

Crawling

DSL 5기_박희경

목차

1

강의 전 준비

2

크롤링이란?

3

Html, x path따는 법

4

실습_ipynb 참고

1

강의 전 준비

가상환경

[가상환경 (Virtual Environments)]

자신이 원하는 Python 환경을 구축하기 위해 필요한 모듈만 담아 놓는 바구니라고 생각하면 됩니다.

[가상환경이 필요한 이유] "독립적인 작업환경에서 작업할 수 있다."

프로젝트를 진행하다보면 여러 라이브러리, 패키지를 다운로드하여서 사용하게 됩니다. 그러다 보면 각 라이브러리들끼리 충돌을 일으키는 문제를 발생시키는 경우가 꽤 있습니다. 또는, 특정 버전과 호환하는 경우가 생겨서 최신 버전과 이전 버전 중 선택해야 하는 상황이 발생합니다. 이러한 문제가 발생한 경우에 있어서 잘못하면 전부 삭제하고 다시 설치해야 하는 경우가 많습니다.

이를 방지하기 위해서 프로젝트 단위로 가상환경을 구성해서 필요한 라이브러리를 설치해서 작업을 진행하면 훨씬 작업이 편해집니다. 또한, 다른 컴퓨터 혹은 다른 환경에서 동일 프로그램을 실행시킬 때, 작업 환경을 고정시켰기 때문에 해당 환경에 맞게 구성하면, 작업환경과 버전 문제로 실행되지 않는 문제를 방지할 수 있습니다. (이런 부분을 보았을 때, Docker의 필요성과의도 이어지겠군요.)

(한 줄 요약) 여러 라이브러리, 패키지 간 충돌 방지.

특히, 크롤링에 필요한 패키지를 가상환경 없이 설치하려고 하면 에러가 발생할 수 있습니다.

1

강의 전 준비

가상환경 설치 : Anaconda Prompt에서 아래의 명령어를 실행해주세요.



Anaconda Prompt (Anaconda3)

안
녕

```
conda update conda
```

```
conda create --name Crawling python=3.7
```

```
conda info --envs
```

```
activate Crawling
```

```
conda install jupyter
```

```
pip install --upgrade
```

```
pip install jupyter
```

```
pip install beautifulsoup4
```

```
pip install lxml
```

```
pip install requests
```

```
pip install selenium
```

```
jupyter notebook
```

중간중간 'Proceed ([y]/n)?'라고 뜨면 'y' 라고 입력하시면 됩니다.

Mac을 사용하시는 경우 조금 다르므로 뒷 페이지들을 따라와주세요.

1

강의 전 준비

가상환경 설치 (1)



Anaconda Prompt (Anaconda3)

안

가상환경 생성하기

```
conda create -n [가상환경 이름] python=[원하는 파이썬 버전]
```

사용하시는 버전으로 에러가 뜬다면 python=3.7 하시는 것을 추천드립니다.

만들어진 가상환경 리스트 확인

```
conda info -envs
```

가상환경 활성화

Windows 경우 conda prompt 창에서

```
activate [가상환경이름]
```

Mac OS 경우 Terminal에서

```
source activate [가상환경이름]
```

1

강의 전 준비

가상환경 설치 (1)



Anaconda Prompt (Anaconda3)

열기

가상환경 생성하기

```
conda create -n [가상환경 이름] python=[원하는 파이썬 버전]
```

사용하시는 버전으로 에러가 뜬다면 python=3.7 하시는 것을 추천드립니다.

만들어진 가상환경 리스트 확인

```
conda info -envs
```

가상환경 활성화

Windows 경우 conda prompt

```
activate [가상환경이름]
```

가상환경이 올바르게 활성화되면 이렇게 됩니다!

