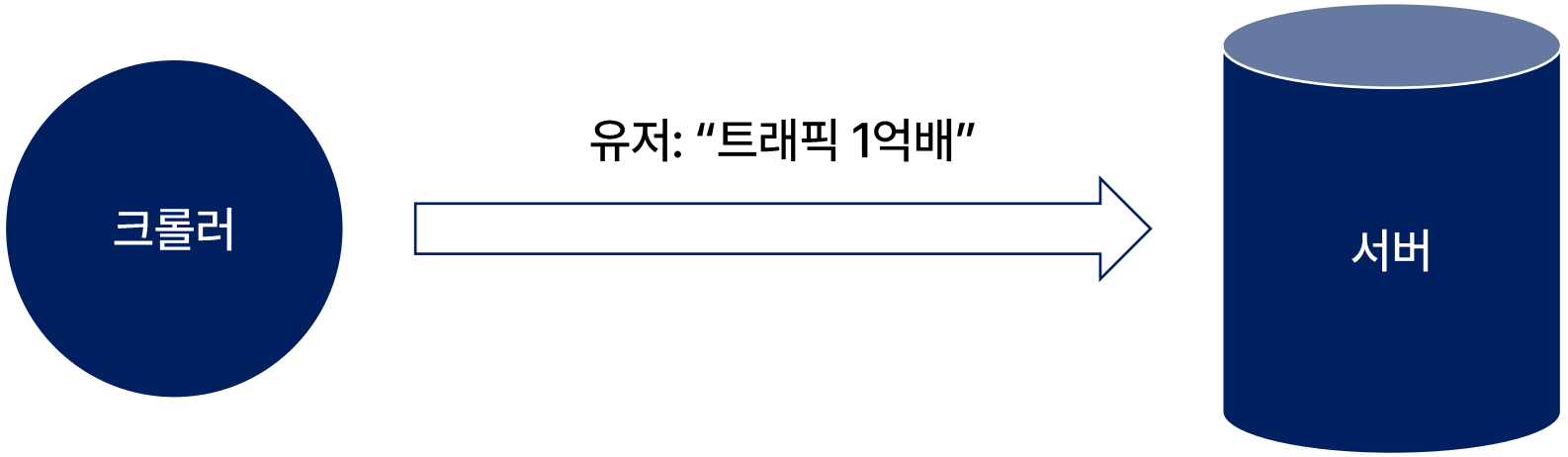
송기태 (kitae040522@gmail.com)
Soongsil Univ. (Computer Science and Engineering)

