

## 프로젝트 #2: DB implementation & query processing

본 프로젝트는 사이트 A의 사용자, 게시물 및 기타 관련 데이터를 바탕으로 데이터에 적합한 데이터베이스 스키마를 설계하여 데이터베이스 테이블을 실제로 생성한 후 데이터를 입력하고 활용하는 프로그램을 구현하는 것을 목적으로 한다. 해당 프로그램은 python과 MySQL을 사용하여 구현하여야 하며, 다음 요구 조건을 만족하여야 한다.

- (R1) 해당 프로젝트에서 주어진 데이터를 바탕으로 데이터 중복을 최소화할 수 있도록 데이터베이스 스키마를 설계하고 설계한 스키마가 3NF를 만족함을 보여야 한다.
- (R2) 프로그램은 MySQL 상에 데이터베이스를 생성한다. 데이터베이스의 이름은 'db2017\_{#}'이다. {#} 부분은 해당 조의 번호로 대체한다.  
ex) db2017\_3
- (R3) 설계한 데이터베이스의 테이블 생성을 수행한다. INTEGER Type은 'INT(11)'로, STRING type은 'VARCHAR(255)'를 이용하여 생성한다. 255자를 넘는 경우 'LONGTEXT'를 이용하여 생성한다. 그 외 날짜시간은 'DATETIME'으로, 날짜는 'DATE'를 통해 생성한다.
- (R4) 생성된 테이블에 데이터를 저장해야 한다. 데이터는 csv파일로 주어져 이틀 직접 변형해선 안 된다.
- (R5) 프로그램 실행 시 데이터베이스 또는 테이블이 이미 존재할 경우 데이터베이스, 테이블의 생성과 데이터 저장을 다시 수행하지 않아야 한다.
- (R6) 사용자의 나이를 10대(10~19), 20대(20~29), 30대(30~39), 40대(40~49), 50대 이상(50 ~)으로 분류했을 때, 업로드 질문 게시물의 총 조회 수가 가장 높은 1명을 나이대 별로 선발하여, 총 5명의 사용자 정보(사용자 고유번호, 평판, 이름, 나이, 계정 생성 날짜시간, 마지막으로 접속한 날짜시간, 홈페이지 주소, 거주 지역, 본인 소개) 및 업로드 질문 게시물의 총 조회 수를 출력하시오. 출력된 결과는 평판의 내림차순으로 정렬되어야 한다.
- (R7) 사용자 계정 생성 날짜시간을 2010년(= 2010/1/1 00:00:00 ~ 2010/12/31 23:59:59), 2011년, 2012년, 2013년, 2014년으로 나뉘었을 때, 각 년도 별 생성 계정 수를 출력하시오. 출력된 결과는 2010년, 2011년, 2012년, 2013년, 2014년 순서이어야 한다. (5 columns)
- (R8) 댓글 수가 10개 이상, 좋아요가 1개 이상인 게시물에 대하여 게시물 고유번호, 게시물이 획득한 좋아요 수, 게시물이 획득한 싫어요 수, 계

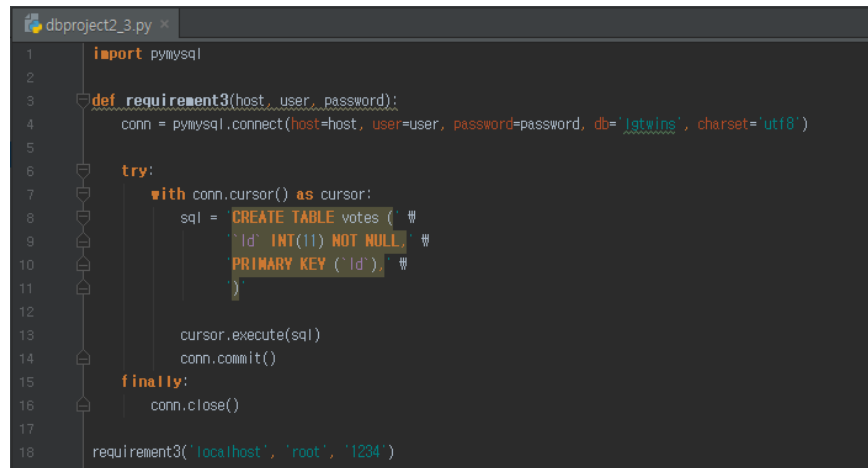
시물이 획득한 점수(좋아요 - 싫어요)로 구성된 결과를 출력하시오. 단, 게시물 점수의 내림차순으로 정렬되어야 한다.