선택 관리자: Anaconda Prompt (Anaconda3) - jupyter notebook

```
Collecting chardet<5,>=3.0.2
  Downloading chardet-4.0.0-py2.py3-none-any.whl (178 kB)
    | 178 kB 3.3 MB/s
Collecting urllib3<1.27,>=1.21.1
  Downloading urllib3-1.26.6-py2.py3-none-any.whl (138 kB)
    | 138 kB 2.2 MB/s
Installing collected packages: urllib3, idna, chardet, requests
Successfully installed chardet-4.0.0 idna-2.10 requests-2.25.1

(Crawling) C:\WINDOWS\system32>pip install selenium
Collecting selenium
  Using cached selenium-3.141.0-py2.py3-none-any.whl (904 kB)
Requirement already satisfied: urllib3 in c:\programdata\anaconda3\lib\site-packages (1.26.6)
Installing collected packages: selenium
Successfully installed selenium-3.141.0

(Crawling) C:\WINDOWS\system32>
```

1

강의 전 준비

가상환경 설치 (2)



Anaconda Prompt (Anaconda3)

명령

새로 만든 가상환경에 필요한 라이브러리들을 설치.

```
conda install jupyter
```

```
pip install --upgrade pip
```

```
pip install jupyter
```

```
pip install beautifulsoup4
```

```
pip install lxml # 구문을 분석하는 parser
```

```
pip install requests
```

```
pip install selenium
```

#아래의 코드를 입력하여 jupyter notebook을 실행

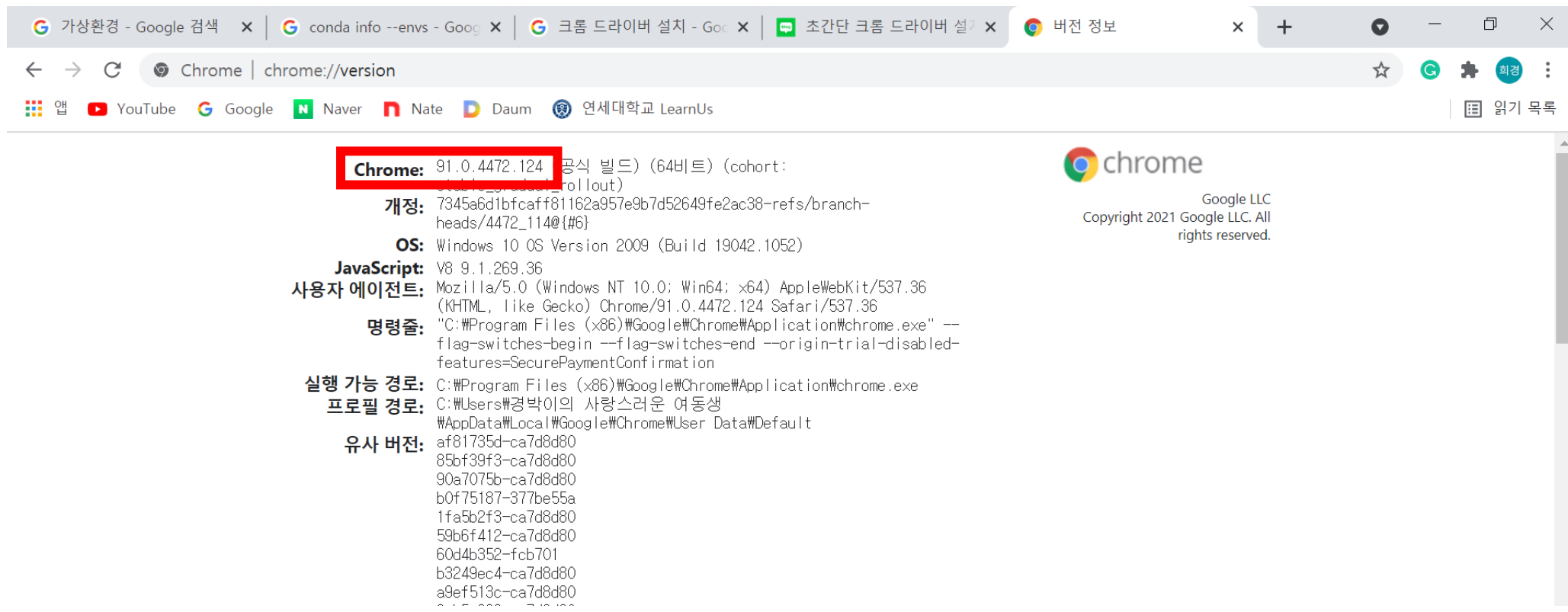
```
jupyter notebook
```

강의 전 준비

크롬 드라이버 설치 : selenium 사용에 필요합니다.

