분석용 데이터 획득 전략

웹 SNS 데이터 확보 기법

학습 목표

학습 목표

- 웹 페이지를 구성하는 방식 중에서 Iframe라는 방식의 의미를 이해하고 Iframe을 사용한 웹페이지에서 Iframe의 정보를 찾는 방법을 학습하며 소스코드에서 해당 Iframe으로 전환하는 방법과 해당 Iframe 부분에 있는 정보를 추출할 수 있다.
- 동일한 URL이지만 내부 구조가 다른 SNS의 경우에 동일 URL로 접속 한 후 태그 이름과 클래스 이름을 활용하여 다양한 세부 유형을 추출하여 해당 블로그의 유형을 파악할 수 있다.
- 추출된 태그명과 클래스명을 활용하여 적절한 타입의 블로그에 접속할 수 있으며 접속 후 각 유형별로 적절한 블로그 정보를 구분하여 상세 항목들을 추출할 수 있다.

학습 내용

- 웹 SNS 데이터 수집하기 - 이론
- 웹 SNS 데이터 수집하기 - 실습

본 강의자료를 무단배포 및 복제하는 것을 금지하며
이를 위반하는 경우 저작권 침해로서 관련법에 따라 처벌될 수 있습니다.

네이버 블로그 수집

1) IFRAME 이해하기

IFRAME이란,

하이퍼텍스트 생성 언어(HTML) 문서에서 글 중의 임의의 위치에 또 다른 HTML 문서를 보여주는 내부 프레임(Inline Frame) 태그
즉, 웹 문서 중간에 다른 웹 문서나 텍스트 파일과 같은 내용을
원하는 크기로 불러들여 보여주는 태그

주의사항

웹 데이터 추출 오류시 IFRAM으로 구성된 데이터인지 확인

(1) IFRAME 찾는 방법

- 데이터가 저장되어 있는 태그를 개발자 도구를 활용하여 찾은 후 위로 올라가면서 IFRAME 태그 탐색
- 주로 IFRAME 속성 중 id나 name 값을 지정하고 있기 때문에 이 값을 찾으면 됨
  - id나 name 값이 없을 경우 XPATH 값 사용
  - 한 페이지에 여러 개 사용 가능

본 강의자료를 무단배포 및 복제하는 것을 금지하며
이를 위반하는 경우 저작권 침해로서 관련법에 따라 처벌될 수 있습니다.

네이버 블로그 수집

1) IFRAME 이해하기

(2) IFRAME 전환하기

1. Selenium이 현재 페이지의 전체 소스코드를 가져오기 전에 먼저 iframe id를 지정해야 함
2. driver.switch_to.frame('iframe id 값') 형태로 전환 후 전체 페이지 소스코드를 가져와야 함

2) 블로그 종류 확인

(1) 블로그 종류 확인 방법

유형1

se-main-container 클래스명 사용

https://blog.naver.com/hy820715/221514204265

유형2

se_component_wrap 클래스명 사용

유형3

postViewArea 클래스명 사용

주의사항

데이터 수집시 각 블로그마다 태그나 명령어가 다름에 주의

본 강의자료를 무단배포 및 복제하는 것을 금지하며
이를 위반하는 경우 저작권 침해로서 관련법에 따라 처벌될 수 있습니다.

블로그 상세 정보 수집하기

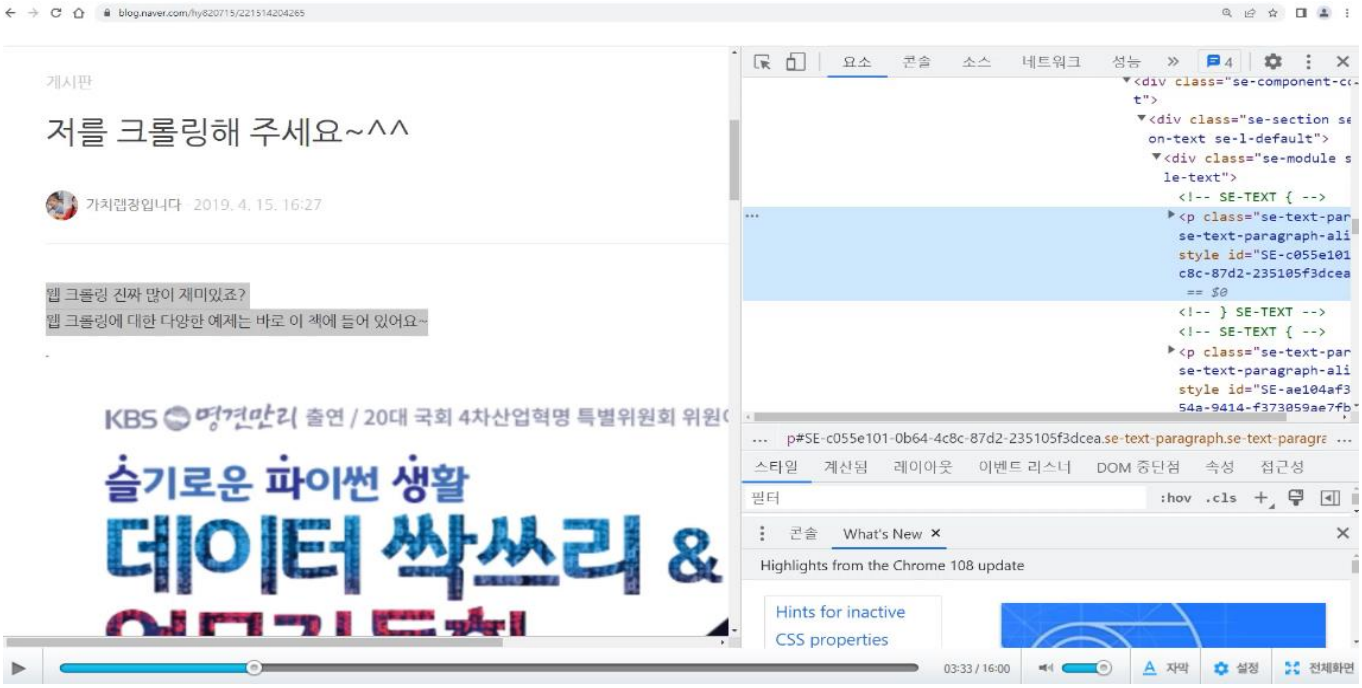
- 블로그 제목, 블로그 주인 이름, 블로그 작성 일자, 블로그 본문 내용