(R9) 획득한 배지가 50개 이상인 사용자들이 1년마다 평균적으로 몇 개의 게시물을 작성하는지 구하시오.(년도의 정의는 (R7) 참고) 출력된 결과는 사용자 고유번호, 획득한 배지 수, 평균 게시물 작성 수 순서로 구성되어야 하며, 획득한 배지 수의 내림차순으로 정렬되어야 한다.

(R10) (자유주제) 주어진 데이터로부터 통계적으로 의미가 있는 결과를 ‘한 가지’ 도출하시오. 단, (R6)~(R9)과 중복되어서는 안 된다.

구현한 프로그램 소스(주석 포함)와 보고서를 함께 제출하여야 한다. 이때 주석에는 작성한 쿼리에 대한 설명이 포함되어야 한다. 제출할 보고서에는 설계한 관계 스키마, 스키마의 정규화 결과, 소스 프로그램 설명, 프로그램 실행 결과와 이에 대한 분석이 포함되어야 한다. 관계 스키마의 경우 integrity constraints, foreign keys, primary keys와 3 domain constraints 들을 명시해야 하며, 3<sup>rd</sup> normal form을 만족하여야 한다. 또한 (R6),(R7),(R8),(R9),(R10)는 nested query를 이용하여 SQL 한 문장 만으로 작성되어야 한다.

(R2)~(R4), (R6)~(R10)은 한 python 프로그램 파일에 존재해야 하며, python 프로그램 파일 이름은 ‘dbproject2\_{#}’ 으로 한다. ({#} 부분은 해당 조의 번호로 대체한다.) 한 requirement 당 한 개의 function으로 정의되어야 하며, 각 function의 이름은 requirement{#} (ex: requirement3) 이고, input parameter로 host명, user명, MySQL server password (총 3개)를 받아야한다. 또한 이를 활용하여 MySQL server에 connect하는 pymysql 객체를 반드시 포함하고 있어야 한다. (R6)~(R10)에 해당하는 function은 실행 시 결과 값이 3분 이내로 출력되어야 한다. 또한 (R2)~(R10)의 function 실행 시 오류가 발생하지 않아야 한다.



```

1  import pymysql
2
3  def requirement3(host, user, password):
4      conn = pymysql.connect(host=host, user=user, password=password, db='lgtwins', charset='utf8')
5
6      try:
7          with conn.cursor() as cursor:
8              sql = 'CREATE TABLE votes ('
9                  'id INT(11) NOT NULL,'
10                 'PRIMARY KEY (id),'
11                 ')'
12
13             cursor.execute(sql)
14             conn.commit()
15         finally:
16             conn.close()
17
18     requirement3('localhost', 'root', '1234')

```

<Figure 1 : 프로그램의 예시>

본 프로젝트의 발표시간은 4분이며 python 프로그램 코드, 보고서와 프레젠테이션 발표자료는 11월 16일 0시까지 ETL에 업로드 해야 한다. 또한 발표날 수업시간에 보고서 및 프레젠테이션 하드 카피를 담당 강의자에게 제출해야 한다. (하드카피의 내용과 ETL에 제출된 내용은 정확히 같아야 한다.)

채점 기준:

- 설계된 DB 및 구현된 프로그램의 requirements 만족 여부: 90%
- 보고서 품질: 10%