Content

쿠팡에서 상품 긁어오기

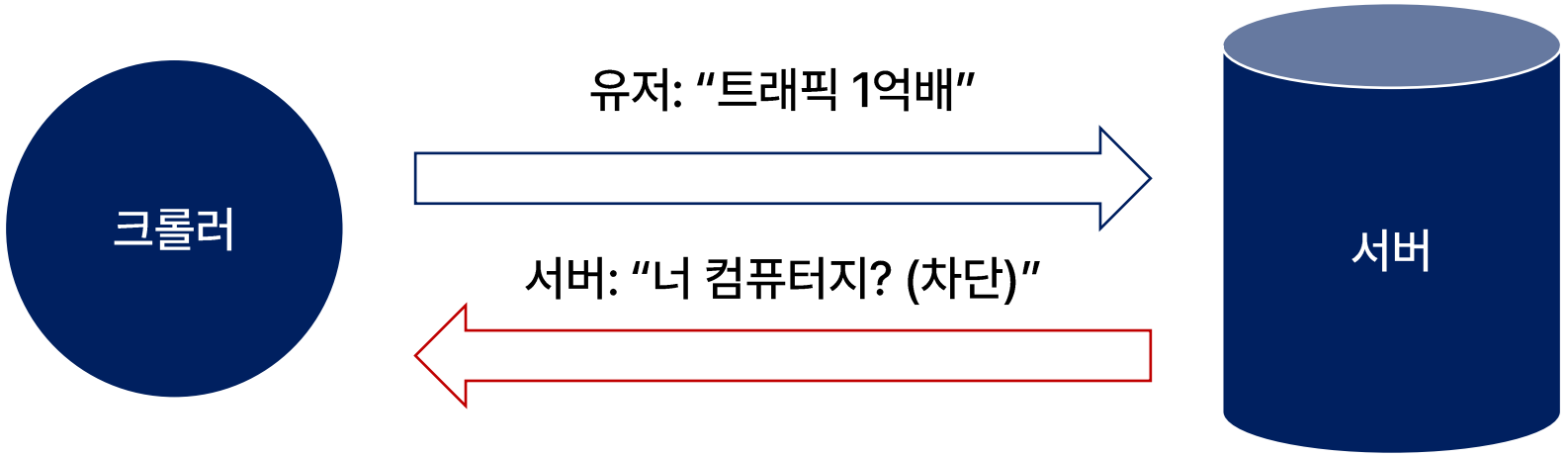
멜론에서 탑100 차트 긁어오기

쿠팡에서 상품 긁어오기



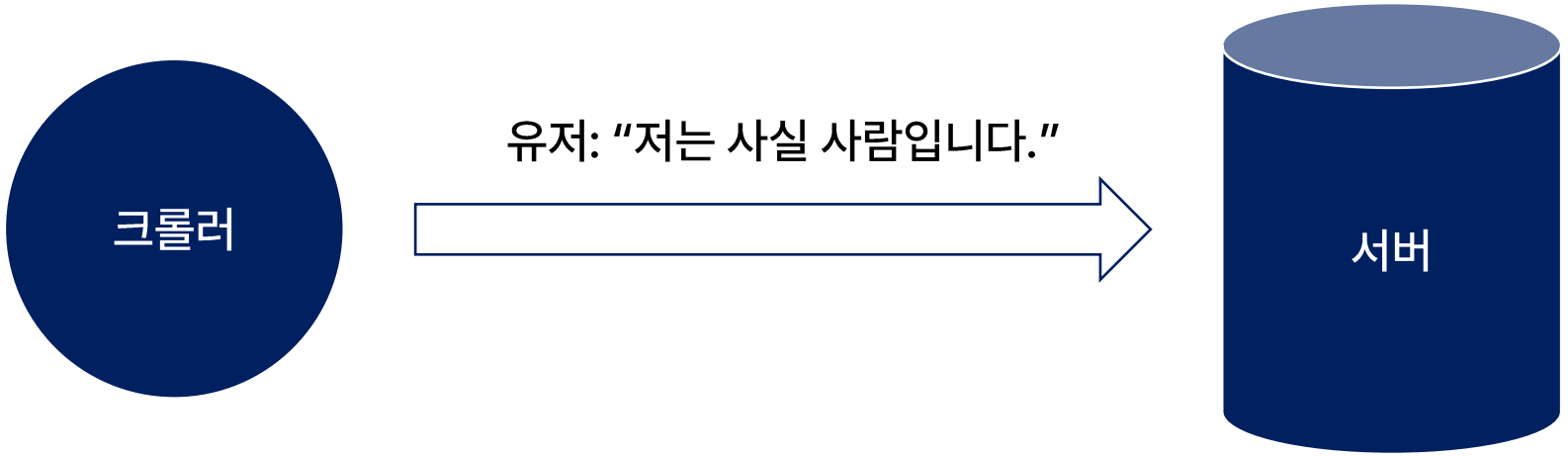
서비스 회사 입장에서는 크롤링 하지말라고 차단할 수가 있음 (자원 낭비하는 일을 왜 허락해줌)

