웹 크롤링 입문: 네이버 IT/과학 뉴스 카테고리별 데이터 수집

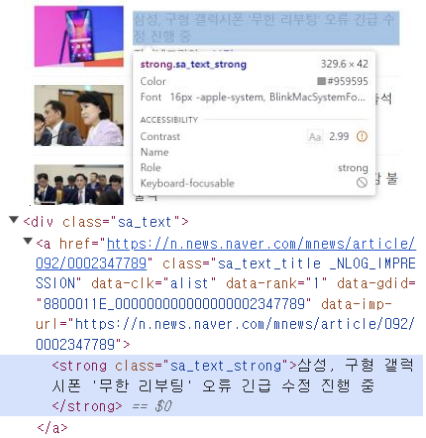
네이버 뉴스 사이트에서 IT/과학 섹션 내의 각 세부 카테고리(모바일, 인터넷/SNS, 통신/뉴미디어 등)에 대한 뉴스를 크롤링하고, 각 카테고리별로 30개의 뉴스 제목과 컨텐츠를 추출하는 과제를 진행하였습니다.

https://www.youtube.com/watch?v=yQ20jZwDjTE 해당 영상의 1:40:00까지 시청하여 웹 스크래핑, 웹 크롤링의 기본에 대해 배우고, 관련 여러 툴 또한 배웠습니다. 비교적 최근 영상이 아니기에 실제와 맞지 않는 부분들이 있었기에 정적 페이지를 대상으로 진행한 크롤링에 대해 찾아보았습니다.

https://www.inflearn.com/course/%ED%8C%8C%EC%9D%B4%EC%8D%AC-%ED%81%AC%EB%A1%A4%EB%A7%81-%EA%B8%B0%EC%B4%88 해당 강의를 통해 정적 페이지 크롤링에 대해 배우고 실습하였습니다. 이제 해볼 수 있겠다는 생각이 들어 네이버 IT/과학 뉴스 카테고리별 데이터 수집 과제를 진행해 보았습니다.

```
'모바일': 'https://news.naver.com/breakingnews/section/105/731',
'인터넷/SNS': 'https://news.naver.com/breakingnews/section/105/226',
'통신/뉴미디어': 'https://news.naver.com/breakingnews/section/105/22
'IT일반': 'https://news.naver.com/breakingnews/section/105/230',
'보안/해킹':'https://news.naver.com/breakingnews/section/105/732',
'컴퓨터':'https://news.naver.com/breakingnews/section/105/283',
'게임/리뷰':'https://news.naver.com/breakingnews/section/105/229',
'과학 일반':'https://news.naver.com/breakingnews/section/105/228'
```

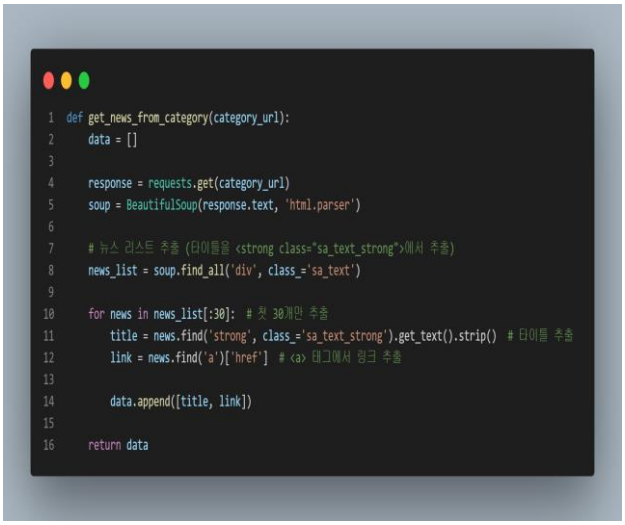
IT/과학 세션에는 총 8개의 카테고리가 있었고 각각에 해당하는 URL을 따주었습니다.



F12를 눌러 뉴스 제목을 확인해 보았습니다. HTML 구조를 확인해보니 타이틀은 안에 있었습니다. 해당 타이틀을 포함하는 링크는 부모태그인 <a>에 있는 것을 볼 수 있었습니다. 이어서 하나의 카테고리에 대해서 한 페이지에 몇 개의 뉴스가 있는지 확인하니 36개의 뉴스가 있었습니다.

페이지를 넘기면서 할 필요 없이 이중에 30개씩만 추려내면 되겠다고 생각을 한 후 진행을 하였습니다.

하나의 url 내에서 제목과 링크를 뽑아내는 함수를 작성하였습니다.



```
def get_news_from_category(category_url):
    data = []

    response = requests.get(category_url)
    soup = BeautifulSoup(response.text, 'html.parser')

    # 뉴스 리스트 추출 (타이틀을 <strong class="sa_text_strong">에서 추출)
    news_list = soup.find_all('div', class_='sa_text')

    for news in news_list[:30]:  # 첫 30개만 추출
        title = news.find('strong', class_='sa_text_strong').get_text().strip()  # 타이틀 추출
        link = news.find('a')['href']  # <a> 태그에서 링크 추출

        data.append([title, link])

    return data
```

제목과 링크를 뽑아내는 함수를 작성하고 나니 언론사, 제목, 간단한 내용 요약도 똑같은 방식으로 다 추출해 낼 수 있겠다는 생각이 들었습니다. 함수에 몇 가지 기능을 더 추가한 후 완성시켰습니다.

```python
def get_news_from_category(category_url):
    data = []

    response = requests.get(category_url)
    soup = BeautifulSoup(response.text, 'html.parser')

    # 뉴스 리스트 추출
    news_list = soup.find_all('div', class_='sa_text')

    for news in news_list[:30]:  # 첫 30개만 추출
        #제목 추출
        title = news.find('strong', class_='sa_text_strong').get_text().strip()
        #링크 추출
        link = news.find('a')['href']  #<a> 태그에서.
        # 언론사 추출
        press = news.find('div', class_='sa_text_press').get_text().strip()
        # 내용 추출
        content = news.find('div', class_='sa_text_lede').get_text().strip()

        data.append([press, title, link, content])  # 언론사, 타이틀, 링크, 내용 저장

    return data
```

추가적으로 제대로 크롤링을 했는지 실행했을시 결과로 출력이 되게 하였습니다.

또한 csv파일 형태로 저장하여 결과를 다시 확인할 수 있게도 하였습니다.

```
Category,Press,Title,Link,Content
모바일,연합뉴스,"[게시판] 티맵모빌리티, 17일 데이터·AI 활용 성장 전략 세미나",https://n.news.naver.com/mnews/article/001
모바일,서울경제,"방통위 김태규 ""구글·애플 반독점 행위, 단호한 규제 필요""",https://n.news.naver.com/mnews/article/011/
모바일,지디넷코리아,"삼성, 구형 갤럭시폰 '무한 리부팅' 오류 긴급 수정 진행 중",https://n.news.naver.com/mnews/article/09
```