프로젝트 공지

406.426B 데이터관리와 분석

2020년 2학기

본 과목의 프로젝트는 총 3회로 이루어져 있다. 1차와 2차 프로젝트에서는 주어진 요구사항들을 만족하는 데이터베이스를 설계 및 구현, 그리고 이를 기반으로 한 DB 마이닝과 추천 시스템 구현을 목적으로 한다. 3차 프로젝트에서는 텍스트 데이터를 대상으로 정보 검색, 문서 분류 및 군집화를 구현하는 것을 목적으로 한다.

프로젝트는 다음과 같다.

Project#1) Conceptual DB design & DB implementation

팀 구성일: 9월 29일, 발표일: 10월 13일

Project#2) DB mining & Recommendation system

팀 구성일: 10월 22일, 발표일: 11월 5일

Project#3) Document search engine & Classification and Clustering

팀 구성일: 11월 24일, 발표일: 12월 10일

본 프로젝트는 팀 별로 진행되며, 각 팀은 매 프로젝트마다 자율적으로 3~5명으로 구성된다. 단, 팀원을 구하지 못하는 경우 충원을 희망하는 팀에 임의 배정하거나 팀을 구하지 못한 인원들끼리 팀을 임의로 꾸린다.

비대면 강의 기간이 길어짐에 따라 기간 중 프로젝트 발표는 프로젝트 결과 제출 시에 4분 이내의 발표 영상을 제작하여 함께 제출하고, 프로젝트 발표 일시에 조교가 zoom 화상 강의를 통해 재생한 후 Q&A 시간을 가지는 것으로 한다. 결과물 제출은 발표일 전날 23시 59분까지 보고서와 발표 자료 및 발표 영상을 ETL에 업로드해야 한다.

프로젝트 #1: Conceptual DB design & DB implementation

사이트 A는 온라인 쇼핑몰 사업자와 앱 개발자가 교류하는 서비스다. 개발자들은 쇼핑몰 운영에 필요한 기능을 가진 앱을 개발하여 사이트 A에 올리고 쇼핑몰 사업자들은 자신의 쇼핑몰에 필요한 앱을 구매하여 사용한다. 본 프로젝트는 사이트 A가 사용하는 DB의 ER diagram 도식화와 DB 구현을 목적으로 한다. 본 프로젝트는 크게 두 부분으로 나뉜다.

PART Ⅰ. ER diagram 도식화

PART Ⅱ. DB 구현 및 데이터 입력

PART Ⅰ. ER diagram 도식화

PART Ⅰ는 사이트 A가 사용하는 DB에 대한 ER diagram을 도식화하는 것을 목표로 한다. 사이트 A의 DB는 아래의 requirement들을 만족해야 한다.

(R1-1) 사이트 A는 사이트에 가입된 개발자에 대한 정보를 저장하고 있다. 개발자는 앱을 개발하고 사용자들이 앱을 사용하고 남긴 리뷰에 답변을 남길 수 있는 권한을 가지고 있다. 개발자에 대한 정보로는 사이트에서 부여한 고유번호, 이름, 프로필 링크 개수, 프로필 이미지 존재 여부로 구성되어 있다. 또한 개발자가 좋아요를 누른 카테고리 수, 개발한 앱의 수, 작성한 답변의 수가 저장되어야 한다.

(R1-2) 사이트 A는 사이트에 가입된 사용자에 대한 정보를 저장하고 있다. 사용자들은 앱에 대한 리뷰를 남길 수 있으며, 사용자끼리 팔로우할 수 있다. 사용자에 대한 정보로는 사이트에서 부여한 고유번호, 이름으로 구성되어 있다. 또한 사용자가 좋아요를 누른 카테고리 수, 작성한 리뷰의 수, 팔로워의 수, 팔로잉 수가 저장되어야 한다. 사용자끼리 팔로우 할 때는 팔로우한 날짜가 저장되어야 한다.

(R1-3) 사이트 A는 앱에 대한 정보를 가지고 있다. 앱에 대한 정보로는 사이트에서 부여한 고유번호, 제목(앱의 이름), 설명, 가격 정보, 카테고리 수, 주요 혜택 수, 가격 정책의 수가 저장되어야 한다. 이 때 가격 정보는 빈 칸일 수 있다. 앱에는 카테고리 설정이 가능한데, 하나의 앱에 여러 가지 카테고리 설정이 가능하고 모든 앱은 하나 이상의 카테고리를 가져야 한다. 또한 앱은 주요 혜택을 표시할 수 있는데, 여기에는 제목, 설명이 저장되어야 한다. 하나의 앱이 여러 개의 주요 혜택을 가질 수 있으며, 한 앱의 주요 혜택들은 모두 다른 제목을 가진다.

(R1-4) 사이트 A는 가격 정책 정보를 가지고 있다. 여기에는 사이트에서 부여한 고유번호, 앱의 id, 제목, 가격이 저장되어야 한다. 이 때, 제목은 빈 칸일 수 있다.

(R1-5) 사이트 A는 카테고리들에 대한 정보를 가지고 있다. 여기에는 사이트에서 부여한 고유번

호, 제목(카테고리 이름), 좋아요를 누른 개발자의 수, 좋아요를 누른 사용자의 수, 카테고리에 속한 앱의 수가 저장되어야 한다.

