크롤링 과제

By. 정은영

'쿠씨' 씨는 '(주) 지나치지마라탕' 의 마케팅공모전에 참여하게 되었다. 공모전을 준비하며 마라탕집이 대학 근처에 많이 위치한다는 것을 알았고, 이에 20대를 겨냥하여 마라탕마케팅을 진행하고자 한다. '쿠씨' 씨는 20대의 마라탕에 대한 이미지에 대해 알고 싶다. 또한 자신은 비올 때마다 마라탕이 먹고 싶었는데, 다른사람들도 이럴 지에 대해 궁금하였고, 만약 날씨와 마라탕소비가 관련있을 경우 '내일날씨예보'를 통해 시기에 따른 마케팅을 기획하고자 한다. 따라서 '쿠씨' 씨가 크롤링을 통해 얻고 싶은 데이터는 다음과 같다.

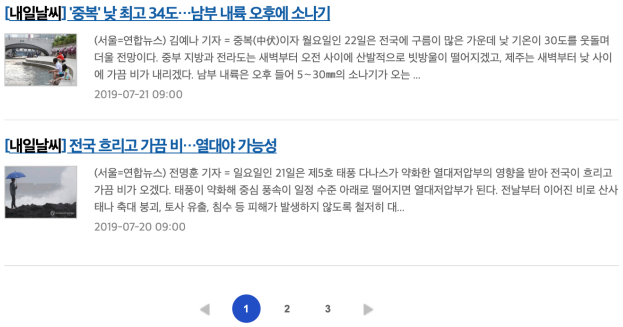
1. **인스타그램에 '마라탕' 해시태그를 검색하여 관련 게시글을 크롤링하여 어떤 단어가 많이 나왔는지 워드클라우드로 보여주어 마라탕과 관련된 20대의 머릿속 이미지를 제시하고 싶다**

   https://www.instagram.com/explore/tags/%EB%A7%88%EB%9D%BC%ED%83%95/?hl=ko

2. **날씨 예보 데이터가 필요하여 기상청 홈페이지에 들어가봤는데, 기상청 홈페이지에서는 과거날씨기록데이터만 있을 뿐 전날에 다음날의 날씨를 예측한 데이터가 없었다. '쿠씨' 씨는 관련 데이터를 모으기 위해 인터넷 기사를 크롤링하여 예측 날씨데이터를 csv 파일로 저장하고 싶다.**

   https://www.yna.co.kr/search/index?query=%EB%82%B4%EC%9D%BC%EB%82%A0%EC%94%A8&from=20190630&to=20190730&period=diy

   위는 관련 뉴스페이지이며, 검색어 '내일날씨'로 검색했을 때, 2019년 6월 30일부터 2019년 7월 30일까지의 검색내역이다. 아래 좌측은 위의 링크로 들어가서 더보기를 눌렀을 때이며 각각 기사에 직접 들어가서 아래로 스크롤하면 우측과 같은 날씨 정보가 제공된다.



다음은 31일 지역별 날씨 전망. [오전, 오후] (최저~최고기온) <오전, 오후 강수 확률>

▲ 서울 : [흐리고 가끔 비, 흐리고 비] (25~29) <70, 70>

▲ 인천 : [흐리고 가끔 비, 흐리고 비] (25~28) <70, 70>

▲ 수원 : [흐리고 한때 비, 흐리고 가끔 비] (25~30) <70, 70>

▲ 춘천 : [흐리고 가끔 비, 흐리고 비] (25~29) <70, 80>

▲ 강릉 : [구름많음, 구름많음] (28~35) <20, 20>

▲ 청주 : [구름많음, 구름많고 한때 비] (26~33) <20, 60>

힌트1> 좌측페이지에서 링크를 긁어서 리스트에 저장한다. (시도해보면 알겠지만 html 태그로 데이터가 보이지 않으며, 따라서 selenium 을 사용해야 할 것이다) + (selenium은 드라이버로 보이는 웹을 그대로 가져오기 때문에 좌측이미지의 페이지를 넘기며 url 을 추가할 수 있다) 좌측 페이지의 output은 아래 형태면 좋다

	date	news_url
0	20190930	http://www.yna.co.kr/view/AKR20190930008000004...
1	20190929	http://www.yna.co.kr/view/AKR20190929004300004...
2	20190928	http://www.yna.co.kr/view/AKR20190928011500004...

```
tmpString = str(listOfsoup[0].find("a").get_text())
tmpString
```

```
'BLT\r\nOld Oak Tap\nRead more '
```

```
# \n or \r\n 으로 들어가는 부분을 구분
```

```
re.split(('\n|\r\n'), tmpString)
```

```
['BLT', 'Old Oak Tap', 'Read more ']
```

힌트2> 위의 형태로 csv로 저장 후, 기사의 서울/인천 … 부분을 긁어오기 위해 ▲가 포함된 문자열을 찾자. 그 후에 필요한 함수는 re.split() 함수이며, 사용방법은 오른쪽 그림과 같다. Output은 아래 형태면 좋다

	기사날짜	지역	오전날씨	오후날씨	최저기온	최고기온	오전강수량	오후강수량	뉴스링크
0	20190630	서울	맑음	맑음	20	29	0	0	http://www.yna.co.kr/view/AKR20190630003200004...
1	20190701	서울	맑음	맑음	20	30	0	0	http://www.yna.co.kr/view/AKR20190701006200004...
2	20190702	서울	맑음	구름많음	20	31	0	30	http://www.yna.co.kr/view/AKR20190702010500004...