Scrapy

1. 크롤링이란?

웹페이지를 그대로 가져와서 데이터를 추출해내는 행위

→ 우리에게 어떻게 적용? : 음원사이트 음원차트나 음원사이트 자체에서 태그(상황, 기분, 날씨, 장소에 맞는 음악)별로 저장해둔 플레이리스트 정보를 추출해옴

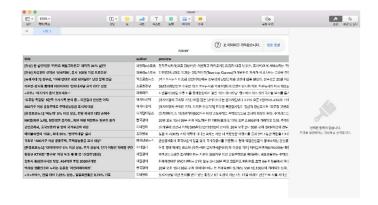
2. Scrapy 란?

https://github.com/scrapy/scrapy.git

- ⇒ Scraping을 도와주기 위한 파이썬 기반 라이브러리
- ⇒ Scrapy를 이용하여 필요한 **페이지로 접속하여 원하는 형태로 데이터 를 가공하여 데이터를 저장**할 수 있도록 도와줌.
- ⇒ 포크 9.8k , 스타 45.1k로 GitHub에서 개발자들 사이에 인기있는 크롤 러.
- ⇒ 크롤링을 할 URL을 입력받아 원하는 데이터를 추출함.
- 3. Scrapy을 이용해 웹 크롤링 예시
- 1) Scrapy을 이용해 네이버 뉴스 크롤링 -> CSV파일로 데이터 추출





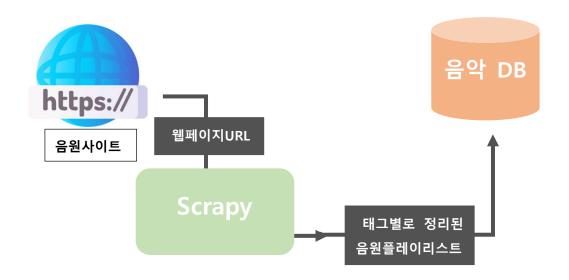


2) Scrapy을 이용해 멜론에서 특정 가수의 발매 곡 목록 크롤링 -> json파일 파일로 데이터 저장

```
Scrapy 파이썬 라이브러리
      class MySpider(scrapy.Spider):
    10
                                             크롤링을 할 URL
          'https://m2.melon.com/cds/artist/mobile4web/artist_songList.htm?startIndex=1&pageSize=20&rowsCnt=20&artistId=686920&fi
    14
    16
17
                                                      크롤링을 통해 데이터를 추출하는 코드
                  'title' : quote.xpath('div[2]/div/a/p').extract_first().replace('<span class=\"sprite title hide\">', '').
'songId' : quote.xpath('div[2]/div/a/@href').extract_first().replace("javascript:goDetail('song','", '').re
'artist' : quote.xpath('div[2]/div/a/span/text()').extract_first(),
'img' : quote.xpath('div[1]/div/a/span[@class="img"]/@style').extract_first().replace('background-image:
   20
21
22
{"title": "\uc9c8\ud22c \ub098\uc694 BABY", "songId": "80490
{"title": "\ub108 \ub54c\ubb38\uc5d0 (Help Me)", "songId": "
{"title": "Oh Boy", "songId": "30179107", "artist": "AOA", "
{"title": "With ELVIS", "songId": "30179108", "artist": "AOA
{"title": "Good Luck", "songId": "8181755", "artist": "AOA",
{"title": "10 Seconds", "songId": "8181756", "artist": "AOA"
{"title": "Cherry Pop", "songId": "8181757", "artist": "AOA"
{"title": "Crazy Boy", "songId": "8181758", "artist": "AOA",
"title". "Still Falle The Rain" "congTd". "8181759" "arti
```

scrapy를 통해 추출한 ison 형식의 데이터

4. Scrapy -> DFD



태그별로 정리된 음원플레이리스트



#카페 #재즈 태그로 유저들이 많이 찾는 곡입니다. *업데이트 요일: 화요일

#탑백 #top100 #인기 #차트