(R1-6) 사이트 A는 사용자들에게 필요할 만한 앱들에 대한 추천 정보를 message로 발신하고 이에 대한 정보를 저장한다. 이 때 사이트에서 부여한 message 고유번호, 수신인 사용자, 보낸 날짜, 포함된 앱의 수를 저장해야 한다.

(R1-7) 사이트 A는 사용자들의 리뷰 정보를 가지고 있다. 여기에는 사이트에서 부여한 고유번호, 작성자 id, 사용자가 앱에 매긴 점수, 본문, 작성 일시, 리뷰가 받은 추천 수가 저장되어야 한다.

(R1-8) 사이트 A는 개발자들의 답변 정보를 가지고 있다. 한 리뷰에 한 개의 답변만 존재할 수 있으며, 답변은 해당 앱의 개발자만 할 수 있다. 이에 대한 정보로는 사이트에서 답변에 부여한 고유번호, 작성자 id, 내용, 작성 일시가 저장되어야 한다.

PART Ⅱ. DB 구현 및 데이터 입력

PART Ⅱ는 이 사이트 A의 데이터에 적합한 데이터베이스 스키마를 설계하여 데이터베이스 테이블을 실제로 생성한 후 데이터 입력까지를 목표로 한다. 해당 프로그램은 Python과 MySQL을 사용하여 구현하여야 하며, 다음의 요구 조건들을 만족하여야 한다. Python에서 mysql-connector-python 외의 별도 라이브러리는 사용할 수 없다.

(R2-1) 사이트 A의 데이터를 활용하기에 앞서 이를 MySQL 상에 저장해야 한다. 이를 위해 먼저 DMA_team##의 이름을 가지는 schema를 생성해야 한다. 예를 들면, 1조의 schema 명은 DMA_team01이다. 이 때 schema가 존재할 경우 생성 과정을 다시 수행하지 않아야 한다.

(R2-2) Schema를 설계한 이후에는 데이터를 저장하기 위한 table을 생성해야 한다. 생성하는 table과 column 이름과 순서는 주어진 데이터셋의 table 및 column과 일치해야 한다. 0 또는 1의 값을 가지는 column은 TINYINT(1)로, INTEGER type는 'INT(11)'로, STRING type는 'VARCHAR(255)'을 이용하여 생성한다. 255자를 넘는 경우 'LONGTEXT'를 이용하여 생성한다. 그 외 날짜시간은 'DATETIME'을 통해 생성한다. 이 때 table이 존재할 경우 생성 과정을 다시 수행하지 않아야 한다. (R2-2)에서는 foreign key 조건을 작성하지 않고 (R2-3)에서 데이터 입력 후 foreign key 조건을 추가한다.

(R2-3) 생성된 테이블에 데이터를 저장해야 한다. 데이터는 csv파일로 주어지며 이를 직접 변형해선 안 된다.

(R2-4) 해당 데이터베이스 스키마에 foreign key 조건들을 반영해주어야 한다.

.

채점 기준(절대평가)

- **PART Ⅰ** ER diagram의 requirement 만족 여부(60 %)
- **PART Ⅱ** 설계한 데이터베이스 스키마와 constraints(20%), requirement 만족 여부(10%)
- 보고서 품질(5%), 발표(5%)

결과물들을 'DMA_project1_team##.zip'파일로 압축하여 발표일 전날인 **10월 12일 23:59까지** ETL에 업로드해야 한다. ETL 상에 문제가 생겼을 경우 changjin9653@snu.ac.kr 로 오류 증명 파일과 함께 해당 일시까지 보내야 한다. 제출해야 할 결과물과 파일명, 파일 확장자는 다음과 같다.

■ 보고서

- 파일명: DMA_project1_team##_보고서.pdf
- 보고서에는 PART Ⅰ에서의 문제 정의, 도식화한 ER 다이어그램의 도식화 과정과 최종 ER 다이어그램, PART Ⅱ에서의 문제 정의, 설계한 스키마와 코드에 대한 설명이 포함되어야 한다. 이 때 스키마 설계 시 constraints들과 이들의 설정 근거가 포함되어야 한다.

■ 발표 자료 및 발표 동영상

- 발표자료 파일명: DMA_project1_team##_발표자료.pdf
- 발표동영상 파일명: DMA_project1_team##_발표동영상.mp4
- 발표동영상은 팀 당 4분 이내로 제작되어야 하며 powerpoint의 녹화기능을 사용한다.

■ Python 프로그램 코드

- Python 코드 파일명: DMA_project1_team##.py
- 함수들의 입력 값들의 의미는 다음과 같다.

  host, user, password: MySQL에 접근하기 위한 계정 정보

  directory_in: 데이터가 저장된 주소(ex. C:/dir/user.txt → 'C:/dir/')

- 뼈대 코드의 주석에 작성된 TODO들에 따라 팀의 번호, MySQL 계정 정보, 데이터(csv)들이 저장된 주소 등을 바꿔야 한다.
- mysql.connector 외의 다른 패키지를 import하여 사용하는 것은 허용되지 않는다.
- PART Ⅱ의 각 requirement(R2-1~R2-4)에 해당하는 Python 코드는 주어진 뼈대 코드의 requirement# 함수로 구현되어야 한다. 예를 들면 R2-1은 requirement1 함수에 구현되어야 한다.