# 네이버 기사 크롤링

- http://news.naver.com/main/main.nhn?mode=LSD&mid=shm&sid1=105
- thread 사용

In [ ]:

In [18]:
```
!pip install selenium
```

```
Requirement already satisfied: selenium in c:\programdata\anaconda3\lib\site-packages (4.28.0)
Requirement already satisfied: urllib3<3,>=1.26 in c:\programdata\anaconda3\lib\site-packages (from urllib3[socks]<3,>=1.26->se
lenium) (2.2.2)
Requirement already satisfied: trio~=0.17 in c:\programdata\anaconda3\lib\site-packages (from selenium) (0.28.0)
Requirement already satisfied: trio-websocket~=0.9 in c:\programdata\anaconda3\lib\site-packages (from selenium) (0.11.1)
Requirement already satisfied: certifi>=2021.10.8 in c:\programdata\anaconda3\lib\site-packages (from selenium) (2024.6.2)
Requirement already satisfied: typing_extensions~=4.9 in c:\programdata\anaconda3\lib\site-packages (from selenium) (4.11.0)
Requirement already satisfied: websocket-client~=1.8 in c:\programdata\anaconda3\lib\site-packages (from selenium) (1.8.0)
Requirement already satisfied: attrs>=23.2.0 in c:\programdata\anaconda3\lib\site-packages (from trio~=0.17->selenium) (24.3.0)
Requirement already satisfied: sortedcontainers in c:\programdata\anaconda3\lib\site-packages (from trio~=0.17->selenium) (2.4.
0)
Requirement already satisfied: idna in c:\programdata\anaconda3\lib\site-packages (from trio~=0.17->selenium) (3.7)
Requirement already satisfied: outcome in c:\programdata\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.3.0.post0)
Requirement already satisfied: sniffio>=1.3.0 in c:\programdata\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.3.0)
Requirement already satisfied: cffi>=1.14 in c:\programdata\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.16.0)
Requirement already satisfied: wsproto>=0.14 in c:\programdata\anaconda3\lib\site-packages (from trio-websocket~=0.9->selenium)
(1.2.0)
Requirement already satisfied: pysocks!=1.5.7,<2.0,>=1.5.6 in c:\programdata\anaconda3\lib\site-packages (from urllib3[socks]<
3,>=1.26->selenium) (1.7.1)
Requirement already satisfied: pycparser in c:\programdata\anaconda3\lib\site-packages (from cffi>=1.14->trio~=0.17->selenium)
(2.21)
Requirement already satisfied: h11<1,>=0.9.0 in c:\programdata\anaconda3\lib\site-packages (from wsproto>=0.14->trio-websocket~
=0.9->selenium) (0.14.0)
```

In [19]:
```
!pip install webdriver_manager
```

```
Requirement already satisfied: webdriver_manager in c:\programdata\anaconda3\lib\site-packages (4.0.2)
Requirement already satisfied: requests in c:\programdata\anaconda3\lib\site-packages (from webdriver_manager) (2.32.2)
Requirement already satisfied: python-dotenv in c:\programdata\anaconda3\lib\site-packages (from webdriver_manager) (0.21.0)
Requirement already satisfied: packaging in c:\programdata\anaconda3\lib\site-packages (from webdriver_manager) (23.2)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\programdata\anaconda3\lib\site-packages (from requests->webdriver
_manager) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\programdata\anaconda3\lib\site-packages (from requests->webdriver_manager)
(3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\programdata\anaconda3\lib\site-packages (from requests->webdriver_manag
er) (2.2.2)
Requirement already satisfied: certifi>=2017.4.17 in c:\programdata\anaconda3\lib\site-packages (from requests->webdriver_manag
er) (2024.6.2)
```

In [20]:
```python
from selenium import webdriver
```

In [21]:
```python
from selenium.webdriver.chrome.service import Service as ChromeService
from webdriver_manager.chrome import ChromeDriverManager
```

In [22]:
```python
from selenium.webdriver.common.by import By
```

In [23]:
```python
article_list = []

def get_article(page):
    ## driver = webdriver.Chrome("C:/Myexam/chromedriver/chromedriver.exe")

    driver = webdriver.Chrome(service=ChromeService(ChromeDriverManager().install()))

    driver.get("https://news.naver.com/section/10" + str(page))
    articles = driver.find_elements(By.CSS_SELECTOR,'#newsct li')


    for article in articles:
        try:
            tmp_elements = article.find_elements(By.CSS_SELECTOR, '.sa_text strong')
            if tmp_elements:
                title = tmp_elements[0].text
            else:
                tmp_elements2 = article.find_elements(By.CSS_SELECTOR, '.ss_text a')
                if tmp_elements2:
                    title = tmp_elements2[0].text
```

```
                else:
                    title = "해당 정보 없음"

            article_list.append(title)
        except:
            print("에러 발생!")

    print("end :", page)

    driver.quit()
```

In [24]:
```
%%time
for page in range(1, 5):
    get_article(page)
```

```
end : 1
end : 2
end : 3
end : 4
CPU times: total: 594 ms
Wall time: 40.7 s
```

