주제분석 1주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 주제분석 1주차 패키지 문제의 조건 및 힌트는 Python을 기준으로 하지만, R을 사용해도 무방합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 5시 00분에 발표됩니다.
- 제출기한은 **4/27 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요.

Chapter 1 정적 웹페이지 크롤링

크롤링의 종류에는 Open API를 호출해서 필요한 데이터를 추출하는 방법도 있지만, API가 제공되지 않는 경우, 직접 웹페이지 내의 정보들을 추출해야 합니다. 웹페이지 유형에 따라 정적/동적 크롤링으로 나뉘며, 대표적으로 BeautifulSoup, Selenium 패키지를 사용합니다. 패키지 사용법과 기본적인 html 태그, 속성을 알면 원하는 웹페이지에 맞추어 필요한 데이터를 추출할 수 있어 매우 용이합니다. 하지만 수집한 데이터를 상업적으로 이용하는 경우, 제약이 많기 때문에 주의하세요.

첨부된 결과파일과 최대한 비슷하게 결과물을 만들어 주시면 됩니다. 크롤링 시점에 따라 데이터가 변하기 때문에 데이터프레임 구조 외의 속성 정보는 달라도 무방합니다. 제출 시에는 만든 소스코드(HTML, PDF형식)를 제출해주세요.

문제 1. 네이버 뉴스 제목을 크롤링해봅시다.

문제 1-1.

https://search.naver.com/search.naver?where=news&sm=tab_jum&query=%EC%84%B1%EA%B7%A0%EA%B4%80%EB%8C%80

위 링크는 “성균관대” 를 키워드로 한 네이버 뉴스 카테고리 url 입니다. 이 url을 호출하여 html이 어떻게 구성되어 있는지 확인 후 ‘html’ 변수로 저장해주세요.

HINT1. GET 방식으로 호출하기 위해서는 requests 라이브러리의 get 함수를 사용합니다.

HINT2. 다음은 html 예시입니다.

```
'<!doctype html> <html lang="ko"> <head> <meta charset="utf-8"> <meta name="referrer" content="always">  <meta name="format-detection" content="telephone=no,address=no,email=no"> <meta name="viewport" content="width=device-width,initial-scale=1.0,maximum-scale=2.0"> <meta property="og:title" content="성균관대 : 네이버 뉴스검색"/> ...
```

문제 1-2. Beautifulsoup 라이브러리를 이용하여 불러온 html 코드를 beatifulsoup 객체로 구조화시키세요.

문제 1-3. 구조화된 객체에 있는 모든 뉴스 제목을 선택하여 저장해주세요.

HINT1. 크롬창을 통해 위 링크로 이동한 후 [F12] 버튼을 누르면 개발자도구 창을 열 수 있습니다.

HINT2. 개발자도구 창에서 [Ctrl+Shift-C]를 누르거나 왼쪽 상단의 마우스모양 버튼을 누르고 기사 제목을 클릭하면 기사 제목이 a태그 내에서 어떤 class를 사용 중인지 알 수 있습니다.

HINT3. select_one 함수는 하나의 태그만 가져오고 select 함수는 모든 태그를 다 가져옵니다. 상황에 따라 필요한 함수를 사용하세요.

문제 1-4. 총 10개의 기사가 리스트에 저장되어 있습니다. 그 중 3번째 기사의 제목과 url을 확인해보세요

문제 2. 위 코드를 참고하여 특정 키워드와 관련된 기사를 원하는 페이지까지 크롤링하여 페이지수, 기사의 제목, 작성시점, 언론사, url 정보를 가진 데이터를 만들어봅시다.

HINT1. 만들어진 데이터프레임의 예시는 다음과 같습니다. (예시는 3페이지까지 크롤링한 결과입니다.)

