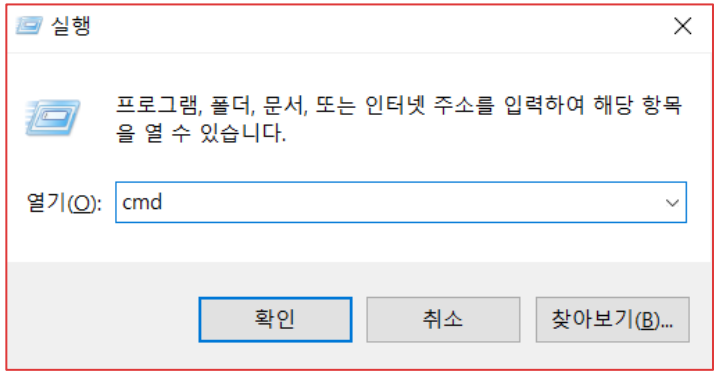
Data 수집하기

Crawling

크롤링 (Crawling)

- 웹페이지의 데이터를 가져와 가공하는 것
- Parsing : 전체 데이터에서 내가 원하는 데이터를 특정 패턴이나 순서로 추출하여 가공하는 것
- 크롤링 순서
    1. 페이지 데이터 가져오기
    2. 페이지의 HTML 코드 구조 분석하여 파싱(parsing)
    3. 추출한 데이터를 사용

외부 라이브러리(패키지) 설치



- 실행(Windows Key + R) → cmd
- pip install PackageName
  - pip : 파이썬에서 제공하는 패키지 관리 시스템
  - 패키지 이름을 정확히 입력 해야 한다.
  - Successfully installed ~~ 가 출력되면 설치 성공!
  - 이미 설치된 경우, already ~~ 관련 내용 출력

- 크롤링 관련 패키지 3개 설치
  - pip install requests
  - pip install beautifulsoup4
  - pip install selenium

PYTHON에서의 크롤링

- requests, selenium, beautifulsoup4 라이브러리 사용
  - requests : 페이지에서 데이터 가져오기
  - selenium : 가상 웹브라우저를 이용하여 웹 자동화 가능
  - beautifulsoup4 : 읽어온 HTML 코드를 파싱
  - 이 외 다른 라이브러리를 사용할 수도 있다.
- requests 의 get() 함수로는 크롬에서 '페이지 소스 보기'를 했을 때의 순수 페이지 코드만 읽어올 수 있다.
- Selenium의 가상 브라우저를 이용하면 실제 브라우저가 페이지에 요청하여 받은 코드를 읽을 수 있다.
  - F12키를 눌렀을 때 나오는 개발자 도구에서 보는 HTML 코드

Selenium 라이브러리

- 가상 브라우저를 구동시켜 마치 사람이 브라우저를 통해 페이지를 접속한 것과 같은 효과를 낼 수 있다.
- Chrome 브라우저를 사용 → PC에 Chrome 설치 필수
- chromedriver.exe 를 실행시키면 자동화된 크롬 브라우저가 실행되고, 우리는 코드를 통해 브라우저를 조작한다.

HTML 기본 구조

- <tag attr1="value" attr2="value">일반텍스트</tag>
  - Tag는 <>열리면 </> 닫히는 구조이다.
    - <tag1> </tag> : 하위 태그를 가지거나 Text를 가졌을 때 닫기
    - <tag1 /> : 하위 태그가 없고 속성만 가진 경우, 한 번에 닫기
  - Tag는 속성(attribute)을 가질 수 있다.
  - Tag의 바깥쪽에는 그 Tag의 Text를 가질 수 있다. (검은색 글씨)
  - Tag는 트리(Tree)구조로 하위 태그를 가질 수 있다.

```
<tag1>
        <tag2 attr="12345"/>
</tag1>
```

