**행사 크롤러 보고서**

**2023.07.03 신 주 용**

1. **Sources Table**

**1.1 테이블 설명**

* **크롤러가 해당 테이블 소스를 참조해 URL과 TAG를 읽고 자동으로 크롤링**

**1.2 테이블 명세서**

| **컬럼** | **데이터 타입** | **컬럼 타입** | **컬럼 키** | **null 여부** | **컬럼 디폴트값** | **자동여부** | **컬럼 내용** | **특이사항 설명** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| source\_id | int | int(20) | PRI | NO |  | auto\_increment | 소스 ID |  |
| event\_type | int | int(1) |  | NO |  |  | 기관 구분 | 0=이벤트, 1=교육 |
| class\_type | int | int(1) |  | NO |  |  | 동식물 구분 | 0=동물, 1=식물 |
| addr | varchar | varchar(100) |  | YES |  |  | 주소지 | 주소가 정해진 URL은 주소를 그대로 불러옴 |
| url | varchar | varchar(900) |  | NO |  |  | url | 크롤링 할 메인 URL 주소 |
| img\_tag | varchar | varchar(900) |  | YES |  |  | 이미지 태그 | 행사의 메인 이미지 |
| enter\_url\_tag | varchar | varchar(900) |  | YES |  |  | 상세 페이지 태그 | 행사의 상세 정보 URL |
| page\_tag | varchar | varchar(900) |  | YES |  |  | 다음페이지 태그 | 다음 페이지로 넘어가기 위한 태그 |
| title\_tag | varchar | varchar(900) |  | YES |  |  | 행사명 태그 | 행사 제목 태그 |
| overview\_tag | varchar | varchar(900) |  | YES |  |  | 행사 내용 태그 | 행사 상세 내용 태그 |
| event\_date\_tag | varchar | varchar(900) |  | YES |  |  | 행사 날짜 태그 | 행사 날짜 태그 |
| playtime\_tag | varchar | varchar(900) |  | YES |  |  | 운영시간 태그 | 행사 운영시간 태그 |
| place\_tag | varchar | varchar(900) |  | YES |  |  | 장소 태그 | 행사 장소 태그 |
| addr\_tag | varchar | varchar(900) |  | YES |  |  | 주소 태그 | 행사 주소 태그 |
| organizer\_tag | varchar | varchar(900) |  | YES |  |  | 주최 태그 | 행사 주최 태그 |
| superviser\_tag | varchar | varchar(900) |  | YES |  |  | 주관 태그 | 행사 주관 태그 |
| sponsor\_tag | varchar | varchar(900) |  | YES |  |  | 후원 태그 | 행사 후원 태그 |
| pay\_tag | varchar | varchar(900) |  | YES |  |  | 비용 태그 | 행사 비용 태그 |
| homepage\_tag | varchar | varchar(900) |  | YES |  |  | 홈페이지 태그 | 행사 담당 홈페이지 태그 |
| tel\_tag | varchar | varchar(900) |  | YES |  |  | 연락처 태그 | 행사 담당 연락처 태그 |
| sns\_tag | varchar | varchar(900) |  | YES |  |  | 기타 연락 태그 | 행사 담당 이메일 태그 |
| cycles | int | int(3) |  | NO | 1 |  | 크롤링 주기 | 0 = 매일, 1 = 매 주 월요일, 2 = 매 월 1일 |
| is\_activated | int | int(1) |  | NO | 0 |  | 활성화 여부 | 크롤링이 가능한지 여부 (0=가능, 1=불가능)  \*만약 URL접속이 안되면 자동으로 1로 변경 |
| insert\_date | datetime | datetime |  | NO |  |  | 생성 날짜 | 소스 데이터 생성 날짜 |
| insert\_name | varchar | varchar(20) |  | NO |  |  | 생성자 | 소스 데이터 생성자 |
| update\_date | datetime | datetime |  | NO |  |  | 수정 날짜 | 소스 데이터 수정 날짜 |
| update\_name | varchar | varchar(20) |  | NO |  |  | 수정자 | 소스 데이터 수정자 |
| crawling\_date | datetime | datetime |  | NO |  |  | 크롤링 날짜 | 해당 URL 크롤링이 돌아간 날짜 |