	page	date	press	title	url
0	1	12시간 전	쿠키뉴스	"대학교 상권 살아나나" 매출 22%↑…성균관대역 가장 큰 폭	http://www.kukinews.com/newsView/kuk202304200019
1	1	10시간 전	뉴스1	성균관대·서울과기대·서울시립대, 서울경제진흥원과 'R&D산학협력'	https://www.news1.kr/articles/5021775
2	1	7시간 전	베리타스알파	성균관대 오정수 교수 연구팀 난자의 특이적 DNA 손상 복구 기전 규명	http://www.veritas-a.com/news/articleView.html...
3	1	4시간 전	내일신문	2024학년도 의·치대 입학전형	http://www.naeil.com/news_view/?id_art=458358
4	1	5시간 전	한국강사신문	성균관대학교, 2023학년도 성균인문동양학아카데미(SAAH) 11기 입학식 성료	https://www.lecturernews.com/news/articleView....
...	...	...	...	...	...
25	3	6일 전	에너지경제	성균관대, 초고민감 과불화옥탄산(PFOA) 검출 센서 개발	https://www.ekn.kr/web/view.php?key=2023041401...
26	3	2023.04.10.	이데일리	성균관대, 챗GPT 부정행위 대응 플랫폼 전국 최초 구축	http://www.edaily.co.kr/news/newspath.asp?news...
27	3	2023.04.11.	연합뉴스	[게시판] 성균관대-아임뉴런, 뉴로바이오테크놀로지 심포지엄	https://www.yna.co.kr/view/AKR2023041109170000...
28	3	2023.04.05.	뉴스1	수원시, 성균관대 등 지역 5개 대학과 첨단기업 유치 '맞손'	https://www.news1.kr/articles/5005794
29	3	3일 전	오늘경제	농협중앙회, 성균관대학교와 '데이터사이언스 프로젝트' 교육 맞손	http://www.startuptoday.co.kr/news/articleView...

30 rows × 5 columns

HINT2. 크롬창에서 직접 페이지를 넘기면서 url의 'start' 파라미터의 수가 어떻게 변화하는지 확인해보세요.

https://search.naver.com/search.naver?where=news&query=성균관대…
Proctocol Domain Path Parameter
'&'로 연결

참고. url을 복사&붙여넣기 하는 경우 키워드 파라미터인 'query' 부분이 한글인 경우 다른 글자로 인코딩 되는데 이는 잘못된 것이 아닙니다. 코드 내에서 키워드를 한글로 작성해도 정상적으로 작동합니다.

Chapter 2 동적 웹페이지 크롤링

우리가 보는 대부분의 웹 페이지는 동적 웹 페이지라고 할 수 있습니다. 사용자의 요청에 따라서 원하는 페이지를 동적으로 생성하여 보내주는 것입니다. 예를 들어, 아래의 문제 1 에서 사용하는 '네이버 쇼핑' 웹 페이지의 경우, 개발자도구를 확인해보면 스크롤을 내림에 따라 추가적으로 제품 정보가 생성됩니다. 이처럼 동적 웹페이지는 사용자의 동작에 따라 html 의 데이터가 변하기 때문에 url 만 불러오는 기존의 방식으로는 원하는 데이터를 추출할 수 없습니다. 따라서 사용자의 동작을 프로그램이 자동으로 수행하도록 해야 하는데 이때 사용하는 것이 Selenium 입니다. Selenium 버전 4 기준으로 설명을 드리며, 처음 설치하시는 분들은 크롬 드라이버 설치 또한 필요합니다. 'webdriver-manager' 패키지를 사용하면 편리하게 자동 설치 가능합니다.

Selenium 간단 설명 – 네이버에 '피셋' 검색하기 예시

크롤링을 처음 접해보는 분들께 selenium이 어떤 식으로 작동되는지 간단하게 설명해드리는 코드입니다.

```
#크롬 자동 업데이트(크롤링 시 새로운 창으로 실행하도록 하는 코드)

from webdriver_manager.chrome import ChromeDriverManager

#창 꺼짐 방지(이 코드를 실행시키지 않으면 브라우저가 바로 꺼짐)

chrome_options = Options()

chrome_options.add_experimental_option('detach', True)

#에러 없애기(불필요한 에러가 출력되지 않게 하는 코드)

chrome_options.add_experimental_option('excludeSwitches',['enable-logging'])

service = Service(executable_path=ChromeDriverManager().install())

driver=webdriver.Chrome(service=service, options = chrome_options)
```

위의 코드는 크롤링 코드를 작성할 때 자주 사용하기 때문에 어떤 역할을 하는지만 숙지하고, 복사해서 사용하시면 편합니다.