In [25]:
```
len(article_list), article_list[:30]
```

Out[25]:
```
(201,
 ['‘트럼프 AI투자’에 강세 보이는 뉴욕증시…S&P500 장중 최고가',
  "국토부 'LCC 특별점검' 회의…안전기준 미달 항공사에 운항증명 정지",
  "[속보] SK하이닉스, 연간 영업익 23.5조…'사상 최대 실적'",
  '반려동물 양육비 月 14.2만원…유기동물 입양 의향 80%',
  '"피벗의 시대, 핵심 사업 중심 포트폴리오 리밸런싱 필요"',
  '기업심리 석 달 연속 하락…펜데믹 이후 최저',
  '작년 성장률 2% 턱걸이…계엄 충격 등에 4분기 0.1% 그쳐',
  '삼성바이오, 연 매출 4조↑…국내 바이오 첫 ‘4조 클럽’',
  'ETF에 꽂힌 연금개미…한투證 "퇴직연금 ETF 투자 1년간 두배↑"',
  "기보, 부산 소외계층 노인에 '설 맞이 특식' 나눔",
  '티웨이항공, 경영권 분쟁 본격화에 52주 신고가 경신',
  'LG이노텍, 지난해 실적 부진에 6%↓…52주 신저가',
  "LS일렉트릭, '호실적+AI 수혜 기대감'에 8%↑",
  "현대건설, 영업적자에도 '빅 배스' 기대감…이틀 연속 상승",
  '해당 정보 없음',
  '해당 정보 없음',
  'HD한국조선해양, 올해 마수걸이…컨테이너선 12척 3조7000억 수주 확정',
  '셀트리온 스테키마, 프랑스·영국 등 유럽 5개국 출시',
  '`3시간 개최 지연` 고려아연 임시주총, 정오쯤 시작될 듯',
  '우이신설선 연장선, HL D&I한라 컨소시엄과 수의계약 추진',
  '‘올해 첫 공모주 상장’ 미트박스, 상장 첫날 장초반 약세',
  '美트럼프 관세부과·IRA폐지에…日자동차 업계 비상',
  '"HBM 내년 물량까지 상반기 계약"…SK하이닉스 역대급 실적 축포(종합)',
  '최대 수천억 받는다고?...K배터리 "구원투수 왔다" 외친 이 법은',
  '티웨이·에어프레미아, 대명소노로 합쳐지면…제3의 FSC 탄생할까',
  '한화그룹, 설 연휴 전 협력사 대금 1700억 조기 지급',
  '‘한국판 IRA’ 조특법 개정한 발의…배터리 업계 “구원투수 될 것” 환영',
  '"조카 선물사러 갈까?"…W컨셉, 설빔 구매 10배 ↑',
  'SK하이닉스 "HBM 장기 성장…내년 공급 이미 논의 중"',
  '여야정 연금개혁 재시동…보험료·소득대체율부터 논의'])
```

pandas사용

In [26]:
```
import threading
import pandas as pd

df = pd.DataFrame(columns=["category","title"])
```

In [27]:
```
def get_article(page):
    #driver = webdriver.Chrome("C:/Myexam/chromedriver/chromedriver.exe")

    driver = webdriver.Chrome(service=ChromeService(ChromeDriverManager().install()))

    driver.get(
        "https://news.naver.com/section/10" + str(page))
    category = driver.find_element(By.CSS_SELECTOR, '.ct_snb_h2_a').get_attribute("innerText")
    articles = driver.find_elements(By.CSS_SELECTOR,'#newsct li')

    for article in articles:
        try:
            tmp_elements = article.find_elements(By.CSS_SELECTOR, '.sa_text strong')
            if tmp_elements:
                title = tmp_elements[0].text
            else:
                tmp_elements2 = article.find_elements(By.CSS_SELECTOR, '.ss_text a')
                if tmp_elements2:
                    title = tmp_elements2[0].text
                else:
                    title = "해당 정보 없음"
        except:
            print("에러 발생!")


        df.loc[len(df)] = {
            "category": category,
            "title": title,
        }
    print("end :", page+1)

    driver.quit()
```

In [28]:
```
%%time
for page in range(0, 6):
    get_article(page)
```

```
end : 1
end : 2
end : 3
end : 4
end : 5
end : 6
CPU times: total: 1.03 s
Wall time: 1min 7s
```

In [29]:
```
df
```

Out[29]:

| | category | title |
|---|---|---|
| **0** | 정치 | [영상] 홍장원, 尹통화내용 공개…"'싹다 정리' 지시에 간첩사건인 줄" |
| **1** | 정치 | 민주, 일타강사 전한길 '부정선거론' 유튜브 영상 신고 |
| **2** | 정치 | 오세훈 "트럼프 행정부와 '핵 잠재력' 향상 논의해야" |
| **3** | 정치 | 권성동 "이재명 '전 국민 25만 원 상품권', 조기 대선 염두 현금살포" |
| **4** | 정치 | 허은아 "이준석, 대표 내놓아라? 그가 했던 대로 똑같이 하겠다" 가처분 예고 |
| **...** | ... | ... |
| **297** | IT/과학 | [사이언스영상] 12시간 정지했다 재개한 틱톡 |
| **298** | IT/과학 | "스타게이트는 허구?"…머스크, 오픈AI 700兆 프로젝트에 의문 제기 |
| **299** | IT/과학 | 구글 딥마인드 "올해 말 AI 신약 개발·임상실험 목표" |
| **300** | IT/과학 | "역시 IT 강국" 세계 최초…ETRI, 200Gbps급 '6G' 무선시연 성공 |
| **301** | IT/과학 | "갤럭시Z폴드 7, S펜 지원 중단…더 얇아질 것" |

302 rows × 2 columns

In [30]:
```
df['category'].value_counts()
```

Out[30]:
```
category
경제        52
IT/과학     52
정치        50
사회        50
생활/문화    50
세계        48
Name: count, dtype: int64
```

In [ ]:

In [ ]: