K-Digital Training 신재생에너지 활용 IoT 과정

강의 내용

- 웹 자동화
- Selenium
- ChromeOptions
- 동적 크롤링 실습

웹 자동화

웹 자동화란?

- 사람이 일반적으로 웹 브라우저를 사용하여 수행하는 작업을 자동으로 수행할 수 있도록 도와주는 기술
- 웹 자동화를 통해 시간을 절약하고 반복적이고 지루한 작업을 자동화하여 생산성을 향상시킬 수 있다.
- 예) 웹 개발, 데이터 수집, 테스트 자동화

동적 크롤링 & 정적 크롤링

- 정적 크롤링
  - 웹 페이지의 소스 코드에 표시된 정적 데이터(정해진 데이터)만을 수집
  - html 소스에서 원하는 데이터를 추출하는 것을 말함
  - javascript로 생성되는 데이터는 수집 불가

- 동적 크롤링
  - javascript와 같은 동적 콘텐츠를 실행하고 해당 데이터를 가져오는 방식
  - 입력, 클릭, 로그인, 페이지 생성 등을 통해 데이터가 바뀌는 것을 동적 데이터라 함
  - 정적 크롤링보다 느림

selenium

selenium

- 웹 애플리케이션의 테스트를 자동화하거나, 웹 사이트에서 데이터를 스크랩하는 데 사용되는 파이썬 라이브러리
- 실제 사용자가 웹 브라우저를 조작하는 것과 유사한 방식으로 웹 페이지의 요소 클릭, 텍스트 입력, 페이지 간 이동을 수행하며, 결과를 확인할 수 있음
- 동적으로 HTML이 생성되는 경우에는 requests를 사용할 수 없다.
  - 스크롤 했을 때 데이터가 생성되는 경우
  - URL 주소는 동일한데 데이터가 변하는 경우

selenium 사용

- selenium 설치

```
!pip install selenium
!pip install webdriver_manager
```

- 실행

```
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from webdriver_manager.chrome import ChromeDriverManager

# 최신 ChromeDriver 설치 및 사용
service = Service(ChromeDriverManager().install())
driver = webdriver.Chrome(service=service)

# 웹사이트 열기
driver.get("https://naver.com")
```

selenium 사용 (수동 설치 방법 1/3)

- selenium 설치
  - !pip install seleium
- chromedriver 설치
  - https://googlechromelabs.github.io/chrome-for-testing/#stable
    - mac book m1, m2, m3 는 mac-arm64 다운로드
    - windows 는 win64 다운로드
  - 다운로드 받아 압축 해제
    - selenium 실행시 파일 경로 기입해야 하므로 위치 기억

selenium 사용 (수동 설치 방법 2/3)

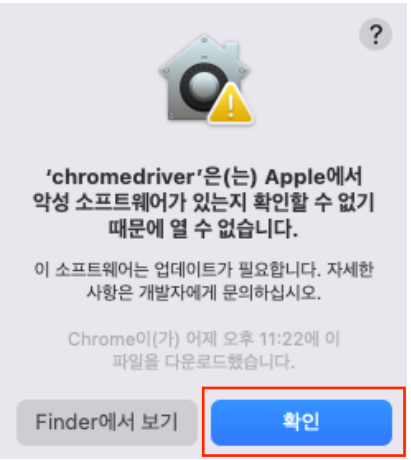
- 실행
  - driver_path 에 다운로드 받아 압축 푼 파일 경로 지정

```
from selenium import webdriver
from selenium.webdriver.chrome.service import Service

# 드라이버 경로
driver_path = '/Users/leeseunghak/Downloads/chromedriver-mac-arm64/chromedriver'
service = Service(executable_path=driver_path)

# 크롬 드라이버 실행
driver = webdriver.Chrome(service=service)
driver.get('https://naver.com')
```

