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]:
```
# 문자열 입력
driver.find_element(By.CSS_SELECTOR, '#q').send_keys('셀레니움')
```

In [9]:
```
# 검색 버튼 클릭
driver.find_element(By.CSS_SELECTOR, '.btn_ksearch').click()
```

In [10]:
```
# 브라우저 종료
driver.quit()
```

텍스트 데이터 가져오기

- TED 사이트 : `https://www.ted.com`

In [11]:
```
# 브라우저를 실행하여 테드 사이트 열기
driver = webdriver.Chrome()
driver.get('https://www.ted.com/talks')
```

In [13]:
```
# 팝업 레이아웃 제거 : x 버튼 클릭
time.sleep(3)
driver.find_element(By.CSS_SELECTOR, '#close-pc-btn-handler').click()
```

In [14]:
```
# CSS Selector를 이용하여 HTML 태그와 태그 사이의 text 데이터 가져오기
driver.find_element(By.CSS_SELECTOR, 'h2.text-textPrimary-onLight').text
```

Out[14]:
```
'TED Talks: Discover ideas worth spreading'
```

In [15]:
```
# 제목 데이터 가져오기
selector = '[test-id='Talk Grid Default'] > div > div:nth-child(2) \
> div > div'
contents = driver.find_elements(By.CSS_SELECTOR, selector)
len(contents)
```

Out[15]:
```
24
```

In [16]:
```
# 가장 처음 텍스트 데이터 가져오기
selector = 'span.text-textPrimary-onLight'
contents[0].find_element(By.CSS_SELECTOR, selector).text
```

Out[16]:
```
'To end extreme poverty, give cash — not advice'
```

In [17]:
```
# 전체 제목 데이터 가져오기
titles = []
selector = 'span.text-textPrimary-onLight'
for content in contents:
    title = content.find_element(By.CSS_SELECTOR, selector).text
    titles.append(title)
titles[:3], len(titles)
```

Out[17]:
```
(['To end extreme poverty, give cash — not advice',
  'The arrest of Telegram CEO Pavel Durov — and why you should care',
  'Can math help repair democracy?'],
 24)
```

In [18]:
```
contents[0].find_element(By.CSS_SELECTOR, 'a').get_attribute('href')
```

Out[18]:
```
'https://www.ted.com/talks/rory_stewart_to_end_extreme_poverty_give_cash_not_advice'
```

In [19]:
```
# 링크 데이터 크롤링 (속성(attribute)값 가져오는 방법)
links = []
for content in contents:
    link = content.find_element(By.CSS_SELECTOR, 'a').get_attribute('href')
    links.append(link)
links[:3], len(links)
```

Out[19]:
```
(['https://www.ted.com/talks/rory_stewart_to_end_extreme_poverty_give_cash_not_advice',
  'https://www.ted.com/talks/eli_pariser_the_arrest_of_telegram_ceo_pavel_durov_and_why_you_should_care',
  'https://www.ted.com/talks/sam_wang_can_math_help_repair_democracy'],
 24)
```

In [20]:
```
df = pd.DataFrame({'title': titles, 'link': links})
df.head(3)
```

Out[20]:

	title	link
0	To end extreme poverty, give cash — not advice	https://www.ted.com/talks/rory_stewart_to_end_...
1	The arrest of Telegram CEO Pavel Durov — and w...	https://www.ted.com/talks/eli_pariser_the_arre...
2	Can math help repair democracy?	https://www.ted.com/talks/sam_wang_can_math_he...

In [21]:
```
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 [22]:
```
# 현재 사용중인 크롬 버전 확인
driver = webdriver.Chrome()
version = driver.capabilities['browserVersion']
print(version)
driver.quit()
```

```
128.0.6613.138
```

In [23]:
```
# headless 사용
options = webdriver.ChromeOptions()
options.add_argument('headless')
driver = webdriver.Chrome(options=options)
driver.get('https://www.ted.com/talks')
text = driver.title
selector = 'h2.text-textPrimary-onLight'
sub_title = driver.find_element(By.CSS_SELECTOR, selector).text
driver.quit()
print(text, sub_title, sep='\n')
```

```
TED: Ideas change everything
TED Talks: Discover ideas worth spreading
```