지난 url 따는 코드(9.10)에서 실수한 부분

```
'
이터 내보내기 (url)
'

ble = pd.DataFrame(data_set.keys(), data_set.values())
ble = table.drop_duplicates()
ble.to_csv('전체기사 url주소({}_{}).csv'.format(first_date_str,end_date_str) , enc
```

이번 url 따는 코드 (9.13)에서 수정된 부분

```
데이터 내보내기 (url)
'''

table = pd.DataFrame(data_set.keys(), data_set.values())
table = table.drop_duplicates()
table.to_csv('전체기사 url주소.csv', encoding = 'cp949')
```

전체 url 웹크롤러 목표날짜 입력란

➔ First_date, end_date 각각에 YYYYMMDD 형식으로 작성

In [ ]:

```
'''1. 목표 날짜 지정  (검색할 날짜)'''
# 검색 시작 날짜
first_date = 20200101




first_date_dt = dt.datetime(int(str(first_date)[:4]) , int(str(fi
first_date_str = first_date_dt.strftime('%Y%m%d')



# 검색 종료 날짜
end_date = 20210910
```

본문 웹크롤러

삭제해도 되는 부분

In [1]:

```
# install pakages  (  -> first time )


!pip install urllib3
!pip install bs4
!pip install pandas
!pip install tqdm


# 기본 세팅

from urllib.error import HTTPError
```

생성되는 파일의 파일명

```
        dataframe.columns = ['date', 'title', 'body', 'company']
        # 본문 내 특수문자 제거
        dataframe['body'] = dataframe['body'].str.replace(pat=r'[
        dataframe['body'] = dataframe['body'].str.replace(pat=r'[
        # 한자 제거
        dataframe['body'] = dataframe['body'].str.replace(pat=r'[
        dataframe['body']= dataframe['body'].str.replace(pat=r'['
        dataframe['body'] = dataframe['body'].str.replace(' b ',
        dataframe = dataframe.reset_index()
    dataframe.to_csv('동아일보 크롤링.csv', encoding = "cp949", i


##################################################
end_time = t.time()
duration = round((end_time - start_time)/60, 2)
```