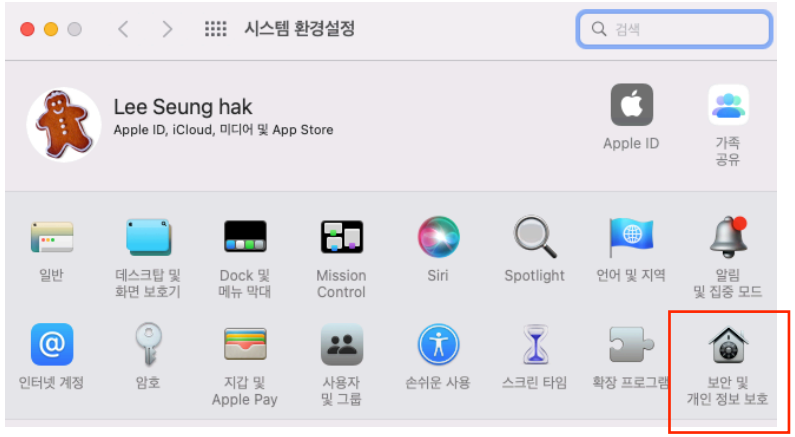
selenium 사용 (수동 설치 방법 3/3)

- Mac book 보안 이슈 발생 시

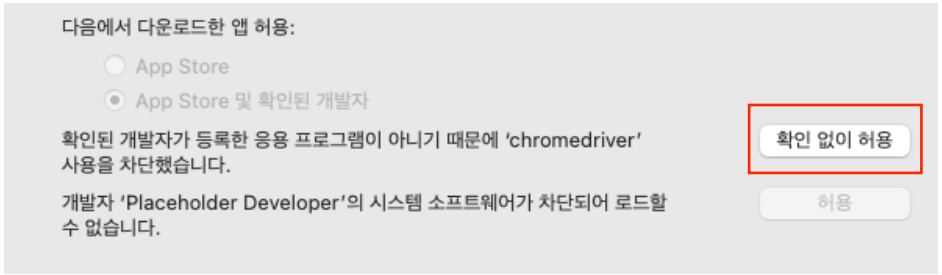
(1)



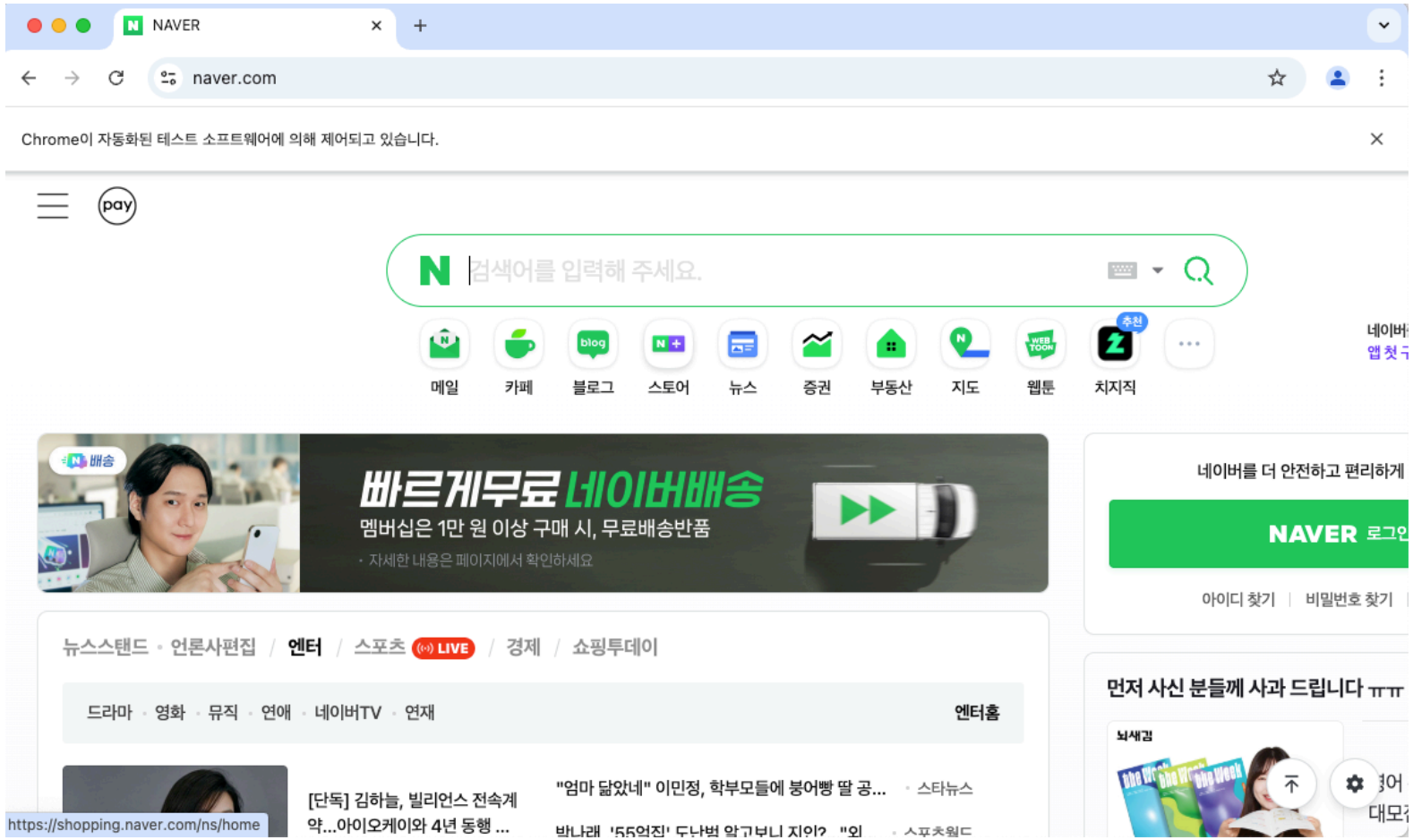
(2)



(3)



selenium 실행


NAVER
naver.com
Chrome이 자동화된 테스트 소프트웨어에 의해 제어되고 있습니다.
pay
검색어를 입력해 주세요.
메일
카페
블로그
스토어
뉴스
증권
부동산
지도
웹툰
치지직
빠르게무료 네이버배송
멤버십은 1만 원 이상 구매 시, 무료배송반품
자세한 내용은 페이지에서 확인하세요
네이버를 더 안전하고 편리하게
NAVER
아이디 찾기
비밀번호 찾기
뉴스스탠드
언론사편집
엔터
스포츠 LIVE
경제
쇼핑투데이
드라마
영화
뮤직
연애
네이버TV
연재
엔터홈
[단독] 김하늘, 빌리언스 전속계약...아이오케이와 4년 동행 ...
"엄마 닮았네" 이민정, 학부모들에 붕어빵 딸 공...
스타뉴스
먼저 사신 분들께 사과 드립니다
https://shopping.naver.com/ns/home

selenium 메서드

- 브라우저 제어 메서드

get(url)	지정된 URL로 브라우저를 이동.
close()	현재 탭을 닫음.
quit()	전체 브라우저를 종료.
refresh()	현재 페이지를 새로 고침.
maximize_window()	브라우저 창을 최대화.
minimize_window()	브라우저 창을 최소화.
back()	브라우저에서 이전 페이지로 이동.
forward()	브라우저에서 다음 페이지로 이동.

실습 1

- 네이버 접속
- 브라우저 창 최대화
- 1초 후에 구글 접속
- 1초 후에 back 하여 네이버로 이동
- 1초 후에 새로고침
- 1초 후에 브라우저 종료

```python
import time

print('wait for 1 second')
# 1초 대기
time.sleep(1)
print('done')
```

참고: 1초 대기 예시 코드

selenium 메서드

- 웹 요소 찾기

find_element(by, value)	지정된 위치에서 첫 번째 요소를 찾음
find_elements(by, value)	지정된 위치에서 모든 요소를 리스트로 반환

- By.ID: ID로 요소 찾기
- By.NAME: name 속성으로 찾기
- By.CLASS_NAME: 클래스 이름으로 찾기
- By.TAG_NAME: 태그 이름으로 찾기
- By.CSS_SELECTOR: CSS 선택자로 찾기
- By.XPATH: XPath로 찾기

```
from selenium.webdriver.common.by import By
```

By.xxxx 를 사용하기 위해서는 By를 import

실습 2

- 네이버의 서비스 바로가기 이름 수집

- 크롬 인스펙터로 요소 패턴 찾기
  - 서비스 바로가기 모음은 ul.shortcut_list
  - 각 서비스는 span.service_name