Step1> 나의 Chrome 버전 확인

크롬 주소 창에 "Chrome://version" 이라고 치면 아래와 같이 크롬 버전이 확인됩니다.



1

강의 전 준비

크롬 드라이버 설치 : selenium 사용에 필요합니다.

Step2>_Chrome driver 다운

<https://sites.google.com/a/chromium.org/chromedriver/downloads>

에서 내 Chrome 버전과 동일한 Chrome driver를 다운 받는다.

ChromeDriver -
WebDriver for Chrome

CHROMEDRIVER
CAPABILITIES & CHROMEOPTIONS
CHROME EXTENSIONS
CHROMEDRIVER CANARY
CONTRIBUTING

Downloads

Current Releases

- If you are using Chrome version 92, please download [ChromeDriver 92.0.4515.102](#)
- If you are using Chrome version 91, please download [ChromeDriver 91.0.4472.101](#)
- If you are using Chrome version 90, please download [ChromeDriver 90.0.4430.24](#)
- If you are using Chrome version 89, please download [ChromeDriver 89.0.4389.23](#)
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

ANDROID
CHROMEOS

저는 91이라 2번째 것을 다운받았습니다.

1

강의 전 준비

크롬 드라이버 설치 : selenium 사용에 필요합니다.







Step3> 컴퓨터 운영체제에 맞게 다운

Home Page x | 크롤링 세션 x | 크롤링 세션 x | 쿠팡! - 노트 x | 4 Selenium x | ChromeDriv x

← → ↺ chromedriver.storage.googleapis.com/index.html?path=91.0.4472.101/

앱 YouTube Google Naver Nate Daum 연세대학교 LearnUs