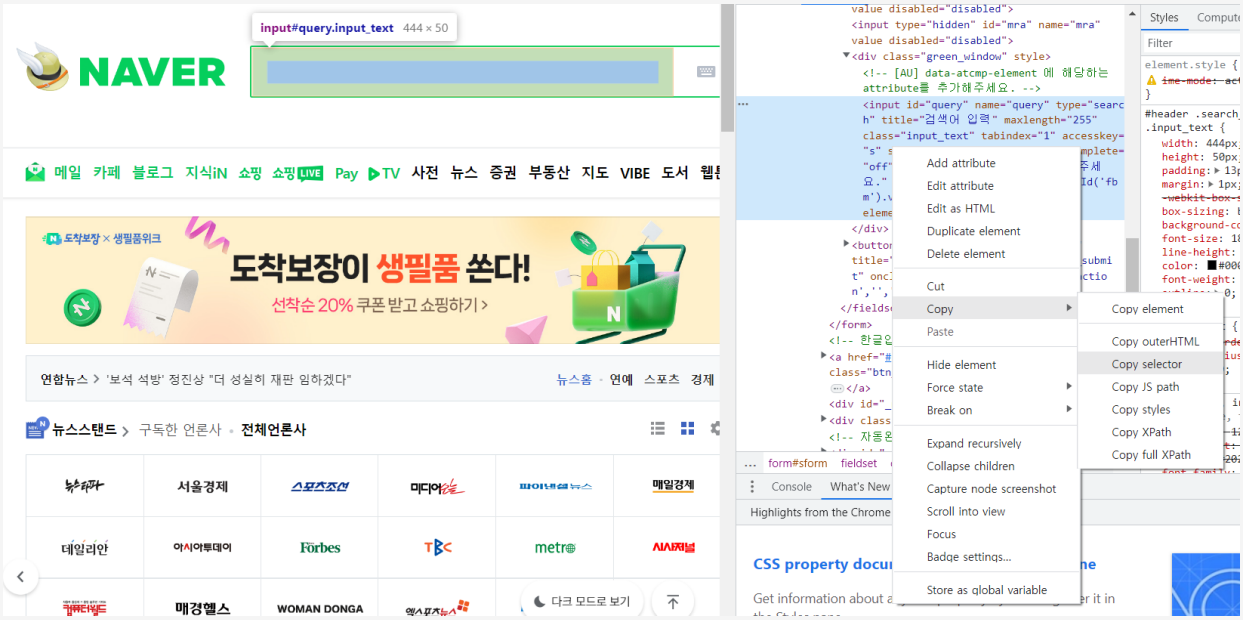
```
driver.implicitly_wait(5)      #브라우저가 로딩될 때까지 5초까지는 기다리도록 명령

driver.maximize_window() #창 최대화 함수

driver.get('https://www.naver.com/')#링크 불러오기
```

위 코드까지 실행해보면 네이버 창이 새로운 브라우저에서 오픈됩니다.

que = driver.find_element(By.CSS_SELECTOR,'#query') #검색창 지정하기



element를 찾는 방식은 class name, path, css 등 다양합니다. 위의 예시에서는 css_selector를 사용했으며 Xpath와 selector의 경우 위의 이미지와 같이 해당 태그에서 우클릭을 하여 복사할 수 있습니다.

que.click() #검색창 클릭하기

que.send_keys('피셋') #'피셋' 입력하기(send_keys함수는 키보드를 조작하는 함수입니다)

btn = driver.find_element(By.CSS_SELECTOR,'#search_btn > span.ico_search_submit') #검색버튼 지정

btn.click() #검색버튼 클릭

검색버튼 코드는 아래 코드로 대체 가능

que.send_keys(Keys.ENTER) #엔터키 치기

문제1. 네이버 쇼핑 웹페이지를 크롤링하여 상품 정보 데이터를 만들어봅시다.

문제1-1. Selenium을 사용하여 https://shopping.naver.com/home 를 호출한 후 원하는 키워드를 검색해주세요.

HINT1. 로딩되는 데에 로딩 시간에 의해 에러가 발생하면, 작동을 지연시키는 time.sleep()함수를 이용합니다.

문제1-2. 스크롤을 페이지 가장 아래로 내리도록 합니다.

HINT1. 해당 웹페이지는 스크롤을 내릴 때마다 데이터가 늘어나기 때문에 데이터가 늘어나지 않을 때까지 스크롤을 내려야 합니다.

HINT2. execute_script을 사용하여 스크롤을 내리고 스크롤 위치를 알 수 있습니다.

문제1-3. 모든 제품 정보를 찾아 저장합니다.

HINT1. find_element는 하나의 태그를 찾고, find_elements는 여러 개의 태그를 찾아줍니다. 상황에 따라 필요한 함수를 사용하세요.

문제1-4. 제품명, 가격, url, 광고여부 데이터프레임을 만들어보세요.