실습 2

- Option 1
  - CSS 선택자 - ul.shortcut_list span.service_name
  - find_elements() 로 해당하는 모든 원소를 찾아 text 출력

```
from selenium.webdriver.common.by import By
driver = webdriver.Chrome(service=service)
```

```
driver.get('https://naver.com')

time.sleep(1)
elements = driver.find_elements(By.CSS_SELECTOR, 'ul.shortcut_list span.service_name')

for element in elements:
    print(element.text)
```

1초 대기 유/무 결과 비교

```
메일
카페
블로그
스토어
뉴스
증권
부동산
지도
웹툰
치지직
```

실습 2

- Option 2
  - (1) 바로가기 모음 찾기
    - CSS 선택자 - ul.shortcut_list
  - (2) 바로가기 모음에서 서비스들 찾기
    - CSS 선택자 - .service_name
    - (1)에서 반환된 결과 객체도 find_elements() 메소드를 제공함

CSS 선택자가 복잡해지는 경우,
Option2 처럼 선택 패턴을
나누어서 찾는 것을 추천

```
driver.get('https://naver.com')
time.sleep(1)

shortcuts = driver.find_element(By.CSS_SELECTOR, 'ul.shortcut_list')
services = shortcuts.find_elements(By.CSS_SELECTOR, '.service_name')
for service in services:
    print(service.text)
```

```
메일
카페
블로그
스토어
뉴스
증권
부동산
지도
웹툰
치지직
```

selenium 메서드

- 웹 요소 상호작용

click()	요소를 클릭
send_keys(keys)	입력 상자에 텍스트를 입력
clear()	입력 상자의 기존 텍스트를 삭제
is_displayed()	요소가 페이지에 표시되는지 여부 확인
is_enabled()	요소가 활성화되어 있는지 여부 확인
is_selected()	체크박스 또는 라디오 버튼이 선택되었는지 확인
get_attribute(attribute)	요소의 특정 속성 값을 반환
get_property(property)	요소의 특정 속성(property)을 반환
text	요소 내부의 텍스트를 반환
submit()	폼(form)을 제출

실습 3

- 네이버의 서비스 바로가기에서 뉴스를 클릭하여 이동
  - 실습 2의 option2 에서 찾은 services에서 text가 '뉴스' 인 것을 click()

```
for service in services:
    if service.text == '뉴스':
        service.click()
```

selenium 메서드

```
from selenium.webdriver.common.keys import Keys
```

- send_keys(키값)

Keys.ENTER	엔터 키	send_keys(Keys.ENTER)
Keys.RETURN	리턴 키	send_keys(Keys.RETURN)
Keys.TAB	탭 키	send_keys(Keys.TAB)
Keys.ESCAPE	ESC 키	send_keys(Keys.ESCAPE)
Keys.BACKSPACE	백스페이스 키	send_keys(Keys.BACKSPACE)
Keys.DELETE	삭제 키	send_keys(Keys.DELETE)
Keys.SHIFT	쉬프트 키	send_keys(Keys.SHIFT + "A")
Keys.CONTROL	컨트롤 키	send_keys(Keys.CONTROL + "a")
Keys.ARROW_UP	위 방향 화살표 키	send_keys(Keys.ARROW_UP)
Keys.ARROW_DOWN	아래 방향 화살표 키	send_keys(Keys.ARROW_DOWN)
Keys.F1 ~ Keys.F12	F1 ~ F12 키	send_keys(Keys.F5)
Keys.SPACE	스페이스 키	send_keys(Keys.SPACE)

실습 4

- 네이버 검색 창에 검색어 "날씨"를 입력하여 검색
  - 힌트
    - 검색어를 입력할 수 있는 요소는 <input>
    - 검색 행위는 send_keys()로 '날씨'를 입력 후, 엔터 입력

selenium 메서드

- 브라우저 상태 및 정보

title	현재 페이지의 제목(title)을 반환.
current_url	현재 URL을 반환.
page_source	현재 페이지의 HTML 소스를 반환.
window_handles	열려 있는 모든 창(탭)의 핸들 ID를 리스트로 반환.
current_window_handle	현재 활성 탭의 핸들 ID를 반환.

