selenium

- https://www.selenium.dev
- 자동화를 목적으로 만들어진 다양한 브라우저와 언어를 지원하는 라이브러리
- 크롬 브라우저 설치
  - 크롬 브라우저 드라이버 다운로드 (크롬 브라우저와 같은 버전)
  - 다운로드한 드라이버 압축 해제
  - chromedriver, chromedriver.exe 생성
  - windows : 주피터 노트북 파일과 동일한 디렉토리에 chromedriver.exe 파일 업로드
  - mac : sudo cp ~/Download/chromedirver /usr/local/bin

In [1]:

```
import time
import pandas as pd
from selenium import webdriver
from selenium.webdriver.common.by import By
```

In [2]:

```
driver = webdriver.Chrome()
```

In [3]:

```
# 페이지 이동
driver.get("https://daum.net")
```

In [4]:

```
# 브라우저 사이즈 조절
driver.set_window_size(200, 600)
```

In [5]:

```
# 브라우저 스크롤 조절
driver.execute_script("window.scrollTo(200, 300);")
```

In [6]:

```
# alert 다루기
driver.execute_script("alert('hello selenium!!!');")
```

In [7]:

```
alert = driver.switch_to.alert
alert.accept()
```

In [8]:

```
!pip list | grep selenium
```

```
selenium                      4.11.2
```

In [9]:

```
# 문자열 입력
driver.find_element(By.CSS_SELECTOR, "#q").send_keys("셀레니움")
```

In [10]:

```
# 검색 버튼 클릭
driver.find_element(By.CSS_SELECTOR, '.inner_search > .ico_pctop.btn_search').click(
```

In [11]:

```
# 브라우저 종료
driver.quit()
```

텍스트 데이터 가져오기

- TED 사이트 : https://www.ted.com

In [12]:

```
# 브라우저를 실행하여 테드 사이트 열기
driver = webdriver.Chrome()
driver.get("https://www.ted.com/talks")
```

In [13]:

```
# CSS Selector를 이용하여 HTML 태그와 태그 사이의 text 데이터 가져오기
driver.find_element(By.CSS_SELECTOR, ".talks-header__title").text
```

Out[13]:

```
'4300+ talks to stir your curiosity'
```

In [14]:

```
# 제목 데이터 가져오기
contents = driver.find_elements(By.CSS_SELECTOR, "#browse-results > .row > .col")
len(contents)
```

Out[14]:

```
36
```

In [15]:

```
# 가장 처음 텍스트 데이터 가져오기
contents[0].find_element(By.CSS_SELECTOR, '.media__message .ga-link').text
```

Out[15]:

```
"What's it like to be a giant sequoia tree?"
```

In [16]:

```
# 전체 제목 데이터 가져오기
titles = []
for content in contents:
    title = content.find_element(By.CSS_SELECTOR, '.media__message .ga-link').text
    titles.append(title)
titles[:3], len(titles)
```

Out[16]:

```
(["What's it like to be a giant sequoia tree?",
  'Which is better for you: "Real" meat or "fake" meat?',
  'The molecular love story that could help power the world'],
 36)
```

In [17]:

```
# 셀렉트 박스를 선택후 데이터 가져오기
# 이벤트 발생 기능(값 입력, 클릭 이벤트등)은 화면에 해당 엘리먼트가 보여야 합니다.
# 한국어 선택
driver.find_element(By.CSS_SELECTOR, '#languages [lang="ko"]').click()
time.sleep(1)
```

In [18]:

```
# 전체 제목 데이터 가져오기
contents = driver.find_elements(By.CSS_SELECTOR, "#browse-results > .row > .col")
titles = []
for content in contents:
    title = content.find_element(By.CSS_SELECTOR, '.media__message .ga-link').text
    titles.append(title)
titles[-3:]
```

Out[18]:

```
['혼잣말은 정상적인 행동일까요?', '무엇이 "좋은 대학"을 만들고, 왜 그것이 중요한가?', '정
신 분열로 얻은 교훈']
```

In [19]:

```
# 링크 데이터 크롤링 (속성(attribute)값 가져오는 방법)
links = []
for content in contents:
    link = content.find_element(By.CSS_SELECTOR, '.media__message .ga-link').get_att
    links.append(link)
links[-3:]
```

Out[19]:

```
['https://www.ted.com/talks/ted_ed_is_it_normal_to_talk_to_yourself?la
nguage=ko',
 'https://www.ted.com/talks/cecilia_m_orphan_what_makes_a_good_college
_and_why_it_matters?language=ko',
 'https://www.ted.com/talks/andy_dunn_lessons_from_losing_my_mind?lang
uage=ko']
```

In [20]:

```
driver.quit()
```

3. Headless

- 브라우져를 화면에 띄우지 않고 메모리상에서만 올려서 크롤링하는 방법
- window가 지원되지 않는 환경에서 사용이 가능
- chrome version 60.0.0.0 이상부터 지원 합니다.

In [21]:

```
# 현재 사용중인 크롬 버전 확인
driver = webdriver.Chrome()
version = driver.capabilities["browserVersion"]
print(version)
driver.quit()
```

```
116.0.5845.96
```

In [22]:

```
# headless 사용
options = webdriver.ChromeOptions()
options.add_argument('headless')
driver = webdriver.Chrome(options=options)
driver.get("https://www.ted.com/talks")
text = driver.find_element(By.CSS_SELECTOR, ".talks-header__title").text
driver.quit()
print(text)
```

```
4300+ talks to stir your curiosity
```