HINT1. 데이터 프레임의 예시는 다음과 같습니다. (키워드: 아이폰13미니)

	제품명	가격	url	광고
0	아이폰 13 mini 자급제 128GB 미드나이트 Apple	874,000원	https://adcr.naver.com/adcr?x=XHbV86+TZi3em2Q1...	o
1	아이폰 13 mini 자급제 128GB 핑크 Apple	874,000원	https://adcr.naver.com/adcr?x=PeLOefdcGMx0Bhba...	o
2	아이폰 13 mini 자급제 256GB 스타라이트 Apple	997,350원	https://adcr.naver.com/adcr?x=7thpv3SbQkVbK6X9...	o
3	Apple 아이폰 13 256GB [자급제]	1,168,480원	https://cr.shopping.naver.com/adcr.nhn?x=Mc6YN...	x
4	Apple 아이폰 13 미니 128GB [자급제]	855,000원	https://cr.shopping.naver.com/adcr.nhn?x=BjUpf...	x
...	...	...	...	...
41	애플 아이폰13mini 미니 128GB / 공기계/자급제	798,000원	https://cr.shopping.naver.com/adcr.nhn?x=KC9hT...	x
42	아이폰 13 mini 자급제 128GB 핑크 Apple	874,000원	https://cr.shopping.naver.com/adcr.nhn?x=rG%2B...	x
43	아이폰13 미니 256G 애플	546,700원	https://cr.shopping.naver.com/adcr.nhn?x=QP50E...	x
44	해외아이폰13미니 128GB 미국판 직구 무음 언락폰공기계자급제 미개봉미활성화 홍콩폰	869,000원	https://cr.shopping.naver.com/adcr.nhn?x=yP2Du...	x
45	아이폰13미니 iPhone13 Mini 128GB 256GB 기가 정품 그린	798,000원	https://cr.shopping.naver.com/adcr.nhn?x=f40EC...	x

46 rows × 4 columns

HINT2. 데이터 웹페이지에 있는 제품 중 판매중단 되어 가격 정보가 삭제된 경우에는 에러가 발생합니다. 이 때는 가격 대신 '판매중단' 값을 갖도록 해야 합니다.

HINT3. '광고'가 붙은 제품과 아닌 제품의 차이점을 찾아 광고 관련 태그를 이용하여 광고여부를 구할 수 있습니다.

문제2. 네이버 항공권 정보를 크롤링하여 데이터프레임을 만들어봅시다.

조건1. https://flight.naver.com/ 를 호출하여 실행해주세요.

조건2. 편도, 오사카행, 6월 15일자에 출발하는 상위 30개의 항공권의 항공사, 출발시간, 소요시간, 가격 데이터를 만들어주세요.

조건3. 할인을 하지 않는 항공권은 셀 색깔로 표시해주세요. (예시 '#FFE8C4' 색 사용)

조건4. 할인을 하는 항공권의 경우 할인 후 가격을 불러오도록 해주세요.

HINT1. 로딩시간이 걸리는 명령에는 time.sleep()을 사용해주세요.

HINT2. 팝업창이 뜨는 경우 창을 지우는 명령도 해야 합니다.

HINT3. 접근 방법은 여러가지가 있기 때문에 본인이 편한 방법대로 해결하시면 됩니다. 단, 클래스 이름을 사용하여 접근하는 경우, 동일한 클래스 이름을 가진 태그가 있기 때문에 이를 고려하여 해결해야 합니다. 날짜 선택 시 일(日), 할인 전/후 가격 등이 모두 카테고리 내에서 같은 클래스 이름을 가집니다.

	항공사	출발시간	소요시간	가격
0	피치항공	15:10	01시간 50분	119,460
1	피치항공	21:00	01시간 50분	119,460
2	에어부산	13:20	01시간 50분	130,100
3	에어부산	15:20	01시간 50분	130,100
4	피치항공	07:30	01시간 50분	130,552
5	티웨이항공	15:40	01시간 40분	140,400
6	에어서울	13:15	02시간 00분	141,600
7	에어서울	07:15	01시간 50분	150,800
8	티웨이항공	12:10	01시간 40분	156,600
9	진에어	07:40	01시간 50분	163,000
10	티웨이항공	08:00	01시간 50분	172,800
11	진에어	13:45	01시간 50분	178,200
12	에어부산	09:30	01시간 50분	179,504
13	제주항공	07:10	01시간 50분	180,000
14	한에어	15:40	01시간 40분	194,448