다나와 크롤링 분석 예제

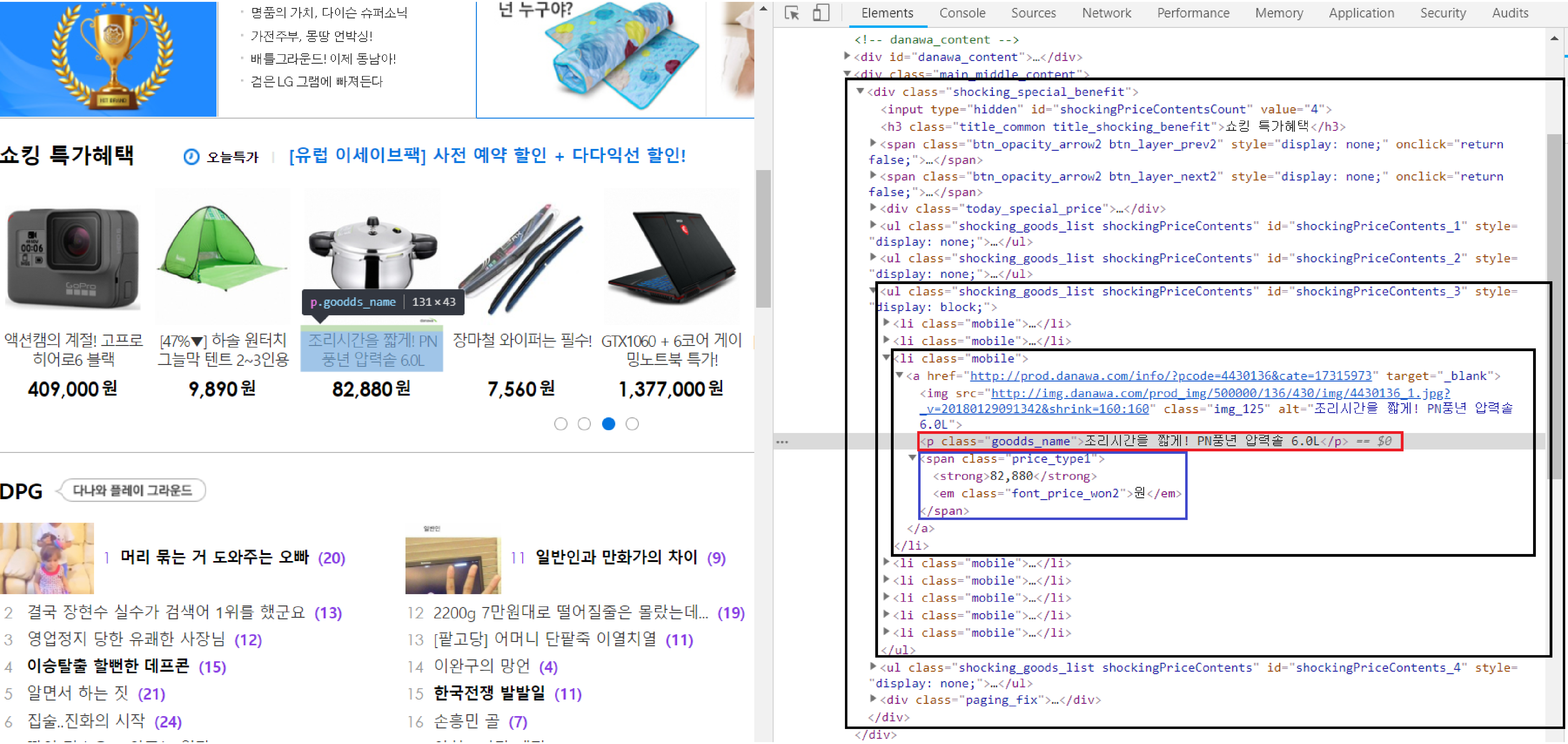
- http://www.danawa.com
- 다나와 메인 페이지의 '쇼킹 특가혜택' 의 32개 제품 크롤링
  - 하나의 tab에 8개의 제품 정보가 있고, tab은 총 4개이다. (8*4=32)

제품명 : <p> 의 text
가격 : 태그 아래 하위 태그들의 text
(반복되는 구간을 찾아야함)

```
<ul>
        <li>
         <a>
          <p> 제품명
          <span> 가격
        </li>
</ul>
```

가 8개 있다 = 하나의 탭에 제품이 8개 있다.
 4개 있다 = 쇼킹특가혜택은 4개의 탭이 있다.
전체 html 코드 안에는 ul, li 태그명은 너무 많다...

→ 태그 하나는 제품 하나의 정보를 가지고 있다.
→ 태그 하나는 제품 8개의 정보(li태그)를 가지고 있다.
→ 전체 html에서 원하는 부분의 코드만 가져온다.



명품의 가치, 다이슨 슈퍼소닉
가전주부, 통땅 언박싱!
배틀그라운드! 이제 동남아!
검은 LG 그램에 빠져든다
넌 누구야?
쇼킹 특가혜택
오늘특가
[유럽 이세이브팩] 사전 예약 할인 + 다다익선 할인!
p.goodds_name 131 × 43
액션캠의 계절! 고프로 히어로6 블랙
409,000원
[47%▼] 하솔 원터치 그늘막 텐트 2~3인용
9,890원
조리시간을 짧게! PN 풍년 압력솥 6.0L
82,880원
장마철 와이퍼는 필수!
7,560원
GTX1060 + 6코어 게이밍노트북 특가!
1,377,000원
DPG
다나와 플레이 그라운드
1 머리 묶는 거 도와주는 오빠 (20)
2 결국 장현수 실수가 검색어 1위를 했군요 (13)
3 영업정지 당한 유쾌한 사장님 (12)
4 이승탈출 할뻔한 데프콘 (15)
5 알면서 하는 짓 (21)
6 집술..진화의 시작 (24)
11 일반인과 만화가의 차이 (9)
12 2200g 7만원대로 떨어질줄은 몰랐는데... (19)
13 [팥고당] 어머니 단팥죽 이열치열 (11)
14 이완구의 망언 (4)
15 한국전쟁 발발일 (11)
16 손흥민 골 (7)
Elements Console Sources Network Performance Memory Application Security Audits
<!-- danawa_content -->
<div id="danawa_content">…</div>
<div class="main_middle_content">
<div class="shocking_special_benefit">
<input type="hidden" id="shockingPriceContentsCount" value="4">
<h3 class="title_common title_shocking_benefit">쇼킹 특가혜택</h3>
…
…
<div class="today_special_price">…</div>
<ul class="shocking_goods_list shockingPriceContents" id="shockingPriceContents_1" style="display: none;">…
<ul class="shocking_goods_list shockingPriceContents" id="shockingPriceContents_2" style="display: none;">…
<ul class="shocking_goods_list shockingPriceContents" id="shockingPriceContents_3" style="display: block;">
<li class="mobile">…
<li class="mobile">…
<li class="mobile">

<img src="http://img.danawa.com/prod_img/500000/136/430/img/4430136_1.jpg?_v=20180129091342&shrink=160:160" class="img_125" alt="조리시간을 짧게! PN풍년 압력솥 6.0L">
<p class="goodds_name">조리시간을 짧게! PN풍년 압력솥 6.0L</p> == $0

82,880
<em class="font_price_won2">원

<li class="mobile">…
<li class="mobile">…
<li class="mobile">…
<li class="mobile">…
<li class="mobile">…

<ul class="shocking_goods_list shockingPriceContents" id="shockingPriceContents_4" style="display: none;">…
<div class="paging_fix">…</div>
</div>
</div>

크롤링 주제 선정

- 가급적 관심 분야의 정보를 크롤링
  - 네이버 스포츠 중 선수나 팀 순위
  - 멜론 뮤직 TOP100 의 가수, 노래명
  - 네이버 웹툰 목록
  - 네이버 뉴스 댓글 많은 순
  - 기타 등등