네이버 기사 크롤링

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

In [1]:
```
from selenium import webdriver
```

In [2]:
```
from selenium.webdriver.chrome.service import Service as ChromeService
from webdriver_manager.chrome import ChromeDriverManager
```

In [3]:
```
from selenium.webdriver.common.by import By
```

셀레니움을 통해 크롬 브라우저를 실행하여 자동으로 네이버 뉴스페이지에서 분야별로 뉴스의 타이틀을 article_list에 넣는 함수

In [4]:
```
article_list = []


def get_article(page):

    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()
```

%%time은 Jupyter Notebook에서 해당 셀의 실행 시간을 측정하는 매직 커맨드

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

```
end : 1
end : 2
end : 3
end : 4
CPU times: total: 406 ms
Wall time: 29.1 s
```

In [6]:
```
# 기사의 총 개수와 일부 타이틀 확인
```

In [7]:
```
len(article_list), article_list[:30]
```

Out[7]:
```
(152,
 ['한성숙 장관 "금융 등 상생협력 범위 확대"…은탑산업훈장 삼성전자',
  "대형 GA 열 중 셋은 내부통제 미흡…사이버사고 관리는 '위험'",
  "배민 '로드러너' 엇갈린 주장…라이더 수익 개선 vs 수익 불안정",
  '엔비디아, 구글에 견제구…“우리가 업계보다 한세대 앞선다”',
  "LS '3세 경영' 속도…구동휘 MnM 대표이사, 사장 승진",
  '“주행 퍼포먼스 강화”…현대차·기아, 미쉐린과 타이어 기술 공동 개발',
  '이재용, ‘인도 최고재벌’ 암바니 회장과 만찬…삼성 AI·6G 첨단기술 총집결',
  '\'또 삼성전자야?\'…"구글 효과에 주가 뛴다" 개미들 \'두근\'',
  '"딸기 시즌 포문 연다" 롯데마트, 팩 딸기 전품목 할인',
  '현대차·기아, 2년 연속 안전관리 최우수연구실 선정',
  '기관 매수세에…코스피, 3900 회복 시도',
  '뒷심 부족했던 코스피, 외인 ‘사자’에도 3850선 강보합',
  '코스피, 美 기술주 훈풍에 3900선 회복',
  '개인·외인 쌍끌이 매도에…코스피, 3850선 붕괴',
  '해당 정보 없음',
  '해당 정보 없음',
  '[단독] 빗썸, 미인가 거래소에서 100억원대 코인 수령… 특금법 위반 소지',
  '“핸드크림이 커피 한 잔 값도 안 되네”…이제 화장품도 무신사에서',
  '환율 폭등 진짜 이유는 \'유동성 증가 속도\'..."美보다 2배 빨라"',
  '“악몽”·“이게 맞나 눈 의심”…‘보름 만에 -22%’ 50만닉스도 위태롭다는데, 왜? [종목Pick]',
  '코스피 3,890선 상승 출발…환율 1,465원대로 내려',
  '엔씨소프트 주가 4거래일째 상승…“수수료율 절감 효과 기대” [종목Pick]',
  '삼성전자 AI 가전, 중남미서 불티…1년새 판매 40% 늘어',
  "[하우머니] 코스피 '전강후약'…반도체 대장주보다 소부장주 랠리?",
  '‘매수심리 회복’ vs ‘미분양 누적’…내년 수도권·지방 집값 더 벌어진다',
  '비트코인, 500일 만에 최장 약세…"증시 출렁이면 또 흔들려"[코인브리핑]',
  "'기관 매수세' 코스피 상승 출발…3900선 회복",
  '삼성생명 부사장에 오성용·이상희·이팔훈 선임',
  '“삼양식품, 자사주 소각 의무화 앞두고 선제 매각…자산 취급 나쁜 선례”',
  '“전세사기·아파트 값 부담↑”…서울 오피스텔 거래량 크게 상승'])
```

pandas 사용

In [8]:
```
import threading
import pandas as pd

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

In [9]:
```
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 [10]:
```
%%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: 953 ms
Wall time: 50.1 s
```

In [11]:
```
df
```

Out[11]:

	category	title
0	정치	아프리카·중동 순방 마친 李…'톱다운 외교'로 AI·원전 협력 물꼬
1	정치	민주당, 오늘 대미투자특별법 발의…관세인하 11월 1일자 소급
2	정치	李대통령 "이재명 흉봐도 좋다"…튀르키예 동포 타운홀미팅
3	정치	"퇴직 대법관 대법사건 수임 제한"…민주, 사법행정 개혁안 발표
4	정치	정동영 "한반도 문제, 美승인 기다려선 해결 못해"…美대사대리 접견(종합)
...	...	...
299	IT/과학	"GPU는 26만장 샀는데 굴릴 사람은?"… 韓 AI의 '5년 승부수'
300	IT/과학	[KT 거버넌스 시험대]⑰ 김태호, 6년만의 재도전…'안전통' 자처
301	IT/과학	[누리호 4차 발사]③엔진·발사대 넘어 '우주 서비스'로…한화가 여는 상업 발사 시대
302	IT/과학	[단독]"180억 썼는데" 마동석도 못 살렸다…LGU+, 콘텐츠 자체제작 철수
303	IT/과학	누리호 주탑재 위성, 오로라 관측한다…우주 바이오 연구도 진행

304 rows × 2 columns

In [12]:
```
df['category'].value_counts()
```

Out[12]:
```
category
경제        52
생활/문화     52
IT/과학     52
정치        50
사회        50
세계        48
Name: count, dtype: int64
```