

크롤링

엔지팀 19기 배지호

크롤링

웹 페이지를 그대로 가져와서 거기서 필요한 정보를 추출하는 작업

Request & Response

웹상에서 클라이언트와 서버 간에 요청/응답으로 데이터를 주고 받을 수 있는 프로토콜

Method : 클라이언트가 서버로 보내는 요청의 방법

Get : 서버에게 조회할 리소스를 요청
요청 내용을 Body에 담지 않고 쿼리스트링을 통해 전송

Post : 서버의 상태를 변경하거나 값을 추가할 때 사용
전송할 데이터를 HTTP Body에 담아서 전송

크롤링 라이브러리

BeautifulSoup

간단한 코드, 빠른 속도

크롤링이 불가능한 상황이 있음
(ex. 동적 페이지)

 Selenium

Chrome을 이용해 실제 페이지를 띄우고 우리가 키보드, 마우스로 하는 동작들을 자동화해주는 라이브러리

-> 동적으로 활용 가능(ex. 검색 사이트)

But, 코드가 복잡하고 속도가 느림

실습

1. 크롬 버전 확인 -> `chrome://version`

Chrome 97.0.4692.99 (공식 빌드) (64비트) (cohort: Stable)

실습

2. 크롬 웹 드라이버 설치







<https://sites.google.com/chromium.org/driver/>

All versions available in Downloads

- Latest beta release: ChromeDriver 98.0.4758.48
- Latest stable release: ChromeDriver 97.0.4692.71

실습

Index of /97.0.4692.71/

| | Name | Last modified | Size | ETag |
|---|---|---------------------|--------|----------------------------------|
|  | Parent Directory | | - | |
|  | chromedriver_linux64.zip | 2022-01-05 05:45:10 | 9.52MB | 156efec320a9760140fcbeac740513d2 |
|  | chromedriver_mac64.zip | 2022-01-05 05:45:13 | 7.89MB | 952c27f9ca42748db82b15f8c0c59d3c |
|  | chromedriver_mac64_m1.zip | 2022-01-05 05:45:15 | 7.48MB | bb0e354876e64b8725b620d482247dff |
|  | chromedriver_win32.zip | 2022-01-05 05:45:17 | 5.89MB | 58ac3bf76466773680a5fe04b69ad1d3 |
|  | notes.txt | 2022-01-05 05:45:22 | 0.00MB | 0fe69c56feb42175dd29cf69e4f38e9d |

과제

실습에서 완성한 코드를 확장 시켜 봅시다!

1. 흥행지수, 출연영화, 랭킹 추가하기(출연영화의 value는 리스트로 만들어주세요)
2. 1~10 페이지까지 크롤링하도록 확장

제출 파일은 .ipynb로 해주시면 됩니다.

```
[{'이름': '김운석',  
  '흥행지수': 114195,  
  '출연영화': ['모가디슈'],  
  '랭킹': '1',  
  '직업': '배우',  
  '생년월일': '1968-01-21',  
  '성별': '남',  
  '신장/체중': '178cm, 68kg',  
  '학교': '동의대학교 학사',  
  '취미': '여행, 낚시',  
  '소속사': '심엔터테인먼트'},  
 {'이름': '조우진',  
  '흥행지수': 100327,  
  '출연영화': ['발신제한', '국가부도의 날'],  
  '랭킹': '2',  
  '다른 이름': '조신제',  
  '직업': '배우',  
  '생년월일': '1979-01-16',  
  '성별': '남'}]
```

예시

| | | | |
|--|----------------------------------|--|--|
|  | 허준호(2편) 흥행지수 79,716 |  모가디슈 |  국가부도의 날 |
|  | 김소진(2편) 흥행지수 66,445 |  모가디슈 |  아마겔돈 |
|  | 진경(1편) 흥행지수 65,677 |  아마겔돈 | |

1 2 3 4 5 6 7 8 9 10



1~10 페이지까지 크롤링!



Thank you