

데이터베이스 기말 프로젝트 개요

2022년 5월

데이터베이스 강의 추후 일정

- 10주 - 12주: 데이터베이스 설계
- 13주 - 14주: 인덱싱
- 15주: 기말 프로젝트 관련
- 16주: 기말고사
 - SQL 이후부터 ~ 배운 것까지

기말 프로젝트

- 제목: 네이버 영화 구축
- 내용: 네이버 영화 사이트(movies.naver.com)로부터 영화 관련 데이터를 수집하여, mySQL에 DB화하고, 검색 인터페이스 제공하는 프로젝트
- 프로젝트 주요 구성요소
 - DB 설계
 - 네이버 영화 데이터 수집(crawling) 및 DB 구축
 - 검색 인터페이스 구현
- 팀구성: 1인 또는 2인

데이터 수집 연습 과제

- 기말 프로젝트 선행 과제임
- 네이버에 현재상영영화 화면에 있는 내용을 크롤링하고 하나의 movie 테이블로 구축할 것
- table name: movie
- columns:
 - title: 영화 제목
 - movie_rate: 영화 등급
 - netizen_rate: 네티즌 평점
 - netizen_count: 네티즌 평점 참여자수
 - journalist_score: 기자평론가 평점
 - journalist_count: 기자평론가 참여자수
 - scope: 개요
 - playing_time: 상영시간
 - opening_date: 개봉날짜
 - director: 감독
 - image: 영화 대표 이미지 주소
 - enter_date: 투플 생성 날짜/시간

python을 통한 데이터 수집(crawling) 참고 자료

- 데이터 수집을 위한 package
 - requests: 주어진 url의 웹페이지를 fetch 할 수 있음.
 - BeautifulSoup: 웹페이지를 파싱해서 특정 부분을 접근하여 가공할 수 있음.
 - Selenium: Javascript로 된 웹페이지를 chrome browser와 동일한 방식으로 접근하여, 수집할 수 있음.
 - Scrapy: 웹페이지에 접속하여 원하는 형태로 데이터를 가공하고 저장할 수 있음.
- 참고 웹사이트
 - <https://machinelearningmastery.com/web-crawling-in-python/>