- IFRAM 정보 → 셀레늄

본 강의자료를 무단배포 및 복제하는 것을 금지하며
이를 위반하는 경우 저작권 침해로서 관련법에 따라 처벌될 수 있습니다.

블로그 상세 정보 수집하기

- 개발자 도구 F12 , 위로 올라가면서 IFRAME 태그 탐색



본 강의자료를 무단배포 및 복제하는 것을 금지하며
이를 위반하는 경우 저작권 침해로서 관련법에 따라 처벌될 수 있습니다.

블로그 상세 정보 수집하기

- IFRAME으로 전환

```
#Step 4. 각 블로그의 상세 결과들 출력하여 파일에 저장하기
blog_addr2=[]
w_name2=[]
w_date2=[]
blog_txt2=[]

gubun = blog_addr.split("/")

if gubun[2] != 'blog.naver.com' :
    print(" 네이버 블로그만 가능합니다")

else :
    driver.get(blog_addr)
    driver.maximize_window( )
    time.sleep(random.randrange(2,5))  # 2 - 5 초 사이에 랜덤으로 시간 선택

    print("\n")
    print("데이터 추출 시작합니다 ==========================================")
    print("\n")

    # iframe 전환하기
    driver.switch_to.frame('mainFrame')

    #전체 소스코드 가져오기
    html = driver.page_source
    soup = BeautifulSoup(html, 'html.parser')

    addr_1 = soup.select('div[class="se_component_wrap sect_dsc __se_component_area"]')
    addr_2 = soup.select('div[class="se-main-container"]')

    if addr_1 :
        img_no = 1
```

주의사항

- ❖ 네이버 블로그의 경우, IFRAME으로 전환해주지 않으면 본문 내용 추출 불가
- ❖ 태그는 블로그 종류에 따라 다름

본 강의자료를 무단배포 및 복제하는 것을 금지하며
이를 위반하는 경우 저작권 침해로서 관련법에 따라 처벌될 수 있습니다.

블로그 상세 정보 수집하기

- 예외 처리

```
    # iframe 전환하기
    driver.switch_to.frame('mainFrame')

    #전체 소스코드 가져오기
    html = driver.page_source
    soup = BeautifulSoup(html, 'html.parser')

    addr_1 = soup.select('div[class="se_component_wrap sect_dsc __se_component_area"]')
    addr_2 = soup.select('div[class="se-main-container"]')

    if addr_1 :
        img_no =  1
        print()
        print("요청하신 블로그는 유형 1형이고 해당 블로그 정보를 수집합니다~~~~~~~~~~~~")

        # 블로그 주소
        print("1.블로그주소: ",blog_addr)
        blog_addr2.append(blog_addr)

        # 작성자 닉네임
        writer = soup.select("div.blog2_container > span.writer")
        try :
            wname = writer[0].get_text( )   # 작성자 닉네임
        except IndexError :
            wname = "작성자 닉네임이 없습니다"
        else :
            wname = wname.replace("\n","")

        int("2.작성자 닉네임: ",wname )
        ame2.append(wname)
```

주의사항

데이터 추출시 태그 이름이 다를 수 있으니, 블로그 유형 반드시 확인

본 강의자료를 무단배포 및 복제하는 것을 금지하며
이를 위반하는 경우 저작권 침해로서 관련법에 따라 처벌될 수 있습니다.