쿠팡에서 상품 긁어오기



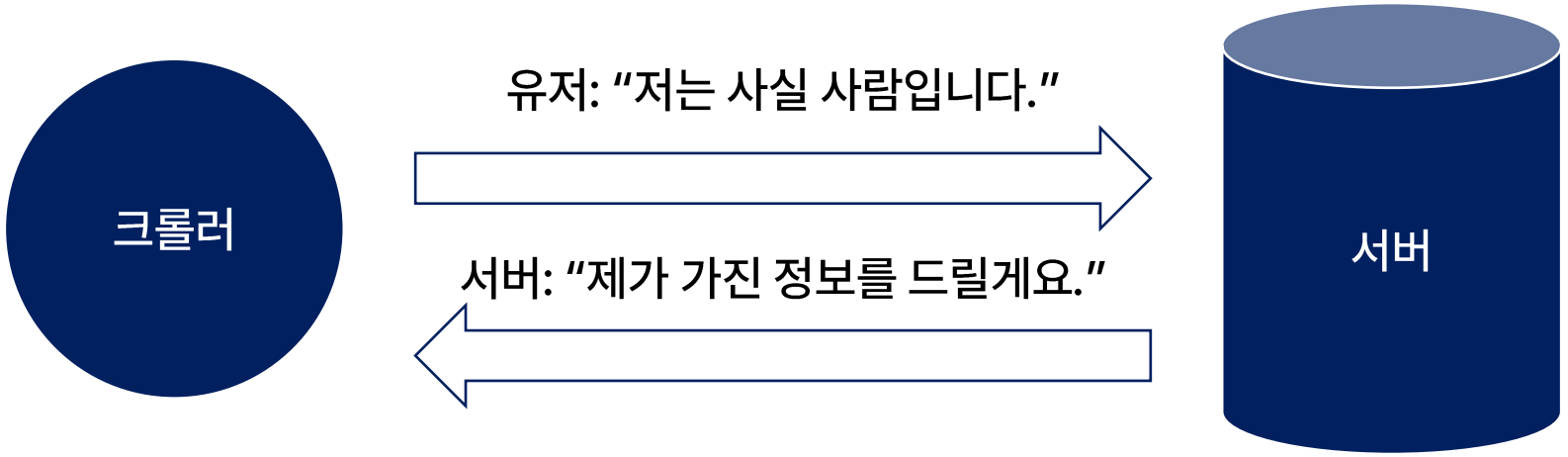
캡챠(cAPTCHA)를 도입하는 이유도 비슷함 (봇들의 요청을 무시하기 위해서)

쿠팡에서 상품 긁어오기



우회하는 방법은 있음. 하지만, 사실 크롤링은 합법적인 방법은 아님...

쿠팡에서 상품 긁어오기



HTTP 요청 헤더에 ‘User-Agent’를 브라우저를 사용하는 유저인 척 꾸며주면 봇 아닌 척 가능.

쿠팡에서 상품 긁어오기

```python
from bs4 import BeautifulSoup
import requests as req

if __name__ == '__main__':
    def main():
        words = '아이폰 14 Pro'
        url = 'https://www.coupang.com/np/search?component=&q=' + words

        headers = {
            'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)'
        }
        html = req.get(url, headers=headers)
        soup = BeautifulSoup(html.content, 'html.parser')

        res = {}

        product_list = soup.find('ul', {'id': 'productList'}).findAll('li')
        for idx, item in enumerate(product_list):
            product = item.find('div', {'class': 'name'}).text
            price = item.find('strong', {'class': 'price-value'}).text
            res[product] = price + '원'
        print(res)

    main()
```

멜론에서 탑100 차트 긁어오기

```
from bs4 import BeautifulSoup
import requests as req

if __name__ == '__main__':
    def main():
        url = 'https://www.melon.com/chart/index.htm'

        headers = {
            'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)'
        }
        html = req.get(url, headers=headers)
        soup = BeautifulSoup(html.content, 'html.parser')
        top100_box = soup.find('div', {'class': 'service_list_song type02 d_song_list'})
        top100_table = top100_box.find_all('tr', {'class': ['lst50', 'lst100']})

        for idx, item in enumerate(top100_table):
            music_title = item.find('div', {'class': 'ellipsis rank01'}).a.text
            music_artist = item.find('div', {'class': 'ellipsis rank02'}).a.text
            print(f'{idx + 1}: {music_title} | {music_artist}')


    main()
```

Thank You!

송기태 (kitae040522@gmail.com)
Soongsil Univ. (Computer Science and Engineering)