1. **Events Table**

**2.1 테이블 설명**

* **크롤러가 크롤링 해온 정보들을 담는 테이블**

**2.2 테이블 명세서**

| 컬럼 | 데이터 타입 | 컬럼 타입 | 컬럼 키 | null 여부 | 컬럼 디폴트값 | 자동여부 | 컬럼 내용 | 특이사항 설명 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| event\_id | int | int(20) | PRI | NO |  | auto\_increment | 이벤트 ID | 행사 ID |
| source\_id | int | int(20) | FK | NO |  |  | 소스 ID | 소스 ID |
| event\_type | int | int(1) |  | NO |  |  | 행사 구분 | 0 = 이벤트, 1 = 교육  ※ 이미 선별 된 정보를 가져오기 때문에 신뢰도 높음 |
| class\_type | int | int(1) |  | NO |  |  | 동식물 구분 | 0=동물, 1=식물  ※ 키워드로 분류하기 때문에  신뢰도 중상 |
| img\_url | varchar | varchar(900) |  | YES |  |  | 이미지 url | 메인 이미지 주소 |
| title | varchar | varchar(900) |  | YES |  |  | 행사명 | 행사 제목 |
| overview | varchar | varchar(900) |  | YES |  |  | 행사 상세 설명 | 행사 상세 설명 |
| start\_date | datetime | varchar(900) |  | YES |  |  | 행사 시작 날짜 | 행사 시작 날짜 |
| end\_date | datetime | varchar(900) |  | YES |  |  | 행사 종료 날짜 | 행사 종료 날짜 |
| playtime | datetime | varchar(900) |  | YES |  |  | 운영 시간 | 행사 운영 시간 |
| place | varchar | varchar(900) |  | YES |  |  | 행사 장소 | 행사를 진행하는 위치 |
| addr | varchar | varchar(900) |  | YES |  |  | 행사 주소 | 행사 주소 |
| organizer | varchar | varchar(900) |  | YES |  |  | 주최 | 행사 주최 |
| superviser | varchar | varchar(900) |  | YES |  |  | 주관 | 행사 주관 |
| sponsor | varchar | varchar(900) |  | YES |  |  | 후원 | 행사 후원 |
| pay | varchar | varchar(900) |  | YES |  |  | 이용 요금 | 행사 참여에 필요한 비용 |
| homepage | varchar | varchar(900) |  | YES |  |  | 행사 홈페이지 | 행사 상세 홈페이지 |
| tel | varchar | varchar(900) |  | YES |  |  | 연락처 | 행사 문의 전화번호 |
| sns | varchar | varchar(900) |  | YES |  |  | 기타 연락 | 행사 문의 이메일, 인스타그램 등 |
| insert\_date | datetime | datetime |  | NO |  |  | 입력 날짜 | 크롤링 된 날짜 |
| crawling\_info | int | int(1) |  | NO |  |  | 크롤링 데이터 정보 | 0 = 크롤링 된 행사정보, 1 = 서비스에 적용 |

1. **테이블 생성**

**3.1 테이블 생성 범위**

* **sources 테이블: columns, rows (tables.xlsx, sources.xlsx 참조)**
* **events 테이블: columns (tables.xlsx 참조)**

| **python3 create\_tables.py** |
| --- |

1. **코드 명세서**

| File | 설명 | 구분 |
| --- | --- | --- |
| create\_tables.py | Sources 테이블과 Events 테이블을 생성하는 파일  \* sources 테이블에 row 값 자동 생성 |  |
| selenium\_crawling | 크롤러 코드 \*events 테이블에 row 값을 넣음 |  |
| db\_info.txt | Mysql과 연동하기 위한 정보 입력 |  |
| err\_err.txt | 크롤러가 URL 접속을 못했을 때, log기록 |  |
| requirements.txt | 관련 라이브러리 이름, 버전 |  |
| sources.xlsx | create\_tables.py를 실행할 때 필요한 엑셀 파일 |  |
| tables.xlsx | create\_tables.py를 실행할 때 필요한 엑셀 파일 |  |

1. **필요한 설치 프로그램**

**5.1 리눅스용 구글크롬**

* **설치 코드**

| sudo yum install epel-release wget https://dl.google.com/linux/direct/google-chrome-stable\_current\_x86\_64.rpm sudo yum install ./google-chrome-stable\_current\_x86\_64.rpm |
| --- |

**5.2 python==3.10.9**

* **설치 코드**

| sudo yum install epel-release sudo yum install centos-release-scl sudo yum install rh-python310 |
| --- |

**5.3 관련 라이브러리**

* **selenium==4.9.1**
* **chromedriver-autoinstaller==0.4.0**
* **pyperclip==1.8.2**
* **pandas==1.5.3**
* **SQLAlchemy==1.4.39**
* **mysqlclient==2.2.0**
* **openpyxl==3.1.2**
* **beautifulsoup4==4.11.1**
* **requests==2.28.1**
* **tqdm==4.65.0**

| pip3 install -r requirements.txt |
| --- |

**6. 크롤러 진행도**