selenium 메서드

- 스크롤 및 뷰 제어

execute_script(script)	자바스크립트를 실행
set_window_size(width, height)	브라우저 창 크기를 설정

- 프레임, 팝업 창, 또는 알림 창과 상호작용

switch_to.frame(frame)	지정된 프레임으로 포커스 이동
switch_to.default_content()	프레임에서 기본 콘텐츠로 포커스 이동
switch_to.alert	알림(alert) 창으로 포커스 이동

selenium 메서드

- 요소 로드나 특정 조건 충족까지 대기

implicitly_wait(time) sleep()보다 유용	모든 요소를 찾기 전에 지정된 시간만큼 암시적으로 대기
WebDriverWait(driver, timeout)	특정 조건이 충족될 때까지 명시적으로 대기

- 스크린 샷

save_screenshot(file_path)	현재 브라우저 화면을 이미지 파일로 저장
get_screenshot_as_file(file_path)	특정 요소의 스크린샷을 이미지로 저장

실습 5

- driver.implicitly_wait(time) 를 호출 시,
  - find_element(), find_elements() 가 요소를 찾을때 까지 최대 time 만큼 대기
- 네이버의 서비스 바로가기 이름 수집을
  - time.sleep(1) 대신, driver.implicitly_wait(1)을 사용하여 작성
- 참고: driver.implicitly_wait()는 한번만 호출하면 이후 계속 적용됨

실습 6

- driver.implicitly_wait(10) 으로 인자를 10으로 변경하여 실습 5 실행
- 대기 시간이 10초가 아닌 짧은 이유는 무엇인가?
- time.sleep()을 사용하는 것 대비 장점이 무엇인가?

ChromeOptions

- add_argument 메서드를 사용하여 브라우저의 실행 동작을 설정

```python
from selenium import webdriver
from selenium.webdriver.chrome.options import Options

# ChromeOptions 객체 생성
chrome_options = Options()

# 옵션 추가
chrome_options.add_argument("--start-maximized")
chrome_options.add_argument("--disable-notifications")
chrome_options.add_argument("--incognito")
chrome_options.add_argument("--headless")

# 드라이버에 옵션 전달
driver = webdriver.Chrome(options=chrome_options)
driver.get("https://www.google.com")
# driver.quit()
```

ChromeOptions

- 자주 쓰는 옵션

--start-maximized	브라우저를 최대화된 상태로 실행
--window-size=width,height	브라우저 크기를 지정 (예: --window-size=1920,1080)
--headless	브라우저 GUI 없이 백그라운드에서 실행 (리소스 절약)
--disable-notifications	알림 창 비활성화
--incognito	시크릿 모드 실행
--disable-gpu	GPU 렌더링 비활성화 (Linux에서 Headless 모드에서 사용 권장)
--no-sandbox	샌드박스 모드 비활성화 (권한 문제 해결에 유용)
--disable-extensions	확장 프로그램 비활성화

실습7. 네이버 지도 검색하기

- 네이버 지도에서 특정 식당 검색하여 리뷰 점수 가져오기
  - ex) 양반집 은평구, 은평구 알소곱창, ...

실습8. github 크롤링

- 로그인 후 사용자 대시보드에서 사용자 이름이나 프로필 관련 정보를 크롤링하세요.

실습9. 쇼핑몰 크롤링하기

- 쇼핑몰에서 특정 검색어(예: "노트북")를 검색하세요.
- 검색 결과에서 가격이 50만 원 이상인 상품의 이름과 가격을 추출하세요.

실습10. 여행사이트 크롤링하기

- 여행 예약 사이트(예: Agoda 또는 Booking.com)에 접속하세요.
- 특정 도시를 검색하고, 캘린더에서 출발 날짜와 도착 날짜를 선택한 후, 첫 번째 검색 결과의 호텔 이름과 가격을 출력하세요.

실습11. 이미지 크롤링하기

- Google 이미지 검색에서 원하는 동물을 검색하세요.
- 무한 스크롤을 통해 10개 이상의 이미지를 로드하고, 이미지를 다운로드하여 로컬 디렉토리에 저장하세요.

codingOn

복.습.철.저

수고하셨습니다