Index of /91.0.4472.101/

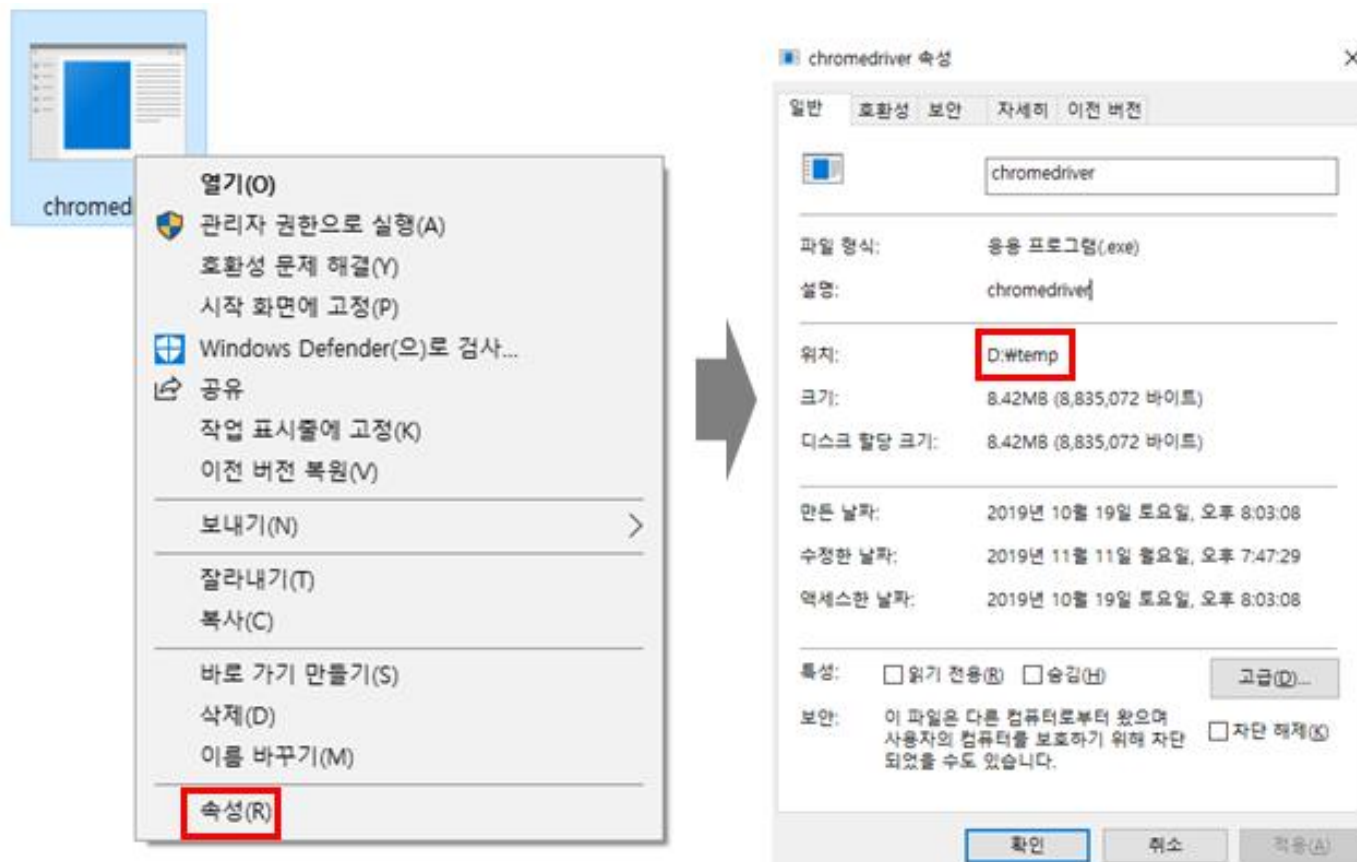
	<u>Name</u>	Last modified	Size	ETag
	Parent Directory	-	-	-
	chromedriver linux64.zip	2021-06-11 10:24:57	5.69MB	cc43ba0babbfff7f22b48165ec8e8c81
	chromedriver mac64.zip	2021-06-11 10:24:59	7.70MB	032eac2d797a0bdc2484cce1843334d1
	chromedriver mac64 m1.zip	2021-06-11 10:25:02	7.03MB	03f777064144a4a76221e5a6d15f9f11
	chromedriver win32.zip	2021-06-11 10:25:04	5.60MB	0300a26b734e93972552a80a6b716100
	notes.txt	2021-06-11 10:25:09	0.00MB	371f90b6fc18478864154ec6e95350ac

1

강의 전 준비

크롬 드라이버 설치 : selenium 사용에 필요합니다.

Step4> 압축해제 후 경로 알아두기



2

크롤링이란?

Crawling





크롤링이란?

Crawling



1. 크롤링이란?

- 웹 상에 있는 데이터들을 긁어오는 기술
- 인터넷에 있는 정보 중 우리가 원하는 것만 골라서 자동으로 수집해주는 기술

2. 필요한 이유?

- 데이터 수집
- 업무 자동화 : 크롬에서 클릭, 아이디 입력, 스크롤 등을 자동화할 수 있음

Home Page - Select or create a x 크롤링 세션 준비 - Jupyter Note x Tryit Editor v3.6 네이버 뉴스

news.naver.com

앱 YouTube Google Naver Nate Daum 연세대학교 LearnUs

읽기 목록

NAVER 뉴스 TV연예 스포츠 뉴스스탠드 날씨 프리미엄

박희경 84 99+


뉴스홈 속보 정치 경제 사회 생활/문화 세계 IT/과학 오피니언 포토 TV 랭킹뉴스

뉴스 검색

신문 헤드라인 저녁 방송 뉴스 팩트체크 언론사 설정 언론사 뉴스 라이브러리

헤드라인 뉴스

헤드라인 뉴스와 각 기사묶음 타이틀은 기사 내용을 기반으로 자동 추출됩니다.



