# 컴퓨터공학개론 기말프로젝트

## { 키워드 맞춤 크롤링 서비스 }

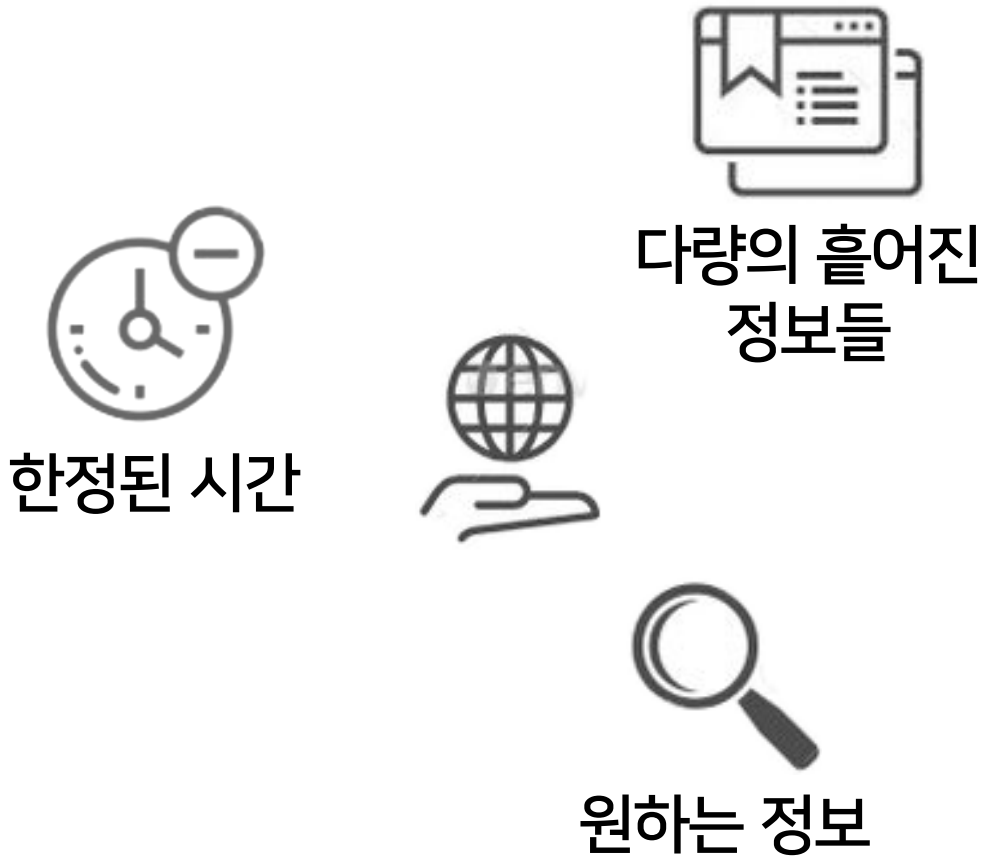
12조

B711222 박조은　　C011153 이신영

# 목차

# 주제 소개

{ 원하는 키워드를 포함한 새 게시글 알림 서비스 }



# 주제 소개

키워드 및
메일 주소
설정

크롤링

게시물
발견

메일 발신

메일 수신

# 소스 코드


from bs4 import BeautifulSoup
HTML 및 XML 문서의 구문 분석 위한 패키지


from selenium import webdriver
웹 애플리케이션 테스트를 위한 포터블 프레임워크

import smtplib
from email.mime.text import MIMEText


import openpyxl from Workbook
엑셀 파일을 다루는 라이브러리

# 소스 코드

```
from selenium import webdriver
import smtplib
from email.mime.text import MIMEText
import time
from bs4 import BeautifulSoup
from openpyxl import Workbook

keyword = input("알림을 받길 원하는 키워드들을 띄어쓰기로 구분하여 입력해 주세요!(예시:코로나 19 등록금 컴공)\n").split()
address = input("모든 키워드가 포함된 글이 올라오면 알림을 받을 이메일 주소를 입력해 주세요!(예시:hongik@hongik.com)\n ")
your_keyword = ', '.join(keyword)

new_wb = Workbook()
new_ws = new_wb.active

driver = webdriver.Chrome('C:\\Users\\dntos\\Desktop\\chromedriver_file\\chromedriver.exe')
url = "https://everytime.kr/login"
driver.get(url)

driver.find_element_by_name('userid').send_keys('아이디 자리입니다!!!!')
driver.find_element_by_name('password').send_keys('비밀번호 자리입니다!!!')
driver.find_element_by_xpath("""//*[@id="container"]/form/p[3]/input""").click()

latest_num = 0

while (True):
    try:
        text_keyword = list()
        com_url = "https://everytime.kr/382283"
        driver.get(com_url)

        res = driver.page_source
        soup = BeautifulSoup(res, "html.parser")

        # 컴공게시판 최상단 게시물 선택
        post = soup.find("article")

        # 게시물 내의 a 태그에 속하는 href 를 find 키워드로 추출
        url_num = post.find("a").attrs['href']
        post_num = url_num.split('/')[3]

        # 게시물 내용 find 로 추출
        data_text = post.find("p", {"class": "medium"}).get_text('\n', strip=True)
        data_time = post.find("time", {"class": "medium"}).get_text()

        # 게시물 링크 저장
        link = "https://everytime.kr" + url_num

        # 추출한 게시물의 전체 내용을 엑셀 시트에 추가
        new_ws.append([data_text, data_time, link])

        # 저장경로 및 파일명 작성
        new_wb.save('C:\\Users\\dntos\\Desktop\\project\\res.xlsx')

    except:
        continue

    # 메일 전송시 게시물 중복 여부 확인 위한 조건
    if latest_num != post_num:
        latest_num = post_num
        for i in range(0, len(keyword)):
            included_keyword = data_text.find(keyword[i])
            text_keyword.append(included_keyword)
        if -1 not in text_keyword:
            m1 = '홍익대학교 컴퓨터공학과 게시판 게시글 업데이트' + '\n' + '키워드 알림을 설정해 주신 키워드인 ' +
대한 게시글이 업데이트되었습니다!\n'
            m2 = '게시글의 내용 : ' + '\n' + data_text + '\n'
            m3 = '작성 시간 : ' + data_time + '\n'
            m4 = '게시글 주소 : ' + link + '\n'
            sending = smtplib.SMTP('smtp.gmail.com', 587)
            sending.starttls()
            sending.login('dntoskwem@gmail.com', 'ppsjorxpbmbbifwq')

            message = MIMEText(f'{m1}{m2}{m3}{m4}')
            message['subject'] = f'{your_keyword}에 대한 키워드 알림!'

            sending.sendmail("dntoskwem@gmail.com", address, message.as_string())
        else:
            print('해당 키워드들이 내용에 모두 들어있지 않습니다 ㅠㅠ')

    print('크롤링 중입니다 현재 게시글은!', latest_num)
    print(link)
    print(data_text)
    time.sleep(60)
```

# 시연 영상

# 감사합니다