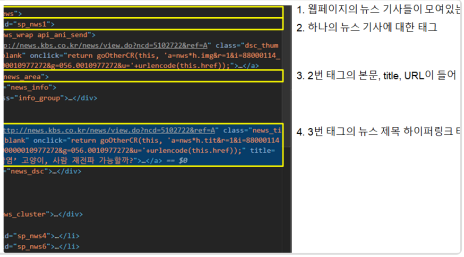
Python News Crawling TIL 1

2022-03-08

[1탄] 쉽게 따라하는 네이버 뉴스 크롤링(python) - 계획 짜기

실제 python 코드를 확인하기를 원하시는 분들은 아래 링크(2탄)을 참고해주세요 :)" 데이터 분석을 공부하면서 python의 여러 패키지를 툴로 사용하였고 그 중 데이터 수집 단계에서 사용되는 크롤링에 대해서

https://everyday-tech.tistory.com/entry/%EC%89%BD%EA%B2%8C-%EB%94%B0%EB%9D%BC%ED%95%98%EB%8A%94-%EB%84%A4%EC%9D%B4%EB%B2%84-%EB%89%B4%EC%8A%A4-%ED%81%AC%EB%A1%A4%EB%A7%81python-1%ED%83%84?category=922285



웹 크롤링이란

웹상에서 보는 정보들(텍스트, URL, 사진 정보 등)을 긁어오는 작업

TIL 1 목표

키워드와 원하는 뉴스기사의 수를 입력해서 관련 네이버 뉴스를 자동으로 엑셀파일로 저장하는 프로그램을 python으로 구현하는 방법 배우기

- input : 검색할 키워드, 추출할 뉴스 기사 수
- output : 뉴스 제목과 URL 담긴 엑셀 파일
- process : 웹 크롤링 알고리즘 (python)

선행 조사

python 웹 크롤링 주요 사용 패키지 조사 및 공부 필요

적용 대상 (네이버 뉴스) 웹 구성 체계 파악

(예 : 제공 정보, 어떤 태그에 속해 있는지, 동적 크롤링 필요 여부)

수행 단계

STEP 1. 소스 조사 : 제공 사이트 조사, 제공 정보 조사, 확보 가능 정보 확인

STEP 2. 웹 구성 체계 확인 : HTML 구조 확인, 제공 정보 확인, 크롤링 가능 여부 확인

STEP 3. 크롤링 진행 : parsing 방안 및 위치 확인, request 방법 확인, data 저장 형태 설계

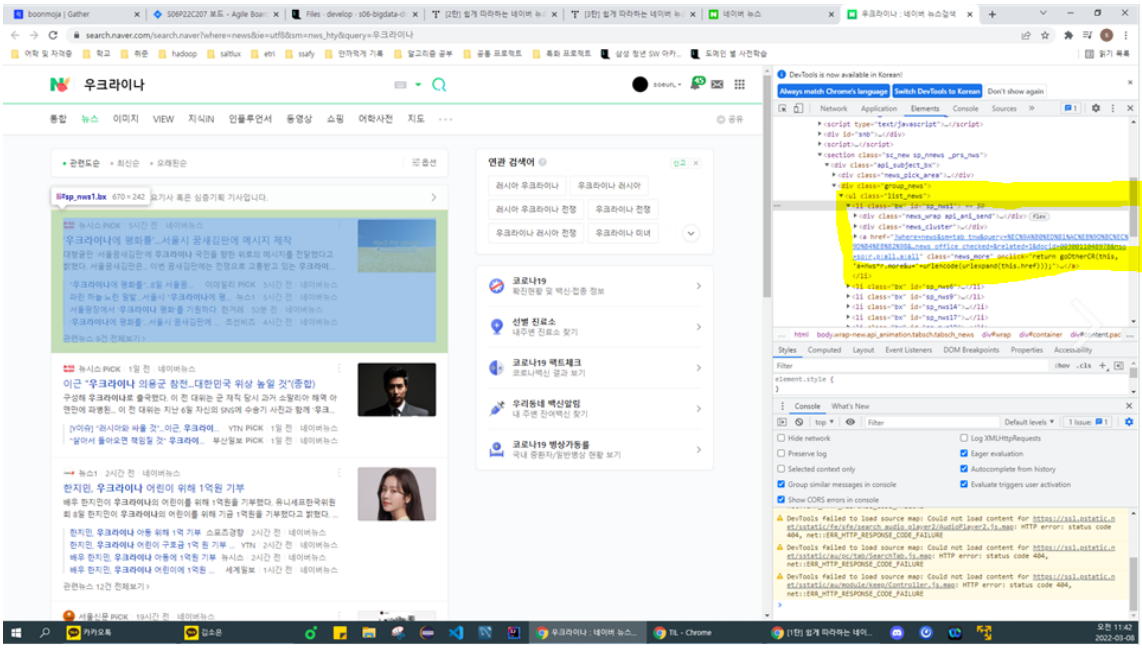
STEP 4. 최종 데이터 생성 : 데이터 저장 형태, 데이터 저장

SETP 1. 소스 조사

- **네이버 뉴스 URL**
  - https://news.naver.com/
  - 네이버 뉴스 검색창에 “우크라이나” 입력 했을 때 https://search.naver.com/search.naver?where=news&ie=utf8&sm=nws_hty&query=우크라이나
  - 네이버 뉴스 DB 검색 시 사용되는 조건문 형태 where=news&ie=utf8&sm=nws_hty&query=우크라이나
- **크롤링 원리**
  - 네이버 서버가 조건에 맞는 뉴스를 HTML 형태로 반환하면 웹 브라우저에서 UI/UX를 적용하여 인터페이싱
  - 웹 페이지에서 원하는 정보가 담긴 HTML 의 위치를 찾아 추출함으로써 원하는 정보 추출

STEP 2. 웹 구성 체계 확인

- **뉴스 기사 title, 뉴스 기사 url 이 위치한 HTML 코드 위치 찾기**



- 뉴스 기사를 감싸고 있는 바운딩 박스 : <li>, id 속성값이 “sp_nws”로 시작
- 각 뉴스마다 id 속성값인 sp_nws 뒷 부분 숫자가 바뀌므로 일반화 조건 검색 필요

- **뉴스 기사 내부의 제목, URL, 본문의 일부를 담고 있는 태그 찾기**

- 뉴스 기사 바운딩 박스 (li 태그) 하위의 div 태그이며 class 속성값이 “new_area”
- div 태그 하위에 최종적으로 뉴스 기사의 title, URL 정보가 담겨있는 a태그 위치
- a태그, class 속성값이 “new_tit”

- **원하는 title, URL 을 해당 태그에서 추출 (해당 속성값 추출)**

- 뉴스 기사 title : title 속성 값
- 뉴스 기사 URL : href 속성값

- **필요한 각각의 정보의 위치를 찾아가는 해당 태그의 위치 최종 정리**

- 노랑색 하이라이트 : 하나의 뉴스 기사에 대한 태그
- 핑크색 하이라이트 : 하나의 뉴스 기사에 대한 title, URL이 들어있는 태그
- 파랑색 하이라이트 : 하나의 뉴스 기사에 대한 뉴스 제목, 하이퍼링크 태그