네이버 기사 크롤링

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

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

In [2]:
```
article_list = []

def get_article(page):
    driver = webdriver.Chrome("C:/Myexam/chromedriver/chromedriver.exe")
    driver.get("http://news.naver.com/main/main.nhn?mode=LSD&mid=shm&sid1=105#&date=%2000:00:00&page=" + str(page))
    articles = driver.find_elements_by_css_selector('#section_body li')

    for article in articles:
        title = article.find_element_by_css_selector('dt:not(.photo) > a').text
        article_list.append(title)

    print("end :", page)

    driver.quit()
```

In [3]:
```
%%time
for page in range(1, 5):
    get_article(page)
```

```
end : 1
end : 2
end : 3
end : 4
Wall time: 20.8 s
```

In [4]:
```
len(article_list), article_list[:30]
```

Out[4]:
```
(80,
 ['성장세 탄 LG전자 전장사업…하반기 인력 늘려 매출 8조 정조준',
  '‘상반기 평균급여’ 카카오>네이버… 배재현(카카오) 81억700만원 수령',
  "쿠팡·GS 참전에 배민 'B마트'도 단건배달...'15분 배송경쟁 퀵스타트'",
  "김태호 연출인데 MBC서 못본다…'새로운 실험' 통할까",
  '영화 속의 호빗 닮은꼴 생명체 등장… 공룡 다음 포유류 시대 시작된 흔적',
  '[팩플]"몸값 100조? 당근 가능" 3조 당근마켓 김용현 승부수',
  '아마존 이어 월마트도 가상화폐 전문가 채용… 일상생활에 자리잡나',
  'S펜 탑재하니 “진짜 폴더블폰이 완성된 느낌” ',
  '상추, 새싹삼, 허브… 지하철역·지하상가서 자란다',
  '게임 채팅창에 ‘ㄱㄱㄱ’ 치면 ‘GO’로 자동 번역… 해외 게이머와 언어장벽 없앤다',
  '韓 연구진, ‘양자역학’ 난제 해결 실마리 찾았다',
  "양자역학 난제 '상보성 원리', 국내 연구진이 실험으로 검증",
  '[줌인]만화책 8천권 모은 ₩'덕후 CEO₩'…"BTS도 웹툰으로 만나요"',
  '게임업계 직원 평균 급여 1위 카카오게임즈…2위 엔씨소프트, 3위 크래프톤',
  '구글, 50만원대 ‘픽셀5a’ 출시…아이폰13은 9월 출시',
  '"이런 신박한 방법이"…갤폴드3 사라진 카메라 만든 기술 공개됐다',
  '쏘카 납치 초등생 성폭행 피의자, 1심서 ‘징역 10년’ ',
  '“200만원 갤폴드3, 4년이나 써야 해?” 48개월 ‘족쇄’ ',
  "수천억 수익도 포기..삼성이 '앱 광고' 삭제 결정한 까닭은?",
  "韓 스타트업, 배터리 양극재 '마의 벽' 넘었다…전기차 가격 싸질까",
  '“1년에 100억도 넘는다” 웹툰 작가, 이렇게 많이 벌어?',
  '[IT애정남] 너무 오래된 노트북, 그래도 업그레이드?',
  '온라인, 오프라인 황제 꺾었다... 아마존, 2년 앞당겨 월마트 추월',
  'BTS 손잡은 네이버웹툰, 1등 플랫폼 굳히기',
  '삼성 중고폰 시세, 저가 샤오미보다 못하다니…',
  '설계 온도 상향, 또 제동…“안전성 우려” ',
  '워밍업 마친 중견SI, 클라우드·AI로 진격한다',
  '“삼성·LG·SK도 못 한 일”…전기차 배터리 용량 16% 늘린 스타트업',
  '모든 기부과정 블록체인에 입력… 순도 100% 기부 플랫폼 온다',
  '"네이버웹툰 슈퍼 IP, 메타버스로 확장할 것"'])
```

thread 사용

In [5]:
```
import threading
import pandas as pd

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

In [6]:
```
def get_article(page):
    driver = webdriver.Chrome("C:/Myexam/chromedriver/chromedriver.exe")
    driver.get(
        "http://news.naver.com/main/main.nhn?mode=LSD&mid=shm&sid1=105#&date=%2000:00:00&page=" + str(page))
    articles = driver.find_elements_by_css_selector('#section_body li')

    for article in articles:
        title = article.find_element_by_css_selector('dt:not(.photo) > a').text
        df.loc[len(df)] = {
            "title": title,
        }

    driver.quit()
```

In [7]:
```
for page in range(1, 5):
    th = threading.Thread(target=get_article, args=(page,))
    print("end :", page)
    th.start()
```

```
end : 1
end : 2
end : 3
end : 4
```

In [11]:
```
import time
sec = 0
while len(df) < 80:
    time.sleep(1)
    sec += 1
    print("{}sec({})".format(sec, len(df)), end=" ")
```

In [9]:
```
df.tail()
```

Out[9]:

	title
75	수천억 수익도 포기..삼성이 '앱 광고' 삭제 결정한 까닭은?
76	김기현 만난 창작자들 “구글 갑질방지법, 여야 이슈 사라졌다”
77	기술개발에 100조 쓰는데 60조 영업비밀 유출로 줄줄 샌다
78	시세 빠지고 글로벌사업자 떠나고...특금법에 韓 가상자산 '흔들'
79	‘입찰담합’ 통신사, 공정위 이어 국세청도 벌금 처분

In [ ]:
```

```