거리두기 개편 시행과 함께 부산 주점서 확진자 ...

김총리 "방역 중대위기...언제라도 거리두기 단계 상향" 39

확진자 폭증에 델타변이까지...서울시 "민노총 1만명 ... 85

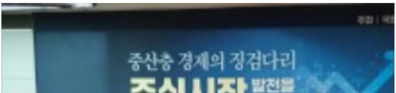
"중국산 백신 격리면제 괜찮나"...싱가포르, 시노백 접... 114

'요양급여 편취' 윤석열 장모 징역 3년...법정구속

[단독]女43% 男29% "성관계 안한다"... 한국인 '섹스...

정치

일반 국회/정당 청와대



중산층 경제의 장검다리

"잘나가던 감사의 이중잣대"...與 대선주자, 尹 장모 판결에... 시사저널

박인호 공군참모총장 "이 중사에게 명복을 빈다, 공군 분골... 파이낸셜뉴스

언론사별 가장 많이 본 뉴스

더보기

오후 3시~오후 4시까지 집계한 결과입니다.

이준석 "한국은 연좌제 없는 나라... 尹에 속았다? 3심 ... 중양일보

옷 벗겨진 시신 잇단 발견..."자살로 위장, 재수사해라" 靑 ... 뉴스1

이재명 "검찰 발표 사실이라면 조국 가족 책임져야" 노컷뉴스

칠판에 동그라미 그리는 여고생..해외서도 난리난 영상 파이낸셜뉴스

윤석열 장모 구속에...조국 "10원 아니다, 22억9천만원" 국민일보

https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=102...

검색하려면 여기에 입력하십시오.

오후 4:23 2021-07-02

HTML, Xpath

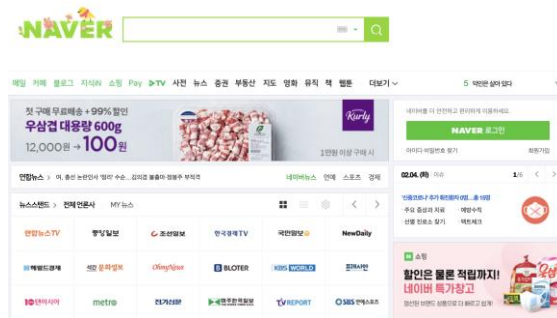
HTML(Hyper Text Markup Language)

네이버에 가고 싶어...

<https://www.naver.com>
가져와!



125.209.222.141



Get!

Html 코드

[illegible]

3

HTML, Xpath

HTML

[웹의 구성]

HTML



Html (집의 뼈대)
웹페이지를 만들 때
사용하는 언어

CSS



css(인테리어)
예쁘게 꾸며줌.

JS



java script
(사람이 들어와 산다.)
살아있게 해줌

페이지는 바뀌므로 왜 이런 코드가 나왔는지 이해하는 것이 필요합니다.

3






HTML, Xpath

HTML

W3 school

Try editor : 간단한 html 코드를 구현할 수 있음

(https://www.w3schools.com/html/tryit.asp?filename=tryhtml_basic)



Run »

Result Size: 429 x 422

Get your own website

```
<!DOCTYPE html>
<html>
<head>
  <meta charset="utf-8">
  <title> DSL</title>
</head>
<body>
  <h1> 글자를 입력할 수 있지 </h1>
  그냥 글을 쓰셔도 됩니다
  <input type = "text" value= "아이디를 입력하세요.">
  <input type = "password">
  <input type = "button" value="로그인">
  <a href = "http://google.com"> 구글로 이동하기 </a>
</body>
</html>
```

글자를 입력할 수 있지

그냥 글을 쓰셔도 됩니다

아이디를 입력하세요.

로그인 [구글로 이동하기](http://google.com)



3

HTML, Xpath

HTML

```
<html>
<head>
  <meta charset="utf-8">
  <title> DSL</title>
</head>
<body>
  <h1> 글자를 입력할 수 있지 <h1>
  그냥 글을 쓰셔도 됩니다
  <input type = "text" value= "아이디를 입력하세요.">
  <input type = "password">
  <input type = "button" value="로그인">
  <a href = "http://google.com"> 구글로 이동하기 </a>
</body>
</html>
```

갈색 → '태그'

태그 안의 빨간색 → 'type'

“ “ 안의 파란색 → 'value(값)'

Type과 value는 '속성' (attribute)라고 하며, 하나의 element를 이룬다.

열고 닫는 개념이 존재함 : <head> 연다. </head>닫는다. 그 사이에 있는 웹의 텍스트

tag간 부모-자식-형제 관계가 존재한다.

3

HTML, Xpath

HTML

```
<html> # 열기
<head> # 웹페이지 제목, 기본 설정
  <meta charset="utf-8"> # 한국어
  <title> DSL</title> # 페이지 제목
</head>
<body>
  <h1> 글자를 입력할 수 있지 <h1> # 글자 입력
  그냥 글을 쓰셔도 됩니다
  <input type = "text" value= "아이디를 입력하세요.">
  <input type = "password"> # 글을 입력하면 암호화
  <input type = "button" value="로그인">
  <a href = "http://google.com"> 구글로 이동하기 </a> # 글자 링크
</body>
</html> # 닫기
```

3

HTML, Xpath

X path : HTML의 경로

```
<학교 이름 ="연세 고등학교">
  < 학년 value="1학년">
    <반 value="1반">
      <학생 value="1 번">유재석</학생>
      <학생 value="2 번">조세호</학생>
      <학생 value="3 번">정우성</학생>
      <학생 value="4 번">이지은</학생>
      <학생 value="5 번">유재석</학생>
    </반>
  < /학년>

  < 학년 value="2학년"/>
```

(연세고등학교 조세호 학생) /학교/학년/반/학생[2]

(ex) /html/body/div[2]/span[3]/a

3

HTML, Xpath

X path : HTML의 경로

```
<학교 이름 = "연세 고등학교">
  < 학년 value="1학년">
    <반 value="1반">
      <학생 value="1번" 학번="1-1-1">유재석</학생>
      <학생 value="2번" 학번="1-1-2">조세호</학생>
      <학생 value="3번" 학번="1-1-3">정우성</학생>
      <학생 value="4번" 학번="1-1-4">이지은</학생>
      <학생 value="5번" 학번="1-1-5">유재석</학생>
    </반>
  < /학년>

  < 학년 value="2학년"/>
```

(연세고등학교- 유재석 학생) //*[@학번="1-1-5"]

(ex) //*[@id="login"]



HTML, Xpath

X path



[Chrome에서 X- path 따는 법]

Chrome의 해당 페이지에서

(방법1) 여러분이 원하는 element에 마우스 우측 클릭 - [검사]- [개발자도구]

(방법2) 오른쪽 위 점 3개- [도구 더 보기]- [개발자 도구]

(방법3) 키보드 [f 12]

화면에서 여러분이 선택한 element의 html코드에서
우클릭- [copy] - [copy x path]

4

실습

lpynb 파일을 참고하시면 됩니다.

1. 네이버 뉴스 크롤링_beautifulsoup
(+a. 네이버 영화 리뷰_beautifulsoup)
2. 연세 포탈 로그인 자동화_selenium
3. 네이버 항공권_selenium

4

실습

lpynb 파일을 참고하시면 됩니다.

1. BeautifulSoup

보편적으로 사용하기 용이한 라이브러리
html 가독성이 좋아 대부분 잘 읽힌다.

2. selenium

스크래파이, lxml, 뷰티풀 스프 등 다른 라이브러리가 하지 못하는 동적할당페이지를 유일하게 크롤링 할 수 있다.

Q) 동적할당페이지란?

페이지를 웹상에서 클릭했을 때 코드가 전혀 보이지 않는 구조를 갖고 있는 페이지
하지만 셀리니움은 속도면에서 새로운 페이지를 띄워야 하니 느리다.

Q & A

THANK YOU

감사합니다